

ASTES

Advances in Science, Technology & Engineering Systems Journal

Special Issue

Advancement in Engineering Technology

2017-18

www.astesj.com
ISSN: 2415-6698

EDITORIAL BOARD
(Special Issue)

Editor-in-Chief

Prof. Passerini Kazmerski
Pritzker School of Molecular Engineering, University of Chicago, USA

Guest Editors

Prof. Saleh Nabi
Mitsubishi Electric Research Laboratories, USA

Prof. Mohamed Abdelaziz Hassan Eleiwa
University of Hail, KSA

Dr. Mohammad Mohiuddin Uzzal
University of New Mexico, USA

Dr. Neeti Ananthaswamy
The Catholic University of America, USA

Dr. Arnab Hazari
University of Michigan, Ann Arbor, USA

Dr. Jules Ferdinand Pagna Disso
Nettitude Ltd, UK

Dr. Zamberi Bin Jamaludin
Universiti Teknikal Malaysia Melaka, Malaysia

Dr. Danielli A. Lima
Universidade Federal de Uberlândia, Brazil

Dr. Chenxiao Cai
Nanjing University of Science and Technology, China

Prof. María Jesús Espinosa Trujillo
Universidad Tecnológica Metropolitana, Mexico

Prof. Nicolae Tudoroiu
John Abbott College, Canada

Dr. Zupan Hu
University of Michigan, Ann Arbor, USA

Dr. Sarhan M. Musa
Prairie View A&M University, USA

Dr. Shahab Tayeb
University of Nevada, Las Vegas, USA

Dr. Bhuvan Modi
Prairie View A&M University, USA

Dr. Dillon Chrimes
Vancouver Island Health Authority, Canada

Dr. Sahar Mohamed Hamed
Canadian International College, Egypt

Dr. Shaddrack Yaw Nusenu
Koforidua Technical University, Ghana

Prof. Ahmad Yusairi Bani Hashim
Universiti Teknikal Malaysia Melaka, Malaysia

Dr. Hongbo Du
Prairie View A&M University, USA

Prof. Jin Billy Li
University of Michigan, Ann Arbor, USA

Dr. Remya Ampadi Ramachandran
University of Illinois Chicago, USA

Dr. Mostafa Fadaee Fathabadi
Concordia University, Canada

Dr. Saurin H. Patel
Clemson University, USA

Dr. Matin Pirouz
University of Nevada, Las Vegas, USA

Dr. Raafat Elfouly
Rhode Island College, USA

Dr. Hossam Mosbah
Dalhousie University, Canada

Dr. Issa Tamer Elmabrouk Elfergani
Instituto de Telecomunicações, Portugal

Dr. Lihong Zheng
Charles Sturt University, Australia

Dr. Eman Hammad
University of Toronto, Canada

Dr. Nanthi Suthikarnnarunai
University of the Thai Chamber of Commerce, Thailand

Editorial

The Special Issue on *Advancement in Engineering Technology (2017–2018)* in the *Advances in Science, Technology and Engineering Systems Journal (ASTES Journal)* presents a comprehensive collection of scholarly contributions that reflect the accelerating pace of technological innovation across engineering disciplines. The period of 2017–2018 marked a significant phase in which emerging technologies began to mature and integrate more deeply into practical engineering applications. This special issue captures that transition, showcasing research that bridges the gap between theoretical development and real-world implementation, and highlighting the transformative impact of advanced technologies on engineering practice.

A prominent theme throughout this issue is the growing role of digitalization and intelligent systems in engineering technology. Contributions emphasize the integration of advanced computational tools, automation, and data-driven methodologies to enhance system design, monitoring, and control. The incorporation of artificial intelligence, machine learning, and Internet of Things (IoT) frameworks has enabled engineers to develop smarter and more responsive systems. These advancements are evident in applications ranging from intelligent manufacturing and process optimization to smart infrastructure and real-time system diagnostics, demonstrating how digital technologies are reshaping the engineering landscape.

Another key focus of this special issue is the advancement of materials and manufacturing technologies. The research presented explores innovative materials with enhanced properties, as well as modern fabrication techniques such as additive manufacturing and advanced machining processes. These developments have significantly expanded the possibilities for design and production, allowing for greater efficiency, customization, and performance. The integration of such technologies into engineering systems not only improves product quality but also contributes to reduced waste and more sustainable production practices.

Sustainability and energy efficiency also emerge as critical considerations within the collected works. As global demand for resources continues to rise, engineering technology must evolve to support environmentally responsible solutions. The studies included in this issue address challenges related to renewable energy systems, energy management, and sustainable design practices. By focusing on efficiency, resource optimization, and environmental impact, these contributions underscore the importance of aligning technological advancement with sustainable development goals.

Interdisciplinary collaboration remains a defining feature of the advancements highlighted in this issue. The complexity of modern engineering challenges requires the convergence of expertise from multiple domains, including mechanical, electrical, civil, and computer engineering. The research demonstrates how such collaboration fosters innovation, enabling the development of integrated systems that are more efficient, reliable, and adaptable. This interdisciplinary approach is particularly evident in areas such as robotics, mechatronics, and cyber-physical systems, where the synergy between hardware and software plays a crucial role.

The editorial team extends its sincere appreciation to all authors, reviewers, and contributors whose dedication and expertise have made this special issue possible. Their collective efforts have resulted in a high-quality compilation of research that not only advances academic understanding but also provides practical insights for industry and society. It is anticipated that

the findings presented herein will inspire further research and development, encouraging continued progress in engineering technology.

The Special Issue on *Advancement in Engineering Technology (2017–2018)* highlights the ongoing evolution of engineering practices in response to rapid technological change. The contributions reflect a shift toward more intelligent, efficient, and sustainable systems, driven by innovation and interdisciplinary collaboration. As engineering technology continues to advance, it will play an increasingly vital role in addressing the complex challenges of the modern world, paving the way for a more connected and resilient future.

Guest Editor

Prof. Saleh Nabi

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Special Issue **August 2017**

CONTENTS

Clean Energy Use for Cloud Computing Federation Workloads
Yahav Biran, Sudeep Pasricha, George Collins, Joel Dubow

The Internet of Things ecosystem: the blockchain and data protection issues
Nicola Fabiano

Cancer Mediating Genes Recognition using Multilayer Perceptron Model- An Application on Human Leukemia
Sougata Sheet, Anupam Ghosh and Sudhindu Bikash Mandal

Impact of Line Resistance Combined with Device Variability on Resistive RAM Memories
Hassan Aziza, Pierre Canet and Jeremy Postel-Pellerin

Design of Cognitive Radio Database using Terrain Maps and Validated Propagation Models
Anwar Mohamed Fanan, Nick Riley, Meftah Mehdawi

Building an Efficient Alert Management Model for Intrusion Detection Systems
El Mostapha Chakir, Mohamed Moughit and Youness Idrissi

Smart Grid Operational functions and Control Challenges by Implementing SSSC Tailored to Optimize performance in between Qatar and KSA on the GCC Electrical-power grid
Tariq Masood, Muhammad Tajammal, Samer Karim Shah, Ghulam Hashmi, Suhail Aftab Qureshi, D. P Kothari

Community Detection in Social Network with Outlier Recognition
Htwe Nu Win and Khin Thidar Lynn

Virtual Memory Introspection Framework for Cyber Threat Detection in Virtual Environment
Himanshu Upadhyay, Hardik Gohel, Alexander Pons and Leo Lagos

A novel beamforming based model of coverage and transmission costing in IEEE 802.11 WLAN networks
Mehdi Guessous, Lahbib Zenkouar

Software and Hardware Enhancement of Convolutional Neural Networks on GPGPUs
An-Ting Cheng, Chun-Yen Chen, Bo-Cheng Lai and Che-Huai Lin

A Test Code Generation Method for Coding Standard Input/Output with Exception Handling in Java Programming Learning Assistant System
Ei Ei Mon, Nobuo Funabiki, Ryota Kusaka, Khin Khin Zaw and Wen-Chung Kao

Framework for the Formal Specification and Verification of Security Guidelines
Zeineb Zhioua, Rabea Ameur-Boulifa and Yves Roudier

A New profiling and pipelining approach for HEVC Decoder on ZedBoard Platform
Habib Smei, Kamel Smiri, Abderrazak Jemai

An Advanced Algorithm Combining SVM and ANN Classifiers to Categorize Tumor with Position from Brain MRI Images
Rasel Ahmmed, Md. Asadur Rahman and Md. Foisal Hossain

Flexible Aperture Tuning Solution for Cellular Main Antenna in Metallic Back Cover Mobile Phone
Yew Choon Mark Tan, Guan Hong Ng, Yew Siow Roger Tay

Improving System Reliability Assessment of Safety-Critical Systems using Machine Learning Optimization Techniques
Ibrahim Alagoz, Thomas Hoiss and Reinhard German

Interference Avoidance using Spatial Modulation based Location Aware Beamforming in Cognitive Radio IOT Systems
Jayanta Datta and Hsin-Piao Lin

A New Identification Approach of MIMO Hammerstein Model with Separate Nonlinearities
Chekib Ghorbel, Zeineb Rayouf, Naceur Benhadj Baraiek

Enhanced Outdoor to Indoor Propagation Models and Impact of Different Ray Tracing Approaches at Higher Frequencies
Muhammad Usman Sheikh, Kimmo Hiltunen and Jukka Lempiainen

ARMA feeding techniques for isoflux coverage from a micro satellite
Ali Siblini ,Hussein Abou Taam, Bernard Jecko, Mohamed Rammal, Eric Arnaud, Bellion Anthony

Numerical Solution of Fuzzy Differential Equations with Z-numbers using Fuzzy Sumudu Transforms
Sina Razvarz, Raheleh Jafari and Wen Yu

TPMTM: Topic Modeling over Papers' Abstract
Than Than Wai and Sint Sint Aung

Selective Electron Beam Melting Manufacturing of Electrically Small Antennas
Saad Mufti, Christopher Smith, Alan Tennant, Luke Seed

MPC-based energy efficiency improvement in a pusher type billets reheating furnace
by Silvia Maria Zanoli, Francesco Cocchioni and Crescenzo Pepe

Design and Simulation of an RF-MEMS Switch and analysis of its Electromagnetic aspect in realtion to stress
by Amna Riaz, Muhammad Umair Javed Ilam Sindhu and Tahir Zaidi
Real-Time Flux-weakening Control for an IPMSM Drive System Using a Predictive Controller
Tian-Hua Liu, Shao-Kai Tseng, Yi Chen, Mao-Bin Lu

A Perfect Ecosystem for Learning? Modern Thoughts for Organizing Higher Education
Pasi Juvonen and Anu Kurvinen

Estimating short time interval densities in a CTM-KF model
Arlinda Alimehaj Rrecaj and Marija Malenkovska Todorova

Discovering Interesting Biological Patterns in the Context of Human Protein-Protein Interaction Network and Gene Disease Profile Data
Rami Alroobi, William Perizzo

Agent Based Fault Detection System for Chemical Processes using Negative Selection Algorithm
Naoki Kimura, Yuya Takeda and Yoshifumi Tsuge

Auto-Encoder based Deep Learning for Surface Electromyography Signal Processing
Marwa Farouk Ibrahim Ibrahim and Adel Ali Al-Jumaily

Modelling and Analysis of Radial Flux Surface Mounted Direct-Driven PMSG in Small Scale Wind Turbine
Theint Zar Htet, Zhengming Zhao, Qing Gu, Jing Li
A Switched-Capacitor Low-Pass Filter with Dynamic Switching Bias OP Amplifiers
Hiroo Wakaumi

Detection of ExoMars launcher during its passage over Europe with Space Surveillance radar breadboard
Stéphane Saillant, Marc Flécheux and Yann Mourot

A Switched-Capacitor Low-Pass Filter with Dynamic Switching Bias OP Amplifiers
Hiroo Wakaumi

A Statistical Approach for Gain Bandwidth Prediction of Phoenix-Cell Based Reflect arrays
Hassan Salti and Raphael Gillard

Performance of Location and Positioning Systems: a 3D-Ultrasonic System Case
Khaoula Mannay, Jesus Urena, Álvaro Hernández and Mohsen

Cognitive Cybernetics vs. Captology
Zdenko Balaž, Davor Predavec

Impact of Crosstalk on Signal Integrity of TSVs in 3D Integrated Circuits
Shadi MS. Harb and William R. Eisenstadt

Constructing Learning-by-Doing Pedagogical Model for Delivering 21st Century Engineering Education
Ghassan Frache, Hector Nistazakis and George Tombras
Automated Text Annotation for Social Media Data during Natural Disasters
Si Si Mar Win and Than Nwe Aung

Control of a three-stage medium voltage solid-state transformer
Claudio Busada, Hector Chiacchiarini, Sebastian Gomez Jorge, Favio Mengatto, Alejandro Oliva1, Jorge Solsona, German Bloch and Angelica Delgadillo

A 3D Full Wave Inversion (FWI) Analysis for Handheld Ground Penetrating Radar (GPR)
Suki Dauda Sule and Kevin Paulson

Mission-Critical Systems Design Framework
Kyriakos Houliotis, Panagiotis Oikonomidis, Periklis Charchalakis and Elias Stipidis

Leaf-shaped solar cell antenna for Energy Harvesting and RF Transmission in ku-band
Chokri Baccouch, Hedi Sakli, Dhaou Bouchouicha, Taoufik Aguili

Modelling of the resistance heating of the moving molybdenum sheet
Miroslav Pavelek, Michal Frivaldsky, Pavol Spanik and Tibor donič

Algorithms for Technical Integration of Virtual Power Plants into German System Operation
André Richter, Ines Hauer and Martin Wolter

Distributing the computation in combinatorial optimization experiments over the cloud
Mario Brcic, Nikica Hlupic, Nenad Katanic

Efficient Limited Feedback Technique for FDD MIMO Systems
Papis Ndiaye, Moussa Diallo and Idy Diop

A Relational Database Model and Tools for Environmental Sound Recognition
Yuksel Arslan, Abdussamet Tan?s, Huseyin Canbolat

Limitations of HVAC Offshore Cables in Large Scale Offshore Wind Farm Applications
Tiago Antunes, Tiago Alexandre dos Reis Antunes, Paulo Jorge da Costa Santos and Armando José Pinheiro Marques Pires

Development of Indicators for Technical Condition Indexing of Power Transformers
Gints Poiss, Sandra Vitolina and Janis Marks

Analysis of Economic Load Dispatch with a lot of Constraints Using Vortex Search Algorithm
Mustafa Saka, Ibrahim Eke, Suleyman Sungur Tezcan, Muslum Cengiz Taplamacioglu

Influence of supply frequency on dissipation factor measurement and stator insulation diagnosis
Cyrille Caironi, Bernhard Fruth, Detlef Hummes and Rudolf Blank

Numerical evaluation of ABS parts fabricated by fused deposition modeling and vapor smoothing
Sung-Uk Zhang
Analysis of Outdoor and Indoor Propagation at 15 GHz and Millimeter Wave Frequencies in Microcellular Environment
Muhammad Usman Sheikh and Jukka Lempiainen

Mealy-to-Moore Transformation – A state stable design of automata
Mustafa Oezguel, Florian Deeg, Sebastian M. Sattler

Improving Patient Outcomes Through Untethered Patient-Centered Health Records
Mohammed Abdulkareem Alyami, Majed Almotairi, Alberto R. Yataco and Yeong-Tae Song

Linear algebra as an alternative approach to the synthesis of digital devices of automation and control systems
Nikolay Chernov, Nikolay Prokopenko, Vladislav Yugai and Nikolay Butyrlagin

A Model for Optimising the Deployment of Cloud-hosted Application Components for Guaranteeing Multitenancy Isolation
Laud Charles Ochei and Christopher Ifeanyichukwu Ejiofor

Data Exchange Strategies for Aggregating Geographical Distribution of Demands for Location-Dependent Information Using Soft-State Sketches in VANETs
Akihiro Yamada, Susumu Ishihara

Fault Tolerant Control Based on PID-type Fuzzy Logic Controller for Switched Discrete-time Systems: An Electronic Throttle Valve Application
Wafa Gritli, Hajer Gharsallaoui, Mohamed Benrejeb

Review on security issues in RFID systems
Mohamed El Beqqal, Mostafa Azizi

A decision-making-approach for the purchasing organizational structure in Moroccan health care system
Kaoutar Jenoui and Abdellah Abouabdellah

Novel Analysis of Synchronous and Induction Generators in Parallel Operation Mode in an Isolated Electric System
Vinicius Zimmermann Silva, Ângelo Jose Junqueira Rezek, Rafael Di Lorenzo Corrêa

Design of an Automatic Forward and Back Collision Avoidance System for Automobiles
Tasneem Sanjana, Ferdus Wahid, Mehrab Masayeed Habib and Ahmed Amin Rumel

The impact of Big Data on the Android Mobile Platform for Natural Disaster Situations
Zijadin Krasniqi and Adriana Gjonaj

A High Efficiency Ultra Thin (1.8 um) CdS/CdTe p-i-n Solar Cell with CdTe and Si as BSF layer
Nahid A. Jahan, Md. Minhaz Ul Karim and M. Mofazzal Hossain

Measurement of the Bridge Surface Deflections Using Near-Field Amplitude of Secondary Radiators System
Oleksandr Poliarus, Yevhen Poliakov, Andrii Lebedynskyi, Viktor Ivanov, Ruslan Pashchenko

Domain Independent Feature Extraction using Rule Based Approach
Sint Sint Aung and Myat Su Wai

A Taxonomy for Enhancing Usability, Flexibility, and Security of User Authentication
Susan Gottschlich

Velocity obstacles for car-like mobile robots: Determination of colliding velocity and curvature pairs
Emese Gincsainé Szádeczky-Kardoss and Zoltán Gyenes

An Aggregation Model for Energy Resources Management and Market Negotiations
Omid Abrishambaf, Pedro Faria, João Spínola and Zita Vale

Medium Voltage Microgrid Test Setup and Procedures Implemented on a Real Pilot Project
Bruno Alberto Pacheco, Marcos Aurelio Izumida Martins, Cesare Quinteiro Pica and Nilo Rodrigues

A Study on Isomorphic Properties of Circulant Graphs
V. Vilfred Kamalappan

Approximate method of analysis of log-periodic antennas with in-phase currents
Boris Levin

Efficiency Assessment of Plasma-Aided Porous Media Surface Finishing
Peter Dineff, Dilyana Gospodinova, Ivaylo Ivanov

Measuring modifiability in model driven development using object oriented metrics
Nwe Nwe and Ei Thu

Constant Envelope DCT based OFDM System with M-ary PAM Mapper over Fading Channels
Rayan Hamza Alsisi and Raveendra Kolarramakrishna Rao

Heuristic Based Approach for Voltage Stability Improvement using FACTS Devices
Hajer Jmii, Asma Meddeb, Souad Chebbi

Enhancement of Voltage Stability through Optimal Location of UPFC
Hajer Jmii, Asma Meddeb, Souad Chebbi

Short CCA-Secure Attribute-Based Encryption
Hiroaki Anada and Seiko Arita

2×1 Microstrip Patch Array Antenna with Harmonic Suppression Capability for Rectenna
Nur Aisyah Amir, Shipun Anuar Hamzah, Khairun Nidzam Ramli
Cyclical Wave Bolt for Sound Waves in a Gas Stream
Vladimir Arabadzhi

A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning
Batoul Haidar, Maroun Chamoun, Ahmed Serhrouchni

Applying Machine Learning and High Performance Computing to Water Quality Assessment and Prediction

Ruijian Zhang, Deren Li

Flexible lengthening-shortening arm mechanism for fishery resource management
Yoshiki Iwamochi, Motoki Takagi, Tasuku Miyoshi

The Use of LMS AMESim in the Fault Diagnosis of a Commercial PEM Fuel Cell System
Reem Izzeldin Salim, Hassan Noura and Abbas Fardoun

Parametric Study of Micro Strip Patch Antenna Using Different Feeding Techniques for Wireless and Medical Applications
Debajyoti Chatterjee and Anjan Kumar Kundu

Simulation and FPGA Implementation of a Ring Oscillator Sensor for Complex System Design
Aziz Oukaira, Idir Mellal, Ouafaa Ettahri, Mohamed Tabaa and Ahmed Lakhssassi

A Dynamic Reallocation Based Window Access Scheme for Enhancing QoS of Vehicular Ad-hoc Networks (VANETs)
Md. Amirul Islam and Hossen Asiful Mustafa

Mission Profile Analysis of a SiC Hybrid Module for Automotive Traction Inverters and its Experimental Power-loss Validation with Electrical and Calorimetric Methods
Ajay Poonjal Pai, Tomas Reiter, Oleg Vodyakho and Martin Maerz

Signal-Based Metamodels for Predictive Reliability Analysis and Virtual Testing
Veit Bayer, Stephanie Kunath, Roland Niemeier and Jurgen Horwege

LabVIEW-based data acquisition system for Diode I-V Characterization
Nor Shaida Mohd Saufi, Nurul Syafiqah Yap Abdullah and Mohd Ikhwan Hadi Yaacob

Innovative design with learning reflexiveness for developing the Hamiltonian circuit learning games
Meng-Chien Yang, Hsuan-Yu and Chiang

Theoretical Investigation of Combined Use of PSO, Tabu Search and Lagrangian Relaxation methods to solve the Unit Commitment Problem
Sahbi Marrouchi, Nesrine Amor, Moez Ben Hessine and Souad Chebbi

Non-rigid Registration for 3D Active Shape Liver Modeling
Nesrine Trabelsi, Mohamed Ali Cherni and Dorra Ben Sellem
Adaptive observer design for a class of nonlinear systems with time delays
Ahlem Sassi, Michel Zasadzinksi, Harouna Souley Ali and Kamel Abderrahim

A Joint Safety and Security Analysis of message protection for CAN bus protocol
Luca Dariz, Gianpiero Costantino, Massimiliano Ruggeri and Fabio Martinelli

Decentralized Control Approaches of Large-Scale Interconnected Systems
Rabeb Ben Amor and Salwa Elloumi

Security Analysis and the Contribution of UPFC for Improving Voltage Stability

Asma Meddeb, Hajer Jmii and Souad Chebbi

The method of correlation investigation of acoustic signals with priority placement of microphones
Bohdan Trembach, Roman Kochan and Rostyslav Trembach

Systematic Tool Support of Engineering Education Performance Management
Aneta George, Liam Peyton and Voicu Groza

Co-designed accelerator for homomorphic encryption applications
Asma Mkhinini, Paolo Maistri, Régis Leveugle and Rached Tourki

A Smart Mobile Application for Assisting Parents in Anti-Drug Support
Tsz Hei Yeung, Chi Kit Ng and Vincent Ng

Actuator Fault Reconstruction based Adaptive Polytopic Observer for a Class of Continuous-Time LPV Systems
Radhia Houimli, Neila Bedioui and Mongi Besbes

Innovative Research on the Development of Game-based Tourism Information Services Using Component-based Software Engineering
Wei-Hsin Huang, Huei-Ming Chiao and Wei-Hsin Huang

Performance Analysis of Regenerative Braking in Permanent Magnet Synchronous Motor Drives
Andrew Adib and Rached Dhaouadi

Tracking and Detecting moving weak Targets
Naima Amrouche, Ali Khenchaf and Daoud Berkani

Structure-Preserving Modeling of Safety-Critical Combinational Circuits
Feim Ridvan Rasim and Sebastian M. Sattler

Hardware Acceleration on Cloud Services: The use of Restricted Boltzmann Machines on Handwritten Digits Recognition
Eleni Bougioukou, Nikolaos Toulgaridis, Maria Varsamou, Theodore Antonakopoulos

An Analysis of K-means Algorithm Based Network Intrusion Detection System
Yi Yi Aung, Myat Myat Min
Computation of Viability Kernels on Grid Computers for Aircraft Control in Windshear
Nikolai Botkin, Varvara Turova, Johannes Diepolder and Florian Holzapfel

Adaptive and Non Adaptive LTE Fractional Frequency Reuse Mechanisms Mobility Performance
Uttara Sawant, Robert Akl

Two-Stage Performance Engineering of Container-based Virtualization
Zheng Li, Maria Kihl, Yiqun Chen, He Zhang

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 6, 1-12 (2017)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Clean Energy Use for Cloud Computing Federation Workloads

Yahav Biran[*,1], Sudeep Pasricha[2], George Collins[1], Joel Dubow[3]

[1]*Colorado State University, Department of Systems Engineering, Fort Collins, 80523, CO, USA*

[2]*Colorado State University, Department of Electrical and Computer Engineering, Fort Collins, 80523, CO, USA*

[3]*Fulcrum Co., Cyber Security, Centreville, 20120, VA, USA*

ARTICLEINFO

Article history:
Received: 01 July, 2017
Accepted: 10 August, 2017
Online: 19, August, 2017

Keywords:
power consumption
resource utilization
cloud computing
cloud federation

ABSTRACT

Cloud providers seek to maximize their market share. Traditionally, they deploy datacenters with sufficient capacity to accommodate their entire computing demand while maintaining geographical affinity to its customers. Achieving these goals by a single cloud provider is increasingly unrealistic from a cost of ownership perspective. Moreover, the carbon emissions from underutilized datacenters place an increasing demand on electricity and is a growing factor in the cost of cloud provider datacenters. Cloud-based systems may be classified into two categories: serving systems and analytical systems. We studied two primary workload types, on-demand video streaming as a serving system and MapReduce jobs as an analytical systems and suggested two unique energy mix usage for processing that workloads. The recognition that on-demand video streaming now constitutes the bulk portion of traffic to Internet consumers provides a path to mitigate rising energy demand. On-demand video is usually served through Content Delivery Networks (CDN), often scheduled in backend and edge datacenters. This publication describes a CDN deployment solution that utilizes green energy to supply on-demand streaming workload. A cross-cloud provider collaboration will allow cloud providers to both operate near their customers and reduce operational costs, primarily by lowering the datacenter deployments per provider ratio. Our approach optimizes cross-datacenters deployment. Specifically, we model an optimized CDN-edge instance allocation system that maximizes, under a set of realistic constraints, green energy utilization. The architecture of this cross-cloud coordinator service is based on Ubernetes, an open source container cluster manager that is a federation of Kubernetes clusters. It is shown how, under reasonable constraints, it can reduce the projected datacenter's carbon emissions growth by 22% from the currently reported consumption. We also suggest operating datacenters using energy mix sources as a VoltDB-based fast data system to process offline workloads such as MapReduce jobs. We show how cross-cloud coordinator service can reduce the projected datacenters carbon emissions growth by 21% from the currently expected trajectory when processing offline MapReduce jobs.

1 Introduction

Over the past decade, cloud-based systems have been required to serve an increasing demand from users work flows and data. Cloud-based systems may be classified into two categories: *serving* systems and *analytical* systems. The former provides low-latency read or write access to data. For example, a user requests a web page to load online or requests video or audio streaming. The latter provides batch-like compute tasks that process the data offline that are later sourced to the serving systems. The service level objectives (SLO) for serving jobs are on the order of fractions of a second, while the SLO for analytical jobs are on the order of hours, sometimes days.

[*]Corresponding Author, ybiran@colostate.edu

https://dx.doi.org/10.25046/aj020601

Today, public cloud service providers (CSP) attempt to process both of these workloads with a rich platform that guarantees cost and SLO to their clients. Cloud computing is an emerging infrastructure with limited regulation and compliance requirements [1]. Recently the Office of Management and Budget issued a Federal Data Center Optimization Initiative that promotes increasing use of Green Energy and increased utilization efficiency for all US Federal datacenters [2]. Specific target numbers are set for the end of fiscal year 2018. This publication addresses how those federal requirements may be attained and how federated cloud computing is a key enabler for attaining those performance targets.

Beginning in 2013, the US government initiated a carbon-tax on IT organizations to encourage major CSPs to pursue green energy opportunities for their datacenters operations [3]. US datacenters are projected to consume approximately 73 billion kWh by 2020 [3] with a corresponding increase greenhouse gases. Green energy generation growth is expected to triple by 2040 [4]. However, there is no cohesive system existing to coordinate the rising datacenter energy demand with rising green energy supply. This chapter utilizes Cloud-Federation as a multi-cloud resource coordination system that matches computational resource demands with available energy supply to maximize the utilization of green energy for processing cloud-workloads.

Most CSPs seek more market share in competition with other CSPs. One outcome of such competition is an ever-growing infrastructure in the form of new datacenters across the globe with no countervailing forces to meet user demand more efficiently and satisfy societal environmental and energy requirements. This sub optimum use of infrastructure increases the carbon footprint attributable to cloud computing services and also drives up costs to CSP's. The following sections investigate two types of workloads, serving and analytical systems. This chapter will focus on two workload types, On-demand streaming as a representation of *serving* systems. It will also investigate *analytical* systems and present a unique model that allows optimal clean energy usage.

1.1 Enabling Green Content Distribution Network by Cloud Orchestration

On-demand streaming constitutes up to 85% of Internet traffic consumption [5]. On-demand streaming content is managed and distributed by content service providers. It then cached and distributed by Content Delivery Networks(CDN) located at the edges of the Internet network close to the consumers. Because streaming constitutes such a large fraction of Internet resource consumption, this paper will, of necessity, focus on methods to employ green energy to better operate CDN instances of on-demand streaming jobs, which include both video and audio content.

Meeting the Federally mandated approach of maximizing the utilization of green energy to operate CDN instances (for government with recommendations for private sector use as well) requires an energy source-demand coupling scheme that insures SLO levels of power availability but is structurally biased towards green energy sources over hydrocarbon fueled energy sources. A system to accomplish this will have to provide seamless failover in the case of sudden interruption of green energy to grid-energy sources or vice versa i.e., fallback from grid-energy to green energy when surplus green energy is available.

User expectations in on-demand streaming requires different service level requirements than other serving systems workloads. Serving systems workloads are comprised of interactive sessions that pivots on minimum latency. However, low latency is less critical in analytical on-demand streaming since application clients use buffering techniques to mitigate long latency effects. Therefore, on-demand streaming workloads fits, more closely than interactive workloads, with the observed intermittent and varying green energy availability characteristics.

Green energy supply is unpredictable and requires a complex, adaptable, resource allocation system to provide CDN services with steady energy supplies while concurrently seeking minimal carbon footprint. This dynamic availability of green energy resources in a smart grid requires real time communication of both short term and predictive energy needs from cloud service providers to green energy providers. The green energy providers need to disclose availability dynamically to CSPs, who, in turn, disclose their changing energy demands for near term computing. SPs can then better maximize the use of green energy for on-demand streaming processing.

This is a classical resource management and coordination problem [6, 7]. The following approach builds upon prior work that was done in this area [8, 9], specifically that done on alleviating the sudden lack of green energy to meet low-latency workloads. The approach herein employs an application-buffering scheme that better allows for opportunistic, green, on-demand streaming processing. It requires an extended, cohesive, federated system that aggregates supply and demand across multiple geographic locations employing the smart grid command and control infrastructure to achieve an optimal dynamic matching of green energy sources and computing loads.

This chapter proposes an implementation that utilizes a control component in a federated cloud that coordinates and optimizes the resource allocation among the participant CDN providers. It treats the volatile nature of green energy resources as a resource allocation problem, the solution of which is a resource orchestration system that is optimized with the goal to operate increasingly near to the limit of supply by green energy sources constrained by SLO reliability requirements. This system will be demonstrated by modeling a prototype that simulates resource allocation in a micro federated cloud eco-system to achieve an energy supply-computation demand match optimized within seconds.

Since the focus of this work is on green energy utilization in a federated cloud, the scheduling algorithms and resource management issues, while important, are discussed only to the extent necessary to help the reader understand the required architecture for heterogeneous energy compute clustering. This work is meant both as a case study in energy utilization, and a presentation of a novel method of coordinating high-velocity data streams, and extended to a unified orchestration system, to optimize the performance of federated cloud systems.

The chapter starts, with the on-demand streaming economics, increased green energy utilization and anticipated smart grid progressions as applied to on-demand streaming. Then it discusses the green energy utilization problem is analyzed. Finally, we present a cloud coordinator prototype that is built on Kubernetes [10, 11], an open source cluster manger, and extend that prototype to discusses the need for and requirements of a unified system that orchestrates cluster compute resources in a federated cloud.

1.1.1 On-Demand Streaming and CDN

Over the last decade, video and audio traffic became the dominant segment of consumer internet traffic. Cloud service-providers such as Netflix, Amazon Instant and YouTube disrupted the prior linear TV data distribution model. Also, video streams delivered by mobile terminals grew as mobile connectivity improved [12]. Video streaming is expected to constitute up to 85% of Internet consumers traffic [5] within a few years. The US portion of video streaming is 14%[1] and the number of US Unique IPv4 connected addresses is 17% [12]. The streaming workload is comprised of live streaming and on-demand streaming, with the relative fractions of 6% and 94% respectively [5]. Other predictions support similar ratios, 12% live-streaming and 88% for on-demand.

A key driver for the rapid expansion of streaming video was the shift from specialized streaming protocols and infrastructures such as RTSP, and RTMP [13] to a simple HTTP progressive download protocol. This led to a shift from proprietary streaming appliances to commodity servers. In turn, this change removed a barrier for CDN's to process on-demand workloads. Most present day, CDN service providers support a seamless integration with cloud-based object storage that pipelines the digital content from the organization site to the CDN instance that runs at the Internet provider edge[2]. Furthermore, the HTTP chunk-based streaming protocol support in a CDN allows the client application sufficient time to detect the optimal CDN instance to handle user workload. The optimal CDN instance assignment is done by the cloud control plane resource manager. The prototype described below will demonstrate such optimal resource allocation.

We used server utilization and power metrics from [3, 12] to design the prototype. Most of these sources we considered have limited utilization rates and server utilization distribution. Also, utilization and power consumption do not scale linearly [14]. However, for clarity, in the interest of maintaining focus on the larger goal of the paper, CDN resource management systems for green energy utilization, we assumed linear relationships and accepted the risk of loss of accuracy in our estimates. High fidelity simulation accuracy is not critical for the goal of this chapter.

1.1.2 Energy Saving Potential in Operating a Distributed CDN Resource Management System

The approach is to aggregate the required traffic for on-demand workload processing, and use standard compute device specification to assess the electrical energy and carbon footprint that will be required by that workload[3].

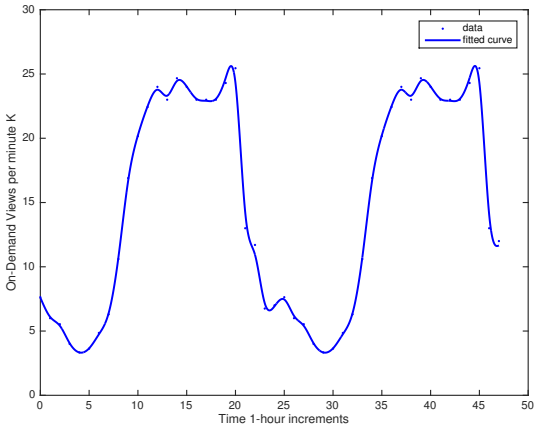


Figure 1: On-Demand Video views observed throughout 48 hours with 1-hour increments. Data was fitted with smoothingspline for curve and surface fitting [5]

The estimated data rate for streaming is given as $S_{total} = 63000PB/mo$ (where PB is Petabytes). Figure 1 shows users workload pattern of 9 busy hours in which the workload spans throughout 13 hours a day which yield $126PB/sec/mo/$ The on-demand streaming portion is estimated as 78% across four main US regions denoted by $k = 4$, the number of region used, for the purposes of this paper although k can be varied depending upon the degree of granularity desired in the simulations. $S_{on-demand}$ denote the on-demand portion.

$$S_{on-demand} = S_{total} \cdot 78\% = 49140PB/mo$$

$n_{hours/day}$ denotes the number of effective hours in a day for streaming.

$$n_{hours/day} = n_{busy} + k = 13$$

$$D_{rate} = \frac{S_{on-demand}}{n_{hours/day}} = 126PB/sec/mo$$

N_{max} denotes the estimated number of required servers in maximum CPU capacity. N_{max} is bounded by the maximum network throughput a single server can ingest. Standard commodity servers can handle up to 8.5Gbps i.e. 1.026GB/sec.

$$N_{max} = \frac{D_{rate}}{T_{max}} = \frac{126 \cdot 10^6 GB/sec}{1.026GB/sec} = 118588235.3$$

u_{opt} denotes the CPU utilization factor so servers has sufficient capacity to handle management tasks. We estimate 60% utilization factor $u_{opt} = 100/60$.

$$N_{opt} = N_{max} \cdot u_{opt} = 118588235.3 \cdot \frac{100}{60} = 197647058.8$$

E_s denotes the midrange server energy consumption for various server types s. We consider three types of servers:(1) compute server(5kWh/server), (2)

[1]Internet Statistics retrieved from https://www.statista.com/chart/2647/global-internet-usage-by-the-numbers/
[2]Cloud Front reference retrieved from https://aws.amazon.com/cloudfront/
[3]OpenCompute Project, Servers Specification guide retrieved from http://www.opencompute.org/wiki/Server

digital storage server(1.7kWh/server) and (3) network server(1kWh/server). Storage server acts the digital storage controller. The network server acts as the router and switch. The compute server is the server that processes the on-demand streaming.

$$E_y = N_{opt} \cdot \sum_{s \in S} E_s \quad (1)$$

$$= N_{opt} \cdot (5kWh + 1.7kWh + 1kWh)$$

$$= 1.521 \cdot 10^6 kWh/mo = 18.26GWh/y$$

The saving potential from running on-demand video streaming using green energy resources is 18.26GWh a year based on current on-demand consumption and expected to grow 89% by 2019 [4]. i.e. 34.5GWh per year for on-demand streaming. The next sections will explain the challenges in utilizing green energy followed by a method that addresses some of these challenges and thereby maximizes the utilization of green energy.

1.2 Coordinating Green Clouds as Data-Intensive Computing

The following paragraph focuses on *analytical* systems workloads that typically comprise 48% of the cloud workload [15]. Also, some of the workloads patterns can be predicted as recurring jobs [16] and the deadlines for *analytical* jobs are more liberal by an order of magnitude than the *serving* jobs. Finally, one of the ways to handle offline workloads is to process the data streams offline through a highly scalable data-parallel frameworks like MapReduce by submitting jobs to a control plane [17].

This chapter proposes a component in a federated cloud that will optimize the resource allocation and coordination among the participant CSPs. It will also focus on addressing the time and power volatile nature of renewable energy resources as a fast date problem. It will present a resource management system with a goal to increase the utilization of data-centers, and to operate increasingly to the limit of supply by renewable energy sources. Finally, a prototype is built to simulate resources allocation in a micro federated cloud eco-system to achieve a supply demand optimized match within seconds.

Clean energy sources are time and power volatile and require complex resource allocation and coordination systems to maintain highly available data-center service with steady energy at a minimal carbon footprint. Further, availability of clean energy resources in a smart grid can be pushed from a variety of energy sources deployed across the nation as can the pull needs of data-service centers. Thus, CSPs will publish their compute resources availability, and service providers (SP), will publish their changing energy demands for near term computing. These streams of data include both high-volume and high-velocity characteristics termed a fast-data problem [18]. Our proposal uses as a base previous work that was done in this area [8, 19, 20]. However, previous work treated clean energy within a single data-center operated by a CSP. The proposal below suggests an extended cohesive system that aggregates supply and demand across multiple geographic locations employing the smart grid sense command and control to achieve an optimal match.

1.2.1 Optimum Clean Energy Utilization is a Fast-Data Problem!

This study suggests that efficient, clean energy utilization requires three consecutive steps: (1) Clean energy resource availability signals. (2) Exploration and analysis of prior years seasonal solar data and current weather reports and evolving green energy capacity availability planning. (3) Acting fast enough based on the predictive analysis. The last step is the key component in solving the optimum allocation problem.

We judge that clean energy utilization is different from the classic big-data problem such as the Hadoop MapReduce Method. Big-data solutions solve the case where the data is at rest, not fully consistent (aka eventual consistency) and require liberal SLO that is later provided as computing trends, and other business intelligence applications. However, our case handles data in motion that requires consistent, real-time aggregation, and transaction processing. That is per-event decision making using both real-time contextual and historical data as dual guides for proper algorithms. Finally, we argue that based on the clean energies' volatile nature, processing streams of compute demand and clean energy supply requires that a need for a compute resources can be addressed with a currently available demand.

1.3 Why is Green Energy Utilization Hard?

The following section describes why utilizing green energy for compute purposes, while a justifiable goal, is limited by SLO reliability. It will present a scenario where balancing time-varying energy generation patterns with changing dynamic energy demands of cloud computing sometimes conflict. The green energy time varying generation patterns considered by us focuses on wind and solar generation. Figure 4 show historical data on dynamic nature of green energy sources and Figure 5 shows the dynamic cloud energy demand.

1.3.1 Frequency Stability in Wind-Power Generation

The daily wind power variation characteristics will be employed as a metric that illustrates the duration and level for a given amount if wind energy availability. The electricity generation process from wind is comprised of a wind turbine extracting a kinetic energy from the air flow. The wind is rarely steady; it is influenced by the weather system and the ground surface conditions, which are often turbulent [21]. Also, the generation process must happen at the same instant it is consumed [22] unless it is stored in grid level battery banks. Unfortunately, grid level energy storage technology is not keeping up with grid level energy generation technology.

Sample wind and power generation data were obtained from NREL [22]. We used datasets from 2006-2012 across different regions in the US and aggregated more than 600

[4] https://www2.deloitte.com/content/dam/Deloitte/in/Documents/technology-media-telecommunications/in-tmt-rise-of-on-demand-content.pdf

observations. Finally, the data were fit using smoothing splines[5]. The usable power generated from a wind turbine is generally described by a Rayleigh distribution [21]. It defines three main points in the wind power generation process: (1) the cut-in is the minimum viable wind speed for electricity generation from a wind turbine. (2) the rated level, describes the point where the power reached its local maximal capacity without adverse effects on the turbine life by too strong a wind. (3) Cut-Off, is the term for the local minimum for the generation cycle. Beneath that speed, there is not enough power for viable electricity generation. Thus, if the wind velocity is too low, the data-server gets no wind energy. Figure 2 shows wind generation variations that crudely fit a Rayleigh distribution with $b = 300$ assuming the form of the Rayleigh Probability Distribution Function is:

$$f(x|b) = \frac{x}{b^2} e^{\frac{-x^2}{2b^2}}$$

The measured generation cycles range between 140 and 180 minutes per cycle. Equation 2 expresses the generated power by a wind turbine, given a wind velocity. The function g describes a viable electricity generation given a wind power. The wind power availability indications will be generated by a wind turbine and fed into the coordinator database as a potential power source to datacenters in a region. Our prototype will assume wind power availability indications as the wind tuple {*region,cut-in,rated,cut-off*} indications.

$$P_{output}(wind_v) = \begin{cases} 0 & : \text{ if } wind_v \leq rated \\ g(wind_v) & : \text{ if } rated < wind_v \leq c_{out} \\ 0 & : \text{ if } wind_v \leq c_{out} \end{cases} \quad (2)$$

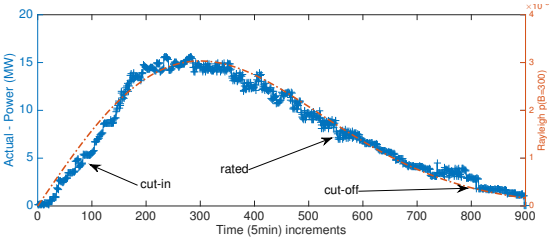


Figure 2: Aggregated wind power measurements between 2007-2012 that fits Rayleigh Distribution with b=300

1.3.2 Efficiency and Daily/Hourly Availability in Solar-Power Generation

Photovoltaic solar (PV) energy availability is defined by the solar power intensity denoted by s(*Watts/m2*), which varies with local daylight hours and the clear or cloudy sky conditions [22]. Moreover, the PV cells are most effective at lower temperatures [23]. The PV cells electrical power generation, defined by Equation 3, is a function of the solar intensity denoted by η_{solar}. Solar power generation also depends on the PV power efficiency denoted by s. It encapsulates both the predicated temperature,the sky conditions, the solar cell efficiency, and the DC to AC inverter efficiency. The solar cell area denoted by $a(m^2)$.

$$P_{output}(s) = \eta_{solar} \cdot s \cdot a \quad (3)$$

Our prototype will assume solar power availability indications as the solar-tuple {*region,power-efficiency*}. Based on the solar generation pattern presented in figure 3, the generation prediction utilizes the the local time in a region and the given power-efficiency

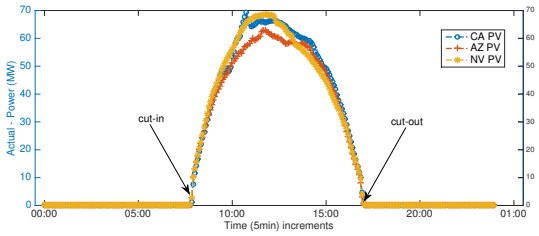


Figure 3: Aggregated solar power generation between 2008-2011 taken in Palm Springs CA, Prescott Airport CPV, AZ and Nevada Solar One, NV, indicating on a stable and fixed solar-based power [22]

1.3.3 Optimum Green Energy Utilization for On-Demand Streaming is a Resource Management Problem!

This study suggests that efficient; green energy utilization for on-demand video streaming workloads has three main requirements: (1) efficient compute resource discovery, (2) efficient load balancing among the provisioned compute resources and (3) smart failover mechanism that mask failover events from *green* CDN-edge instance to *grid* CDN-edge instance while end-users stream on-demand video [24]. These are discussed below.

Compute Resources Discovery. This assessment comprises both internal and external discovery. Internal discovery refers to CDN-edge instances that run in compute pods that must be able to be easily discovered and connected to control-plane endpoints consistently regardless of which cloud-service-provider is hosting the CDN-edge. External discovery refers to the ability of end-users discovering CDN-edge instances through DNS services for HTTP(S) on-demand video streaming.

Optimal Load Balancing is the seeking of the "best" CDN-edge, based on optimization criteria, for any given the workload processing. After initial discover and connection, clients should be served by the optimal instances based on proximity from the end-users, current load factor, and the availability of green energy resources. e.g., session requests originated from New Jersey should be served by US East as oppose to US West to avoid latency and signal loss.

Efficient Failover is a main component for on-demand video streaming based on green energy. If the endpoint becomes unavailable, in this case due to a sudden lack of green energy, the system must failover the client to another available endpoint that manages the streamed content. Also, failover must be completely automatic i.e. the clients end of the connection remains intact, and the end-user oblivious to the failover event, which means that the end-user's client software requires no support handling failover events. Finally, multiple CDN-edge instances co-located in a region should be accessible by end-users through Domain Naming

[5]Matlab Smoothing Splines retrieved from https://www.mathworks.com/help/curvefit/smoothing-splines.html?requestedDomain=www.mathworks.com

Service (DNS), as most clients-streaming (browsers) software supports DNS resolutions for finding available CDN-edges.

The green energy utilization model for processing on-demand video streaming is different than the classic scheduling problem where classical optimal resource allocation techniques are applied [25, 26]. We argue that based on the green energies' volatile nature and the on-demand video streaming workload characteristics, the optimal resource allocation approach should be opportunistic. It requires an effective resource management system for processing on-demand video streaming workloads. Our prototype will employ a Kubernetes flavor "Ubernetes" that implements the three main requirements above.

2 Clean Energy Mix Evaluation for Online On-Demand System

In the following section we evaluate a compute load coordination system component that harmonizes *on-demand streaming* job demands with available compute resources, with priority given to those powered by green energy sources. Such resources will be published to the coordination system through a resource availability tuple {*region,cut-in,rated,cut-off,power-efficiency*}, where *region* indicates the geographic availability region. *power-efficiency* indicates solar or wind based energy power efficiency. *cut-in, rated* and *cut-off* the values appropriate to those energies.

On-demand streaming job demand includes the specific *region, total-job workload, load-factor*, as well as *contract deadline* SLA. The *load-factor* indicates the required number of CPU cores per the *total-job-workload*. The geographic *region* indication will be used to optimize the match between the supply and demand. Also, the *total-job-workload* and the *deadline* will be compared against the *cut-in,rated, cutoff* time for wind or *power-efficiency* for solar, based on the published *load-factor*.

We suggest a hybrid datacenter that does not deviate from the common datacenter architecture. The core difference lies on an automatic transfer switch (ATS) that switches between different available power sources: generator, grid or green energy when available. In both cases the datacenter design does not change and requires incremental changes only by adding green energy power sources to the datacenter's ATS's (Figure 6).

We suggest two types of compute clusters, green-clusters powered by green energy and grid-clusters powered by the electrical grid. Figure 6 shows a simplified datacenter power distribution that supports green energy sources. In such datacenter, both *serving* and *analytical* systems deployed in grid clusters. Further, for incoming *analytical* workloads, few clusters use green resources when there are a viable green energy and mostly standby. As a mitigation strategy, a compute live migration procedure will be available in case of unpredicted lack of renewable resources during a workload processing which presents a risk for SLO violation.

2.1 Experiment Planning

Below is simulation of a cross-regional platform that is comprised of control-plane, workload-plane and coordinating components. This will be embodied in a resource allocation system (Kubernetes). This system will: (1) provision resources to be neared users; (2) optimize utilization by prioritizing the use of underutilized resources; and (3) seamlessly remove malfunctioning hardware from the system. The control-plane will enable an effective compute resource provisioning system that spans across different public cloud providers and regions. The coordinating components will accept user-workload demands as well as green energy availability from various regions and opportunistically seek to process streaming workloads using compute resources provisioned by green energy resources. The workload-plane will be comprised of edge streaming servers that process the end-user on-demand video streaming. It will built of standard Apache HTTP[6] servers that runs on the edge location.

The control-plane software infrastructure is based on Kubernetes [10], it facilitates internal discovery between CDN instances so instances can connect across different cloud boundaries and regions. Further, end-users can discover the optimal CDN-edges that are (1) nearby, (2) less loaded and (3) healthy. Finally, the Kubernetes automation framework allows the failover mechanism with no dependency upon the end-user client. In particular, we will exploit the *livenessProbe* option that automatically removes green-compute pods, a set of CDN-edge instances, in case of a sudden lack of green energy.

The coordinator component accepts incoming supply and demand traffic, calculates a potential match, within minutes, and notifies back the CSP and the SP for transaction completion. We use Redis[7] as the in-memory data store as the database that stores the system supply and demand calls originated by the end-user workload. The workload is generated by Jmeter instances[8]. The workload generated based on the on-demand video views observed by [12] depicted in Figure 1. Green energy availability simulated based on the known regional patterns depicted by Figure 2 and Figure 3.

We count the number of matches i.e. on-demand video streaming processed by CDN-edge instances operating on green energy. Also, we measure the false-positive cases where a match was suggested but did not met the SLO's deadline due to a violation that caused by a sudden lack of green energy resources. We use the data to extrapolate the possible energy (kWh) that could be generated by the using green CDN-edge instances depicts in Equation 1.

2.2 Execution - The Preparation

The prototype experiment included the setup of three virtual datacenters deployed in different regions: (1) Central US, (2) West US and (3) East US. The clusters were sized

[6] Apache Web Server reference retrieved from https://httpd.apache.org
[7] http://redis.io
[8] http://jmeter.apache.org

based on US population distribution[9] by regions i.e. 20% for West US, 40% for East US and 40% Central US. The cluster sizes for West US, Central US, and East US are 3, 7 and 7 machines respectively. Each machine is standard 2-CPU cores with 7.5GB of memory. Also, the user demand simulation will rely on the US population distribution. Finally, the green energy supply simulation will be based on wind or solar availability observed in the various regions.

The control-plane is comprised of docker API server and controller-manager. The controller coordinator component will need to allocate resources across several geographic regions different cloud providers. The API server will run a new federation namespace dedicated for the experiment in a manner such that resources are provisioned under a single system. Since the single system may expose external IPs it needs to be protected by an appropriate level of asynchronous encryption[10]. For simplicity, we use a single cloud provider, Google Container Engine, as it provides a multi-zone production-grade compute orchestration system. The compute instances that process the user workloads are deployed as Docker containers that run Ubuntu 15 loaded with Apache HTTP server. For simplicity, we avoid content distribution by embedding the video content to be streamed in the Docker image. We run 52 Docker containers that span across the three regions and act as CDN-edges. Green CDN-edge instances differ from grid CDN-edge instances by Kubernetes labeling. The simulation of the hybrid datacenter is depicted in Figure 6.

A coordination database system that aggregates green energy, solar or wind, availability, was built in software. When energy sources manifest the cut-in patterns depicted by Equation 2 and Equation 3, the coordination system starts green CDN edges in the availability regions. Also, when green energy availability reaches cut-off rates, the coordination system turns off green CDN edge instances.

2.3 Baseline and Variability of Workloads

The baseline execution included data populations of both green energy availability and user demand for video streaming. The data population was achieved by the Kubernetes-based Jmeter batch jobs. The loader jobs goal is to populate the coordinator database with green energy supply based on using a Weibull distribution, which is a generalization of the Rayleigh distribution described above for wind and a normal distribution for solar. Also, the user demand was populated according to the observed empirical patterns depicted by Figure 1.

We simulated the availability and unavailability of green energy using Jmeter-based workload plan against the coordination system. Our implementation starts green CDN-edge instances opportunistically upon green energy availability. Once a CDN-edge instance declares its availability it processes live workloads.

We use the Kubernetes *livenessProbe* for communication between CDN-edge instance pool and its load-balancer that divert traffic to its pool members. Finally, another workload Jmeter-based simulator generates on-demand streaming calls. This workload simulates end-user demand. It includes HTTP progressive download calls to pre-deployed video media in the CDN-edges.

2.4 Main Execution

In each of the three regional CDN-edge clusters the Kubernetes Jmeter batch jobs that generated green availability traffic to the coordination component were executed. The simulation is comprised of availability indication that are based on Figure 2 and Figure 3. We randomized solar production by using a factor of $\alpha = 0.2$ based on collected data between 2008-2011 in Palm Springs CA, Prescott Airport CPV, AZ and Nevada Solar One, NV [3]. Also, we randomized the wind production by a factor of $\beta = 0.4$ based on collected data between the years 2007-2012 [22]. The demand simulations included a set of calls to the coordinator component spread across 48 hours. The calls originated from three different timezones. The supply simulations consist of wind and solar-based energy time and power windows.

The experiment executions generated two main data traces that we used for the resulting generation computation. The first trace is the simulators logs. The simulator logs includes the demand and supply records. Demand records stored in the Redis key-value store under the key "DemandEvents" followed by timestamp, region and the required compute capacity. The supply calls were stored in the Redis key-value store under the key "SupplyEvents" follows by timestamp, region and supply phase i.e. cut_{in}, cut_{out} or *rated*. For query simplicity the loader ingested three types of records for each supply and demand the by the keys: (1) supply or demand (2) timestamp, and (3) by region. This approach optimized the coordinator queries by timestamp and regions for green CDN-edge instances allocation.

The second trace is the actual allocation logs. It is generated by the coordination system that invokes the Kuberenetes command for green CDN-edge instances initialization and disposal. This was used to determine the green energy utility translated into energy (kWh) that did not use grid energy sources.

2.4.1 Limitations

Every supply and demand was recorded three times to ease the query process. This approach was used since Redis provided limited query abilities by different keys. This approach might suffer data inconsistency issues where a supply metric was successfully committed to one key recode but missing on other key. Production systems should add extra safety gates when ingesting data. We used Redis because of its popularity in the Kubentese community. However, our approach is not limited to Redis or other database systems for that matter.

When measuring the green energy overall utility, we used the container initialization and disposal as indication that green energy utility was used. Specifically, we used the *'kubectl logs POD'* command based on the assumption that the coordination system invocation commands are tightly

[9] US Population Distribution retrieved from https://www.census.gov/popclock/data

[10] Simulation code and data retrieved from https://github.com/yahavb/green-content-delivery-network

coupled with green energy availability. It is likely that collecting the actual video streaming traces through the various Apache access logs of the CDN-edges will be more accurate.

In the case of a sudden lack of green energy while streaming video a failover occurs. Such failover event relies on domain naming services (DNS), the impact of DNS caching was not included since that might cause streaming delay on the user side. Also, when the coordinator algorithm determines there is enough green energy available it will take grid pods down and activate green pods up in a controlled fashion e.g., one at a time so that no requests are lost during the transition phase. For simplicity, the algorithm avoid that.

2.5 Analysis

The green energy supply simulation plotted in Figure 4 shows the energy generation in MW for both wind and solar sources. The simulated amounts were adjusted to the amount observed in the traces between 2007-2012 [22].

The user workload simulation plotted in Figure 4 follows the observed user patterns depicts in Figure 1. Also, it shows the aggregated green energy availability for each region. The cloud-coordinator uses these data sets to determine if there is enough green energy available before provisioning green-pods and possibly taking grid-pods down.

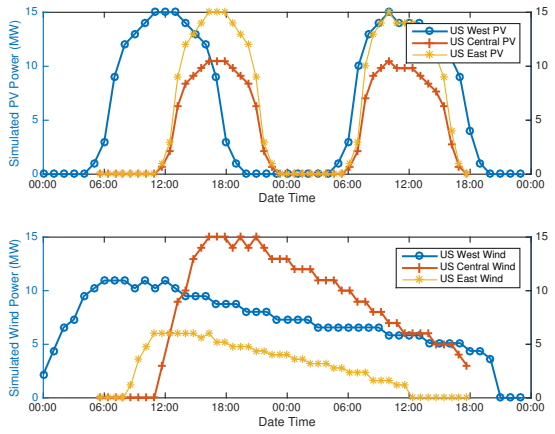


Figure 4: Green energy availability simulated in MW across three regions. Amounts are adjusted to NERL measurements i.e., wind generation in West US moderate, Central US outstanding, East US fair. For solar generation in West US strong, Central US moderate-high, and East US low.

The utility of the green energy was calculated based on the cases where sufficient green energy was available to run the green-CDN pods within the same region. Otherwise cross-regional latencies might degrade the on-demand video experience. The measurements in Figure 5 were adjusted to the estimates of required energy (kWh) for operating the green compute pods. The case where there was negative green energy available it was considered as a miss in the overall utility reckoning.

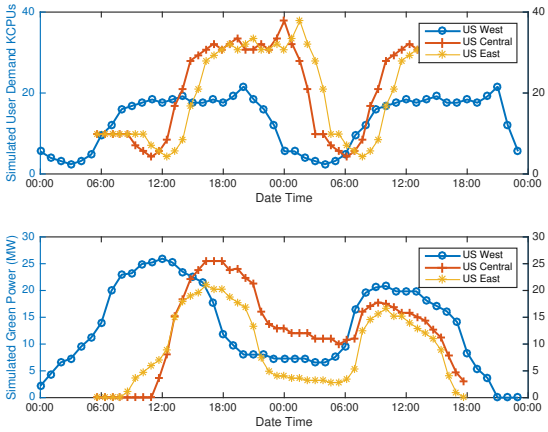


Figure 5: Video on-demand user workload per region adjusted with user population opposed to aggregated green energy availability, solar and wind energy.

2.6 Discussion on Energy Mix when Processing CDN Workloads

The scenario described above simulated the usage of green 801.3 *kWh* out of total 3642 *kWh* to process the video on-demand streaming workload. i.e. 22% by opportunistic matching. When counting the utility per region West US used 42% of the green energy. Central US used 28% of the simulated green energy. East US utilized only 18% as the initial ratio between user demand and green energy availability was relatively low. Although West US reached 40% utilization it contributed nationally only to the 20% portion it contains from the entire experiment test set. By way of comparison, Jeff Barr of AWS noted that their data centers utilize a 28% cleaner power mix[11]. Extrapolating the simulation results to the initial assessment in Equation 1 yields to a saving of:

$$(18.26(GWh) \cdot 1.89) \cdot 0.1(\$/kWh) \cdot 22\% = \$759,250.8/year$$

3 Clean Energy Mix Evaluation for Offline System

In the following section we evaluate a coordination component that harmonizes *analytics* jobs demands with available compute resources powered by green energy resources. Such resources will be published to the coordination system through a resource availability tuple {*region,cut-in,rated,cut-off,power-efficiency*}, where *region* and *power-efficiency* indicates solar or wind based energy and *region, cut-in, rated* and *cut-off* of those energies.

Analytics job demand includes the specific region, total-job workload, load-factor, as well as contract deadline stermed tuple. The *load-factor* indicates the required number of CPU cores per the *total-job-workload*. The region indication will optimize the match between the supply and demand. Also, the *total-job-workload* and the *deadline* will be checked against the *cut-in-rated, cutoff* time for wind or *power-efficiency* for solar, based on the published *load-factor*.

[11] Amazon Web Services Sustainability reference retrieved from https://aws.amazon.com/about-aws/sustainability/

We will suggest a hybrid data center structure that does not deviate from the common data-center architecture. The core difference lies on an automatic transfer switch (ATS) that switches between different available power sources: generator, grid or clean-energy when available. In both cases the data-center design does not change and requires incremental changes only by adding clean-energy power sources to the datacenter's ATS's (Figure 6).

Figure 6 shows a simplified data-center power distribution that supports clean energy sources. In a data-center with available clean-energy resources for both *serving* and *analytical* systems deployed in brown clusters. Further, few clusters use green resources when there are a viable clean energy and standby for incoming *analytical* workloads. As a mitigation strategy, a compute live migration procedure will be available in case of unpredicted lack of renewable resources during a workload processing with a risk for SLO violation.

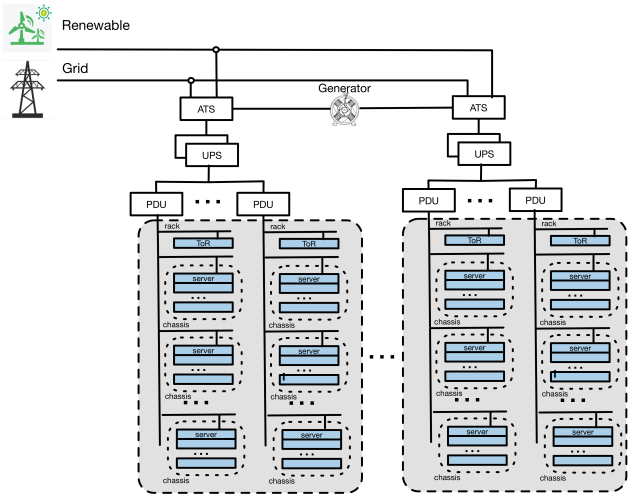


Figure 6: Datacenter Architecture - Compute server clusters aggregated by racks and chassis. Racks are fed by separate Power Distribution Unit (PDU) that connects to Uninterruptible Power Supply (UPS) that connects to the primary power source through Automatic Transfer Switch (ATS).

3.1 Experiment Planning

We wish to simulate an isolated group of computing resources so it can operate by various energy resources, especially available green ones. We use a group of leased resources from existing cloud providers to form a virtual-data-centers set that operates in a federated scheme. Each virtual-data-center includes with internal arbitrator component that collects and aggregates internal signals about its utilization and availability. The arbitrator then reports to the central coordination system. We use Apache Mesos[12] for the virtual-data-center abstraction.

Also, we build a highly available coordination component that accepts incoming supply and demand traffic, calculates a potential match, within minutes, and notifies back the cloud-service-provider and the service-provider for transaction completion. We use VoltDB[13] as the database and application server for the coordinator component.

[12]http://mesos.apache.org
[13]http://voltdb.com

Finally, we simulate customer's demand for compute resources through client simulator with a Java-based application that generates pseudo demand traffic to the coordinator service. The coordinator service will run on separate resources pool than the virtual-data-centers and the client simulator.(Figure 7)

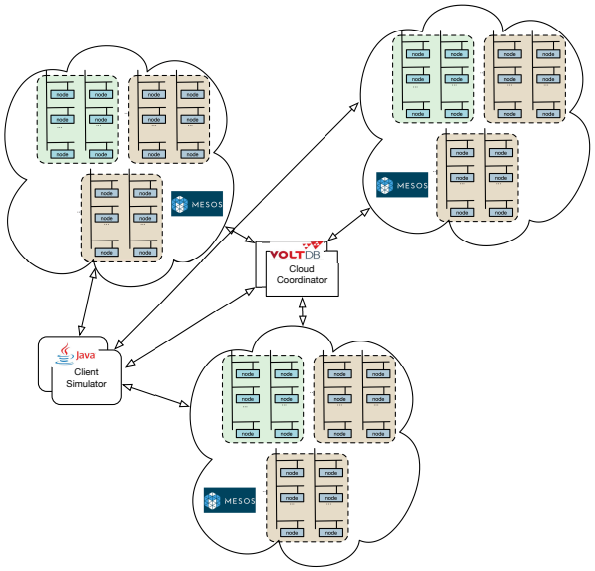


Figure 7: Experiment Architecture - Three virtual clusters deployed on a public cloud services located in a different regions grouped by Apache Mesos. A single VoltDB instance as the coordinator service. Finally, Client simulator that generates pseudo demand that sends data to process

The data collected includes customer workload for jobs processing and jobs deadline. Also, the number of matches found by the coordination component and processed by green clusters without SLO. Finally, we measure the false-positive cases where a match was suggested but not met the SLO's deadline due to a violation that caused by a sudden lack of clean energy resources. We use the data to extrapolate the possible carbon-footprint that could be generated by the used virtual clusters.

The research goal is to show a significant improvement in the carbon emission generation by data-centers. Let $MtCO_2e$ denote the carbon emission. Electrical usage can be consumed by the $Consumers$ set: {cooling, storage, servers, CPU, power} systems and denoted by EU measured in kWh. The average regional carbon dioxide emissions measured in lbs/kWh and denoted by RE_{co_2}. Therefore, the total electrical usage is:

$$EU_{Total}(kWh) = \sum_{c_i \in Consumers} EU(c_i) \tag{4}$$

and the carbon footprint generated by the workload is:

$$MtCO_2e = \frac{EU_{Total}(kWh) \cdot RE_{co_2}(lbs/kWh)}{2,204.6(lbs)} \tag{5}$$

As the experiment uses virtual data-centers, we do not have access to the power consumption by the cooling, power and storage system. The results capture compute jobs durations

measure in server $\frac{CPUcore}{time}$ and use the average Thermal Design Power (TDP) of 200W per core (0.2kW). c_i and t_i denotes the number of cores and time used per job respectively. The electric usage by server ($EU_{servers}$) will be the sum of the electric utilization of the executed jobs that was matched by the coordinator component (Equation 6)

$$EU_{servers}(kWh) = \sum_{i=1}^{jobs} c_i \cdot 0.2 \cdot t_i \quad (6)$$

We assume a linear relation between the power utilization of the cooling, power and storage systems with the servers power utilization. (Equation 4)

3.2 Execution - The Preparation

The Preparation for the execution included the setup of the three virtual data-centers each with 6 machines D4 series with 8 Intel(R) Xeon(R) CPU E5-2660 0 @ 2.20GHz cores, 10Gbps NICs, 28 GB memory and run Ubuntu 15.0. Each virtual data-center is located in a different geo-location. The virtual data-centers were simulated by Mesosphere cluster with three jobtracker(Mesos masters) and three tasktracker(Mesos slaves). We installed Hadoop Cloudera CDH 4.2.1-MR1 on the takstrackers. The Coordinator Database run on a separate resource pool with 1 machine D series similar to the virtual data-centers specification. Finally, the loaders run on a 3 D2 series machines with 2 cores Intel(R) Xeon(R) CPU E5-2660 0 @ 2.20GHz cores, 10Gbps NICs, 7 GB memory and run Ubuntu 15.0. Apache Hadoop ships with a pre-built sample app, the ubiquitous WordCount example. The input data file was created using */dev/urandom* on the takstrackers hosts[14]. The input data file was copied to the HDFS directory that was created as part of the Hadoop preparation (/user/foo/data).

Anticipated Required Deviations from the original job plan. The original plan was to simulate the scenario where the customer keeps his data at a different location than it might be processed. We plan on using a job migration scheme that was originally designed for workload migration across different geo-location[20]. The suggested method optimizes the bandwidth costs of transferring application state and data over the wide-area network. Our experiment generated the data file at the loader host and did not include the job migration. We believe that including the job migration aspect could impact the presented results. However, the job migrations proven efficiency and later studies minimize that deviation.

3.3 Baseline and Variability of Workloads

Baseline. The execution baseline included a load that runs without the coordinator component i.e. loaders generated load to the virtual data-centers resources for 48 hours. The load scenario included a single file generation that was submitted to one of the tasktrackers. The output of each executed jobs included the CPU time spent for each execution. The data collection included the execution log of the command: Hadoop jar loader.jar wordcount /user/foo/data /user/foo/out.

Variability of Workloads. The workload comprises of data files with words that need to be counted using the Hadoop WordCount. The load complexity depends on two factors, the file randomness level, and its size. We rely on the native operating system randomness, and our virtual data-centers and loaders are homogeneous. Therefore, the size is the remaining factor for differentiating workload types. We evaluated the federated-cloud coordinator by generating three load types simultaneously. The three types intend to cover the following cases (1) A match was found between workload and sufficient green energy resources. (2) A match was found, but there was not enough power to complete the job with no SLO deadline violation. (3) Like the former but with SLO deadline violation. The three types will be uniquely distributed across wind and solar based virtual data-centers.

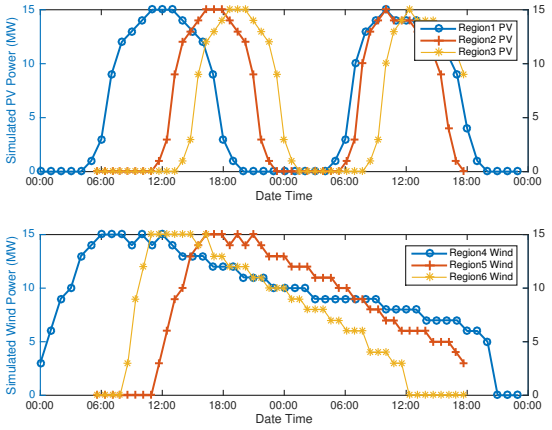


Figure 8: Green Energy availability simulated across six different timezones, three for PV power and the bottom three for wind. These indications are fed, in near-time, to the coordinator database that runs a temporal stored procedure that seeks for match with pending registered jobs.

3.4 Main Execution Issues

In each jobtracker host in a virtual data-center, we executed a simulator that generated green availability traffic to the coordination component. The simulation comprises of availability indication that are based on figure 2 and figure 3. We randomized solar production by using a missing factor of $\alpha = 0.2$ based on collected data between 2008-2011 in Palm Springs CA, Prescott Airport CPV, AZ and Nevada Solar One, NV [22]. Also, we randomized the wind production by a missing factor of $\beta = 0.4$ based on collected data between the years 2007-2012 [22]. The demand simulations included a set of calls to the coordinator component spread across 48 hours triggered by a Monte Carlo simulation inspired by [27]. We built a Rayleigh-based distribution model in the Monte Carlo simulation using Matlab Statistics and Machine Learning Toolbox[15]. The calls originated from three different timezones. Each call comprises of the tuple {region,total-job-workload, load-factor, contract deadlines}. The supply simulators are comprised of

[14]https://github.com/yahavb/GreenCloudCoordination/
[15]https://www.mathworks.com/products/statistics.html

the wind and solar-based energy time and power windows. Supply call to the coordinator includes the tuple {region,cut-in,rated,cut-off,power-efficiency}. The coordinator runs a temporal stored procedure that find a match between the supply and demand data. When a match found, the coordinator generates an assignment call to the jobtracker in the corresponded virtual data-center to execute the jobs. The execution will fetch the data to be processed and report temporal statuses to the coordinator. Every status call generates a check against the current demand levels in the particular region. If the demand changed, and the request cannot be fulfilled, the job considered as false-positive. If the jobs are completed successfully, the request counted as success along with the Core/hour saved.

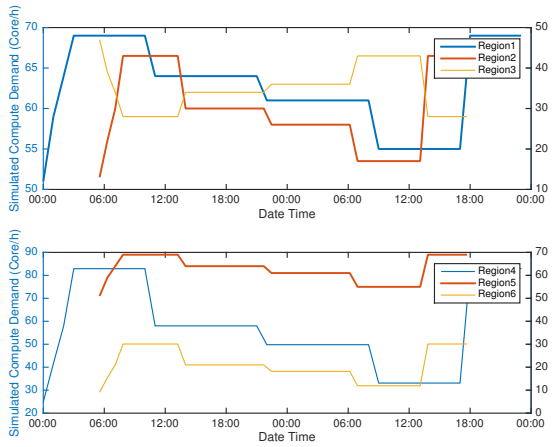


Figure 9: A Monte-Carlo-based simulated compute demand across six different timezones, three for PV power and the bottom three for wind. These indications are fed, in near-time, to the coordinator database that runs a temporal stored procedure that seeks for match with available green energy.

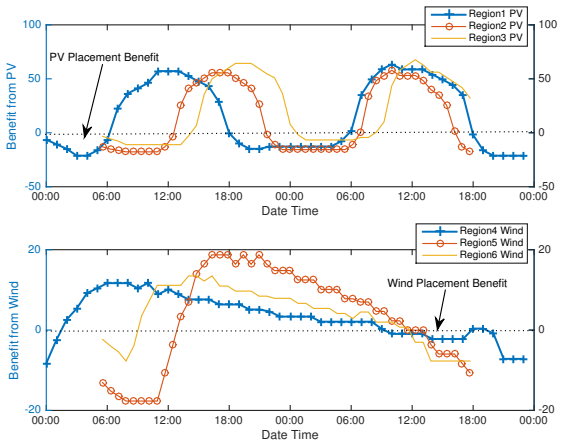


Figure 10: Job placement ranges denoted by the doted line in both the PV and the Wind plots. Any values above the zero levels indicate on potential benefits. However, assigned jobs might not be able to fully processed when unpredicted drop in the availability.

3.5 Analysis

The experiment executions generated three core data logs that we used for the result generation. The first data set is the simulators logs. The simulator logs comprise of the demand records. Demand records stored in the Demand VoltDB table. The supply calls were stored in the Supply VoltDB table. The match transactions stored in the Jobs VoltDB table with status and the number of cores/hour used for the job. Status values can be Success or False-Positive. Figure 9 shows the two supply indications. Both signals were generated based on Equation 2 and Equation 3. The demand data was generated based on known usage patterns that are spread across three timezones.

3.6 Discussion on Energy Mix when Processing Offline Workloads

The experiment simulated the usage of green 708 *kWh* out of total required 3,252 *kWh* for *analytical* systems workload processing i.e. 22% less carbon emission (Equation 5). 1822 *kWh*, 50% of the total workload consumption, was processed by brown energy because of false-positive events i.e. the coordinator assigned a job with no sufficient green recourses to process the job. We believe that optimizing the coordinator algorithm can improve the footprint reduction up to 50%.

Figure 10 shows the job placement ranges denoted by the dotted line in both the PV and the Wind plots. Any values above the zero levels indicate on potential benefits. However, such cases are subject to false positive events that can occur when an unpredicted drop in the availability. Further, such cases utilize the hybrid datacenter power scheme describes in figure 6.

4 Conclusions

The future growth of cloud computing will increase its energy consumption as a fraction of grid power and will cause a significant addition to the ever growing carbon emission since 70% of US power is generated by hydrocarbon fired power plants. Using rapidly emerging green energy for processing cloud computing workloads can limit the anticipated carbon emission growth. However, balancing time varying green energy utilization with time varying energy demands of cloud computing is a complex task that requires sophisticated command and control prediction algorithms beyond the scope of this paper but are emerging in the form of a smart grid system of systems [28]. Our study shows that green energy utilization for on-demand streaming workload is best described as a resource management problem. The solution presented demonstrates real time balance of green resource supply and cloud computing workload demand and utilizes Ubernetes, an open source container cluster manager. The results approximate within 21% those observed in a single cloud instance in the field. Our study also shows that green energy utilization for offline workload processing is a fast data problem. Its solution best utilizes VoltDB, an in-memory database, to allow near-time response for green resources supply and workload demand. Our future work will focus on optimizing the false-positive ratio, to further reducing the cloud computing carbon footprint by up to 50%.

Conflict of Interest The authors declare no conflict of interest.

References

[1] Mell, P.; Grance, T. *Department of Energy* **2011**,

[2] Rao, L.; Liu, X.; Xie, L.; Liu, W. Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment. INFOCOM, 2010 Proceedings IEEE. 2010; pp 1–9.

[3] Shehabi, A.; Smith, S.; Sartor, D.; Brown, R.; Herrlin, M.; Koomey, J.; Masanet, E.; Horner, N.; Azevedo, I.; Lintner, W. *Data Center* **2016**,

[4] Outlook, A. E. *Department of Energy* **2010**, *92010*, 1–15.

[5] Krishnan, S. S.; Sitaraman, R. K. Understanding the effectiveness of video ads: a measurement study. Proceedings of the 2013 conference on Internet measurement conference. 2013; pp 149–162.

[6] Liu, X.; Dobrian, F.; Milner, H.; Jiang, J.; Sekar, V.; Stoica, I.; Zhang, H. A case for a coordinated internet video control plane. Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication. 2012; pp 359–370.

[7] Biran, Y.; Collins, G.; Dubow, J.; Pasricha, S. Coordinating Green Clouds as Data-Intensive Computing. 2016: 3rd Smart Cloud Networks & Systems Conference (SCNS'16). 2016; pp 1–8.

[8] Lim, N.; Majumdar, S.; Ashwood-Smith, P. Engineering resource management middleware for optimizing the performance of clouds processing mapreduce jobs with deadlines. Proceedings of the 5th ACM/SPEC international conference on Performance engineering. 2014; pp 161–172.

[9] Biran, Y.; Collins, G.; Liberatore, J. Coordinating Green Clouds as Data-Intensive Computing. 2016 IEEE Green Technologies Conference (GreenTech). 2016; pp 130–135.

[10] Burns, B.; Grant, B.; Oppenheimer, D.; Brewer, E.; Wilkes, J. *Communications of the ACM* **2016**, *59*, 50–57.

[11] Biran, Y.; Collins, G.; M, B. J.; Dubow, J. *Advances in Science, Technology and Engineering Systems Journal* **2017**, *2*, 553–561.

[12] Nygren, E.; Sitaraman, R. K.; Sun, J. *ACM SIGOPS Operating Systems Review* **2010**, *44*, 2–19.

[13] Maier, G.; Feldmann, A.; Paxson, V.; Allman, M. On dominant characteristics of residential broadband internet traffic. Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. 2009; pp 90–102.

[14] Barroso, L. A.; Clidaras, J.; Hölzle, U. *Synthesis lectures on computer architecture* **2013**, *8*, 1–154.

[15] Abolfazli, S.; Sanaei, Z.; Ahmed, E.; Gani, A.; Buyya, R. *IEEE Communications Surveys & Tutorials* **2014**, *16*, 337–368.

[16] Ferguson, A. D.; Bodik, P.; Kandula, S.; Boutin, E.; Fonseca, R. Jockey: guaranteed job latency in data parallel clusters. Proceedings of the 7th ACM european conference on Computer Systems. 2012; pp 99–112.

[17] Dean, J.; Ghemawat, S. *URL: http://static. googleusercontent. com/media/resear ch. google. com (diunduh pada 2015-05-10)* **2004**, 10–100.

[18] Lam, W.; Liu, L.; Prasad, S.; Rajaraman, A.; Vacheri, Z.; Doan, A. *Proceedings of the VLDB Endowment* **2012**, *5*, 1814–1825.

[19] Liu, Z.; Lin, M.; Wierman, A.; Low, S. H.; Andrew, L. L. *ACM SIGMETRICS Performance Evaluation Review* **2011**, *39*, 62–66.

[20] Buchbinder, N.; Jain, N.; Menache, I. *NETWORKING 2011* **2011**, 172–185.

[21] Deshmukh, M.; Deshmukh, S. *Renewable and Sustainable Energy Reviews* **2008**, *12*, 235–249.

[22] Energy, G. *NREL Subcontract Report* **2010**, 1–13.

[23] Tiwari, A.; Sodha, M. *Solar Energy* **2006**, *80*, 751–759.

[24] Abdelbaky, M.; Diaz-Montes, J.; Parashar, M.; Unuvar, M.; Steinder, M. Docker containers across multiple clouds and data centers. Utility and Cloud Computing (UCC), 2015 IEEE/ACM 8th International Conference on. 2015; pp 368–371.

[25] Beloglazov, A.; Abawajy, J. *Future generation computer systems* **2012**, *28*, 755–768.

[26] Zhang, Q.; Zhu, Q.; Boutaba, R. Dynamic resource allocation for spot markets in cloud computing environments. Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on. 2011; pp 178–185.

[27] Feijoo, A. E.; Cidras, J.; Dornelas, J. G. *IEEE transactions on Energy Conversion* **1999**, *14*, 1582–1588.

[28] Biran, Y.; Collins, G.; Azam, S.; Dubow, J. Federated Cloud computing as System of Systems. Computing, Networking and Communications (ICNC), 2017 International Conference on. 2017; pp 711–718.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 2, 1-7 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

The Internet of Things ecosystem: the blockchain and data protection issues

Nicola Fabiano*

Studio Legale Fabiano, 00179, Roma (Italy)

ARTICLEINFO

Article history:
Received: 12 November, 2017
Accepted: 19 February, 2018
Online: 08 March, 2018

Keywords:
Internet of Things
Legal issues
Data Protection and privacy
Law
Security
Blockchain
Risks
Legal framework
Privacy standard

ABSTRACT

The IoT is innovative and important phenomenon prone to several services and applications such as the blockchain which an emerging phenomenon. We can describe the blockchain as blockchain as a ser- vice because of the opportunity to use several applications based on this technology. We, indeed, should take into account the legal issues related to the data protection and privacy law to avoid breaches of the law. In this context, it is important to consider the new European General Data Protection Regulation (GDPR) that will be in force on 25 May 2018. The contribution describes the main legal issues related to data protection and privacy focusing on the Data Protection by Design approach, according to the GDPR. Furthermore, I resolutely believe that is possible to develop a global privacy standard framework that organisations can use for their data protection activities.

1 Introduction

To define the Internet of Things (IoT) could be a challenge [1] due to its technical and conceptual complexity [2]. The IoT is a phenomenon founded on a network of objects linked by a tag or microchip that send data to a system that receives it. The IoT includes every connection among objects, so we have machine-to-machine (M2M) systems, where each machine talks with other machine(s), communicating real-time data and information. Nowadays we are faced with several devices but mainly such as smartphones and apps, sensors, chip, and any other electronic system. We read [3]

> *The concept was simple but powerful. If all objects in daily life were equipped with identifiers and wireless connectivity, these objects could communicate with each other and be managed by computers.*

In 2012 the Global Standards Initiative on Internet of Things (IoT-GSI) the Internet of Things (IoT) defined the IoT as "the infrastructure of the information society[1]." Not that the IoT phenomenon is realised only when two or more objects are linked to each other in a network such as the Internet. Apart from this kind of connection, an object could also be indirectly linked to a person, thereby setting up a ring network among objects and people. Its very simple, for example, to imagine a ring network that could link a person with one or more objects (a clock, a chair, a lamp, etc.) equipped with a technological system (RFID, near field communication NFC, etc.). However, the IoT is a virtual reality that reproduces exactly what happens in the real world. Lets imagine that our clock, chair, and lamp all contain chips and are used by a person with special needs. From a medical point of view, it may be crucial, for instance, to know how many times he uses the chair. At the same time, it is necessary to help him by automatically turning on the lamp when he sits in the chair. Using chips, it is possible for the objects to communicate among themselves (e.g., the lamp turning on when the chair sends data that the man is sitting down) and at the same time send data over the Internet for, say, medical analysis. The information provided by each object can be aggregated, thereby creating a profile for him. The profile may contain sensitive information about the man, which raises the possibility of

*Corresponding Author Nicola Fabiano, ROME (ITALY), info@studiolegalefabiano.eu

[1] The Internet of Things (IoT) has been defined in Recommendation ITU-T Y.2060 (06/2012) as a global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies [4].

https://dx.doi.org/10.25046/aj030201

his being monitored. This is a very important point for privacy. This scenario could present a lot of legal issues related to privacy and data protection law. The main goal is to evaluate the impact of the IoT phenomenon on the fundamental rights such as the right to respect for private and family life according to the European Convention on human rights [5]. There are others legal aspects to take into account developing a project on IoT. It is quite important to highlight that there are differences between data protection and privacy especially in Europe where the Charter of Fundamental Rights of the European Union identifies as fundamental rights privacy (article 7) and data protection (article 8).

2 The correct approach: privacy is not security

The main focal point to address a correct approach to any evaluation of data protection and/or privacy risks, in general, is to understand the differences between security and privacy. The correct equation is the following one:

$$security \neq privacy \tag{1}$$

where security is different from privacy. In fact, according to this principle, it is possible to adopt very high-security measures, but this can not mean to respect privacy law either protect users privacy. Often this concept is every indication that it is necessary to intervene on the security systems to be compliance with privacy law. Obviously, this is a big misunderstanding and could create confusion on the privacy approach and its consequences. Adopting security measures is certainly a value, but it is not the correct way to deal with privacy issues. To address privacy and data protection correctly, it is necessary to start from the privacy by design (or da- ta protection by design and by default) approach as further and better clarified below. Privacy is embedded into design2. More clearly Privacy, having been embedded into the system before the first element of information being collected, extends throughout the entire life-cycle of the data involved, from start to finish.

2.1 Protecting privacy through the privacy by design approach

The Internet of Things represents a global revolution: the objects that people use in the real world can talk to other objects and at the same time to the data subjects themselves. This scenario goes from the Internet of People (IoP) - because of the connection among people - to the Internet of Everything. This consciousness is the real engine that has pressed politicians and regulators to intervene in the IoT realm. In fact, there is a growing desire to create a general, comprehensive, and structured legal framework for the Internet of Things to protect users and consumers. In October 2010, the 32nd International Conference of Data Protection and Privacy Commissioners adopted a resolution on Privacy by Design (PbD) [6] that is a landmark and represents a turning point for the future of privacy. Instead of relying on compliance with laws and regulations as the solution to privacy threats, PbD takes the approach of embedding privacy into the design of systems from the very beginning.

The primary goal is to draw up two concepts: a) data protection and b) user. Regarding privacy, we have always thought in term of compliance with laws, failing to evaluate the real role of the user (and his or her personal data). To develop an adequate data protection and privacy approach, we must start any process with the user the person who has to be protected putting him or her at the centre. That means that during the design process, the organisation always has to be thinking of how it will protect the users privacy. By making the user the starting point in developing any project (or process), we realise a PbD approach.

This methodological approach is based on the following seven foundational principles [7]:

1. **Proactive not reactive; preventative not remedial;**
2. **Privacy as the default setting;**
3. **Privacy embedded into design;**
4. **Full functionality positive-sum, not zero-sum;**
5. **End-to-end security full lifecycle protection;**
6. **Visibility and transparency keep it open;**
7. **Respect for user privacy keep it user-centric.**

We can see why the Privacy by Design approach is so important in the IoT environment. In fact, the Internet of Things should adopt the PbD principles and statements, always placing the user at the centre. The European Data Protection Supervisor (EDPS) has promoted PbD, touting the concept in its March 2010 Opinion of the European Data Protection Supervisor on Promoting Trust in the Information Society by Fostering Data Protection and Privacy [8] as *a key tool for generating individual trust in ICT*. It was not long after this endorsement that the 32nd International Conference of Data Protection and Privacy Commissioners adopted the PbD concept as well. In Europe, this approach became *"Data Protection by Design and by Default"* (DPbDabD) in the EU Regulation 679/2016 [9] and indeed establishing this concept in the law is a welcome development. Nevertheless, it is kind of interesting to notice that the EU legislator used a different expression (i.e., data protection by design and by default) from the one adopted in the international context (i.e., Privacy by Design). These two expressions represent two different methodological approaches. Privacy by Design is structured in a trilogy of applications (information technology, accountable business practices, physical design) and the seven principles quoted above. The EU formulation is more descriptive and not based on a method; also, the by default

concept is autonomous, whereas the PbD approach embeds the same concept into by design. According to the text of Article 25 of the Regulation 679/2016, it is clear that the EU legislator considers by design and by default as different concepts, even though the words by design comprehend the concept by default, making the latter phrase redundant. The EU formulation is more descriptive and not based on a method; also, the by default concept is autonomous, whereas the PbD approach embeds the same concept into by design. Furthermore, the EU Regulation 679/2016 seems to pay a lot of attention to the technical and security aspects instead of the legal concerns, as seen in highlighting of the term "security". Hence, the Internet of Things should adopt the Privacy by Design principles and statements, always placing the user at the centre.

3 IoT evolution and its applications: a challenge

The IoT phenomenon makes to spring several applications in different sectors (Personal, Home, Vehicles, Enterprise, Industrial Internet) [10]. This is a continuously evolving system, and we see to the development of many application in each sector. In the last few years, it arose the interest (the needing) to guarantee highest security levels both for the Industries and the users.

The fields of Big Data and Blockchain are the leading emerging phenomena in the IoT ecosystem, but people paid attention more to the technical and security issues than the privacy ones.

Certainly, the security aspects are relevant to avoid or reduce the risks for data privacy. However, from a privacy point of view, we cannot dismiss the right approach, according to the PbD principles. In the first phase of analysis, any project has to be evaluated also thinking how to protect privacy data and personal information applying the PbD principles. In concrete, after the evaluation process, the project has to comply with the law and not after starting it. Once the project starts, it does not need any process of compliance with the law because, according to the PbD principles, the same project has to be already in compliance with the privacy law before starting it. In this case, (during the life cycle of the project) it is not required any evaluation of compliance with the law. In fact, any assessment it is necessary during the design phase of the project, just for the nature of the approach "by design", applying the PbD principles correctly.

Several IoT applications have been developed in the field named "smart", such as smart grid, smart city, smart home, smart car, etc. This indeed represents what is the IoT evolution that it will continue to grow and develop creating a lot of fields of action. In the "smart context," we cannot dismiss from the privacy and security risks related to the communication among objects especially in the case of processing personal data.

The main questions are:

- "Where are the users' personal data stored?"
- "Who manage the users' personal data?"
- "What kind of security measures has been adopted?"
- "Can it be considered a smart system compliance with the privacy law?"

The answers depend on the design model used during the developing preliminary phase. In fact, the "design" is a fundamental topic as we illustrate in the following considerations.

However, in the IoT echo-system are emerging two relevant and complex aspects in part closely related between them: **Big Data** and **Blockchain**. In the last few years, these have been items of interest in the IoT phenomenon, intensifying the interest by whom deals with it, especially because of the implications both from the side of the developers and from the users. Hence, Big Data and Blockchain represent the new challenge and the new applications of the IoT phenomenon.

This scenario entails the need to deepen these aspects, especially regarding the privacy and data protection issues.

3.1 Big Data: privacy issues and risks in the Internet of Things

Despite its many potential benefits, the Internet of Things poses important privacy and security risks because of the technologies involved.

According to the Gartner Newsroom [11], 6.4 Billion Connected Things will be in use in 2016, up 30 percent from 2015 and the device online are estimated to reach 20.8 billion by 2020. This represents a scenario to be monitored not only for the big data phenomenon but also for threats and risks to privacy and security.

A recent study on the threats to our privacy, security and safety, under the "Cyberhygiene" project [12], carried out the report (not yet published) named *"Understanding end-user cyber hygiene in the context of the Internet of Things: A Delphi-study with experts"*. This report, in the beginning, says that *"This study aimed to establish expert consensus concerning the 1) key malicious IoT threats, 2) key protective behaviours for users to safeguard themselves in IoT environments, and 3) key risky user behaviours that may undermine cyberhygiene in IoT environments"*[2]. In conclusion, this report says *"There was consensus on the need to consider behaviours across IoT lifecycles. By considering behaviour across each lifecycle, we have been able to identify key behaviours that users need to adopt when using IoT devices. Furthermore, we have been able to identify key threats that can, for example, put users sensitive information at risk and risky behaviours that may lead users to be at risk of a successful attack"*.

[2] See also "Review of Cyber Hygiene practices" - ENISA - https://www.enisa.europa.eu/publications/cyber-hygiene

No doubt, therefore, that even in the IoT ecosystem there are important risks and threats to privacy and it should take appropriate precautions.

On the one hand, we can control devices such as vending machines and stereo speakers with our smartphones, manage devices in our homes (domotics for energy saving, security, comfort, communication) by remote control, and use smartphone apps to book reservations or purchase services. Larger-scale IoT applications might include public security systems or warehouse inventory control systems. It is evident the acceleration of the technological evolution in the last few years and the IoT phenomenon it is not exempt[3]. IoT considers the pervasive presence in the environment of a variety of things, which through wireless and wired connections and unique addressing schemes can interact with each other and cooperate with other things to create new applications/services and reach common goals. In the last few years IoT has evolved from being simply a concept built around communication protocols and devices to a multidisciplinary domain. Devices, Internet technology, and people (via data and semantics) converge to create a complete ecosystem for business innovation, reusability, interoperability, that includes solving the security, privacy and trust implications.

On the other hand, we have seen the fast and exponential data growth, data traffic and, hence, another paradigm well-known as Big Data[4]. Big data implies data analysis and data mining procedures but working on big data values[5]. Nowadays it is very simple to develop apps that, by accessing to data, can execute data mining activities with every possible consequence. In this context, the primary goal is to protect data because of their highest value. Among the main risks we can indeed present the following:

- ***Identification of Personal Information***

The IoT system allows you to transfer data on the Internet, including personal data. Personal information may be transmitted only when the object in which the microchip installed is linked to a person. This connection may be direct or indirect.

We could have a direct link when the user is aware of the possible transmission of his or her personal data and gives consent. Alternatively, let us suppose that a person buys something. Alternatively, the connection may be indirect when the object is not linked directly to a person but only indirectly through the use of information that belongs to that person. For example, if we have x objects linked together by the Internet, I might know information about object nr. 1, but I cannot know to whom this information belongs. I can know, however, that objects nr. 2, 3, and so on are connected among themselves and to a person named Jane. In this way, it is possible to link every piece of information provided by the objects (2, 3, etc.) to Jane. Furthermore, if I know that it is possible to link object nr. 1 to the others (2, 3, etc.), I can also indirectly know that the information provided by object nr. 1 likewise belongs to Jane.

- ***Profiling***

There are several risks and threats in the Internet of Things, but the main one is probably profiling [13], [14]. If objects are linked to a person, it will be possible to obtain personal information about that person through the information transmitted over the Internet by each of those objects. Furthermore, these transmitted data may be stored in one or more servers. When a person can be identified through the use of credit or loyalty cards, its very simple to know the types of products purchased and so on to profile the person, learning about his or her habits and lifestyle. The person may have previously provided consent for the dissemination of data related to his or her purchases for advertising purposes. Regarding privacy, is it possible to protect a person? Who manages the personal data? Where will this data be stored? Profiling can also be an issue with the movement toward smart grids and cities, a phenomenon that is close in nature to the Internet of Things. For some years now, there has been an interest in modernising the existing electrical grid by introducing smart meters, which can communicate a consumers energy consumption data to the relevant utilities for monitoring and billing purposes. From a legal perspective, there is the need to consider the privacy issues arising from these initiatives, such as consumer profiling, data loss, data breach, and lack of consent (consent is mandatory by law).

- ***Geolocation***

Geolocation is another risk because nowadays, by our device (first of all the smartphones) it is very simple to find precise details on the location, for instance, digital photos. Inside each picture file there are some fields among them EXIF and GPS that contain the technical information about the photo and also the location where the picture was taken. If the user has not previously deactivated the geolocation service in the camera or smartphone, and the pictures have been published on a website or social network, anyone who views the photo can know exactly where the picture was taken and see who was there.

In this way, privacy could be compromised. When smartphones and other mobile devices are connected

[3]IoT is a concept and a paradigm with different visions, and multidisciplinary activities

[4]Big data is a term for datasets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualisation, querying, updating and information privacy. The term "big data" often refers only to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. *"There is little doubt that the quantities of data now available are indeed large, but thats not the most relevant characteristic of this new data ecosystem."* (Wikipedia)

[5]Is well-known the Four Vs of Big Data: Volume, Velocity, Variety and Veracity (IBM). Considering data as value it is possible extend the approach to 5 V (last V as value).

to the Internet, as they typically are, they contribute every day to the Internet of things, sending data ready to be used by other people.

- ***Liability for Data Breaches***

In Europe, there are numerous national and European Community (EC) laws relating to personal data breaches. Hence, the Internet of Things also has effects on liability in cases where the data being collected and transmitted lacks the appropriate security measures. For example, Directive 2002/58/EC [15] states that: *In case of a particular risk of a breach of the security of the network, the provider of a publicly available electronic communications service must inform the subscribers concerning such risk and, where the risk lies outside the scope of the measures to be taken by the service provider, of any possible remedies, including an indication of the likely costs involved.*

Another risk is the loss of data during processing. The consequences entail, of course, liability for the data controller and data processor related to each particular situation. In fact, because the processing of personal data involves risks to the data in question (such as the loss of it), the EU Regulation n. 679/2016 on data protection contains an article requiring data controllers to conduct a data protection impact assessment (DPIA)[6] an evaluation of data processing operations that pose particular risks to data subjects.

According to the Article 35, paragraph 1, of the EU Regulation n. 976/2016 (GDPR) *"Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data. A single assessment may address a set of similar processing operations that present similar high risks"*.
This is the law prescription on the need to conduct a data protection impact assessment (DPIA) in some cases. This preventive action could avoid or reduce risks for the fundamental rights such as data protection and privacy. This demonstrates as is crucial to pay attention to security and privacy in any project development.

3.2 Blockchain: what about privacy?

The blockchain *"is a shared, immutable ledger for recording the history of transactions"* [16]; it is a ledger of records. The blockchain was imagined by Satoshi Nakamoto [17]. Blockchain works as a distributed database, and its structure guarantees any modification or alteration due to the strong link and timestamp among each block.

The blockchain was theorised thinking to the web of trust principles and, hence, based on the consensus according to the proof of work concept related to the computational power made available by each one. Subsequently, the participation in the blockchain processes has been revised in terms of proof of stake according to the amount of cryptocurrency hold of each one. Given that, nowadays we can have different kind of participation in the blockchain consensus (proof of work - proof od stake), and according to the solution adopted, it will determined the blockchain governance. Furthermore, it is possible to distinguish three different kinds of blockchain:

1. Public blockchain;
2. Consortium blockchain;
3. Private blockchain [18]

The main differences are essentially related to the permission to write the blocks and participate in the consensus processes. The blockchain can be used to provide services, such as the guarantee about the personal identity. It is possible to implement an application for the electronic identification (eID) through which to store, securely in each block of the block- chain, the personal information that could be used to know with certainty the identity of a person. The blockchain can be also used for public services by the Public Bodies or to storage securely and without risks of modification electronic documents or any other resource.

The blockchain poses problems related to data protection and privacy, considering them as two different concepts (but they are rights). It is a field well-known by the engineers and developers, but often ad- dressed only increasing the security measures instead of considering the data protection and privacy law. Regarding privacy, Satoshi Nakamoto [17] argues that *"privacy can still be maintained by breaking the flow of information in another place: by keeping public keys anonymous"*. However, the author says also that *"The risk is that if the owner of a key is revealed, linking could reveal other transactions that belonged to the same owner"*. That represents a significant chink in the privacy perspective. Ensuring privacy and data protection is one of the main aims of any project which has to address "by design", not leaving any possibility to compromise personal data and/or personal information. Axon [19] argues that privacy issues can be dealt with "privacy-awareness" enabling *"two levels of anonymity: total anonymity, and anonymity to the neighbour group level"*. However, "privacy-awareness" do not seem a valid solution because this way it is not enough to be compliance with the EU GDPR, according to the Article 25 (Data protection by design and by default).

In another technical contribution [20] you read *"Maintaining privacy on the blockchain is a complicated issue"*. The authors propose *"A couple of ways to mitigate but not completely eliminate this issue, if privacy is important for the considered application"*. Privacy is certainly important on the blockchain, and for this reason, it would be better to address the issue finding a "legal" solution to be compliance with the law.

[6] In the rest of the world this is well-known as Privacy Impact Assessment (PIA).

Other authors [21] say *"Despite the benets provided by these services, critical privacy issues may arise. That is because the connected devices (the things) spread sensitive personal data and reveal behaviours and preferences of their owners. Peoples privacy is particularly at risk when such sensitive data are managed by centralised companies, which can make an illegitimate use of them . . . "*. It is very appreciable these authors' approach [21] because they propose a technical solution presenting it in terms of "private-by-design IoT" [7]. Despite the fact that this proposed solution highlights the concept "by design", from a legal point of view it does not seem to take on the issue related to the obligation required by the EU GDPR.

This short scenario shows how on the blockchain there are certainly privacy issues addressed only providing technical solutions, without any legal reference. Apart from the high technical solution, hence, we cannot dismiss the law obligations, where they are applicable, like in Europe, according to the GDPR mentioned above. This panorama confirms the equation according to security is different from privacy; a system could be very secure but not compliance with the privacy law. On the contrary, a system could be compliance with the privacy law and, hence, very secure (obviously if it has been adopted the security measures).

This is an obligation for the controller. Giving the structure of the blockchain, it seems that any subject or person or owner (as defined by Nakamoto) should be a controller and consequently bound to respect the EU GDPR. From this scenario arise many consequences for the "owner" regarding law obligations.

In fact, according to the GDPR, it is mandatory to "implement appropriate technical and organisational measures" [8]

In certain cases, if there are high risks to the rights and freedom of natural person, it is mandatory to provide a Data Protection Impact Assessment (DPIA) to carry out, prior to the processing, *an assessment of the impact of the envisaged processing operations on the protection of personal data* (Article 35). Another aspect is related to the consent according to the GDPR. In fact, according to the Article 6, paragraph 1.a, *the data subject has given consent to the processing of his or her personal data for one or more specific purposes*. Moreover, according to the Article 7, paragraph 3, *the data subject shall have the right to withdraw his or her consent at any time*. Against this background, it is clear that the technical infrastructure of the blockchain prevents the concrete withdrawal of the consent by the data subject because this solution implies the node deletion. The consent issue represents undoubtedly another legal point related to the blockchain and the data protection law.

According to the technical structure of the blockchain, all the system has not any controller because of the lack of a central controller (a general "supervisor") who is responsible for all the nodes. Each owner, hence, is a controller for the data processing of his node. In this perspective each owner, apart from the general securities profiles of the blockchain, has to respect the law and he is himself is a data processing controller. Due to the blockchain technical configuration, in the event of a node was compromised, it is possible to amount to a controller's liability and, in this case, there are certainly other consequences for the owner's node.

4 Conclusion

The Internet of Things involves all stakeholders from companies to consumers. Focusing on the user (consumer) is particularly important to guarantee a level of confidentiality that will earn the users trust. This solution is made possible by adopting the maximum level of security through the Privacy by Design (PbD or DPbDabD) approach and performing PIAs to evaluate the privacy risks of data collection and processing.

The industries may be wary of efforts to regulate the Internet of Things, as it regards the IoT phenomenon as a source of enormous business opportunities. For example, changes in lifestyle such as the use of more technological services like domotics applications can certainly increase the consumers quality of life (and industrys profits). It will be up to consumers, regulators, and privacy professionals to convince the business sector that understanding the risks related to the IoT will produce the same business opportunities to protect privacy and increase the quality of life.

As I hope I have shown, it is crucial to set up a privacy standard to facilitate a methodological approach to privacy and data protection. With the Internet of Things reaching ever more deeply into peoples lives, it would be beneficial to have an international privacy standard for processing personal data in the same way throughout the world using the forward-looking PbD (or DPbDabD) approach.

From a legal point of view, the main difficulty in setting up and using a privacy standard relates to existing laws, which are different in each nation (and even in different states and provinces within those nations). It is possible to develop a standard privacy framework that organisations can use for their data protection activities, adapting it to national legislation while keeping the central framework for all nation-

[7] You read "With the purpose of preventing this situation, the goal of our research is to encourage a decentralized and private-by-design IoT, where privacy is guaranteed by the technical design of the systems. We believe that this can be achieved by adopting Peer-to-Peer (P2P) systems."

[8] Article 25, par. 1, says *"Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects"*.

states.

Since the Privacy by Design (or DPbDabD) approach is the foundational methodological approach to privacy protection, the privacy standard should be adopted according to PbD principles and statements. At the moment we have no record of international privacy standard model.

A Privacy Management System (PMS) could be a reference model or a software system working on the PbD principles. To develop a PMS confers a benefit to all the stakeholders because in this way it is possible to automate every process guaranteeing a good data protection level, by reducing the privacy and security risks. Furthermore, it is feasible to use the Artificial Intelligence and Machine Learning principles to develop a software based on a PMS to facilitate professionals, public body, Industries and Organizations in their activities.

References

[1] Hahn Jim: The Internet of Things (IoT) and Libraries - Library Technology Reports; Chicago 53.1 (Jan 2017): 5-8,2.

[2] AA.VV.: River Publishers, Digitising the Industry Internet of Things Connecting the Physical, Digital and Virtual Worlds, 2016

[3] Cisco.com. San Francisco, California: Lopez Research, An Introduction to the Internet of Things (IoT), November 2013. Retrieved 23 October 2016 http://www.cisco.com/c/dam/en_us/solutions/trends/iot/introduction_to_IoT_november.pdf

[4] ITU - Global Standards Initiative on Internet of Things (IoT-GSI): The Internet of Things (IoT) - http://www.itu.int/en/ITU-T/gsi/iot/Pages/default.aspx

[5] European Convention on human rights - http://www.echr.coe.int/Documents/Convention_ENG.pdf

[6] Resolution on Privacy by Design. 32nd International Conference of Data Protection and Privacy Commissioners, Jerusalem - https://edps.europa.eu/sites/edp/files/publication/10-10-27_jerusalem_resolutionon_privacybydesign_en.pdf

[7] 7 Foundational Principles. "Privacy by Design" - https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

[8] EDPS: Opinion of the European Data Protection Supervisor on Promoting Trust in the Information Society by Fostering Data Protection and Privacy. European Data Protection Supervisor (EDPS) - https://edps.europa.eu/sites/edp/files/publication/10-03-19_trust_information_society_en.pdf

[9] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) - http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

[10] What's The Big Data? - Internet of Things Market Landscape - https://whatsthebigdata.com/2016/08/03/internet-of-things-market-landscape/

[11] Gartner: Gartner Says 6.4 Billion Connected "Things" Will Be in Use in 2016, Up 30 Percent From 2015 - http://www.gartner.com/newsroom/id/3165317

[12] Cyberhigiene project - https://www.petrashub.org/portfolio-item/cyberhygiene/

[13] Ann Cavoukian: Springer, Identity in the Information Society. Identity in the Information Society, 2010

[14] Mireille Hildebrandt: FIDIS. Behavioural Biometric Pro ling and Transparency Enhancing Tools, 2009

[15] Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector - http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32002L0058&from=en

[16] IBM, Understand the fundamentals of IBM Blockchain - https://www.ibm.com/blockchain/what-is-blockchain.html

[17] Satoshi Nakamoto, Bitcoin: A peer-to-peer electronic cash system - https://bitcoin.org/bitcoin.pdf

[18] Vitalik Buterin, On Public and Private Blockchains. Retrieved from https://blog.ethereum.org/2015/08/07/on-public-and-private-blockchains/

[19] Louise Axon, University of Oxford - Privacy-awareness in Blockchain-based PKI (2015) - https://ora.ox.ac.uk/objects/uuid:f8377b69-599b-4cae-8df0-f0cded53e63b

[20] Konstantinos Christidis and Michael Devetsikiotis, Blockchains and Smart Contracts for the Internet of Things - http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7467408

[21] Conoscenti Marco, Vetr Antonio; De Martin Juan Carlos - Peer to Peer for Privacy and Decentralization in the Internet of Things - In: 39th International Conference on Software Engineering, Buenos Aires (AR), May 20-28, 2017. pp. 1-3 - http://porto.polito.it/2665723/1/peer_to_peer_for_privacy_and_decentralization_in_the_internet_of_things.pdf

[22] Ahmed Kosba, Andrew Miller, Elaine Shi, Zikai Wen, Charalampos Papamanthou - Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts (2016) - https://eprint.iacr.org/2015/675.pdf

[23] Castelluccia, Claude in: Privacy, Accountability and Trust - Challenges and Opportunities - https://www.enisa.europa.eu/publications/pat-study

[24] Ann Cavoukian, Jules Polonetsky, Christopher Wolf: Identity in the Information Society, Smart Privacy for the Smart Grid: Embedding Privacy into the Design of Electricity Conservation.

[25] Guy Zyskind, Oz Nathan, Alex Sandy Pentland - Enigma: Decentralized Computation Platform with Guaranteed Privacy (2015) - https://arxiv.org/pdf/1506.03471.pdf

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 2, 8-20 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Cancer Mediating Genes Recognition using Multilayer Perceptron Model- An Application on Human Leukemia

Sougata Sheet[*,1], Anupam Ghosh[2], Sudhindu Bikash Mandal[1]

[1]*A. K. Choudhury School of Information Technology, University of Calcutta, Kolkata, 700098, India*

[2]*Department of Computer Science & Engineering, Netaji Subhash Engineering College, Kolkata, 700152, India*

ARTICLEINFO

Article history:
Received: 14 November, 2017
Accepted: 22 January, 2018
Online: 08 March, 2018

Keywords:
Artificial Neural Network
Multilayer perceptrons
p-value
t-test

ABSTRACT

In the present article, we develop multilayer perceptron model for identification of some possible genes mediating different leukemia. The procedure involves grouping of gene based correlation coefficient and finally select of some possible genes. The procedure has been successfully applied three human leukemia gene expression data sets. The superiority of the procedure has been demonstrated seven existing gene selection methods like Support Vector Machine (SVM), Signal-to-Noise Ratio (SNR), Significance Analysis of Microarray (SAM), Bayesian Regularization (BR), Neighborhood Analysis (NA), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) is demonstrated, in terms of the affluence of each Go attribute of the important genes based on p-value statistics. The result are properly validated by before analysis, t-test, gene expression profile plots. The proposed procedure has been capable to select genes that are more biologically significant for mediating of leukemia then those obtained by existing methods.

1 Introduction

In the body, cancer is the uncontrolled enhancement of unusual cell. When the bodys normal control procedure stop the working, then cancer are build-up [1]. Old cells do not expire and in its place develop out of control, establishing fresh, abnormal cells. These additional cells may create a mass of tissue, called a tumor [2]. Leukemia is a cancer which starts in blood forming tissue, usually the bone marrow [3]. It leads to the over production of abnormal white blood cells [4]. However, the abnormal cells in leukemia do not function in the same way as normal white blood cells [5]. There are several types of leukemia, based upon how speedily the disease is growing up and the type of abnormal cells are generate. Peoples are may be affected at all stage by cancer [6]. In 2015, 54,270 people are prospective to be diagnosed with leukemia. The overall five-year correlative anointing rate for leukemia has more than four fold since 1960. From 1960 to 1965, the five-year correlative anointing rate among whites (only data available) with leukemia was 14%. From 1975 to 1980, the five-year correlative anointing rate for the total population with leukemia was 34.20% [7], and from 2004 to 2010, the overall correlative anointing rate was 60.30% [8]. From 2005-2010, the five-year relative correlative anointing overall were Chronic Myeloid Leukemia (CML) is 59.90% [9], Chronic Lymphocytic Leukemia (CLL) is 83.50%, Acute Myeloid Leukemia (AML) is 25.40% overall and 66.30% for children and adolescents younger than 15 years and Acute Lymphocytic Leukemia (ALL) is 70% overall [10], 91.80% for children and adolescents younger than 15 years, and 93% for children younger than 5 years. In 2015, 24,450 people are expected to die from leukemia with 14,040 males and 10,050 females [11]. In Us 2007-2011, leukemia was the fifth most common cause of cancer deaths in men and the sixth most common in women.

In the field of microarray data analysis, one of the most challenging remittance is gene selection [12]. In gene expression data normally contains a large number of variables genes compared to the numbers of samples [13]. The conventional data mining methodology cannot be directly used to the data due to this identity problem [14]. In this reason analysis of gene expression data used dimension reduction procedure [15]. The gene selection which deducts the genes

[*]Corresponding Author: Sougata Sheet & sougata.sheet@gmail.com

https://dx.doi.org/10.25046/aj030202

extremely related to the pattern of every types disease in order to escape such problems [16]. Parametric and non-parametric tests are the statistical approach [17]. For example *t*-test and Wilcoxon rank sum test have been thoroughly used for searching differentially revealed genes since they are instinctive to understand and implement [18]. But they have a restriction to propagate, if more than two classes and require time swallowing coordination to solve the problem of multiple testing [19]. For three or more groups, the Kruskal-Wallis test can be used. But it may be created biased result because of reliance on the number of samples, when it is used to microarray data whose sample size are normally unbalanced [20]. Many diseases, they are reason by the problems such as chromosomal disequilibrium and gene mutations, which give away abnormal gene expression patterns. These pattern get the information about underlying genetic process and states of several types disease. If these patterns can be analyzed appropriately, they can be effective for recognize the disease sample and identifying the extent to which a patient is affliction from the disease and which can be help in the supervision of disease.

The microarray gene expression data have been collected to underlying biological process of a number of diseases [21]. It is very essential to narrow down from thousands of genes to a few disease genes and gene ranking [18]. In microarry data analysis genes selection is most important phase. For classification of data analysis, several forms of technique have been proposed for gene ranking [22]. These are classified into three several types: filter, wrapper and embedded process [23]. Each of these categories has its personal advantages and disadvantages. For example, filter process are computationally useful and simple but minor performance than the other process. On the other hand, wrapper and embedded procedure are comparatively much complicated and computationally costly but it usually gives excellent classification performance as they mainly apply classifier characteristics in gene ranking. Filter procedure include T-score, which is *t*-statistic standardized interrelation between input and output class labels [15]. On the other hand, wrapper and embedded procedure include Support Vector Machine (SVM) and its variants, random forest-RFE, elastic net, guided regularized random forest [24], balanced iterative random forest [25] etc . Main distinction of filter process and wrapper or embedded procedure is how they behave samples when ranking genes. For example, in filter procedure, all the samples are usually used for gene ranking but the quality and relevance data samples are ignore [26]. On the other hand wrapper or embedded procedure, classifier such as boosting algorithm, logistic regression, Support Vector Machine (SVM) etc., are used to gene ranking [27].

For complicated data analysis, we can used an Artificial Neural Network (ANN) model [28]. An ANN was externally used for solve the problem such as diagnosis of different types of cancer such as speech recognition, breast cancer [28] [29]and cervical cancer. Several types of effective researches on blood cells using neural network model has been committed [30]. Ongun *et al* enhancement a completely automated classification of bone marrow and blood using several types of way such as neural network and support vector machine (SVM). The best performance of SVM with accuracy of 91.05% as parallelism to Multilayer Perceptron (MLP) network using Conjugate Gradient Descent, Linear Vector Quantization (LVQ), and k-Nearest Neighbors (KNN) classifier which generate accuracy of 89.74%, 83.33% and 80.76% respectively.

In this paper, we proposed a method based on neural network models for identify a set of possible genes mediating the development of cancerous growth in cell. We said this model is Multilayer Perceptron Model-1 (MLP-1) and Multilayer Perceptron Model-2 (MLP-2) [31]. At first we form group of genes using correlation coefficient, then we select the most important group. Finally, we present a set of possible genes get by this method, which may be responsible for cancerous growth in human cell. In this article, we consider three human leukemia gene expression data sets. The usefulness of the procedure, along with its excellent result over several others procedure, has been demonstrated three cancer related human leukemia gene expression data sets. The results have been compared seven existing gene selection methods like Support Vector Machine (SVM), Signal-to-Noise Ratio (SNR), Significance Analysis of Microarray (SAM), Bayesian Regularization (BR), Neighborhood Analysis (NA), Gaussian Mixture Model (GMM), and Hidden Markov Model (HMM). The results are appropriately validated by some previous investigations and gene expression profiles, and compared using *t*-test, *p*-value, and number of enriched attributes. Moreover, the proposed procedure has get more number of true positive genes than the existing ones in identifying responsible genes.

2 Related Work

In this article, we have proposed procedure based neural network models for identification of cancer mediating genes. On existing gene selection methods, we have made a survey for comparative analysis [32]. Among them, we have select some existing gene selection methods like SVM, SNR, SAM, BR, NA, GMM, HMM.

SVM is a one type of machine learning procedure which is differ two classes by maximizing the margin between them [33]. For cancer classification, support vector machines (SVMs) is used to identify important genes [34]. The Lasso (L1) SVM and standard SVM are often considered using quadratic and linear programming procedure. A recurrent algorithm is used to solve the Smoothly Clipped Absolute Deviation (SCAD) SVM efficiently. Almost all the cases, it is noticed that with smaller standard errors the SCAD-SVM selects a smaller and a more stable number of

genes than the L1-SVM. Another algorithm of gene selection using the weight magnitude as ranking criterion is Recursive Feature Elimination (RFE) SVM [35]. The SVM-RFE procedure ranks all the genes according to some scoring operation, and remove one or more genes with the lowest score values [36]. When the maximal classification accuracy is achieved, then the procedure will be stop [37].

Signal-to-noise ratio (SNR) is applied to rank the correlative genes according to their discriminative power. The procedure starts with the evaluation of a single gene, and frequently finding for other genes based on some statistical criteria [38]. The high SNR genes scores are select as the significant ones. Measurement of SNR score are affected by the size of variables [39]. When there are more than variables, the average and disunity of the other variables of another classes are dependent on the number of variables and data dispersion, which effect SNR ranking of the important variables due to the enhancement in noise in the data. The procedure is more efficient of finding and ranking a smaller number of important variables, when the number of variables can be decrease significantly.

Significance analysis of microarray (SAM) is a one type of gene selection procedure which use a set of gene specific t-test and identify genes with statistical significant changing in expression values [40]. The basis of change in the gene expression values, every gene is assigned a score value. If the gene score value the greater than a threshold value which indicate potentially significant [41]. False Discovery Rate (FDR) is the percentage of such genes identified by chance [42]. In order to calculate FDR, insignificance genes are identified by analyzing layout of the measurements. Identify the smaller or larger sets of genes can be adjusted by using threshold value, and FDRs are computed for every set. The main problem is that, in permutation step where entire gene group are put in one group for evaluation. This needs an expensive computation and it likely confuses the analysis because of the noise in gene expression data.

A simple Bayesian procedure has been used to remove the regularization parameter [43]. The value of a regularization parameter is determined by degree of sparsely, which is get an optimal result. Normally this include a model selection step, and calculate the minimization of cross validation error based on intensive search.

Neighborhood analysis (NA) is a procedure for clustering multivariate data analysis based on a given distance metric over the data. Functionally, purpose of NA and K-nearest neighbor algorithm are same [44]. Any significant correlation cannot detect and this is the main disadvantage. This problem may be due to the few number of genes and also likely that the phenotype is too complicated to be connected with a cluster of genes, and a more extend relationship may present in gene expression.

A Gaussian mixture model (GMM) is represented as a weighted sum of gaussian component densities which is based on a parametric probability density function [45]. For parameter selection GMM has been used. We have designed GMM on microarray gene expression data for gene selection. On the other hand, for genes identification, we have designed Hidden Markov Model (HMM) on microarray gene expression data sets. Normally HMMs provide an intuitional framework for representing genes with their several functional properties, and proficient algorithms can be creating to use these models to identify genes.

3 Methodology

Let us assume a set $G = (g_1, g_2, ..., g_i)$ of i genes are known which is hold the normal samples for first m expression values and diseased samples for subsequent n expression values. Now correlation coefficient of gene based normal samples are calculated. Therefore, correlation coefficient R_{qr} within q^{th} and r^{th} is given by [46] [47]

$$R_{qr} = \frac{\sum_{l=1}^{m}(g_{ql} - y_q) * (g_{rl} - y_r)}{\sqrt{(\sum_{l=1}^{m}(g_{ql} - y_q)^2) * \sqrt{(\sum_{l=1}^{m}(g_{rl} - y_r)^2)}}} \quad (1)$$

Here y_q and y_r are the mean of expression values of q^{th} and r^{th}, gene respectively in normal samples.

Similarly, for diseased samples the iteration coefficient R_{qr} between q^{th} and r^{th} genes is given by

$$R^{'}_{qr} = \frac{\sum_{l'=1}^{n}(g^{'}_{ql'} - y^{'}_{q}) * (g^{'}_{rl'} - y^{'}_{r})}{\sqrt{(\sum_{l'=1}^{n}(g^{'}_{ql'} - y^{'}_{q})^2) * \sqrt{(\sum_{l'=1}^{n}(g^{'}_{rl'} - y^{'}_{r})^2)}}} \quad (2)$$

Using equation 1 and 2 each pair of genes is computed. The genes are located in the similar group if $R_{qr} \geq 0.5$. Now we have used interrelation coefficient to narrow down hearted the invention space by searching genes of a comparable behavior in terms of related expression patterns. The set of responsible genes mediating certain cancers are recognized in this procedure. The choice of 0.50 as a threshold value has been done through extensive experimentation for which the distances among the cluster center have become maximize [48]. The main set of genes is obtained in this pathway [49].

An extremely simplified model of biological structures is an Artificial Neural Networks that imitative the conduct of the human brain. A huge number of interrelated processing components of the layered structure is composed and intended to imitative biological neurons. MLP model is the one of the most popular ANN types with back-propagation algorithm. Figure 1, show the architecture of MLP model containing of interconnected neurons of three layers [50]. This three layers are input layer, hidden layer, and output layer [51]. This model is represented as $n \times p \times q$. Where input layer consist of n number of neurons, hidden layer consist of p number of neurons and output layer consist of q number of neurons. In the higher layer every neuron in every layer is completely attached to all neurons and every link has a weight connected with it. In interconnected neurons, these

weights are define the nature and strength of the influence. The output signals from one layer are conducted to the consequent layer by using links that amplify the signals based on the interconnected weights [52]. Exception of the input layer neurons, the total input of every neuron is the sum of weighted outputs of the previous layer neurons. In hidden and output layers neurons, we can calculate the output by using sigmoid logistic function such as an activation function.

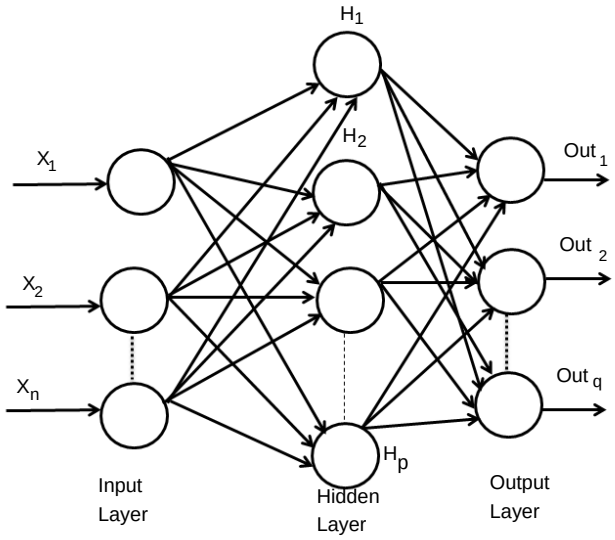


Figure 1: Architecture of an MLP network

Hidden layer neurons execute a non-linear alteration that qualifies the MLP model to simulate a more difficult and non-linear structure within the constraints of a three layer MLP model and more than one hidden layer are used. By employing an incremental adaptation approach, the MLP model has generalized curve fitting capability. Input patterns and output patterns was carried out by MLP model and specified proportion is randomly selected. A learning procedure to correct the linking weights recurrently and to reduce the system error mechanism by every forward processing of the input signal by using back-propagation algorithm. In the initial stage of the learning procedure, to create a input pattern from the input layer to the output layer. The error of every output neuron was calculated from the difference between the calculated and desired outputs. $\varepsilon(p)$ is the system error of p^{th} training pattern, which is defined as

$$\varepsilon(p) = \frac{1}{2}\sum_{k=1}^{q}(D_k(p) - x_k(p))^2 \tag{3}$$

Where $D_k(p)$ and $x_k(p)$ are the k^{th} element of the desired output and calculated output respectively. On the other hand number of neurons in output layer is q. Readjustment of the weights in the hidden layers and output layers is the next step by using a generalized delta rule which is minimizing the distinction between the desired outputs and calculated outputs. Every interconnection weight of the incremental correction can be calculated by

$$\Delta\omega_{kj} = -\gamma\frac{\partial\epsilon(p)}{\partial\omega_{kj}} + \rho\Delta\omega_{kj}(p-1) \tag{4}$$

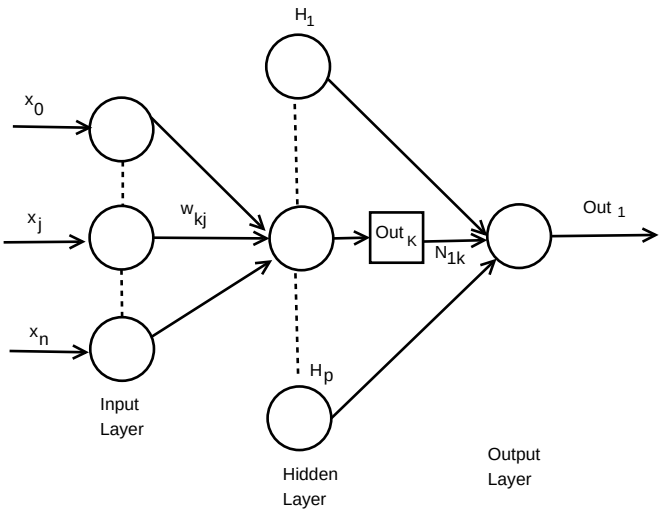


Figure 2: $n \times p \times 1$ MLP model architecture.

The incremental correction of the interconnection weight between j^{th} neurons and k^{th} neuron is $\Delta\omega_{kj}(p)$, in the last iteration, the incremental correction is $\Delta\omega_{kj}(p-1)$ learning rate is γ and the range of $0 < \gamma < 1$, momentum factor is the ρ and range is $0 \leq \rho < 1$. By using learning rate the updated weight is control and improve of the effectiveness of learning procedure by using momentum. Training set carried when the squared errors of average sum are over and all training patterns was globally reduced. On training period completion, the testing period was conduct those input pattern which was not present in the training set. Now we consider a MLP model with $n \times p \times 1$ network. Input layer, hidden layer and output layer are n, p, 1 respectively. Show in Figure 2. For both hidden and output neurons was select sigmoid function as the activation function. The mapping function of input-output are realize by the network Out_1 can be computed as

$$Out_1 = \phi(U_1) = \frac{1}{1 + exp^{(}-U_1)} \tag{5}$$

$$Out_k = \phi(u_k) = \frac{1}{1 + exp^{(}-u_k)} \tag{6}$$

Where

$$U_1 = \sum_{k=0}^{p} N_{1k} Out_k \tag{7}$$

$$u_k = \sum_{j=0}^{p} \omega_{kl} x_j \tag{8}$$

Algorithm 1 Training procedure - MLP-1

Step 1: Initialize all weight and bias.
Step 2: While terminating condition is not satisfied;
Step 2.1 For each training tuple; propagate the input forward,
Step 3: For each output layer of unit q,
Step 3.1: Output of an input unit, which is actual input values,
Step 4: For every hidden layers or output layers,
Step 4.1: Compute the net input of unit q with respect to the previous layer p;
Step 4.2: Compute the output of q, layer.
Step 5: Compute the error of output layer of unit q;
Step 6: Compute the error with respect to the next higher layers.
Step 7: Increment the weight of every layers.
Step 8: Update the weight of every layers.
Step 9: Increment the bias of every layers.
Step 10: Update the bias of each layers.
Step 11: After some iteration, when all $\Delta\omega_{pq}$ in the previous epoch were so small as to bellow specified threshold 0.05, then the procedure will be stop.

Algorithm 2 Training procedure - MLP-2

Step 1: Initialize the weight and bias.
Step 2: Perform a vector of row which is the weight interconnection between the hidden layers (p nodes) and output layer.
Step 2: Perform $n \times p\times$ matrix and interconnection weights between the input nodes (n) and hidden nodes (p) is ω.
Step 3: Row vector are computed.
Step 4: Relative weight of input node are computed.
Step 4: After some iteration, when all weight values in the previous epoch were so small as to bellow specified threshold 0.05, then the procedure will be stop.

Interconnection weight between the k^{th} hidden neuron and output neuron is N_{1k} interconnection weight between the k^{th} hidden neuron and j^{th} input neuron is ω_{kj}. In the j^{th} hidden neuron the output value is Out_k. x_j is the input of the j^{th} input neuron, where $Out_0 = -1$. The case belongs to class 1, where $Out \geq 0.05$ and case belongs to class 0 where $Out_1 < 0.05$. A node can recognized an output of that node given an input or set of inputs by using activation function. A linear access can be produce 1 or 0 output, but non-linear process the activation function can be generate the output in the specific limit. The activation function can be accepted large forms construct on the data sets. A set of training samples, comparing the networks prediction for each sample with the actual known class level are frequently generating by using back-propagation learning method. The weights are changes as to minimize the mean squared error between the actual class and the prediction network for each sample. The weight and every bias are initialized to small random number of the network. Now we can calculate the total input and output of each unit in the output layer and hidden layer. At first the sample are fed to the input layer of the network. Note that for unit q in the input layer, its output is equal to its input, that is $Out_q = \eta_q$ for input q. Given an unit q in a hidden layer or output layer, the net input η_q, to unit q is

$$\eta_q = \sum \omega_{pq} Out_p + \beta_q \tag{9}$$

Where ω_{pq} is the weight of the connection from unit p in the previous layer to unit q; Out_p is the output of unit p from the previous layer and β_q is the bias of the unit. Given the net input η_q to unit q and output of unit q is computed as

$$Out_q = \frac{1}{1 + \lambda^{-\eta_q}} \tag{10}$$

The error is created backwards by updating the weight and bias in the network. For a unit q in the output layer, the error κ_q is computed by

$$\kappa_q = Out_q(1 - Out_q)(\chi_q - Out_q) \tag{11}$$

Where Out_q is the actual output of unit q and χ_q is the true output. The error of hidden layer of unit of unit q is

$$\kappa_q = Out_q(1 - Out_q) \sum \kappa_r \alpha_{qr} \tag{12}$$

Where α_{qr} is the weight of the connection from unit q to unit r and κ_r is the error of unit r. Now the weight of every layer are incremented and update the weight value of each layer.

$$\Delta\omega_{pq} = (\mu)\kappa_q Out_q \tag{13}$$

$$\omega_{pq} = \omega_{pq} + \Delta\omega_{pq} \tag{14}$$

Now the bias of every layer are incremented and update the all bias value of each layer.

$$\Delta\beta_q = (\mu)\kappa_q \tag{15}$$

$$\beta_q = \beta_q + \Delta\beta_q \tag{16}$$

After some iteration, when all $\Delta\omega_{pq}$ in the previous epoch were so small as to bellow specified threshold 0.05, then the procedure will be stop. The procedure will be describe in algorithm 1 and algorithm 2.

4 Description of the Data sets

In this work, we can select three types of leukemia gene expression data sets. The data sets ID is GDS-2643, GDS-2501, GDS-3057 and title of the data set is Waldenstrom's macroglobulinemia (B lymphocytes and plasma cells), B-cell chronic lymphocytic leukemia, and Acute Myeloid Leukemia respectively. The data base web link is http://ncbi.nlm.nih.gov/projects/geo/.

The data set (GDS-2643) consists of 22,283 numbers of genes with 56 samples. Among them, there are 13 normal samples which consist of 8 normal for B lymphocytes and 5 normal plasma cells and 43 diseased samples which consist of 20 Waldenstrom's macroglobulinemia, 11chronic lymphocytic leukemia, 12 multiple myeloma samples.

The dataset (GDS-2501) data set consists of 22,283 numbers of genes with 16 samples. In this sample analysis of B-cell chronic lymphocytic leukemia (B-CLL) cells that express or do not express zeta-associated protein (ZAP-70) and CD38. The prognosis of patients with ZAP-70-CD38- B-CLL cells is good, those with ZAP-70+CD38+B-CLL cells is poor.

The dataset (GDS-3057) content 26 Acute Myeloid Leukemia (AML) patients with normal hematopoietic cells at a variety of different stages of maturation from 38 healthy donors The total data set consist of 22,283 number of genes with 64 samples. Among them, there are 38 normal samples which contents 10 normal for bone marrow, 10 normal samples for peripheral blood, 8 normal samples for bone marrow CD34 plus and 10 normal samples for Primed peripheral blood hematopoietic stem cells (PBSC) CD34 plus. On the other hand there are 26 leukemia samples which contents 7 bone marrow and 19 peripheral bloods.

5 Comparative Performance Evaluation of the Models

In this section, the usefulness of the procedure has been demonstrated three types of human leukemia gene expression data sets. Now we can apply comparative analysis with seven existing methods like SVM, SNR, SAM, BR, NA, GMM and HMM. We have applied this procedure on the gene expression data sets for selecting important genes. We have found two classifier groups. One is normal class and another is disease class. After some iteration, we have found normalized value of every gene. Here we have considered a threshold value which is 0.05. After normalization if the gene value is grater then 0.05, then which types of gene is normal gene. After normalization if the gene value is less than 0.05, then which types of gene is disease gene. We consider several genes that are most significant of our experiment. The gene expression values are significantly changes from normal samples to diseased samples. Applying this process on the first data set (GDS-2643), we have found that genes like CYBB, TPT1 and PRDM2 among the most important genes which are over the expressed the diseased samples. On the other hand CRYAB minimize the expression value and fully significant in diseased samples. The gene are recognize as an under expressed gene. In order to limited size of manuscript, we have showed only the profile plots of genes of GDS-2643 data set (depicted in Figure 3). In the case of GDS-2501 data set, the genes like ACTB, CENPN, ALCAM, PXN have changed their expression values for normal samples to diseased samples. Similarly, the dataset GDS-3057, like TDG, CTIF, LLS, NAB2 have changed their expression values for normal samples to diseased samples. The usefulness of the methodology has been shown three types of leukemia gene expression data. We have applied the methodology on the aforesaid gene expression data sets for selecting some important gene intercede diseases. Now we have applied the procedure on the previous gene expression data sets for selecting some important diseases mediating genes. In this procedure, at first based on correlation values the genes are placed into group. For GDS-2643, we have found six groups which holding 1869, 1131, 1033, 601, 537 and 1208 number of genes (Table 5). The most important group by both MLP-1 and MLP-2 has been selected 1869 number of group genes. Now we have considered two classes for both MLP-1 and MLP-2. One is normal samples class and other is diseased samples class. ω is a weight coefficient value which is initialized by random numbers. Both MLP-1 and MLP-2, we have found 28 and 30 number of genes respectively, when grouping of genes and most important groups selection are completed. Now we have found 20 numbers genes that are present in both procedure. Among them, based on their ω values we have selected 18 number of genes.

Similarly in GDS-2501, we have found eight groups which containing 652, 1031, 1217, 1301, 816, 539, 741 and 912 number of genes. Similarly in GDS-3057, eight group of genes which contain 2521, 2624, 2241, 2341, 2471, 803, 2191, and 2238. Now we have found 1869 genes for GDS-2643, 1217 genes for GDS-2501 and 2624 genes for GDS-3057 respectively by applying both MLP-1 and MLP-2. Finally we have found 24, 21 and 20 number of most important genes corresponding to the three data set by using MLP-1. On the other hand we have found 25, 21 and 19 number of most important genes corresponding to the three data set by using MLP-2. The number of genes which is found by both MLP-1 and MLP-2 are 18, 21 and 17 to these data sets. Table (5) show that for different sets of genes the number of functionally enriched attributes corresponding to these methods. It has been found that both procedure MLP-1 and MLP-2 performed the best results for all data sets. These results show that

Data set ID	Selected group	No of selected groups from selected group	Group	No genes in each group
GDS-2643	1	18	1	1869
			2	1131
			3	1033
			4	601
			5	537
			6	1208
GDS-2501	2	21	1	652
			2	1031
			3	1217
			4	1301
			5	816
			6	539
			7	741
			8	912
GDS-3057	4	17	1	2521
			2	2624
			3	2241
			4	2341
			5	2471
			6	803
			7	2191
			8	2238

Table 1: Selection of groups and genes for different data set

the both MLP-1 and MLP-2 procedure has been able to select the more number of important genes responsible for mediating a disease than the other seven existing procedure.

5.1 Validation of the Result

In this part, we have analysis the results which is founded by different types of procedure including MLP-1 and MLP-2. In this comparison we have using p-value, t-test, biochemical pathways, F-test and sensitivity.

5.1.1 Statistical Validation

Prosperity of every GO attributes of every genes has been computed by its p-value. When the p-value is low, it means that the genes are biologically significant. Now we have create a comparative analysis, by using some other procedure like SVM, SNR, SAM, BR, NA, GMM and HMM. For different sets of genes Table 2 show that number of functionally enriched attributes corresponding to these procedure. For all data set we have been show that MLP-1 and MLP-2 performed the best result. The more important genes are select by these procedure are show in these results. Both NFM-1 and NFM-2 are accomplished of getting more number of true positive genes in terms of identifying the GO attributes and cancer attributes with respect to other existing procedure are depicted in Figure(4). Now we have 478 GO attributes of 375 gene set are identified among them 102 GO attributes of cancer related are identified.

Now we validate the results statistically, we have performed t-test on the genes identified by DLM on each data sets. t-test is the statistical significance which indicate whether or not the difference between two groups average most likely reflects an original difference in the population from which the group wear sampled. The t-value show the most significant genes (99.9%) which p-value <0.001. For this three types of data set we can apply t-test and we get corresponding t-value. We have identify some important genes like IARS (4.78), MMP25 (5.68), TYMS (4.96), HPS6 (5.24), MLX (4.12), CALCA (4.32), HIC2 (4.12), ANP32B (5.16), TFPI (4.51), CRYAB (3.08), NCF1C (3.71), HNRNPH1 (3.12), etc. The number in the bracket shows t-value of the corresponding gene. The t-value of this genes exceeds the value for $p = 0.001$. This means that this gene is highly significant (99.9% level of significance). Similarly genes like ERCC5 (3.52), PRDM2 (3.45), PRIM2 (2.54), TPT1 (3.35), RPS26 (2.41), EFCAB11 (3.71), PRPSAP2 (3.43), PRKACA (2.44), etc exceed the t-value for $P = 0.01$. It indicated that this gene is significant at the level of 99%. Similarly genes like MED17 (2.55), MAPK1 (2.35), PIK3CB (2.45), NMD3 (2.15), ARG2 (2.16), EXOC3 (2.56), WHSC1 (2.71), RFC4 (2.35), GLB1L (2.71), HNF1A (2.41) etc exceeds the value for $P = 0.05$. It indicate that this genes significant at the level of 95%. Similarly genes like FLG (1.77), TXNL1 (1.24), RIN3 (1.34), CYBB (2.31), ZNF814 (1.45), KLF4 (1.18) etc exceeds the value for $P = 0.1$. It indicate that this type of genes significant at the level of 90%.

In GDS-2643, we have found cancer pathway for non-small cell and small cell leukemia. In these two

Data set	Gene Set	MLP-1	MLP-2	SVM	SNR	SAM	BR	NA	GMM	HMM
GDS-2643	First 5	60	58	59	10	12	15	10	27	20
	First 10	72	73	65	17	18	20	13	32	25
	First 15	77	79	71	21	22	26	12	41	29
	First 20	79	92	76	30	26	30	16	45	34
GDS-2501	First 5	83	78	85	15	55	48	33	39	48
	First 10	88	85	88	24	62	55	41	44	60
	First 15	101	98	95	31	68	72	52	51	55
	First 20	107	106	103	39	84	65	66	68	77
GDS-3057	First 5	45	52	37	21	24	26	24	19	20
	First 10	63	62	46	35	30	29	33	25	28
	First 15	67	65	57	37	29	31	39	27	35
	First 20	69	70	55	41	31	35	42	36	39

Table 2: Result of several sets of genes on number of attributes

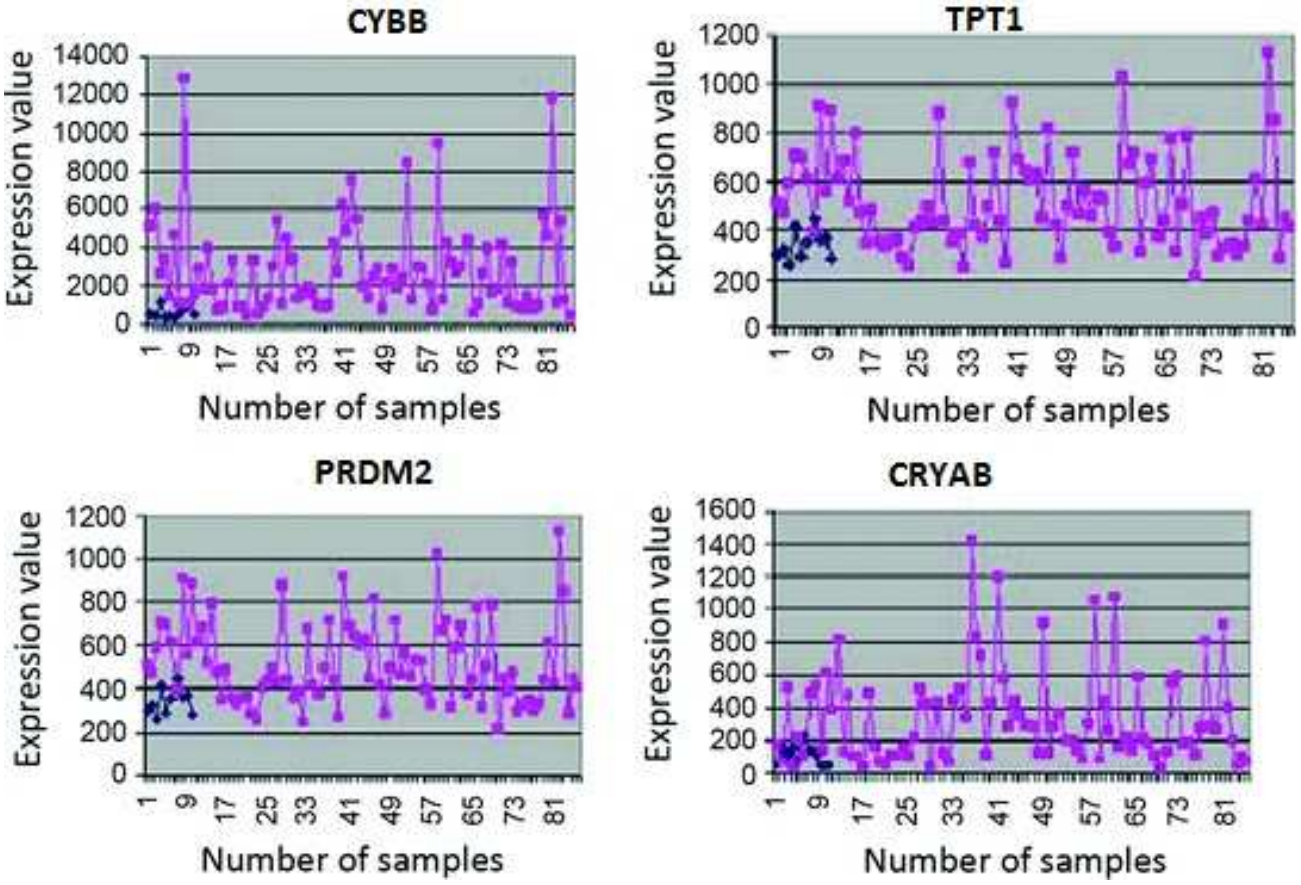


Figure 3: Expression profiles of some over-expressed genes (CYBB, TPT1, PRDM2) and under-express (CRYAB) in normal (shown by blue points) and disease (shown by red points) sample of human leukemia expression data.

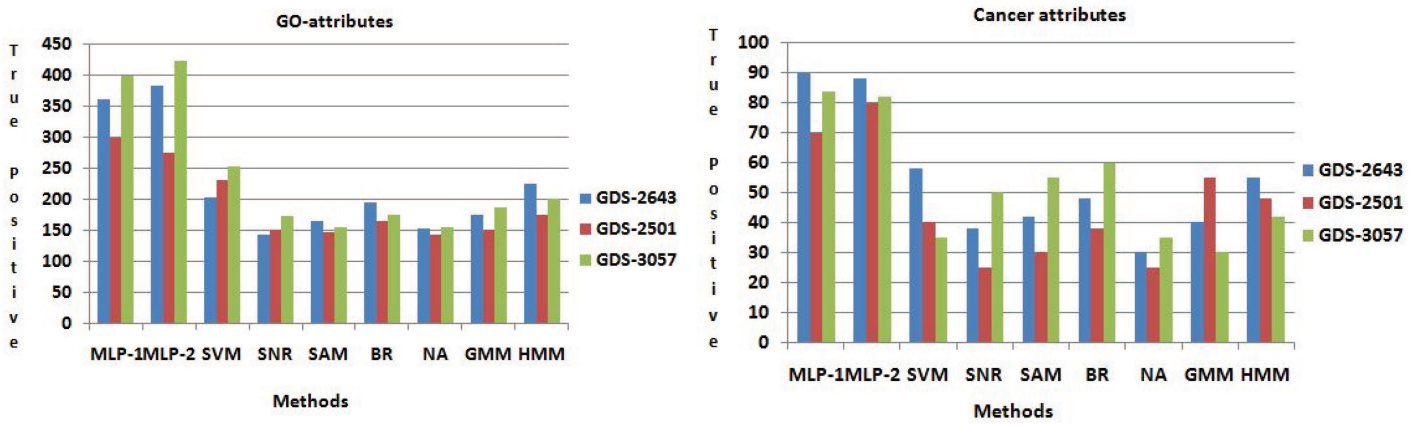


Figure 4: Identification of GO attributes and Cancer attributes the Performance comparisons of MLP-1 and MLP-2 with other seven existing methods.

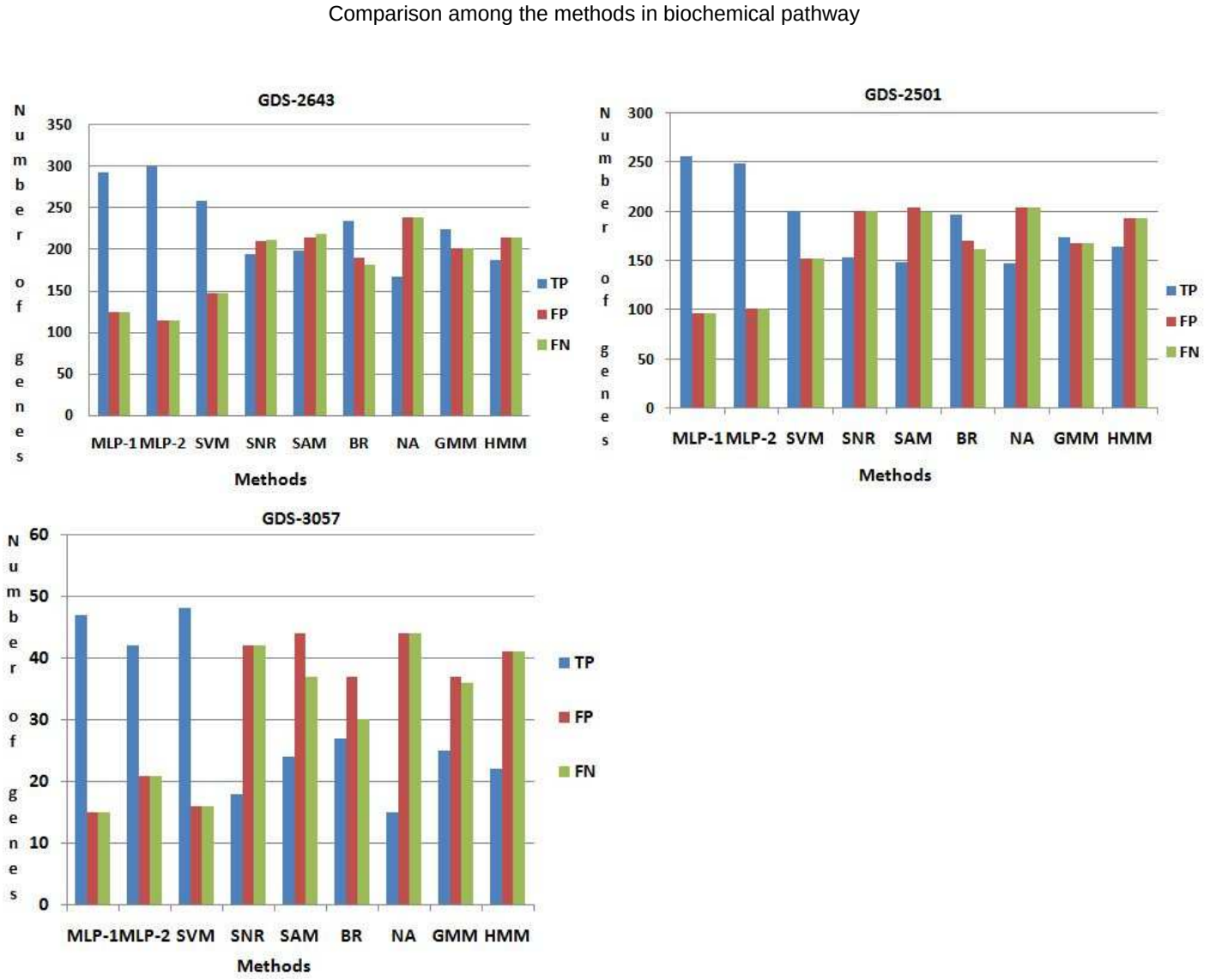


Figure 5: Comparison among the methods. Here TP, FP, FN indicate the number of *truepositive*, *falsepositive*, *falsenegative* respectively.

pathway we have found 407 number of genes. Now this set of genes, are compared with those obtained by seven existing procedures. Here 293 and 300 number of genes are identified which is common in database information and the results of both MLP-1 and MLP-2, respectively. These genes are called $truepositive(TP)$ genes. In top rank 407 genes, we have found 116 and 110 number of genes are present respectively in both procedure MLP-1 and MLP-2 but not present these pathway. These genes are called $falsepositive(FP)$ genes. Similarly we have found 116 and 110 number of genes are $falsenegative(FN)$ for both MLP-1 and MLP-2 procedure respectively. Now we have calculate the number of $truepositive(TP)$, $falsepositive(FP)$ and $falsenegative(FN)$ genes for other seven existing methods. From Figure (5) show that both NFM-1 and NFM-2 procedure have been able to identify more number of $truepositive$ genes, but less number of $falsepositive$ and $falsenegative$ genes compared to all the other procedures. Similarly in GDS-2501 and GDS-3057, we have been able to identify more number of $truepositive$ genes, but less number of $falsepositive$ and $falsenegative$ genes compared to all the other procedures and result are show in Figure (5).

5.1.2 Biological Validation

The disease mediating gene list corresponding to a specific disease can be founded by a gene database namely NCBI (http://www.ncbi.nlm.nih.gov/Database). We have found several set of genes for GDS-2643, GDS-2501 and GDS-3057 respectively. For GDS-2643, by using both MLP-1 and MLP-2, we have identified 351 numbers of genes. Now we have compared this set of genes with 351 genes from NCBI and both MLP-1 and MLP-2 can be identified 247 and 241 number of genes respectively, which is common both sets. These genes said $truepositive(TP)$. On the other hand, (351-247) = 104 and (351-241) = 110 number of genes are not present the list of genes which is found from NCBI for both procedure MLP-1 and MLP-2 respectively. These numbers of gene are called $falsepositive(FP)$ genes. Similarly (351-247) = 104 and (351-241) = 110 number of genes are present NCBI database but not present in the set of genes which is found by both procedure MLP-1 and MLP-2 respectively. These numbers of gene are called $falsenegative(FN)$ genes. Likewise other procedure, viz., SVM, SNR, SAM, BR, NA, GMM, HMM are compared our result on GDS-2643, GDS-2501 and GDS-3057 respectively. Figure (6) show that both MLP-1 and MLP-2 found the more number of true positive (TP) genes then the other existing procedure for GDS-2643, GDS-2501 and GDS-3057 respectively.

Now we can order to validate our result and calculate the Sensitivity of gene expression data sets. The Sensitivity is computed by using following formula:

$$Sensitivity = \frac{TP}{TP + FN} \tag{17}$$

At first we have computed the total number of true positive genes for every data set for every procedure. As our result, Sensitivity of both MLP-1 and MLP-2 procedure is more than the other existing procedure. Figure (7) show that Sensitivity of both MLP-1 and MLP-2 has performed the best for GDS-2501 data set.

6 Conclusion

In this article, we have proposed a model based on multilayer perceptron, which will select the genes that have been changed quite significantly from normal stage to disease stage. Base on vale of correlation, the different types of genes can be obtained and we have found most important groups. The gene of these groups are evaluated by using both procedure MLP-1 and MLP-2. The most important genes are gained by the procedure have also been corroborated by p-values of genes. The best result of the procedure compression too few standing once has been demonstrated. The output have been corroborated using biochemical pathway, p-value, t-test, sensitivity and some existing result expression profile plots. It has been obtained that the procedure has been capable to the genes are most significant.

Appendix

F-score: F-score are a statistical method for determining accuracy accounting for both precision and recall. The formula for traditional F-score is, F-score= $2 * (precision * recall/precision + recall)$. Where precision = $TP/TP + FP$, Recall = $TP/TP + FN$.
True Positive: A true positive test result is one that detect the condition when the condition is present. True positive rate = $TP/(TP + FN)$.
False Positive: A false positive is an error in some rating method in which a condition tested for is badly found to have been detected. A false positive test result is one that detect the condition when the condition is absent. False positive value = $FP/(FP + TN)$.
False Negative: A result that appears negative when it should not. A false negative test result is one that does not detect the condition when the condition is present. False negative value = $FN/(TP + FN)$.
True Negative: A true negative test result is one that does not detect the condition when the condition is absent. True negative value = $TN/(TN + FP)$.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work has been supported by the University Grant Commission (UGC) RGNF-2014-15-SC-WES-69906, Govt of India and TEQIP-II Project, University of Calcutta.

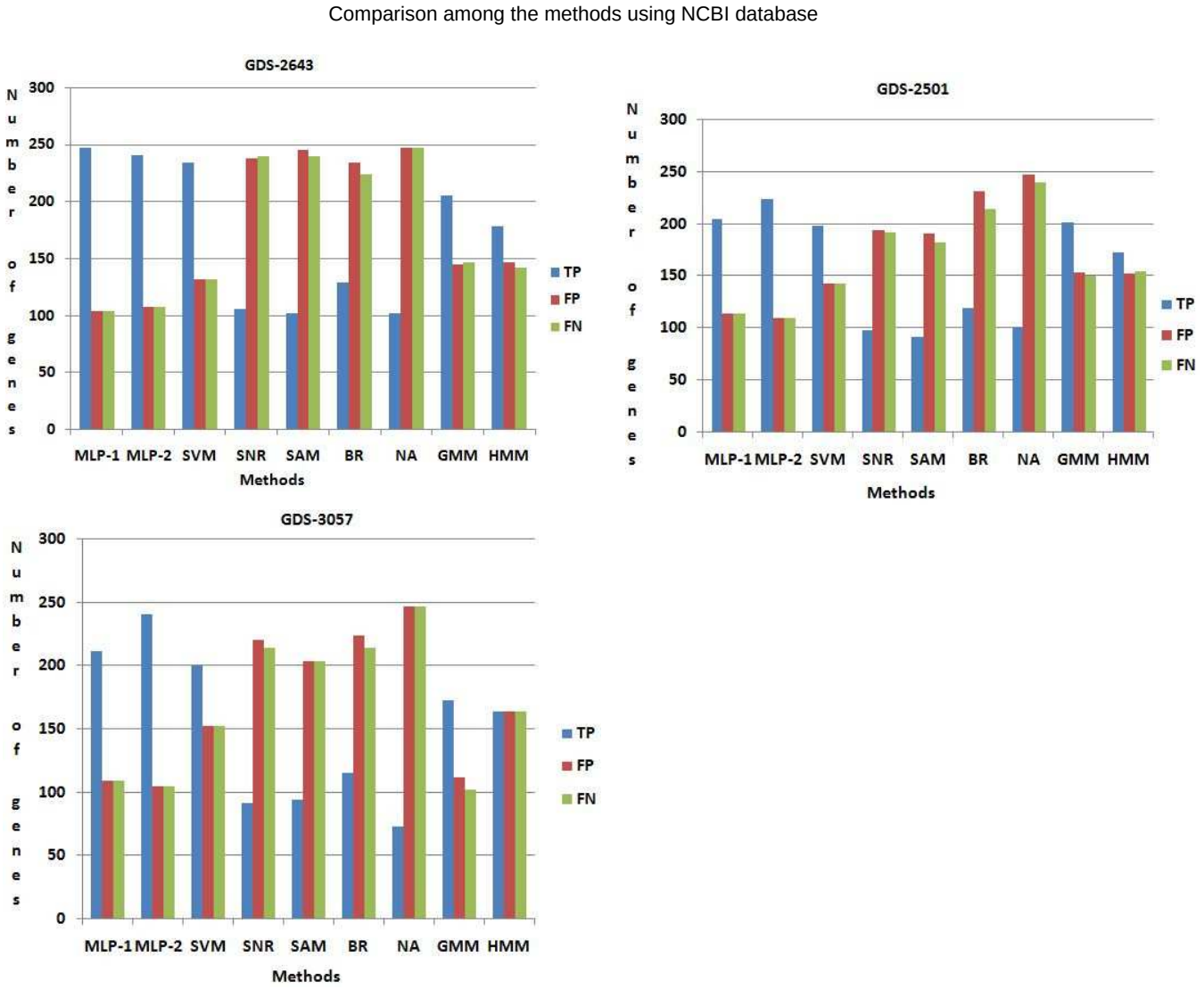


Figure 6: Comparison among the methods using NCBI database. Here TP, FP, FN indicate the number of $truepositive$, $falsepositive$, $falsenegative$ respectively.

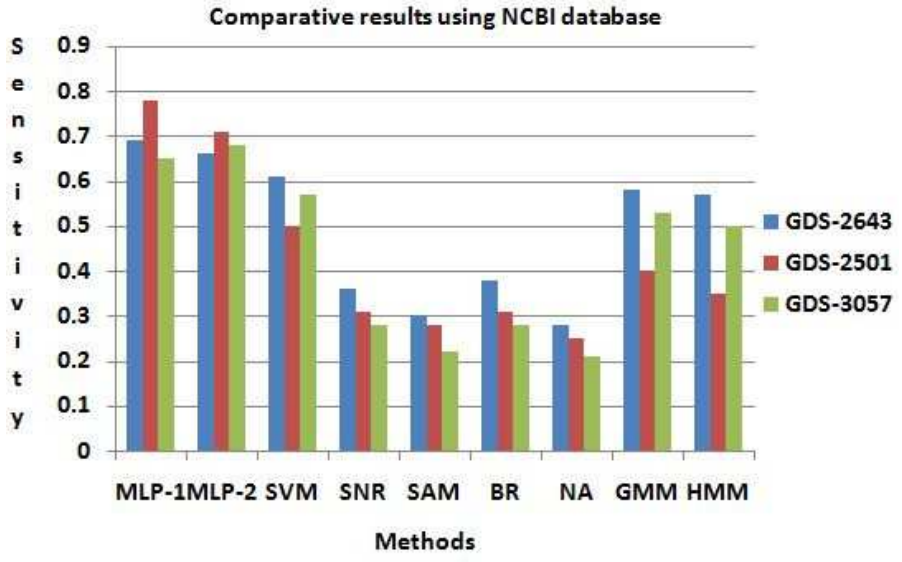


Figure 7: Comparison among the methods using NCBI database.

References

[1] S. K. Veettil, K. G. Lim, N. Chaiyakunapruk, S. M. Ching, M. R. Abu Hassan, "Colorectal cancer in Malaysia: Its burden and implications for a multiethnic country, Asian Journal of Surgery, 2016. 10.1016/j.asjsur.2016.07.005

[2] L. Seewald, J. W. Taub, K. W. Maloney, E. R.B. McCabe, "Acute leukemias in children with Down syndrome, 107(1), 25-30, 2012. 10.1016/j.ymgme.2012.07.011

[3] J. F. Zeidner, J. E. Karp, "Clinical activity of alvocidib (flavopiridol) in acute myeloid leukemia, Leukemia Research, 39(12), 1312-1318, 2015. 10.1016/j.leukres.2015.10.010

[4] R. L. Sielken Jr, C. V.-Flores, "A comprehensive review of occupational and general population cancer risk: 1,3-Butadiene exposureresponse modeling for all leukemia, acute myelogenous leukemia, chronic lymphocytic leukemia, chronic myelogenous leukemia, myeloid neoplasm and lymphoid neoplasm, Chemico-Biological Interactions, 241, 50-58, 2015. 10.1016/j.cbi.2015.06.009

[5] T. Ripperger, B. Schlegelberger, "Acute lymphoblastic leukemia and lymphoma in the context of constitutional mismatch repair deficiency syndrome, European Journal of Medical Genetics, 59(3), 133-142, 2016. 10.1016/j.ejmg.2015.12.014

[6] G. Mezei, M. Sudan, S. Izraeli, L. Kheifets, "Epidemiology of childhood leukemia in the presence and absence of Down syndrome, Cancer Epidemiology, 38(5), 479-489, 2014. 10.1016/j.canep.2014.07.006

[7] S. Izraeli, "The acute lymphoblastic leukemia of Down Syndrome Genetics and pathogenesis, European Journal of Medical Genetics, 59(3), 158-161, 2016. 10.1016/j.ejmg.2015.11.010

[8] H. Suzuki, A. Shigeta, T. Fukunaga, "Death resulting from a mesenteric hemorrhage due to acute myeloid leukemia: An autopsy case, Legal Medicine, 16(6), 373-375, 2014. 10.1016/j.legalmed.2014.07.003

[9] J. F. Zeidner, J. E. Karp, "Clinical activity of alvocidib (flavopiridol) in acute myeloid leukemia, Leukemia Research, 39(12), 1312-1318, 2015. 10.1016/j.leukres.2015.10.010

[10] 10. P. H. Lin, C. C. Lin, H. I. Yang, L. Y. Li, L. Y. Bai, C. F. Chiu, Y. M. Liao, C. Y. Lin, C.Y. Hsieh, C. Y. Lin, C. M. Ho, S. F. Yang, C. T. Peng, F. J. Tsai, S. P. Yeh, "Prognostic impact of allogeneic hematopoietic stem cell transplantation for acute myeloid leukemia patients with internal tandem duplication of FLT3, Leukemia Research, 37(3), 287-292, 2013. 10.1016/j.leukres.2012.10.005

[11] L. Seewald, J. W. Taub, K. W. Maloney, E. R.B. McCabe, "Acute leukemias in children with Down syndrome, Molecular Genetics and Metabolism, 107(1), 25-30, 2012. 10.1016/j.ymgme.2012.07.011

[12] A. Ghosh, R. K. De, "Fuzzy Correlated Association Mining: Selecting altered associations among the genes, and some possible marker genes mediating certain cancers, Applied Soft Computing, 38, 587-605, 2016. 10.1016/j.asoc.2015.09.057

[13] A. Ghosh, R. K. De, "Development of a fuzzy entropy based method for detecting altered gene-gene interactions in carcinogenic state, Journal of Intelligent & Fuzzy Systems, 26, 2731-2746, 2014. 10.3233/IFS-130942

[14] A. Ghosh, R. K. De, "Linguistic Recognition System for Identification of Some Possible Genes Mediating the Development of Lung Adenocarcinoma, Inf. Fusion, 10, 260-269, 2009. 10.1016/j.inffus.2008.11.007

[15] P. A. Mundra, J. C. Rajapakse, "Gene and sample selection using T-score with sample selection, Journal of Biomedical Informatics, 59, 31-41, 2016. 10.1016/j.jbi.2015.11.003

[16] A. Ghosh, R. K. De, "Identification of certain cancer-mediating genes using Gaussian fuzzy cluster validity index, Journal of Biosciences, 40(4), 741-754, 2015. 10.1007/s12038-015-9557-x

[17] S. Tabakhi, A. Najafi, R. Ranjbar, P. Moradi, "Gene selection for microarray data classification using a novel ant colony optimization, Neurocomputing, 168, 1024-1036, 2015. 10.1016/j.neucom.2015.05.022

[18] S. Saha, D. B. Seal, A. Ghosh, K. N. Dey, "A Novel Gene Ranking Method Using Wilcoxon Rank Sum Test and Genetic Algorithm, Int. J. Bioinformatics Res. App, 12, 263-279, 2016. 10.1504/IJBRA.2016.078236

[19] G. Ongun, U. Halici, K. Leblebicioglu, V. Atalay, M. Beksac, S. Beksac, "Feature extraction and classification of blood cells for an automated differential blood count system, "Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on, 4, 2461-2466, 2001. 10.1109/IJCNN.2001.938753

[20] H. H. Zhang, J. Ahn, X. Lin, C. Park, "Gene Selection Using Support Vector Machines with Non-convex Penalty, Bioinformatics, 22, 88-95, 2006. 10.1093/bioinformatics/bti736

[21] A. Ghosh, R. K. De, "Interval based fuzzy systems for identification of important genes from microarray gene expression data: Application to carcinogenic development, Journal of Biomedical Informatics, 42, 1022-1028, 2009. 10.1016/j.jbi.2009.06.003

[22] A. Ghosh, B. C. Dhara, R. K. De, "Comparative Analysis of Cluster Validity Indices in Identifying Some Possible Genes Mediating Certain Cancers, Molecular Informatics, 32(4), 347-354, 2013. 10.1002/minf.201200142

[23] V. Elyasigomari, M.S. Mirjafari, H.R.C. Screen, M.H. Shaheed, "Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization, Applied Soft Computing, 35, 43-51, 2015. 10.1016/j.asoc.2015.06.015

[24] R. D. Uriarte, S. A. de Andres, "Gene selection and classification of microarray data using random forest, BMC Bioinformatics, 7(1), 1-13, 2006. 10.1186/1471-2105-7-3

[25] H. Deng, G. Runger, "Gene selection with guided regularized random forest, Pattern Recognition, 46(12), 3483-3489, 2013. 10.1016/j.patcog.2013.05.018

[26] A. Anaissi, P. J. Kennedy, M. Goyal, D. R. Catchpoole, "A balanced iterative random forest for gene selection from microarray data ,BMC Bioinformatics , 14(1), 1-10, 2013.10.1186/1471-2105-14-261

[27] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning, 46(1), 389-422, 2002. 10.1023/A:1012487302797

[28] L. A. Menndez, F. J. de Cos Juez, F. S. Lasheras, J. A. A Riesgo, "Artificial neural networks applied to cancer detection in a breast screening programme, Mathematical and Computer Modelling, 52, 983-991, 2010. 10.1016/j.mcm.2010.03.019

[29] M. C.Sharma, G. P.Tuszynski, M. R.Blackman, M. Sharma, "Long-term efficacy and downstream mechanism of anti-annexinA2 monoclonal antibody (anti-ANX A2 mAb) in a pre-clinical model of aggressive human breast cancer, Cancer Letters, 373(1), 27-35, 2016. 10.1016/j.canlet.2016.01.013

[30] T. V. da Silva, R. V. A. Monteiro, F. A. M. Moura, M. R. M. C. Albertini, M. A. Tamashiro, G. C. Guimaraes, "Performance Analysis of Neural Network Training Algorithms and Support Vector Machine for Power Generation Forecast of Photovoltaic Panel, IEEE Latin America Transactions, 15(6), 1091-1100, 2017. 10.1109/TLA.2017.7932697

[31] P. Daz-Rodrguez, J. C. Cancilla, G. Matute, D. Chicharro, J. S. Torrecilla, "Inputting molecular weights into a multilayer perceptron to estimate refractive indices of dialkylimidazolium-based ionic liquidsA purity evaluation, Applied Soft Computing, 28, 394-399, 2015. 10.1016/j.asoc.2014.12.004

[32] M. Dashtban, M, Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts, Genomics, 109(2), 91-107, 2017. 10.1016/j.ygeno.2017.01.004

[33] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, N. Cristianini, "Gene selection for cancer classification using support vector machines, Machine Learning, 389-422, 2002.

[34] C. D. A. Vanitha, D. Devaraj, M. Venkatesulu, "Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection, Procedia Computer Science, 47, 13-21, 2015. 10.1016/j.procs.2015.03.178

[35] S. Mishra, D. Mishra, "SVM-BT-RFE: An improved gene selection framework using Bayesian T-test embedded in support vector machine (recursive feature elimination) algorithm, Karbala International Journal of Modern Science, 1(2), 86-96, 2015. 10.1016/j.kijoms.2015.10.002

[36] Y. Tang, Y. Q. Zhang, Z. Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(3), 365-381, 2007. 10.1109/TCBB.2007.70224

[37] W. H. Chan, M. S. Mohamad, S. Deris, N. Zaki, S. Kasim, S. Omatu, J. M. Corchado, H. Al Ashwal, "Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme, Computers in Biology and Medicine, 77, 102-115, 2016. 10.1016/j.compbiomed.2016.08.004

[38] Y. Tang and Y. Q. Zhang and Z. Huang and X. Hu and Y. Zhao, "Recursive Fuzzy Granulation for Gene Subsets Extraction and Cancer Classification, Trans. Info. Tech. Biomed., 12(6), 723-730, 2008. 10.1109/TITB.2008.920787

[39] D. Du, K. Li, X. Li, M. Fei, "A novel forward gene selection algorithm for microarray data, Neurocomputing, 133(6), 446-458, 2014. 10.1016/j.neucom.2013.12.012

[40] V. de Schaetzen, C. Molter, A. Coletta, D. Steenhoff, S. Meganck, J. Taminau, C. Lazar, R. Duque, H. Bersini, A. Nowe, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(4), 1106-1119, 2012. 10.1109/TCBB.2012.33

[41] S. Chakraborty, "Simultaneous cancer classification and gene selection with Bayesian nearest neighbor method: An integrated approach, Computational Statistics & Data Analysis, 53(4), 1462-1474, 2009. 10.1016/j.csda.2008.10.012

[42] Y. Tang, Y. Q. Zhang, Z. Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis, IEEE/ACM Trans. Comput. Biol. Bioinformatics, 4(3), 365-381, 2007. 10.1109/TCBB.2007.70224

[43] Y. Tang, Y. Q. Zhang, Z. Huang, X. Hu, Y. Zhao, "Recursive Fuzzy Granulation for Gene Subsets Extraction and Cancer Classification, IEEE Transactions on Information Technology in Biomedicine, 12(6), 723-730, 2008. 10.1109/TITB.2008.920787

[44] P. A. Mundra, J. C. Rajapakse, "SVM-RFE With MRMR Filter for Gene Selection, IEEE Transactions on NanoBioscience, 9(1), 31-37, 2010. 10.1109/TNB.2009.2035284

[45] F. Ojeda, J. A.K. Suykens, B. De. Moor, "Low rank updated LS-SVM classifiers for fast variable selection, Neural Networks, 21(2), 437-449, 2008. 10.1016/j.neunet.2007.12.053

[46] A. Ghosh, B. C. Dhara, R. K. De, "Selection of Genes Mediating Certain Cancers, Using a Neuro-fuzzy Approach, Neurocomput., 133, 122-140, 2014.

[47] S. Sheet, A. Ghosh, S. B. Mandal, "Selection of Genes Mediating Human Leukemia, Using Boltzmann Machine, Advanced Computing and Communication Technologies: Proceedings of the 10th ICACCT, 83-90, 2018. 10.1007/978-981-10-4603-2-9

[48] S. Sheet, A. Ghosh, S. B. Mandal, "Identification of influential biomarkers for human leukemia - An Artificial Neural Network approach, International Journal of Soft Computing & Artificial Intelligence, 4, 27-32, 2016.

[49] A. Ghosh, R. K. De, "Neuro-fuzzy Methodology for Selecting Genes Mediating Lung Cancer, Proceedings of the 4th International Conference on Pattern Recognition and Machine Intelligence, 388-393, 2011.

[50] S. Sheet, A. Ghosh, S. B. Mandal, "Selection of genes mediating human leukemia, using an Artificial Neural Network approach, Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2010-2014, 2017. 10.1109/AEEICB.2017.7972415

[51] A. Narayanan, E.C. Keedwell, J. Gamalielsson, S. Tatineni, "Single-layer artificial neural networks for gene expression analysis, Neurocomputing, 61, 217-240, 2004. 10.1016/j.neucom.2003.10.017

[52] H. Q. Wang, H. S. Wong, H. Zhu, T. T.C. Yip, "A neural network-based biomarker association information extraction approach for cancer classification, Journal of Biomedical Informatics, 42(4), 654-666, 2009. 10.1016/j.jbi.2008.12.010

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 11-17 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Impact of Line Resistance Combined with Device Variability on Resistive RAM Memories

Hassan Aziza*, Pierre Canet, Jeremy Postel-Pellerin

Aix Marseille Univ., CNRS, IM2NP UMR 7334, 60 rue F. Joliot-Curie, 13453 Marseille Cedex 13, France

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 13 December, 2017
Online: 18 January, 2018

Keywords:
Resistive RAM arrays
Variability
voltage drop (IR drop)

ABSTRACT

In this paper, the performance and reliability of oxide-based Resistive RAM (ReRAM) memory is investigated in a 28nm FDSOI technology versus interconnects resistivity combined with device variability. Indeed, common problems with ReRAM are related to high variability in operating conditions and low yield. At a cell level ReRAMs suffer from variability. At an array level, ReRAMs suffer from different voltage drops seen across the cells due to line resistances. Although research has taken steps to resolve these issues, variability combined with resistive paths remain an important characteristic for ReRAMs. In this context, a deeper understanding of the impact of these characteristics on ReRAM performances is needed to propose variability tolerant designs to ensure the robustness of the technology. The presented study addresses the memory cell, the memory word up to the memory matrix.

1. Introduction

Because data storage and processing solutions are so central to modern technology, many research works are dedicated to pursue new types of computer memory. One of the major goal is to develop a universal memory (i.e. a storage medium that would combine the high speed of RAM with the non volatility of a Flash drive). Additionally, embedded memory is a fundamental component of any electronic system including high-performance System-on-Chip (SoC) and Internet of Things (IoT) devices. In this context, the presented work proposes a reliability analysis of the Resistive RAM memory which considered as a potential universal memory candidate. This study is an extension of work originally presented in [1], with an extension to a whole memory matrix. Moreover, a multilevel storage capability of ReRAM cells is demonstrated at a matrix level.

According to ITRS [2], embedded Non Volatile Memories (NVMs) are occupying a major part of the area of a typical (SoC). Although Flash memory is widely used today, it needs high voltage for Write and Erase operations and has reliability issues that are hard to handle, increasing the cost of circuit design and process integration. Thus, the industry is trying to find a good alternative NVM that can replace Flash memories. Possible candidates include Magnetic RAM (MRAM), ReRAM, Phase Change Memory (PCM), Ferroelectric RAM (FeRAM), etc. Compared to MRAM and PCM technologies, ReRAM technology is still in an emerging phase [3][4]. However, a considerable technological effort is currently driven worldwide to push this technology to prototype level. ReRAM advantages comprise non volatile data storage at low power and latency and high memory density while maintaining device performance and reliability [5]. Moreover, the 3D-staking technology developed for Flash memories can be transferred to ReRAM and the multilevel cell operation scheme of Flash memories can also be achieved in ReRAMs [6].

However, the continuous push for scalability to obtain high-density chips makes the ReRAM technology extremely sensitive to variability, physical defects and environmental influences that may severely compromise its correct behavior [7]. At sub 32nm node, size reduction increases the resistivity of interconnects, inducing a voltage drop along the memory matrix lines, which can cause reliability issues. Indeed, as ReRAM data is stored as two resistance states of the resistive switching device, these memories are sensitive to resistive paths [8]. On top of that, common problems with ReRAMs are related to high variability in operating conditions [9].

In memory devices relying on resistance change such as ReRAMs, complex physical mechanisms are responsible for reversible switching of the electrical conductivity between high and low resistance states. This resistivity change is generally attributed to the formation/dissolution of conductive paths between metallic electrodes [10]. A typical ReRAM device consists of two metallic electrodes that sandwich a thin dielectric layer serving as permanent storage medium making its leakage current close to

*Corresponding Author: Hassan Aziza, Email: hassen.aziza@univ-amu.fr

zero. Oxide-based Resistive Random Access Memory (so-called OxRAM) use transition metal oxides as a dielectric layer. In this study, an HfO_2 Oxide-based ReRAM stack is considered [11].

OxRAM cell operation is depicted in Figure 1. After an initial electroforming process, the memory element may be reversibly switched between two distinct resistance states. Electroforming stage corresponds to a voltage-induced resistance switching from an initial very high resistance state (virgin state) to a conductive state. After FORMING, resistive switching corresponds to an abrupt change between a High Resistance State (HRS or OFF state) and a Low Resistance State (LRS or ON state). This resistance change is achieved by applying specific voltage (i.e. V_{SET} and V_{RES}) to SET and RESET the memory cell.

It is important to note that the FORMING stage is the first and most critical step as it determines the switching characteristics during the future operation of the memory cell. Thus, the Forming Resistance State (R_{FRS}) which characterizes the filament creation is a key parameter in terms of reliability. Besides, the forming step requires high voltage levels (more important than V_{SET} and V_{RES}).

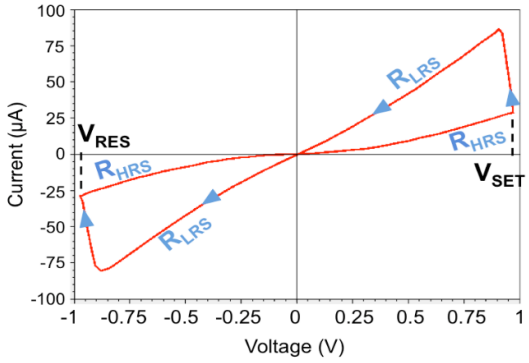


Figure 1. Typical I-V characteristic of a bipolar OxRAM cell

In this paper, the reliability of a ReRAM memory array is investigated versus interconnects and device variability. Section II presents the OxRAM model used for simulations. Section III address a 32-bit memory word. In section III, the study is extended to the memory matrix and the multilevel storage capability of ReRAM cells is demonstrated. Section IV concludes the paper. For each section simulation results based on a 28nm Fully-Depleted Silicon-On-Insulator (FDSOI) are proposed.

2. ReRAM elementary cell model

The proposed OxRAM modeling approach relies on electric field-induced creation/destruction of oxygen vacancies within the switching layer. The model enables continuously accounting for both SET and RESET operations into a single master equation in which the resistance is controlled by the radius of the conduction pathways (r_{CF}) [12]:

$$\frac{dr_{CF}}{dt} = \left(r_{CF\max} - r_{CF}\right)\cdot 10^{\beta_{RedOx}}\cdot e^{\frac{Ea - q\cdot\alpha_{red}\cdot V_{cell}}{k_b\cdot T}} - r_{CF}\cdot 10^{\beta_{RedOx}}\cdot e^{\frac{Ea + q\cdot\alpha_{ox}\cdot V_{cell}}{k_b\cdot T}} \quad (1)$$

Where β_{RedOx} is the nominal oxide reduction rate, E_a is the activation energy, α_{red} and α_{ox} are the transfer coefficients (ranging between 0 and 1 and representing the pathways creation/destruction dynamic), k_b is the Boltzmann constant, r_{CFmax} is the maximal size of the conductive filament radius, T is the temperature and V_{cell} the voltage across the cell.

Moreover, the model makes assumptions of a uniform radius of the conduction pathways, a uniform electric field in the cell and temperature triggered acceleration of the oxide reduction reactions ("redox"). Finally, the total current in the OxRAM includes two components, i.e. one is related to the conductive species (I_{CF}) and the other to the conduction through the oxide (I_{OX}).

$$I_{CF} = \frac{V_{Cell}}{L_x}\cdot\left(\pi\cdot r_{CF}^2\cdot\left(\sigma_{CF} - \sigma_{OX}\right) + \pi\cdot r_{CF\max}^2\cdot\sigma_{OX}\right) \quad (2)$$

$$I_{OX} = A_{HRS}\cdot S_{Cell}\left(\frac{V_{Cell}}{L_x}\right)^{\beta_{HRS}} \quad (3)$$

where L_x is the oxide thickness, S_{Cell} is the total area of the device, σ_{Ox} the oxidation rate and σ_{CF} the reduction rate. To take into account I_{OX} trap assisted current (Poole-Frenkel, Schottky emission, Space Charge Limited Current (SCLC)), a power law between the cell current and the applied bias is considered with two parameters A_{HRS} and β_{HRS}. Finally, the total current flowing through the cell is:

$$I_{Cell} = I_{CF} + I_{OX} \quad (4)$$

I_{CF} is the main contributor to LRS current (I_{LRS}) and I_{OX} is the main contributor to HRS current (I_{HRS}).

The memory cell compact model is calibrated on silicon. The model was confronted to quasi-static and dynamic experimental data before its implementation in electrical circuit simulators. As presented in Figure 2a (current voltage characteristic in logarithmic scale), after calibration, the model satisfactorily matches quasi-static and dynamic experimental data measured on actual HfO_2-based memory elements (TiN/Ti/HfO_x/TiN stack). In Figure 2b, the evolution of SET voltages (V_{app}) as a function of the programming ramp speed is presented. The model implementation focused on this dependence which is crucial for the model to be confidently implemented in circuit simulators.

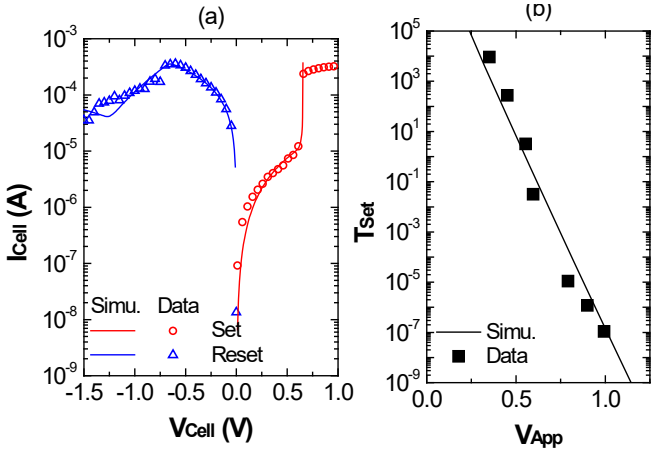


Figure 2. (a) I-V characteristic measured on HfO2-based devices and corresponding simulation using a bipolar OxRAM physical model. (b) SET voltage as a function of the programming ramp

Due to the stochastic nature of the switching process in OxRAMs, leading to large variability, the OxRAM model features a variability dependency. The variation is chosen to fit experimental data as presented in Figure 3. The model behaviour (lines) is consistent with experimental data (symbols). The cell variability modelling is based on OxRAM card model parameters variation. Variability is introduced through specific model parameters (*i.e.* β_{RedOx} and β_{HRS} parameters presented respectively in Equation 1 and Equation 3). Moreover, at sub 32nm node, MOSFET mismatch in the transistor subsystem (digital and analog blocks) increases inherent variability of OxRAM circuits, increasing the overall variability.

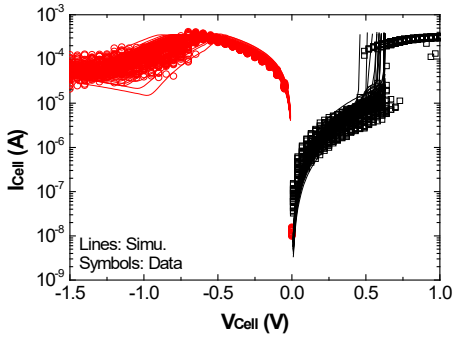


Figure 3. Measured and corresponding simulated I-V characteristic obtained from TiN/Ti/HfO2/TiN devices showing strong variation on R_{LHS} and R_{HRS}

3. ReRAM memory word

3.1. Memory word architecture

A 1T1R ReRAM cell (one MOS Transistor in series with one Resistor) placed in a 32-bit word is considered and presented in Figure 4. The word selection is achieved through the Word Line (WL) before any operation. Word programming is performed in 2 steps, considering that the FORMING operation occurs only once in the product lifetime. Once the word is selected, all cells are RESET in parallel (logical "0") through the Reset Word Line (WL_R), then memory cells are all SET (logical "1") through the Bit Line (BL). The WL_R line drives the whole word RESET current, making it sensitive to line resistivity.

Indeed, during a memory word program operation, a voltage drop occurs along the WL_R line, which can be critical in terms of programming efficiency. To monitor the programming efficiency, High Resistance State (R_{HRS}) and Low Resistance State (R_{LRS}) resistances are extracted after RESET and SET operations respectively.

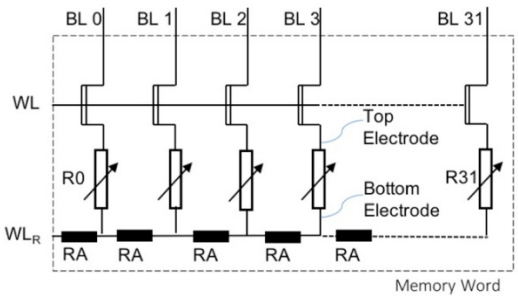


Figure 4. 32-bit ReRAM memory word

The memory word is simulated using the OxRAM model presented in section 2. The model is calibrated on silicon for a 28nm FDSOI technology. With respect to the programming conditions, typical values for LRS and HRS are R_{LRS}=10kΩ and R_{HRS}=130kΩ for the considered technology. The WL_R line resistance per cell (including vias) is called R_A (see Figure 4) and is evaluated to 1.6Ω [8].

3.2. Simulation results

Impact of line resistances on R_{HRS} and R_{LRS} is presented in Figure 5. The voltage drop induced by line resistances results in a weak FORMING (i.e. high resistive path between ReRAM electrodes). SET and RESET programming levels are thus impacted resulting in a higher R_{HRS} (+5%) and R_{LRS} (+30%) values for cells located at the end of the memory word.

Impact of line resistances combined with variability is presented in Figure 6. After 300 Monte Carlo runs, R_{HRS} and R_{LRS} distributions are extracted for 1 cell over 4 and displayed in box-plot forms. Figure 6a shows that LRS distribution spread increases along the word line. Besides, anomalous LRS residual cell populations (far from their typical LRS values, represented by dashed lines) are visible from cell 7. This effect is related to cells not properly formed due to line resistances. In Figure6b, HRS distributions are presented. One can notice the HRS mean distribution shift to higher values with the word length.

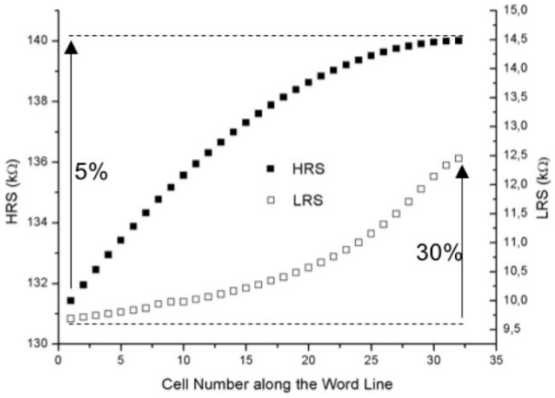


Figure 5. Resistances (HRS & LRS) along the WL

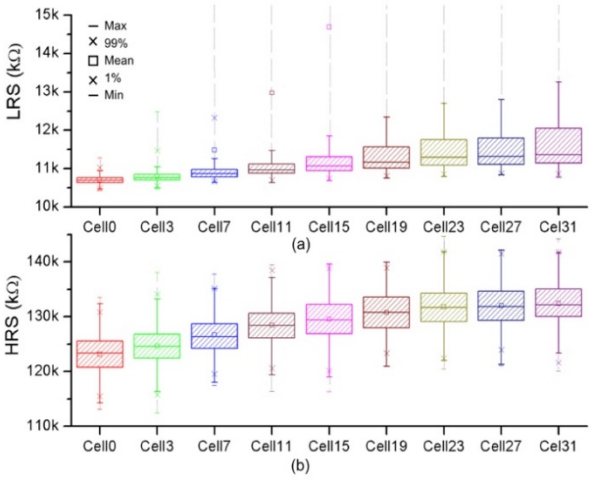


Figure 6. (a) LRS and (b) HRS distributions along the WL

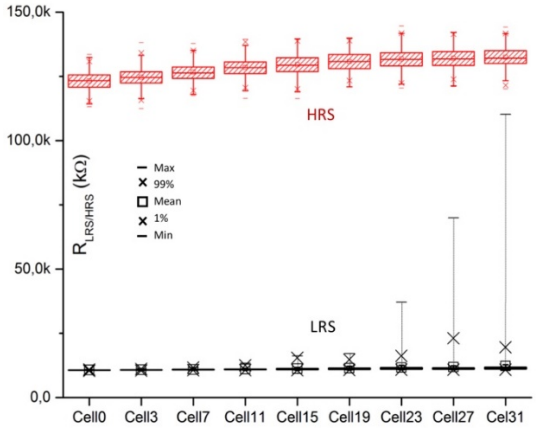


Figure 7. LRS and HRS distributions overlaps along the WL

For comparison purposes, Figure 7 presents the full range variation (from min to max) of each box-plot (HRS and LRS states). Along the WL, HRS/LRS distributions are getting closer and closer due to marginal LRS cells, affecting the memory reliability (i.e. HRS distribution generally larger than LRS distribution is much more degraded by resistive paths).

4. ReRAM memory array

4.1. Memory array architecture

Figure 8 presents the elementary array used for simulation which is constituted by a 3×3 1T1R cell matrix, a row decoder, a column decoder and a sense amplifier for the read operation. The memory cell is modeled by the OxRAM cell model presented in section 2.

The variability analysis is conducted through Monte Carlo simulations. Each Monte Carlo simulation targets specific elements of the circuit: the resistive element, the select transistor, the decoding blocs and the sense amplifier. As a result, R_{ON} and R_{OFF} distribution spreads are extracted. Thus, the contribution of each element in terms of memory performance degradation is demonstrated. A 28-nm Fully Depleted Silicon-On-Insulator (FDSOI) technology is considered for simulations [13].

Memory array cells are first placed in a virgin state. Then, the memory array programming is done in 2 cycles. First, all memory cells are set (logical "1"), then the memory array is reset (logical "0"). R_{ON} value is extracted after the SET operation and R_{OFF} after the RESET operation for each Monte Carlo run.

Variability introduced in the resistive element is chosen to feet experimental date whereas variability of the transistor subsystem is given by the considered technology.

4.2. Simulation results

In this study, 4 different configurations are considered: the "ALL" configuration considers variability in the memory cell (resistive element and select transistor) and the peripheral circuits. In the "CELL" configuration, variability affects only the resistive element. In the "SELECT" configuration, variability affects only the OxRAM select transistor. In the "NO CELL" configuration, variability affects only the peripheral circuits (transistor subsystem, including the select transistor). In this study, the impact of the select transistor variability is subject to special attention. Indeed, the select transistor compliance allows the control of the maximum available current during the set transition, impacting directly the ON/OFF resistance. Table 1 summarizes the 4 configurations to simulate.

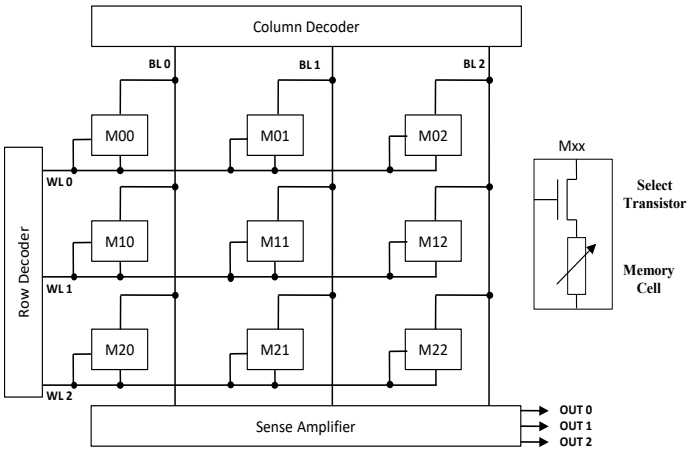


Figure 8. 3×3 OxRAM elementary memory array

Table 1: Simulated Configurations

Configuration	RUNS	Elements under variability
« ALL »	400	All circuit elements
« CELL »	400	Resistive element
« SELECT »	400	Select transistor
« NO CELL »	400	All elements excluding the resistive element

Figure 9 presents the impact of variability of the whole circuit elements (resistive element and transistor subsystem: "ALL" configuration) on R_{ON} and R_{OFF} distributions. Note that R_{OFF} distribution is much larger than R_{ON} distribution. In order to discriminate the contribution of each element of the circuit on ON/OFF resistances, a set of Monte Carlo simulations are also performed with "CELL", "SELECT" and "NO CELL" configurations.

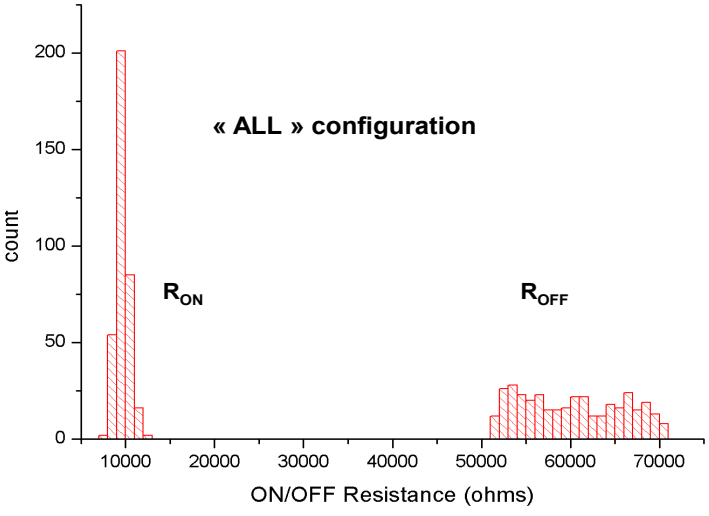


Figure 9. ON/OFF resistance distributions for the "ALL" configuration after 400 Monte Carlo simulations.

Figure 10 presents the impact of variability on R_{OFF} distributions (after RESET) for the 4 configurations. By comparing configuration "ALL" with configuration "CELL", it appears clearly that the impact of cell variability is more important than the impact of the transistor subsystem variability. The third and fourth configurations ("NO CELL" and "SELECT") shows that impact of transistors variability on R_{OFF} is negligible. Besides, the contribution of the select transistor variability is dominant in the peripheral circuit elements. Table 2 proposes a synthesis of results obtained in Figure 10 (mean values, standard deviation and standard deviation reported to the maximum standard deviation value).

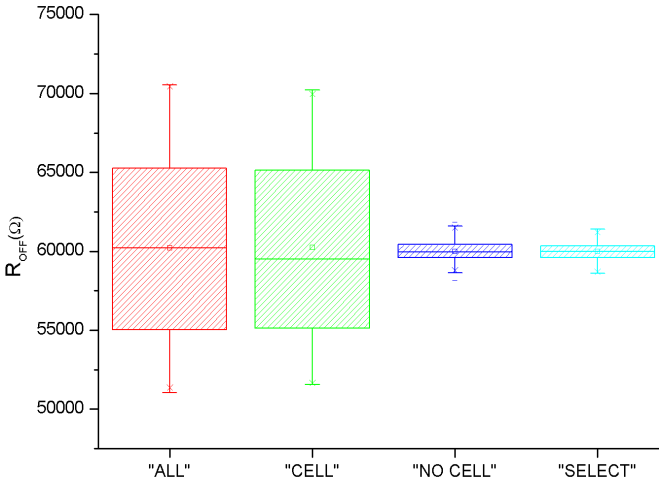


Figure 10. R_{OFF} distributions after 400 Monte Carlo simulations.

Table 2: R_{off} Distribution parameters

Configuration	Mean (Ω)	σ (Ω)	σ (%)
« ALL »	60200	5598	100
« CELL »	60240	5617	99
« NO CELL »	60004	598	9.6
« SELECT »	59990	536	10.6

Figure 11 presents the impact of variability on R_{ON} distributions (after SET) for the 4 configurations.

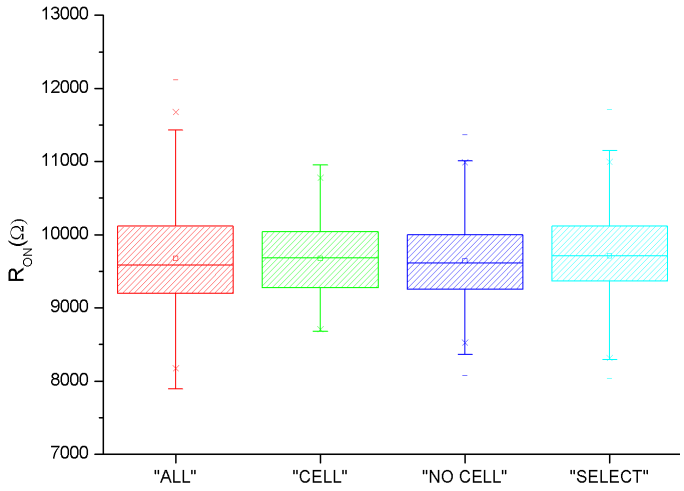


Figure 11. R_{ON} distributions after 400 Monte Carlo simulations.

Table 3: R_{on} Distribution Parameters

Configuration	Mean (Ω)	σ (Ω)	σ (%)
« ALL »	9674	724	100
« CELL »	9679	513	70.8
« NO CELL »	9642	550	78.7
« SELECT »	9713	570	75.9

Compared to Figure 10, it appears clearly that the cell variability impact is less pronounced for low resistive states (LRS), which is a common feature of all OxRAM technologies [14-15]. Moreover, the contribution of transistors variability is comparable to the contribution of the memory cell variability (see "CELL" and "NO CELL" configurations in Figure 11). Here again, simulation results show that the impact of the select transistor is not negligible. Table 3 proposes a synthesis of results obtained in Figure 11.

Figure12 shows a comparison between R_{ON} and R_{OFF} distributions. Based on the considered technology, it is shown that resistance variability in RESET state is much more important than variability in the SET state. This variability is mainly due to the memory cell. In addition, the impact of the select transistor is non-negligible. This effect is visible in the SET state where the impact of the memory cell variability is much less important, making the select transistor variability critical.

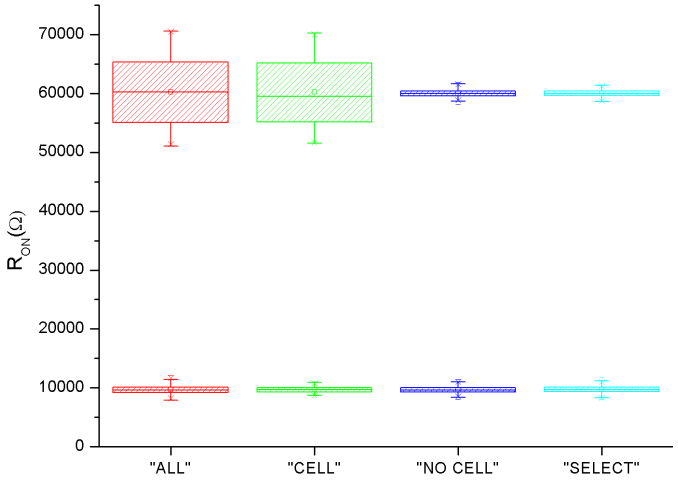


Figure 12. R_{ON}/R_{OFF} distributions after 400 Monte Carlo simulations.

4.3. Multi level approach

When a voltage is applied across a ReRAM cell, depending upon the voltage polarity, one or more Conductive Filaments (CFs) made out of oxygen vacancies are either formed or ruptured. Once the CFs are formed inside the metal oxide to bridge the top and bottom electrodes, current can flow through the CFs, and the cell is in a low resistance state. The larger the size of the CFs, the lower the resistance. Conversely, the rupture of the CFs disconnects the top electrode from the bottom electrode, resulting in a high resistance state (HRS) of the cell [16].

According to Equation 1, CF size is directly linked to the voltage across the cell, thus, multilevel LRS can be achieved in the 1T1R RRAM by modulating the SET voltage. In this study, the amplitude of the RESET voltage remains unchanged during the

cell programming operation. As HRS variation is generally larger than LRS variation, multilevel HRS is not considered. The HRS state is associated with an initial value, restored before each SET operation. Finally, to obtain intermediate SET states, programming is done in 2 steps: a RESET operation is first performed to switch the memory cell in its HRS initial state, then a SET operation sets the desired LRS level.

At a circuit level, different implementations of Multi-Level Cell (MLC) operation can be adopted. Multilevel operation can be achieved:

- By applying an increasing number of identical voltage pulses through the SET decoder. In this case, LRS level is function of the number of pulses [17].
- By modulating the gate voltage (WL) of the memory cell Select Transistor to control the SET current [18].
- By modulating directly the SET voltage generated from the SET decoder. In this case, LRS levels are achieved with different SET voltage values.

The first approach requires a pulse generator circuitry while the two others require different voltage levels generated from a specific circuitry. In this study, the third approach is considered as it is the most effective MLC approach for the considered technology. Four different LRS levels are generated by changing the amplitude of the SET voltage. Table 4 presents the different SET and RESET voltages values with the corresponding resistance nominal values.

Table 4: Programming Voltages and corresponding Resistance Values

	SET				RESET
Logical value	**00**	**01**	**10**	**11**	-
$V_{SET/RESET}$ (V)	1.4	1.8	2.2	2.8	-2.8
$R_{ON/OFF}$ (kΩ)	104	77	46	30	196

To assess the robustness of the analog resistance values, a Monte Carlo (MC) analysis is conducted using a 28-nm Fully Depleted Silicon-On-Insulator (FDSOI) technology. As a result, R_{ON} and R_{OFF} distribution spreads are extracted. The programming protocol is comparable to the one used in section 4.2. The memory array programming is done in 2 cycles. First, all memory cells are RESET (with an more important RESET voltage level), then the memory array is SET with specific SET voltage levels.

Figure 13 presents the impact of the memory cell variability on R_{ON} and R_{OFF} distributions after 300 MC simulations. Note that R_{OFF} distribution is much larger than R_{ON} distribution, preventing a HRS MLC implementation. R_{ON} distributions are narrower with a tendency to spread for small SET voltage values. In order to discriminate the different memory states during the read operation, no distribution overlap is permitted.

Figure 14 focuses on R_{ON} distributions and shows distinct R_{ON} resistance levels. Distributions are displayed in box-plot forms. The full range variation (from min to max) of each box-plot confirms the clear separation between each resistance level.

Figure 13 and Figure 14 results are presented for cell variability parameters included in the range ±15% of the standard deviation of a normal distribution (i.e. $\sigma_{Variability}$ = 15%). This variability is consistent with silicon data (see Figure 3) which means that a MLC approach can be considered for this technology.

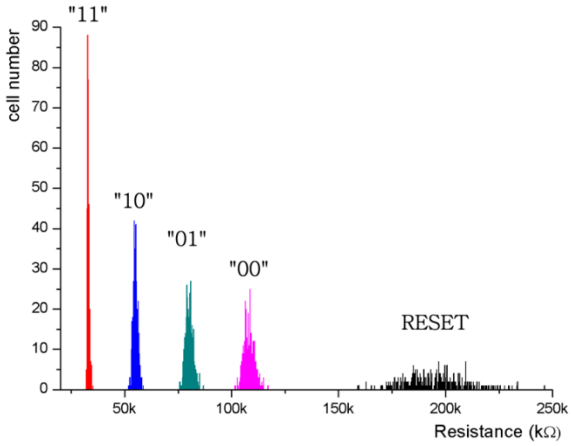


Figure 13. R_{ON} and R_{OFF} distributions versus cell variability ($\sigma_{Variability}$ = 15%) after 300 runs (SET and RESET states)

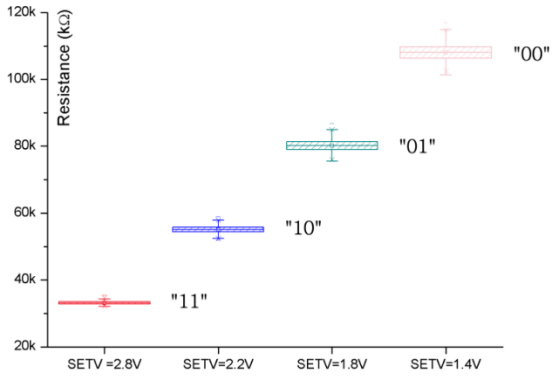


Figure 14. R_{ON} distribution box-plots versus cell variability ($\sigma_{Variability}$ = 15%) after 300 runs (SET states)

To evaluate the robustness of the technology, Figure 15 and Figure 16 results are provided for $\sigma_{Variability}$ = 20% and $\sigma_{Variability}$ = 25%, which is the worst case. An important spreading of R_{ON} is observed. The initial R_{ON} distribution spread increases as variability increase. Even for the worst case ($\sigma_{Variability}$ = 25%), no overlap between LRS distributions is observed which confirms the robustness of MLC approach. In Table 5, R_{ON} (SET) and R_{OFF} (RESET) distribution parameters (mean value and standard deviation) are reported for the different variability parameter values.

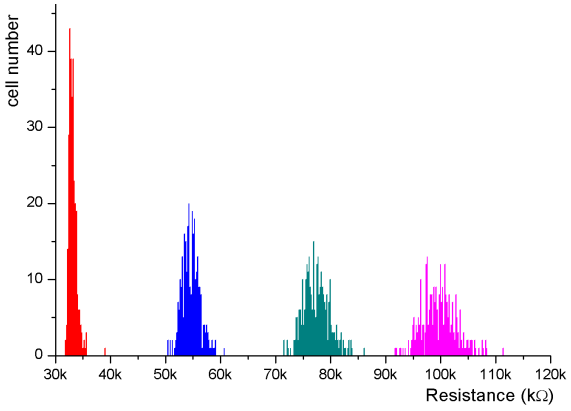


Figure 15. R_{ON} distributions versus cell variability ($\sigma_{Variability}$ = 20%) after 300 runs (SET states)

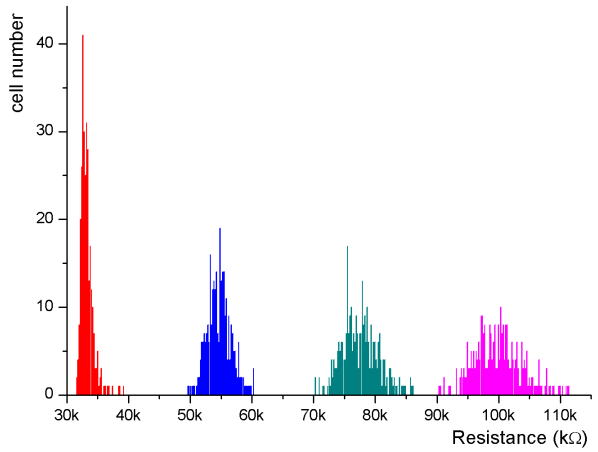


Figure 16. R_{ON} distributions versus cell variability ($\sigma_{Variability}$ = 25%) after 300 runs (SET states)

Table 5: Simulated Configurations

		SET				RESET
$\sigma_{Variability}$	Distributions	**11**	**10**	**01**	**00**	
15%	Mean (kΩ)	33.159	55.111	80.227	108.119	195.350
	σ (kΩ)	0.501	1.153	1.8550	2.5520	14.447
20%	Mean (kΩ)	33.239	54.681	77.402	99.672	196.639
	σ (kΩ)	0.768	1.614	2.408	3.174	19.520
25%	Mean (kΩ)	33.342	54.761	77.529	99.844	198.031
	σ (kΩ)	1.049	1.973	2.978	3.928	24.882

5. Conclusion

Variability combined with voltage drops due to resistive paths present a major challenge for ReRAM memory fabrication process and design engineers. In the proposed study, impact of variability combined with resistive paths is evaluated based on a 1T1R ReRAM elementary memory array. A specific attention is provided to the ReRAM memory word. To assess the technology reliability, the HRS/LRS resistance ratio is extracted at a memory word and at a memory array level to provide a valuable feedback for designers during ReRAM memory array design. At a word level, impact of line resistances combined with variability can affect the memory word operation from a word size greater than 8 bits. At a memory array level, resistance variability in RESET state is much more important than variability in the SET state. Additionally, a stable MLC operation of 2-bits/cell is demonstrated.

References

[1] H. Aziza, P. Canet, J. Postel-Pellerin, M. Moreau, JM. Portal, M. Bocquet, "ReRAM ON/OFF resistance ratio degradation due to line resistance combined with device variability in 28nm FDSOI technology", Ultimate Integration on Silicon (EUROSOI-ULIS), 2017. DOI: 10.1109/ULIS.2017.7962594.

[2] Semiconductor Industry Association, "International Technology Roadmap for Semiconductors".

[3] E. Shahrabi, B. Attarimashalkoubeh, J. Sandrini, and Y. Leblebici, "Towards chip-level reram-cmos co-integration", in International Conference on Memristive Materials, Devices and Systems (MEMRYSIS), 2017. DOI: 10.1109/PRIME.2016.7519497.

[4] H. Aziza, H. Ayari, S. Onkaraiah, M. Moreau, JM. Portal, "Multilevel operation in oxide based resistive RAM with SET voltage modulation", in International Conference on Design and Technology of Integrated Systems in nanoscale Era, 2016. DOI: 10.1109/DTIS.2016.7483892.

[5] Waser, R., Nanoelectronics and Information Technology: John Wiley & Sons, 2012.

[6] Chen, F., Seok, J.Y. and Hwang, C.S., Integration Technology and Cell Design, in Resistive Switching, Wiley-VCH Verlag GmbH & Co. KGaA, 573-596, 2016. DOI: 10.1002/9783527680870.ch20.

[7] Molas, G., et al. "Functionality and reliability of resistive RAM (RRAM) for non-volatile memory applications." 2016 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA). IEEE, 2016. DOI: 10.1109/VLSI-TSA.2016.7480520.

[8] Liang & al., "Effect of wordline/bitline scaling on the performance, energy consumption, and reliability of cross-point memory array", ACM Journal on Emerging Technologies in Computing Systems, Vol. 9, No. 1, 2013. DOI: 10.1145/2422094.2422103.

[9] E.I. Vatajelu et al., "Nonvolatile memories: Present and future challenges", Design & Test Symposium (IDT), 61-66, 2014. DOI: 10.1109/IDT.2014.7038588.

[10] Hyo-Shin Ahn1 et al., "Pairing of cation vacancies and gap-state creation in TiO_2 and HfO_2," Appl. Phys. Lett., Vol. 90, Iss. 25, p. 252908 - 252908-3 (2007). DOI: 10.1063/1.2749858.

[11] C. Lien et al., "The highly scalable and reliable hafnium oxide ReRAM and its future challenges", Solid-State and Integrated Circuit Technology (ICSICT), 2010. DOI: 10.1109/ICSICT.2010.5667553.

[12] M. Bocquet et al., "Compact Modeling Solutions for Oxide-Based Resistive Switching Memories (OxRAM)", Journal of Low Power Electronics and Applications, 4 (1), 1-14, 2014. DOI: 10.1109/FTFC.2013.6577779.

[13] C. Mazure, et al., "FDSOI: From substrate to devices and circuit applications", ESSCIRC, 45-51, 2010. DOI: 10.1109/ESSCIRC.2010.5619767.

[14] F. Nardi, et al, "Resistive switching by voltage-driven ion migration in bipolar RRAM Part I: Experimental study", IEEE Transactions on Electron Devices, 59(9), 2461-2467, 2012. DOI: 10.1109/TED.2012.2202319.

[15] S. Larentis, et al., "Resistive switching by voltage-driven ion migration in bipolar RRAM Part II: Modeling", IEEE Trans. on Electron Devices, 59(9), 2468-2475, 2012. DOI: 10.1109/TED.2012.2202320.

[16] T. Diokh, "Investigation of the Impact of the Oxide Thickness and RESET conditions on Disturb in HfO2-RRAM integrated in a 65nm CMOS Technology" in International Reliability Physics Symposium, 3-6, 2013. DOI: 10.1109/IRPS.2013.6532043.

[17] Y.C. Huang et al., "Using binary resistors to achieve multilevel resistive switching in multilayer NiO/Pt nanowire arrays", NPG Asia Materials, 6 (2), e85, 2014. DOI:10.1038/am.2013.81.

[18] Y. S. Fan et al., "High Endurance and Multilevel Operation in Oxide Semiconductor-Based Resistive RAM Using Thin-Film Transistor as a Selector", ECS Solid State Letters, 4(9), Q41-Q43, 2015. DOI: 10.1149/2.0061508ss.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 13-19 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Technology

ASTESJ

ISSN: 2415-6698

Design of Cognitive Radio Database using Terrain Maps and Validated Propagation Models

Anwar Mohamed Fanan*, Nick Riley, Meftah Mehdawi

School of Engineering, University of Hull, Cottingham Road, Hull, HU6 7RX, United kingdom.

ARTICLE INFO

Article history:
Received: 10 July, 2017
Accepted: 16 August, 2017
Online: 04 September, 2017

Keywords:
Geolocation database
DTED
Propagation Model
TVWS
Cognitive Radio

ABSTRACT

Abstract — Cognitive Radio (CR) encompasses a number of technologies which enable adaptive self-programing of systems at different levels to provide more effective use of the increasingly congested radio spectrum. CRs have potential to use spectrum allocated to TV services, which is not used by the primary user (TV), without causing disruptive interference to licensed users by using appropriate propagation modelling in TV White Spaces (TVWS). In this paper we address two related aspects of channel occupancy prediction for cognitive radio. Firstly, we continue to investigate the best propagation model among three propagation models (Extended-Hata, Davidson-Hata and Egli) for use in the TV band, whilst also finding the optimum terrain data resolution to use (1000, 100 or 30 m). We compare modelled results with measurements taken in randomly-selected locations around Hull UK, using the two comparison criteria of implementation time and accuracy, when used for predicting TVWS system performance. Secondly, we describe how such models can be integrated into a database-driven tool for CR channel selection within the TVWS environment by creating a flexible simulation system for creating a TVWS database.

1. Introduction

As radios in future wireless systems become more flexible and reconfigurable and available radio spectrum becomes scarce, there is the possibility of using TV white space devices (WSDs) as secondary users in the Broadcast Bands without causing harmful interference to licensed incumbents. Currently, one candidate method could be to utilise a geolocation database approach. The white space device should be able to determine available channel opportunities for a given location by accessing a database of TV White Space (TVWS) channels including data on each transmitter and each site, variable channels, transmitter power, and time of validation [1]. Therefore, the TV channel can be protected from harmful interference by accurate prediction of TVWS using an appropriate propagation model. Design of any wireless network depends on accurate prediction of radio propagation, which impacts deployment and management strategies. In this paper we extended the previous work by investigating the best propagation model among three propagation models (Extended-Hata, Davidson-Hata and Egli), using various terrain data resolutions (1000, 100 and 30 m) and comparing with the real measurements taken around Hull UK using two the performance comparison criteria of implementation time and accuracy. Agreement between the measured and predicted values of path loss has investigated, using MATLAB to analyse and compare the variation of path loss between the measured and predicted values. The terrain profile was extracted from terrain database Global1 and then taken into account in selected propagation models. The flexible cognitive TVWS database system was built using different propagation models to calculate available channels in each pixel of the selected area.

2. Digital Terrain Elevation Data (DTED)

DTED was developed by the US Defense Mapping Agency (DMA) and can be used to improve signal detection accuracy. Currently, the paper has selected only three DTED levels of spatial resolution, which are available to the public [2].

Table 1: Resolution levels of DTED.

DTED Level	Post Spacing	Ground Dist	Row x Column
0	30 arcsecond	~ 1 km	121 x 121
1	3.0 arcsecond	~ 100m	1200 x 1200
2	1.0 arcsecond	~ 30 m	3600 x 3600

*Corresponding Author: Anwar Mohamed Fanan, School of Engineering, University of Hull, Cottingham Road, Hull, HU6 7RX, UK,
Email: Anuarfanan@yahoo.com

The elevation data of the resolution from 30 arcsecond (arcsec) to 1 arcsec (level 0,1 and 2) are avaliable for public use and can be downloaded as different DTED extension files in each resolution. The USGS web application has been used to define the desired research region by specifying latitude and longitude as shown in Figure 1, which identifies the required terrain tiles. Figure 2 classifies all tiles, identified by the web application earthexplorer.usgs.gov .

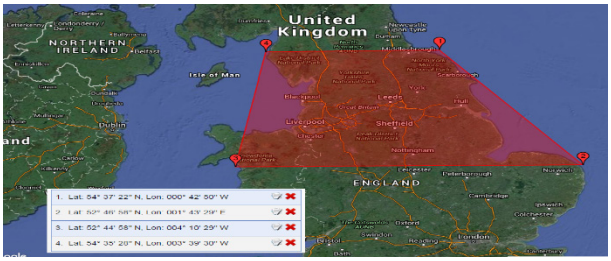


Figure 1. Region selected for terrain data acquisition

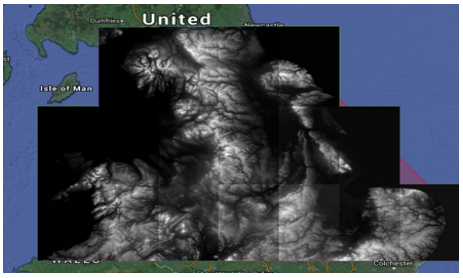


Figure 2. Research region of terrain elevation data by NASA's shuttle radar topography mission (SRTM) with resolution 1 arc sec (30 m)

3. TVWS Geolocation Database

The use of a TV white space geolocation database enables the most effective detection method for prediction of available channels and calculation of TV coverage maps for each pixel in the selected region by using an appropriate propagation model, selected for accuracy and efficiency. The technique can avoid signal detection problems caused by fading effects and shadowing. Construction of the geolocation database requires primary user information including frequency of operation, transmitted power, location, transmission time and height and type of transmit antenna. This information will protect spectrum incumbents from interference from secondary users who will access the database by sending a query to obtain available channels in a given area at a certain time. Furthermore, the geolocation database might have proxy to make queries and identify available channels for WSD [3].

4. Propagation Models

When planning wireless communication systems and designing wireless networks, the accuracy of the prediction of propagation characteristics of each environment should be taken into account. One of the most significant parameters, which can be provided by propagation prediction, is large-scale path loss, which affects directly the coverage of a base station placement and its performance. However, using field measurements to obtain these parameters without depending on propagation models is time-consuming and costly. The following subsections provide a brief explanation of several appropriate empirical propagation models [4] including the Extended Hata, Davidson-Hata and Egli models.

4.1. Extended Hata Model

The Extended Hata model was derived from Hata-Okumura which is widely used for signal prediction in urban areas, wholly based on measured data that have been collected in Tokyo Japan. Also, it does not have any analytical explanation, but includes empirical factors that depend on the type of environment. This model can be used in different environment by adding some correction factors to meet the requirements of ITU-R and also extending range up to 100 km as shown in the following equation:

$$PL_{Ex_Hata} = A + B\log(d)^b + C \quad (1)$$

$$A = 69.55 + 26.16\log_{10} f_c - 13.82\log_{10} h_{tx} - a(h_{rx})$$

$$B = 44.9 - 6.55\log_{10} h_{tx}$$

Where f_c represents the carrier frequency (150 to 1500 MHz), h_{tx} is the height of the base station antenna (m) and h_{rx} is the height of receive antenna (m). The distance from transmitter to receiver is d km. The value of the correction factors a(h_{rx}) and C depend on the type of environment. In small and medium sized cities and metropolitan areas the value of C will be 0, while in other environments such as suburban and rural areas, different equations are used [5]. The factor b denotes an extended range up to 100 km as shown in the following equation.

$$b = \begin{cases} 1, & ,d_{Km} \le 20 \\ 1 + (0.14 + 0.000187 f_c + 0.00107 h_t')(\log_{10} d_{Km}/20))^{0.8} & ,d_{Km} > 20 \end{cases} \quad (2)$$

$$h_t' = \frac{h_t}{1 + 7\times 10^{-6} h_t^{\,2}}$$

4.2. Davidson-Hata Model

$$PL_{David} = PL_{Hata} + A(h_t, d_{km}) - S_1(d_{km}) - S_2(h_t, d_{km}) - S_3(f_{MHz}) - S_4(f_{MHz}, d_{km}). \quad (2)$$

$$PL_{Hata} = 69.55 + 26.16\log_{10} f_{MHz} - 13.82\log_{10} h_t$$

$$- a(h_r) + (44.9 - 6.55\log_{10} h_t)\log_{10} d_{Km} \quad (3)$$

where A is a factor extending distance up to 300 Km, S1 and S2 are correction factors for extending transmitter height up to 2500m, while the factors S3 and S4 extend the frequency range over 30 to1500 MHz [6]. *a(h$_r$)* is a correction factor for receiver antenna height.

4.3. Egli Model

The Egli model was introduced by John Egli in 1957. The model includes a terrain model and was derived from real measurements of UHF and VHF television transmissions conducted in several large cities. It then used a point-to-point model for predicting the total path loss [7]. Thus, this model is commonly used for point to point communications to predict path loss in an urban or rural area, where transmission has to go over an irregular terrain between a fixed transmitter and receiver in the frequency range 40 to 900 MHz. The following equation illustrates the path loss calculation of the Egli model:

$$PL_{Egli} = \begin{cases} 20\log_{10} f_{MHz} + p_0 + 76.3, & h_2 < 10 \\ 20\log_{10} f_{MHz} + p_0 + 83.9, & h_2 < 10 \end{cases} \quad (4)$$

$$p_0 = 40\log_{10} d_{Km} - 20\log_{10} h_t - 10\log_{10} h_r \quad (5)$$

where h_t denotes the height of the transmitter antenna (m), h_r is the height of the receiver antenna (m), the distance between transmitter and receiver is denoted by d km and the transmission frequency is f (MHz) [8].

5. Field Measurement and Data Collection

The main goal of selection of various positions at which to conduct measurements is to examine the signal strength behavior in different environments at various distances from the transmitter and to observe how the terrain affects the received signal. The measurements have been taken at 23 locations distributed randomly around Kingston-upon Hull, UK as shown in Figure 3.



Figure 3. Measurement equipment and geographical location in Hull and surrounding areas.

The measurement equipment used includes an omnidirectional antenna (covering the frequency range 174 to 230 MHz (VHF) and 470 to 790 MHz (UHF) with gain of 3.5 dBi) and spectrum analyser (Agilent E4407B, frequency range 9 kHz to 26.5 GHz) which was connected with a laptop computer by using a general purpose interface bus. A Matlab program on the laptop received raw data and stored them in (bin) files. In addition, the measurement locations were determined by using a mobile GPS application.

6. ANALYSIS of Models' Performance

6.1. Propagation Path Loss Analysis

The main criterion for model assessment is path loss. A simulation program was implemented in Matlab, using channel 33 to conduct the comparison between the three propagation models and the measured results. In order to compare the real measurements with different propagation models, the path loss should be extracted from the real measurements by using the following equation in each location [9].

$$PL = TX + TX_{Gain} + RR_{Gain} - RP \quad (7)$$

Where TX denotes the transmitted power, transmitting antenna gain is represented as TX_{gain}, PL is the path loss, receiving antenna gain is denoted as PR_{gain} and RP is the received power, dBm.

To evaluate the propagation models against real measurements, several parameters might be used to identify the most accurate propagation model. The error between predicted and measured path loss values was calculated by Equation 8 and the average error calculated by Equation 9.

$$e_i = E_{pi} - M_{pi} \quad (8)$$

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i \quad (9)$$

In which e_i is denoted as the difference between the calculated path loss E_{Pi} and measured path loss M_{pi} derived from measured received power in each location.

Equations 8 and 9 are then used to calculate the standard deviation, Equation 10, whilst Root Mean Square Error (RMSE) is calculated by Equation 11, which also depends on the average error calculated in Equation 9.

$$\sigma = \sqrt{\frac{1}{N-1}\left(\sum_{i=1}^{N} e_i^{\,2} - N.(\bar{e})^2\right)} \quad (10)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n} |e_i|^2} \quad (11)$$

To remove the influence of dispersion from the overall error, a significant metric can be derived from RMSE by subtracting the standard deviation of the absolute value of the error to obtain the Spread Corrected Root Mean Square Error (SCRMSE), as illustrated in Equations 12 and 13:

$$E_i = |e_i| - \sigma \quad (12)$$

$$SCRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n} |E_i|^2} \quad (13)$$

6.2. Diffraction Factor Based on Terrain Profile Database

A common phenomenon that exemplifies the wave property of EM waves and light is diffraction, which is the bending of EM waves around obstacles. Diffraction is considered as a non-line of sight (NLOS) propagation mechanism which may occur when the propagation path is obscured by a barrier such as a mountain or hill or man-made obstacles including buildings. Diffraction can

be the cause of significant signal weakness at the reception site, due to the presence of some of the aforementioned barriers between the transmitter and receiver. There are two types of diffraction. "Shadow diffraction", occurs when the received signal is blocked by obstacles and the received field strength will be decreased when the reception site is within the shadowed area. The second case, which occurs if the impediment is underneath the LOS, is called "lit diffraction," and commonly leads to multi-path interference. Shadow diffraction is the one of the main reasons for increased the path-loss. One of the common diffraction models is the single knife edge model, explained by Huygens's Principle, which states that when an electromagnetic wave is obstructed by a natural or man-made obstruction, the obstruction acts as a secondary source for creating a new wavefront which then propagates into the geometric shadow region of the obstruction [10].

6.3. Terrain Profile-based Diffraction Model

One of the main effects of the terrain profile is to cause diffraction or bending of EM waves around obstacles such as mountains, hills or man-made structures which obscure the direct path.

In previous work [11], we considered terrain resolutions of 1 and 30 arcsec and determined that each of them had advantages and disadvantages in terms of accuracy and calculation time.

In this work, we attempt to improve these results by investigating a third resolution value between 1 and 30 arcsec to improve the compromise between accuracy and implementation time. For example, whilst calculating diffraction using the three different resolutions and investigating its effect on the received signal, we noticed that in the location approximately 38 km along the path shown in Figure 4, in the 30 arcsec resolution the elevation value is 100 m, whilst when using 3 arcsec and 1 arcsec resolution, the elevation values are approximately 86 and 83 m respectively.

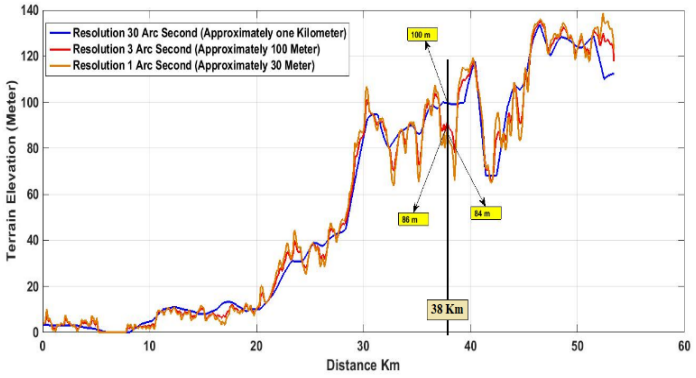
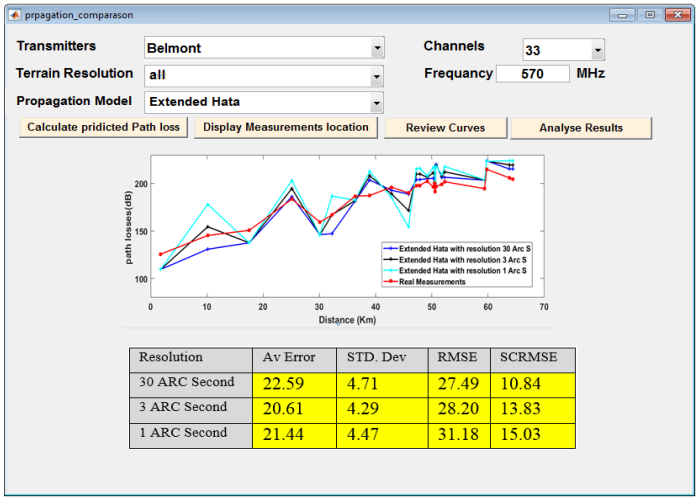


Figure 4. Terrain elevation data for the path from University of Hull to Belmont TV Transmitter in different resolutions

Using 30 arcsec resolution takes a short time for the implementation process but has less accuracy. When using 1 arcsec, we have good accuracy but a long time for the implementation process. However, using 3 arcsec resolution produces the best results in terms of the compromise between accuracy and implementation time.

7. Comparison and Results

The measurement study covered the area around the city of Hull, which was represented to measure the UHF TV band from 470 to 790 MHz with consideration of all radio and TV stations that feed the whole Hull area. Most of the channels transmitted into the area originate from the Belmont and Emley Moor transmitters (see Figure 3). The results of comparison of predicted path loss with measurements for two cases (excluding and including terrain modelling) are presented by using the previously defined criteria average error, standard deviation, RMSE and SCRMSE. Figure 5 shows our Graphical User Interface (GUI) of expected results including path loss curves for each propagation model and a table of calculated parameters.



Resolution	Av Error	STD. Dev	RMSE	SCRMSE
30 ARC Second	22.59	4.71	27.49	10.84
3 ARC Second	20.61	4.29	28.20	13.83
1 ARC Second	21.44	4.47	31.18	15.03

Figure 5. GUI showing propagation model comparison in the selected measurement locations

This analysis may be undertaken for all selected measurement points, by flexible selection of the transmitter name, terrain resolution, propagation model and transmitted channel. Results corresponding to all measurement locations and comparison of the three propagation models with real measurements along with parameter analysis, are discussed and classified in the following sections.

7.1. Influence of Terrain Resolution on Results from the Extended Hata Model.

In this and in the following sections, terrain profile databases with various spatial resolutions and equivalent single knife edge diffraction have been used to calculate the diffraction factors and then evaluate their impact on the performance of the propagation models. The results in Table 2 indicate that 30 arcsec resolution used with the Extended Hata model can be considered the best fit to the measured data with low error when applying the diffraction factor at different terrain resolutions. It can be clearly seen in Figure 6 that the behavior of path loss was influenced by diffraction, when compared with the path loss derived from measured data.

Thus, the propagation behaviour has been affected in most measurement locations when applying the terrain variation with 30 arcsec resolution. The impact of the propagation model is obvious after the third measurement point, where the first three points might be situated within the line of sight and the 1 km resolution results might have missed terrain features situated

along the path which might cause destructive or constructive diffraction. Thus, using 1 km resolution might not give accurate results. Also when using 1 arcsec resolution, the terrain variation becomes worse in this model, which additionally requires a significantly longer processing time. However, the 3 arcsec resolution gives only slightly different results and does not need a long time for calculation of terrain profile.

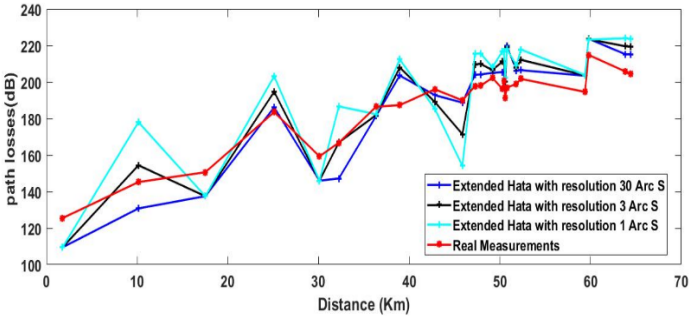


Figure 6. Impact of terrain resolution on results from the Extended Hata model

In Table 2 we can observe how the error statistics have been impacted by the diffraction factor and how the SCRMSE value are decreased in the selected terrain resolution. The results indicate that the 30 arcsec and extended Hata model has the best results of the SCRMSE, at 10.84 dB. On the other hand, the 3 arcsec result is seen to have less error compared with 1 arcsec by about 1.2 dB.

Table 2: Fitted Extended Hata model in different terrain resolution

Resolution, arcsec	Av Error, dB	STD. Dev, dB	RMSE, dB	SCRMSE, dB
30	23.34	4.87	23.52	11.07
3	21.43	4.47	22.82	9.95
1	22.23	4.64	25.02	9.81

7.2. Influence of Terrain Resolution on Egli Model Results.

Due to the nature of the terrain profile near the transmitter sites, which includes rough terrain and hills, the use of 1 arcsec and 3 arcsec resolutions will clearly affect the Egli propagation predictions, as illustrated in Figure 7. Here it may be seen clearly that there are large changes in the path loss at the distance of 48 km and that at other locations such around 52 km there is less variation where the receiver might be in line of sight of the transmitter.

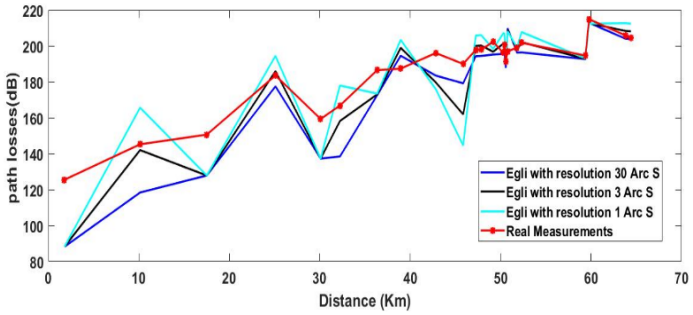


Figure 7. Impact of terrain resolution on the Egli model results

The error of the SCRMSE for the both 1 arcsec and 3 arcsec are slightly different by about 0.14 dB as can be seen in Table 3, whereas the SCRMSE values for 30 arcsec resolution have increased by 1.12 dB.

Table 3: Fitted Egli model in different terrain Resolutions

Resolution, arcsec	Av. Error, dB	STD. Dev, dB	RMSE, dB	SCRMSE, dB
30	22.59	4.71	27.49	10.84
3	20.61	4.29	28.20	13.83
1	21.44	4.47	31.18	15.03

7.3. Influence of Terrain Resolution on Davidson Model Results.

The Davidson model is clearly affected by terrain resolution in a similar manner to the Extended-Hata model at all selected resolution values, as shown in Figure 8.

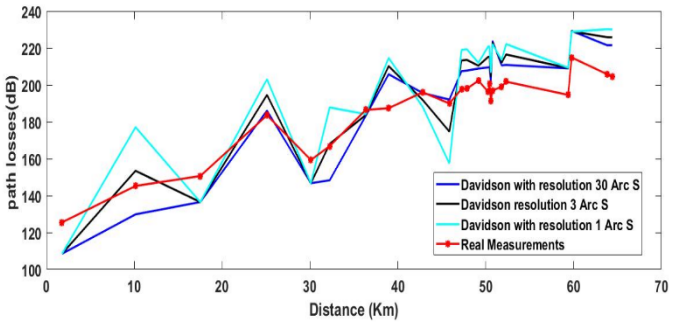


Figure 8. Impact of different terrain resolution on the Davidson model

It can be observed in the statistics of Table 4 that the SCRMSE is increased while the resolution value is decreased as Table 2 of the Extended-Hata behavior.

Table 4: Fitted Davidson model in different terrain Resolutions

Resolution, arcsec	Av Error, dB	STD. Dev, dB	RMSE, dB	SCRMSE, dB
30	24.46	5.10	30.70	12.48
3	22.44	4.68	31.50	15.60
1	22.99	4.79	34.38	16.94

According to the advantage and disadvantage of both previous results in terms of accuracy and processing time, we observed that 30 arcsec has short time and less accuracy, while 1 arcsec has high accuracy and long processing time. Therefore, the 3 arcsec resolution in the Egli model gives the best result when taking these chosen criteria into account when comparing SCRMSE results between different terrain resolutions.

8. Design of Flexible System for Creating TVWS Database by Using Different Propagation Models

Based on the previous results, which indicate that the Egli model is the best among the models that have been chosen for

comparison with the real measurements, a flexible system has been built that performs many functions related to propagation modelling and calculation of signal strength in each pixel. Among these tasks, it is possible to determine any geographic area based on latitude and longitude between two concentric points. In addition, it can be determined that the size of each pixel will affect the implementation time and propagation accuracy. Also, the system can perform three major operations at the same time to create a database for a selected geographic region that can be easily used when connected to the white space devices (WSD). In this work, for illustration, we considered only six transmitters, but more can be added using the "Add Transmitter Detail" button, as shown in Figure 9.

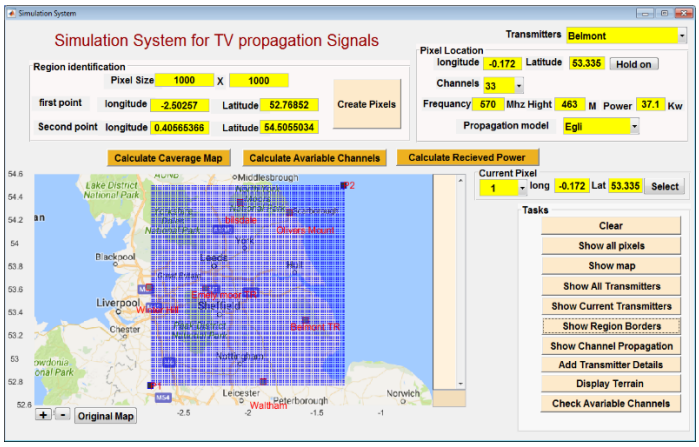


Figure 9. Display of all pixels in the selected region in the flexible simulation system for creating TVWS database.

8.1. Methodology for Calculation of Received Power

The second methodology to be implemented after creating the pixel file is to calculate the receiver power in each pixel in the frequency range 470 to 790 MHz, by considering the selected transmitters. The processing time depends on the number of pixels, propagation model and also the terrain resolution level. The process will be conducted only once to create a complete database of all the predicted TV signals in each pixel, as shown in Figure 10, which can be used for the next stages.

8.2. Methodology for Calculation of Available Channels

The main goal of the system is to calculate available channels with high accuracy and then store all available channels of each pixel in the database, in a way which makes it easy to retrieve the data from WSDs. All of the transmitter information, such as height, channels and transmitted power, is stored previously in the database. The process takes into account all channels of the selected transmitters that might be received in a specific pixel, considering the weak signals as well.

8.3. Methodology for Calculation of Coverage Map

This methodology must be used to translate the database that has been stored to show as a visual map of different levels of signal strength in the selected region for each transmitter, which are then stored in different files in the database as shown in Figure 11.

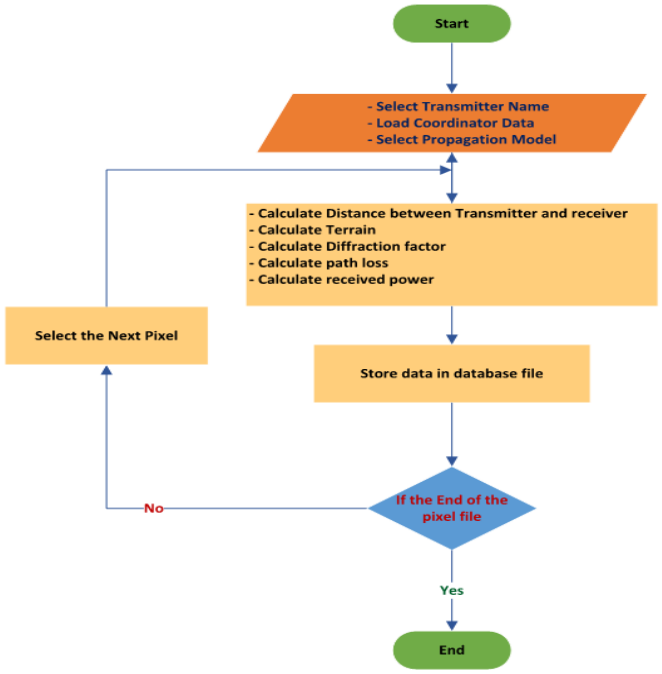


Figure 10. Algorithm of the received power calculation

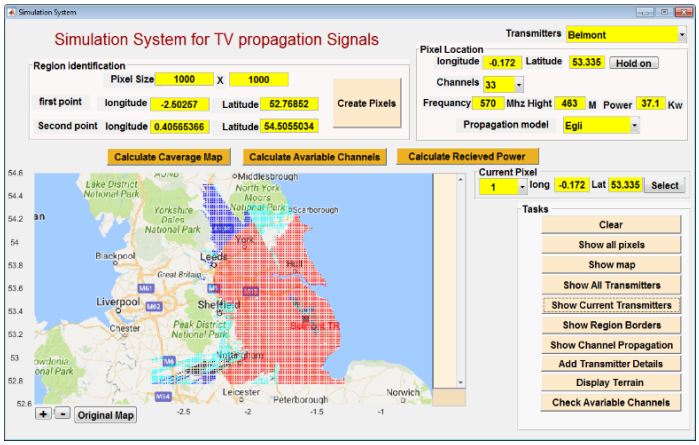


Figure.11 Display of the propagation signals for channel 33 using Egli model.

9. Conclusion

In this paper, TV signal strengths are calculated using various propagation models and then compared with real measurements that have been conducted in various locations. Using a single knife edge model to calculate the diffraction factor with consideration of terrain profile data at different resolutions, we investigate and prove how the terrain data resolution impacts the accuracy and implementation time of the propagation models. We have improved and extended results that have been published in our conference paper in 2016 [11]. RMSE and SCRMSE are the main criteria taken into account to assess the propagation models in different terrain resolutions. The results show that the Egli model still gives the best results when account is taken of the terrain profile data at a resolution of 3 arcsec (100m), providing SCRMSE of 0.14 dB, with shorter computation time and similar accuracy as compared with 1 arcsec. On the other hand, RMSE

and SCRMSE values of the extended Hata and Davidson models are still increasing and have poor performance when terrain data resolution is decreased. Therefore, 3 arcsec is considered the best resolution that can be used to calculate diffraction factor. The results also show that the Egli model is the best model giving a consistently good fit to measured data among other selected models and, with appropriate terrain data, will provide useful input to a system for facilitation of the cognitive radio decision process. In addition, the main benefits for designing the flexible system is to create a TVWS database for a specific area, by selecting the optimum pixel size, adding appropriate transmitter information and choosing a suitable propagation model. In future work, the system will be developed to use additional propagation models at various terrain data resolutions, providing a clear understanding of the differing results between propagation models taking into account the terrain resolution.

Acknowledgment

The lead author is grateful to his supervisor for his unlimited supporting to develop this study, by providing the suitable solutions and suggestions.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Gurney, D., Buchwald, G., Ecklund, L., Kuffner, S.L. and Grosspietsch, J., 2008, October. Geolocation database techniques for incumbent protection in the TV white space. In New Frontiers in Dynamic Spectrum Access Networks, 2008. DySPAN 2008. 3rd IEEE Symposium on (pp. 1-9). IEEE.

[2] National Imagery and Mapping Agency, "Digital Terrain Elevation Data," 30 May, 1990. 'www.globalsecurity.org/ intell/systems/dted.htm.

[3] Nekovee, M., 2009, October. A survey of cognitive radio access to TV white spaces. In 2009 International Conference on Ultra Modern Telecommunications & Workshops (pp. 1-8). IEEE.

[4] Iskander, M.F. and Yun, Z., 2002. Propagation prediction models for wireless communication systems. Microwave Theory and Techniques, IEEE Transactions on, 50(3), pp.662-673.

[5] P. Pardeep, P. Kumar and B. S. Rana, "Performance evaluation of different path loss models for broadcasting applications," *American Journal of Engineering Research (AJER)* , 3(4), pp.335-342, 2014.

[6] K. Stylianos, P. I. Lazaridis, Z. D. Zaharis, A. Bizopoulos, S. Zettas, and J. Cosmas, "Comparison of Longley-Rice, ITU-R P. 1546 and Hata-Davidson propagation models for DVB-T coverage prediction, " In BMSB, pp. 1-4, 2014.

[7] J. J. Egli, "Radio propagation above 40 MC over irregular terrain. Proceedings of the IRE, " , 45(10), pp.1383-1391,1957.

[8] M. Hope, A. B. Bagula, M. Zennaro, and G. Lusilao-Zodi, "On the impact of propagation models on TV white spaces measurements in Africa," *International Conference on. IEEE*, pp.148-154, May 2015.

[9] P. Prajesh, and R. K. Singh, "Investigation of outdoor path loss models for wireless communication in Bhuj," *International Journal of Electronics and Communication Engineering & Technology (IJECET),* Volume 3, Issue 2, pp. 171-178, September 2012.

[10] J. You-Cheol, "Diffraction analysis and tactical applications of signal propagation over rough Terrain," *Air Force INST of Tech Wright-Patterson AFB OH School of Engineering*, no. 97J-01, Dec 1997.

[11] A. M. Fanan, N. G. Riley, M. Mehdawi, and M. Ammar, 2016, November. Comparison of propagation models with real measurement around Hull, UK. In Telecommunications Forum (TELFOR), 2016 24th (pp. 1-4). IEEE.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 18-24 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Building an Efficient Alert Management Model for Intrusion Detection Systems

El mostapha Chakir[*,1], Mohamed Moughit[2,3], Youness Idrissi Khamlichi[4]

[1]*IR2M Laboratory, FST, Univ Hassan 1, Settat, Morocco*

[2]*IR2M Laboratory, ENSA, Univ Hassan 1, Settat, Morocco*

[3]*EEA&TI Laboratory, FST, Univ Hassan 2, Mohammedia, Morocco*

[4]*LERS Laboratory, ENSA, Univ Sidi Mohamed Ben Abdellah, FES, Morocco*

ARTICLE INFO

Article history:
Received: 14 November, 2017
Accepted: 13 December, 2017
Online: 18 January, 2018

Keywords:
Intrusion detection system,
False positive
Risk Assessment
Alerts prioritization
KDD'99
Pattern matching

ABSTRACT

This paper is an extension of work originally presented in WITS-2017 CONF. We extend our previous works by improving the Risk calculation formula, and risk assessment of an alert cluster instead of every single alert. Also, we presented the initial results of the implementation of our model based on risk assessment and alerts prioritization. The idea focuses on a new approach to estimate the risk of each alert and a cluster of alerts. This approach uses indicators such as priority, reliability and asset value as decision factors to calculate alert's risk. The objective is to determine the impact of alerts generated by Intrusion detection system (IDS) on the security status of an information system, and also improve the detection of intrusions using snort IDS by classifying the most critical alerts by their levels of risk. Thus, only alerts that present a real threat will be displayed to the security administrator. The implementation of this approach will reduce the number of false alerts and improve the performance of the IDS.

1. Introduction

This paper is an extension of work originally presented in International Conference on Wireless Technologies, Embedded and Intelligent Systems, EEE WITS-2017 [1]. Based on our work in [2], the goal is to improve the intrusion detection system (IDS) with a Risk Assessment method that can help to prioritize the generated alerts by their importance.

IDS systems generate enormous number of alerts. Often, there are duplicative events from various systems, and other alerts that could be characterized as noise (False Positive). A False Positive is normal events being classified as attacks. This is a major problem for many organizations [3,4]. An attack may in fact be happening, and the network administrator needs to be able to properly identify it, thus, he need to analyze each IDS alert manually, whether it is a false or true positive. So, it is a quite time consuming. Since the number of false positives is high, so alerts of real attacks are hidden among them. The optimal way to deal with this problem is to use an IDS solution that has the ability to prioritize alerts, calculate the risk of each one and correlate them, thereby to help the network administrator focus the efforts on detecting actual threats [5-7]. Therefore, an automated operation is needed to handle alerts and give a good perdition to the security administrator.

In this work, we propose the new Risk Assessment model as a method of prioritizing alerts according to the risk level of each one, and also evaluate the risk of a cluster of alerts. The risk is evaluated as a combination of certain parameters extracted from each alert.

This paper is organized as follows: Section 2 discusses the related work in risk assessment and alert prioritization; Section 3 presents the proposed model and the indicators that are used to support it; Section 4 discusses the implementation of the model and the experimental results to validate the proposed approach; and finally, in Section 5, we summarize the conclusions derived from this work and indicate possible future works.

2. Related Works

Intrusion detection system has an important role in the security and perseverance of active defense system against intruder attacks. They evolved from packet sniffers, used as a network troubleshooting tool to locate malfunctioning equipment and

[*]Corresponding Author: El Mostapha Chakir, IR2M Laboratory, FST, Univ Hassan 1, Settat, Morocco | Email: e.chakir@uhp.ac.ma

software by creating logs showing the activity of network packets [8,9]. Prior to the advent of network switches, IDS products could be connected to any port on a network hub and had a good chance of monitoring network packets on a local area network segment. Many researchers have proposed and implemented various models for IDS, but they often generate too many false positives due to their simplistic analysis [10].

Attacks are presented to a security administrator as alerts generated by IDSs. An IDS generates a large number of alerts and with this large number, security administrators are overwhelmed and it becomes difficult to manually distinguish between the real attacks and the false ones. To deal with this problem, two solutions have been proposed. The first one focuses on the monitoring device by enhancing its detection mechanism, optimizing its signatures, and choosing the right location [11]. Although this solution promises to reduce the number of alerts, it requires prior security administrator knowledge of detection mechanism. The second solution focuses on the sensor's outputs. Several IDS alert management techniques fall into this category and include aggregation, clustering, correlation and fusion [12].

Generally, reducing the number of false positives and prioritizing the most critical attacks are the main objectives of IDS alert management approaches. Furthermore, these techniques help the security administrators in understanding the situation revealed by the IDS.

In our previous work [2] we presented a new model to handle Intrusion detection system alerts based on a stateful pattern matching algorithm. In this paper we improve that model by proposing a new method based on risk assessment and alert prioritization using parameters extracted from alerts generated by IDS.

Risk assessment is the process of identifying security risks related to a system and determining their probability of occurrence, their impact, and the safeguards that would mitigate that impact [13]. Risks can be defined as the potential that a given threat will exploit vulnerabilities of an asset or group of assets to cause loss or damage to the assets. The main objective of risk assessment is to define appropriate controls for reducing or eliminating those risks.

Researchers have been proposed many approaches to prioritize alerts. In [13], the authors proposed a model that estimates the risk index for every single alert based upon indicators and input obtained from asset environments and attributes within the alerts themselves. The model uses a combination of two decision factors: impact on asset and likelihood of threat and vulnerability.

In [14], the authors proposed a method that evaluates IDS alerts based on a number of criteria. They used a Fuzzy Logic Inference mechanism in order to score alerts and developed a rescoring technique that enabled them to rescore alerts to show the early steps of the attackers, and applied their approach to the alerts generated by scanning DARPA 2000 LLDOS 1.0 dataset and they successfully prioritized the most critical alerts along with their preparation steps.

In [15], the authors applied a fuzzy system approach in accessing relative potential risks by associating potential vulnerabilities like the National Vulnerability Database (NVD) and Common Vulnerabilities and Exposures (CVE) with computer network assets.

In [16], the authors estimated risks by associating three criteria: computer network assets, attacks and vulnerabilities.

In [17], the authors proposed a model called "M-Correlator", an alert ranking system based on the likelihood of attacks to succeed, the value of targeted assets and the amount of interest in the type of attack.

All discussed approaches have the ability to priorities alerts, but they also have limitations, especially in the technical aspects of the methods adopted. They used multiple factors to estimate the risk, but do not consider different weightings based upon the importance of different decision factors.

3. The proposed Model

3.1. Overview

In [1] and [2] we have proposed the New System Alert Management for IDSs based on a stateful pattern matching algorithm, which can classify alerts by their importance and reduce number of false positives considerably. In order to improve our system, we propose in this paper more efficient method to prioritize alerts generated by IDSs by evaluating each risk. We assess the risk as a composition of indicators extracted from alerts itself and target assets, and then apply these results of the risk assessment to filter alerts produced by the IDS as **High Risk, Medium Risk** or **Low Risk.**

In our proposed system (Figure 1), we've used binary traffics files of KDD'99 which is used in our previous work. Snort [18] is used to produce alerts of KDD'99 dataset network traffics. Snort is an open source signature based IDS which gets KDD'99 online traffic and then generates alert log files; these files are entered into our proposed system as the inputs. A pattern matching algorithm is used to filter alerts and classify them to different form.

3.2. System architecture

Our approach encompasses three phases for processing events: **Pre-processing phase, Collection phase** and **Post- processing phase**. Each phase provides a level of abstraction to the following one. Figure 1 shows the three abovementioned units.

Pre-processing Phase: In this phase Snort analyzes KDD'99 binary traffic and generates alert files. These alert files are entered in our proposed model as Inputs.

Collection phase: We call this phase also "aggregation and normalization phase", For all data are received from Snort at one location. Aggregation aims at unify alerts in a single format on just one location. Normalization requires a parser familiar with the types and formats of generated alert from Snort after processing them. Snort list files contain information about all packets in KDD'99 dataset [19]. Using this method, we will be able to observe all alerts in the same format.

Post-processing phase: In this phase, once we have all the data in one place, we can implement mechanisms that will improve detection sensitivity and reliability.

In our model, we use three post-processing methods:

- **Classification and Filtering:** In this unit we extract the needed information, such as: Date, IP source, IP Destination, Attack name, etc., and we store them into a database. This information is extracted by parsing the alert file using Perl and regular expression as we will see in section 4.
- **Risk assessment:** Each alert is evaluated in relation to its attack type, and the target host. Several parameters make it possible to qualify the level of danger (Risk) of an alert. It is important to understand their significance in order to be able to manage correctly the alarms according to their level of importance (Table 1).
- **Prioritization:** We prioritize alerts received automatically after calculating the Risk. The priority of an alert depends on the architecture and inventory of the organization's devices. Prioritization is one of the most important steps in filtering alerts received from the Snort output. It means the evaluation of an alert's importance in relation to the organization's environment.

To calculate the Risk, we use the parameters described in Table 1. Each parameter has a value. For the Alert Priority and Alert Reliability, these values are stored in a MariaDB database. For device values, the security administrator must add all the organization's devices, including Servers, Firewalls, Switches, Access Points, Network Printers, etc., and must assign to each equipment a value between 1 and 5 according to the value and the criticality of the device. On the other hand, the other parameters are related to each type or classification of attack by Snort, these values are stored… in a MariaDB database for later use.

We calculate the risk using the three previous indicators by the following formula (3.7).

Alert Priority (P) = {1-5} **(3.1)**

Device Value (D) = {1-5} **(3.2)**

Alert Reliability (R) = {1-10} **(3.3)**

$$\textbf{Risk Assessment(RA)} = \frac{(\mathbf{P}) * (\mathbf{D}) * (\mathbf{R})}{\mathbf{X}} \quad \textbf{(3.4)}$$

The risk should be between 1 and 10 as we will see in Table 2, so the X value and obtained by calculating the Risk using the Maximum Value of each parameter, for example:

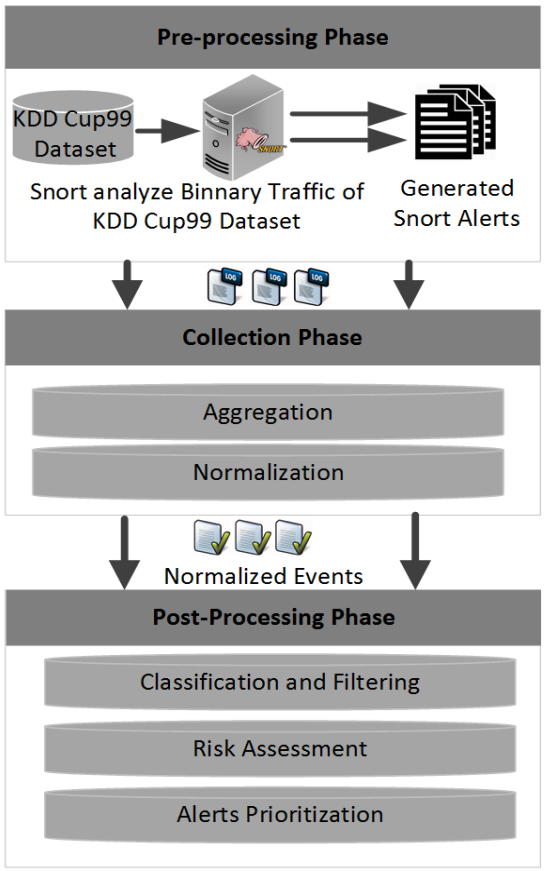


Figure 1: Proposed alert management system using real-time risk assessment and alert prioritization

Table 1. Description of parameters to assess the Risk of Alerts

Parameter	Description
Priority	Priority defines the order in which the action should be taken. A classification type rule assigns default priority that may be overridden with a priority rule [18]. In our model, we categorized priority into three types: **This should be Low = 1-2; Medium = 2-3; High = 4-5** These values are associated with each classification type of Snort IDS, and is stored into a MariaDB Database.
The value of the Destination Device associated with the event	This is a value to define the importance of a machine on the network. A DNS or Web server are more valuable resources for an organization than a network printer. As we will see later, these specifications will be taken into account when calculating the risk of each alert. This value must be between 1 and 5 (**1 = machine less important, 5 = very important machine**). This value is stored into a MariaDB Database for each device of the organization.
The likelihood that the event will occur **Reliability**	In terms of the risk, this parameter could be called "Reliability". This is defined for each independent event; an event may be a set of many alerts. The term reliability can be translated by the reliability that an event is not a *false positive*. The value of this parameter is between 1 and 10 **(equivalent to 1% =** this is a false **positive** and **100% =** it is not a false **positive).** This value is stored into a MariaDB Database and it is associated with an independent type of event (Alert Classification [18]).

Max(P)=5, Max(D)=5 and Max(R)=10

$$\mathbf{RiskAssessment(RA)} = \frac{\mathbf{5 * 5 * 10}}{\mathbf{X}} = \mathbf{10} \quad \textbf{(3.5)}$$

$$\mathbf{X} = \frac{\mathbf{250}}{\mathbf{10}} = \mathbf{25} \quad \textbf{(3.6)}$$

$$\mathbf{Risk\ Assessment(RA)} = \frac{\mathbf{(P) * (D) * (R)}}{\mathbf{25}} \quad \textbf{(3.7)}$$

The proposed model estimates the Risk for each alert. As we see in Table 1, the model uses a combination of three decision parameters.

Using the Risk Assessment, the Total Risk of an attack can be calculated according to the alert rate. Latter is calculated using formula (3.8) presented in [2], the Total Risk of an Attack (TRA) is used to evaluate the Risk of an attack type in a Meta-Alerts. Meta-Alerts can be generated for the clusters that contain all relevant information whereas the amount of data (i.e. alerts), for example all alerts to a specific host with the same attack type, or all alerts from the same source to the same destination and with the same attack type, etc., thus, the TRA is calculated using formula (3.9).

$$\mathbf{AR} = \frac{\mathbf{Number\ of\ Alerts\ byAttack}}{\mathbf{Tolal\ Alerts}} * \mathbf{100} \quad \textbf{(3.8)}$$

$$\mathbf{TRA} = \frac{\mathbf{(RA) * (AR)}}{\mathbf{100}} * \mathbf{10} \quad \textbf{(3.9)}$$

The resulting value can be mapped to the following Risk Categories, Table 2:

Table 2. Risk Assessment Categories

Risk Value	Signification
1-4	Low
5-7	Medium
8-10	High

3.3. Proposed Algorithm

The Algorithm can be explained as follows:

Algorithm: Filtering, classifying and prioritizing alerts according to the Risk Level

Input: Snort Log File (Generated by analyzing KDD'99 Dataset)
Output : Alerts with high Risk

1. Initializes the program
2. Processes the configuration and log files
3. While the number of alerts in log file is not reached
4. Extracts and records details of each alert into database
5. Correlate and classify alerts into many classes (attacks types)
6. Calculate the Alert Rate
7. For each alert in log file
8. Calculate the Risk of alert using (3.7)
9. End For
10. For each Meta-alert in log file
11. Calculate the Total Risk on an Attack TRA of Meta-Alerts using (3.9)
12. End For
13. Prioritize Alerts according to the Risk Assessment
14. Generate alarms if the Risk ≥ 70 %
15. End While

A basic flowchart diagram for the proposed algorithm is shown below (Figure 2):

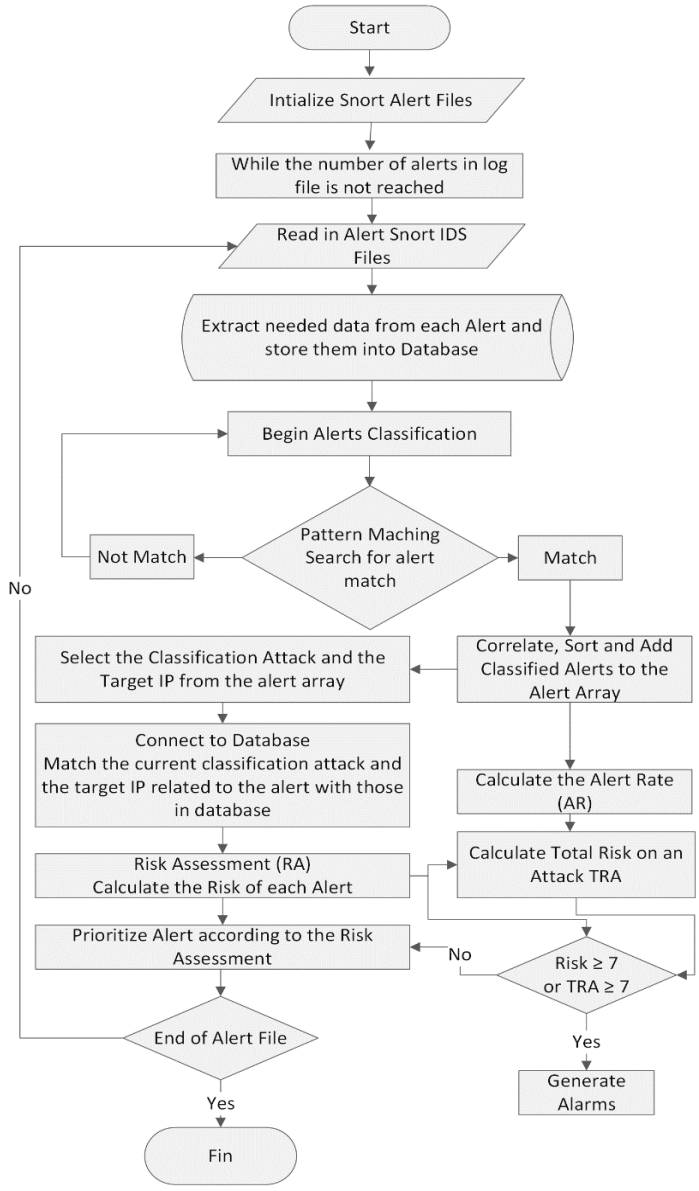


Figure 2: Flowchart of the proposed approach

4. Implementation and results

In order to test the effectiveness of our proposed model, we had it implemented and tested against the KDD'99 dataset. We also used Snort to scan the binary traffic of the dataset. Alerts generated by Snort were analyzed by our model using a pattern matching algorithm. To examine each alert, we wrote a Perl program using regular expression that parse the alerts and extract specific

parameters such as source IP, destination IP, attack name, priority etc... Finally, we use the formula (3.7) to assess the Risk of each alert and (3.9) to evaluate the Total Risk of attacks.

The example of alert below is just one sample of many alerts that we can find in one alert file. In this alert we can find many useful data, such as the Attack Name, Attack Classification, Target IP, Destination IP etc., to extract this useful information we use the Pattern matching rule below using Regular Expression (RE).

An example of alert to deal:

```
[**] [3:19187:7] PROTOCOL-DNS TMG Firewall Client long host entry exploit attempt [**] [Classification: Attempted User Privilege Gain] [Priority: 5] 03/19-16:01:43.762260 10.0.0.254:53 -> 172.16.2.11:1575 UDP TTL:64 TOS:0x0 ID:0 IpLen:20 DgmLen:201 DF Len: 173
```

These variables are stored into a MariaDB Database, thereby we can calculate the Risk for each generated Alert using these parameters and the formula (3.7), after that, using (3.9) we estimate the TRA.

Tables 3 and 4 below presents our experiment results using the output of Snort IDS and KDD'99 Dataset (with 99503 Alerts) that contains different 34 attempted attacks, 1056 Source IP, 485 Destination IP.

The Associated Pattern Matching rule:

```
^(?P<delimeter1>\[\*\*\])\s(?P<SigID_Rev>\[[0-
9:]+\])\s(?P<Attack_Name>[^\"]+)\s+(?P<delimeter2>\[\*\*\])\s\[Classification\:(?P<Attack_Class>[^\]]*)\]\s\[Priority\:(?P<Attack_Priority>[^\]]*)\]\s+(?P<Date>\d+\/\d+\-\d+\:\d+\:\d+)\.\d+\s(?P<SrcIP>\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})\:(?P<SrcPORT>\d+)\s\->\s(?P<DstIP>\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})\:(?P<DstPORT>\d+)\s(?P<Protocol>\w+)\s+[^\]]*.*$
```

Extracting Data to variables:

```
Attack_Name = PROTOCOL-DNS TMG Firewall Client long host entry exploit attempt
Attack_Class = Attempted User Privilege Gain
Attack_Priority =
SrcIP = 10.0.0.254
DstIP = 172.16.2.11
DstPORT=53
Protocol= UDP
```

Tables 3. Classifying alerts by Detection Rate

Attack Name	Nr. of Events	Alert Rate %
(http_inspect) NO CONTENT-LENGTH OR+E3:E19 TRANSFER-ENCODING IN HTTP RESPONSE	45600	45.82 %
(http_inspect) INVALID CONTENT-LENGTH OR CHUNK SIZE	30524	30.67 %
(spp_sdf) SDF Combination Alert	12356	12.41 %
Consecutive TCP small segments exceeding threshold	7052	7.08 %
(http_inspect) UNESCAPED SPACE IN HTTP URI	1159	1.16 %
(http_inspect) LONG HEADER	809	0.81 %
(http_inspect) SIMPLE REQUEST	640	0.64 %
ET CHAT IRC PRIVMSG command	425	0.42 %
ET CHAT IRC PING command	276	0.28 %
ET CHAT IRC PONG response	127	0.12 %
ET CHAT IRC USER command	101	0.10 %
ET CHAT IRC NICK command	99	0.10 %
ET CHAT IRC JOIN command	84	0.08 %
ET POLICY Outbound Multiple Non-SMTP Server Emails	74	0.07 %
Reset outside window	35	0.03 %
(http_inspect) UNKNOWN METHOD	29	0.03 %
(ftp_telnet) FTP bounce attempt	16	0.01 %
ET SCAN Potential SSH Scan OUTBOUND	15	0.01 %
(http_inspect) NO CONTENT-LENGTH OR+E3:E19 TRANSFER-ENCODING IN HTTP RESPONSE	12	0.01 %
(http_inspect) INVALID CONTENT-LENGTH OR CHUNK SIZE	11	0.01 %
(spp_sdf) SDF Combination Alert	11	0.01 %
Consecutive TCP small segments exceeding threshold	11	0.01 %
(http_inspect) UNESCAPED SPACE IN HTTP URI	8	0.01 %
(http_inspect) LONG HEADER	6	0.00 %
(http_inspect) SIMPLE REQUEST	5	0.00 %
ET CHAT IRC PRIVMSG command	5	0.00 %
ET CHAT IRC PING command	3	0.00 %
ET CHAT IRC PONG response	3	0.00 %
ET CHAT IRC USER command	3	0.00 %
ET CHAT IRC NICK command	2	0.00 %
ET CHAT IRC JOIN command	2	0.00 %

Our platform for the experiment are described as follows:

- Processor: Intel (R) Core (TM) i7-6500U CPU @ 2.50GHZ 2.59 GHZ.
- Memory: 4 GB
- System (OS): Linux Ubuntu Server 16.04 64-bit

Table 3 present the classification of alerts according to the alert Rate; Alert Rate is measured by calculating the number of alert for each attack [2].

Table 4 shows the process of evaluating the risk of alerts. After classifying them into three categories, the security administrator will see just alerts with medium and high risk. The rest with low risk can be considered as false positives, this is related to the values of the target hosts and the reliability and priority of the attack. As we can see in Table 4, the risk is evaluated for each alert. For example, in the first row in the table, we notice that the attack "**(http_inspect) NO CONTENT-LENGTH OR+E3:E19 TRANSFER-ENCODING IN HTTP RESPONSE**", has been generated **45600** times to the target "**192.168.11.52**". The **RA** is **10** which is **High**. It is evaluated using the three parameters related to the alert: the priority, the reliability and the device target value. In other ways, the Total Risk of an Attack is evaluated according to the number of events of each attack in relation to the Alert Rate of this Attack.

The specific and complex characteristics of the network system environment make the implementation of Intrusion Detection System more difficult with the multitude of alerts and the huge number of false positives. Therefore, the new approach for detection and analysis of malicious activities is needed in order to check the effectiveness of the current security controls that protect information data.

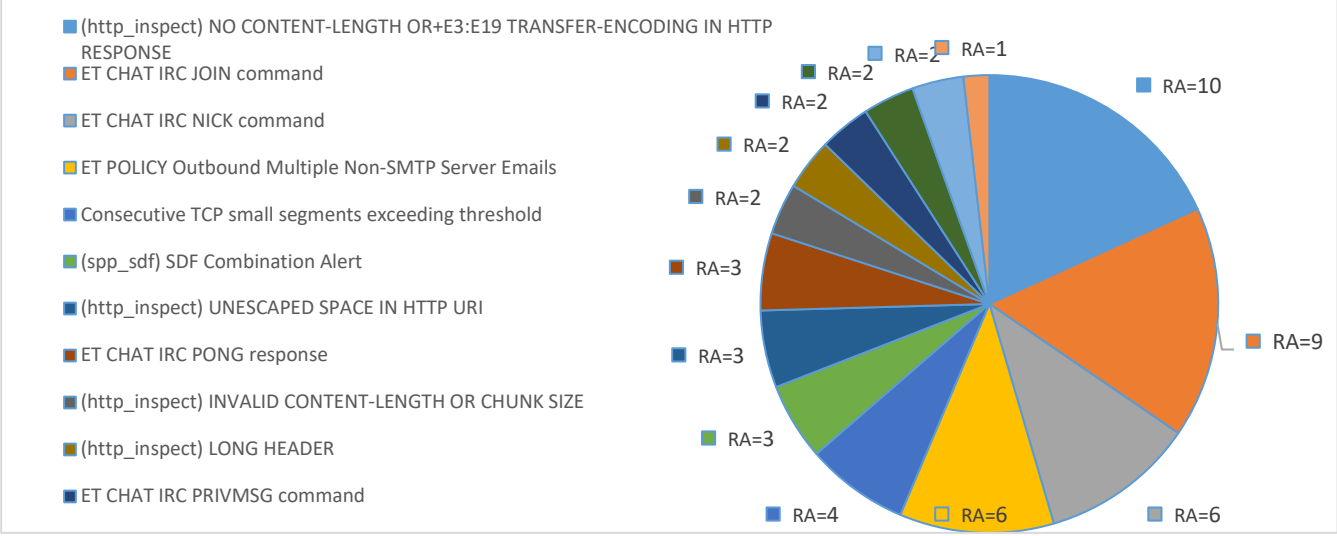


Figure 3: Alerts prioritization after risk assessment

Table 4. Risk Assessment according to the alert and target host parameters

No. of events	To Target	Attack Name	P	R	D	RA	RA Ranking	AR %	TRA %
45600	192.168.11.52	(http_inspect) NO CONTENT-LENGTH OR+E3:E19 TRANSFER-ENCODING IN HTTP RESPONSE	5	10	5	10	**High**	45.82 %	**4.6%**
30524	192.168.2.12	(http_inspect) INVALID CONTENT-LENGTH OR CHUNK SIZE	3	5	4	2	**Low**	30.67 %	**0.6%**
12356	192.168.11.5	(spp_sdf) SDF Combination Alert	3	5	5	3	**Low**	12.41 %	**0.4%**
7052	172.16.2.56	Consecutive TCP small segments exceeding threshold	3	7	5	4	**Low**	7.08 %	**0.3%**
1159	192.168.2.100	(http_inspect) UNESCAPED SPACE IN HTTP URI	3	7	3	3	**Low**	1.16 %	**0.03%**
809	172.16.16.123	(http_inspect) LONG HEADER	3	5	3	2	**Low**	0.81 %	**0.01%**
640	172.16.16.52	(http_inspect) SIMPLE REQUEST	3	5	2	1	**Low**	0.64 %	**0.06%**
425	192.168.11.101	ET CHAT IRC PRIVMSG command	3	5	4	2	**Low**	0.42 %	**0.008%**
276	192.168.2.58	ET CHAT IRC PING command	3	5	4	2	**Low**	0.28 %	**0.005%**
127	10.10.101.2	ET CHAT IRC PONG response	3	7	4	3	**Low**	0.12 %	**0.003%**
101	10.222.1.1	ET CHAT IRC USER command	3	5	4	2	**Low**	0.10 %	**0.002%**
99	10.22.1.231	ET CHAT IRC NICK command	5	8	5	6	**Medium**	0.10 %	**0.006%**
84	172.16.2.112	ET CHAT IRC JOIN command	5	9	5	9	**High**	0.08 %	**0.007%**
74	172.16.16.52	ET POLICY Outbound Multiple Non-SMTP Server Emails	5	8	4	6	**Medium**	0.07 %	**0.004%**

In this work we demonstrated how the implementation of risk assessment reduces the number of false positive. With such an approach, the security network administrator will see just the alerts with a medium and high-risk level that presents a real threat to the organization. The rest of alerts will be considered as a false positive and will not be sent as alarms. Thus, the network security administrator can check the effectiveness of the current security controls that protect the organization's assets (Figure 3). In the next step of this work, we will set up a knowledge base for all the false positives to be compared further and to see if they are indeed.

5. Conclusion and future works

A novel approach that evaluates intrusion detection system alerts using a new risk assessment and alert prioritization is presented. We proposed a model that recognizes and analyzes malicious actions by calculating the risk related to attack pattern and qualify the level of dangerousness of each attack, thus prioritizing alerts generated by IDS. The implementation has demonstrated the efficiency of our approach in both decreasing the huge number of false positives that can reach over 95 % of alerts in the usual cases with a normal IDS, using our model we can control the rate of false positives; thus, we increase the effectiveness of the IDS system.

In the next step of this work, we will focus on the implementation of our approach with other IDSs. Moreover, we will improve our model by using advanced functions as well as more sophisticated algorithms such as machine learning algorithms to classify the attacks according to their dangerousness.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] E.M. Chakir, M.Moughit, Y.I. Khamlichi, "An efficient method for evaluating alerts of Intrusion Detection Systems" in 2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), 2017. Pages: 1 - 6, https://doi.org/10.1109/WITS.2017.7934678

[2] E.M. Chakir, Y.I. Khamlichi, M.Moughit, "Handling alerts for intrusion detection system using stateful pattern matching" in 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), 2016, Pages: 139 - 144, https://doi.org/10.1109/CIST.2016.7805031

[3] M. Ahmed, A.N. Mahmoud,J. Hu, "A survey of network anomaly detection techniques", 2016. J. Netw. Comput. Appl. 60, 19– 31.

[4] N. Moustafa, J. Slay, " The evaluation of network anomaly detection systems: statistical analysis of the unsw-nb15 data set and the comparison with the kdd99 data set", Perspect. 25 (1 –3), 18– 31. 2016. Inf. Secur. J.: a Glob.

[5] A. Shameli-Sendi, M. Chariet, A. Hamou-lhadj, "Taxonomy of Intrusion Risk Assessment and Response System", ELSEVIER Volume 45, , Pages 1–16. September 2014.

[6] H.M. Bhuyan, D.K. Bhattacharyya, JK. Kalita, "An effective unsupervised network anomaly detection method" In International conference on advances in computing, communications and informatics, no. 1; p. 533–9. 2012.

[7] S. Wang, Z. Zhang, and Y. Kadobayashi," Exploring attack graph for cost-benefit security hardening: A probabilistic approach," Computers & Security, vol. 32, pp. 158-169, 2013.

[8] Zhang, Z., Ho, P. and He, L. "Measuring ID S-estimated attack impacts for rational incident response: A decision theoretic approach", Computers & Security , Vol. 28 No. 7, pp. 605-614, 2009.

[9] K. Scarfone and P. Mell: Guide to Intrusion Detection and Prevention Systems. Technical report, NIST: National Institute of Standards and Technology, U.S. Department of Com-merce, 2007.

[10] Hung-Jen Liao, Ch.-H. Richard Lin, Y.-Ch. Lin, K.-Y. Tung,"Intrusion detection system: A comprehensive review", Journal of Network and Computer Applications 36, pp: 16–24. 2013.

[11] N. B. Anuar, H. Sallehudin, A. Gani, and O. Zakaria," Identifying False Alarm for Network Intrusion Detection System Using Hybrid Data Mining and Decision Tree," Malaysian Journal of Computer Science, ISSN 0127- 9084, pp. 110-115. 2008.

[12] H. Debar, A. Wespi, "Aggregation and correlation of intrusion-detection alerts" , Recent Advances in Intrusion Detection, 2001.

[13] N. B. Anuar, H. Sallehudin, A. Gani, and O. Zakaria," A risk index model for security incident prioritisation" Proceedings of the 9th Australian Information Security Management Conference, Edith Cowan University, Perth Western Australia, 5th -7th December, 2011.

[14] Alsubhi, K., Al-Shaer, E. and Boutaba, R. "Alert prioritization in intrusion detection systems", Proceedings of the IEEE Network Operations and Management Symposium, Salvador, Brazil, pp. 33-40, 2008.

[15] Dondo, M.G., "A vulnerability prioritization system using a fuzzy risk analysis approach", Proceedings of the 23rd International Informa tion Security Conference, Milano, Italy, pp. 525-539, 2008.

[16] Lee, B. Chung, H. Kim, Y. Lee, C. Park, and H. Yoon," Real-time analysis of intrusion detection alerts via correlation," Computers & Security, vol. 25, no. 3, pp. 169-183, 2006.

[17] P.A. Porras, M.W. Fong, and A. Valdes, "A Mission-Impact-Based Approach to INFOSEC Alarm Correlation, Proceedings of the 5th International Symposium on Recent Advances in Intrusion Detection (RAID 2002)" 95–114, 2002.

[18] The Snort Project, Snort user's manual https://www.snort.org/downloads, 2017.

[19] KDDCup '99 dataset, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 20-27 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Technology

ASTESJ

ISSN: 2415-6698

Smart Grid Operational functions and Control Challenges by Implementing SSSC Tailored to Optimize performance in between Qatar and KSA on the GCC Electrical-power grid

Tariq Masood[*1], Muhammad Tajammal[2], Samer Karim Shah[3], Ghulam Hashmi[4], Suhail Aftab Qureshi[5], D. P Kothari[6]

[1]*Asset Integrity, Department Qatar Petroleum, Dukhan Operations, Qatar*

[2]*University of Hail, Hail, Kingdom of Saudi Arabia*

[3]*Maintenance Department, QAFAC, Doha Qatar*

[4]*Power System Engineering Department, Saudi Aramco, Dhahran, KSA*

[5]*University of Engineering and Technology, Lahore Pakistan*

[6]*J. D. College of Engineering & Management, Nagpur India*

ARTICLE INFO

Article history:
Received: 19 August, 2017
Accepted: 06 September, 2017
Online: 30 September, 2017

Keywords:
Reactive voltage
Voltage regulator
Power flow

ABSTRACT

This research work is novel technique to control and optimize SSSC (Subsynchronous Series Controller) functions with degree of precision in between Qatar and Kingdom of Saudi Arabia. The SSSC model developed and simulated in order to identify and determine its control and functioning parameters by introducing new tuning parameters based on that the SSSC can be adjusted stringently to witness desired results lead to address outstanding reactive power management issue. The proposed new parameters are contributing significantly to control SSSC functions in multiple directions in a power system network in between QATAR and Kingdom of Saudi Arabia at different time-based transmission contingencies on the GCC Electrical-power grid. Strategically, the SSSC capacity and capability can be utilized fully in between Qatar and Kingdom of Saudi Arabia by introducing and optimizing its control and tuning parameters more tangibly under both steady and dynamic states

1. Introduction

The SSSC has operational controllability results are clearly indicating that introduction of SSSC in between Qatar and Kingdom of Saudi Arabia power network will equitably improve the power system loadability, curtailing the losses and value-added sustainability of the power system enactment by addressing control and operational issue throughout the GCC Electrical-power grid.

Hereafter, new SSSC optimization technique can thus be magnificently expended for this type of power system process optimization. This work published in 13th the IET International AC/DC Conference held in Manchester, February 14-16, 2017. Whereas a SSSC determines and validates the three control and effective limits which have been made-to-order at (minimum (+/-) medium (+/-), and maximum (+/-) compensation. These rheostats functioning limits are regulated by constituting their consequent PI control-values by exercising D.J. Cooper PID regulator Performance to achieve the following [1].

(1) Augmented Power flow
(2) Developed consistency & controllability
(3) Augmented angle and voltage constancy.

Therefore, FACTS Controllers are convincing candidate technology options centered on that following benefits can be perceived.

[*]Corresponding Author: Tariq Masood, Asset Integrity, Department Qatar Petroleum, Dukhan Operations, Qatar | Email: t.masood.dr@bath.edu

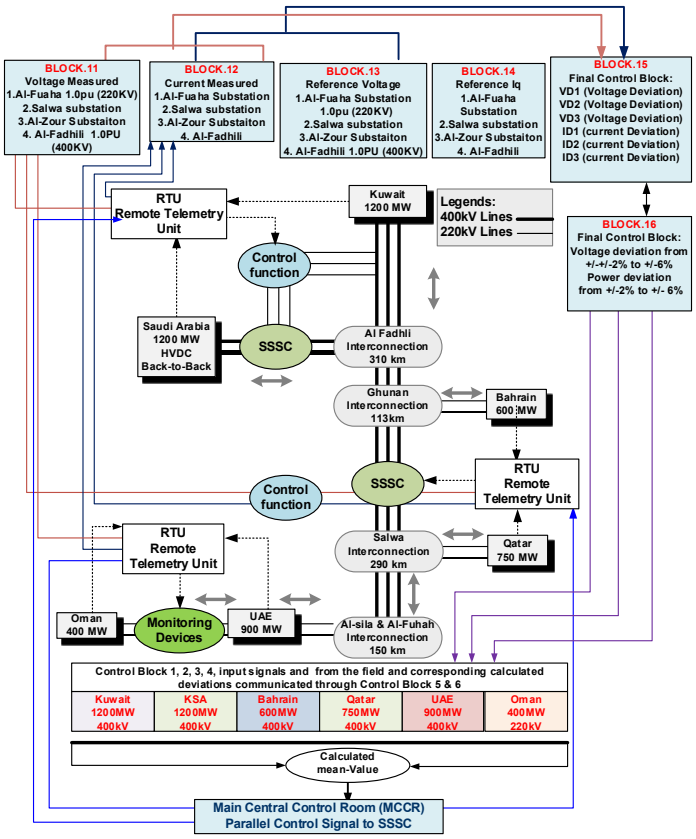


Figure 1. SSSC structure in between Qatar and Kingdom of Saudi Arabia

Figure 1 shows the GCC power network by retaining SSSC at augmented whereabouts by exercising Wideband-Delphi-Technique.

- Model reveals how much notional concentrated power can be dispensed in between QATAR and Kingdom of Saudi Arabia.
- At maximum conjectural power, what is borderline voltage
- Model reveals how much reactance must be retained by the orthodox series capacitance to double the power conveyance in between Qatar and Kingdom of Saudi Arabia if vital and prerequisite.
- Compute the introduced voltage by the series capacitor in between QATAR and Kingdom of Saudi Arabia
- By SSSC how much introduced and instilled/injected reactive voltage Vr is upheld at diverse compensation of SSSC in between Qatar and Kingdom of Saudi Arabia and Compute the determined power supplied at persistent introduced voltage Vr
- If the load angle decreases from 71 to 0 how much power will be produced by SSSC in between Qatar and Kingdom of Saudi Arabia.

2. GCC Electrical-power grid background

The GCC power-grid is constituted and employed into eight premeditated power system operational directions in order to meet national and industrial customers necessities in a consistent and viable demeanors at the GCC Electrical-power grid-network as shown in Figure 1 as considered and described with facts and figures, this is also published in [1] .

3. SSSC Operational Analyses

Principally, the SSSC-series counterbalance device location is not very much perilous it can be employed anyplace on the GCC Electrical-power grid. In this case the SSSC has been retained at the borderline to decrease the power-line transmission impedance unnaturally by SSSC which is controlling to upsurge the power flow in the power transmission system in (5) and (6) determine how much line impedance (X_C) has been Controlled and Managed after employing series but stable capacitor's K_{series} factor.

SSSC encompasses of capacitors and reactors to diverge the power-line impedance on domineering need basis vigorously. Predominantly, the SSSC injects the voltage in a series into the power transmission network at the midpoint of Kingdom of Saudi Arabia and Qatar whereas instilled/injected the voltage at midpoint Vr has been computed in (7). The SSSC stipulates or expends reactive-power in capacitive/inductive control-mode of functions on the GCC power-grid, wherein the SSSC instilled/injected voltage in phase quadrature with the line current of the power network.

Figure 2 denotes maroon trend in the graphical presentation of the SSSC processes in capacitive and inductive mode of functions. Practically, blue trends denote how much real power transferred can transferred when electrical-power transmission network compensation carried-out capacitive, inductive mode of operations or neutral without any compensation factor. Figure 3 signifies that a SSSC has a main and great influence on power steadiness and load flow, but ithas limited impact on voltage profile enhancement as exhibited in (28) that V_m, V_s, V_r has negligible or no effect with SSSC [2].

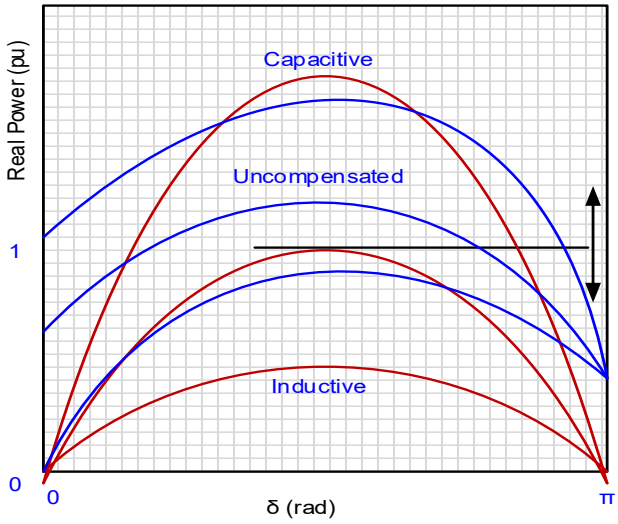


Figure 2. Signify the compensation factor by SSSC [2]

Principle of Operating parameters	SSSC Controller	Impact in between Qatar and KSA		
		Load Flow	Stability	Voltage Quality
Series Compensation: To Vary the power transmission line impedance	FSC: Fixed Series Compensation	o	ooo	o
	TPSC: Thryrister protected Series Compensation	o	ooo	o
	TCSC: Thryrister controlled Series Compensation	oo	ooo	o
Benchmark	o: Low or no Impact	o o: Medium		o o o: Strong

Figure 3. Denotes the SSSC impact.

3.1. By employing Series Fixed Capacitor to upsurge the Power on the GCC Electrical-power grid

Herein determined conjectural power can be dispensed and distributed without conservative or non-conventional series compensation, the voltage structured transitorily but not vigorously and unvaryingly which is outfitted at 800 km long electrical-power transmission network from Kingdom of Saudi Arabia to Qatar or vice versa along with its attendant substations. Now if series capacitor is connected at midpoint voltage as stated in Figure 3 to resource and expend sufficient reactive power to counterbalance as prerequisite. Equation (5) expended to compute series compensation factor to convalesce the power system steadiness and double the power at power transmission lines reactance X_C= 136.626 Ohms as intended in (6) and (7) [3].

$$K_{series} = \frac{X_c}{2 \times Z_a} cos\frac{\theta}{2} \tag{5}$$

$$X_c = 2 \times K_{series} \times Z_a tan\frac{\theta}{2} = 136.626\ Ohms \tag{6}$$

$$P_{max} = \frac{V^2}{Z_a - X_C \times Sin(43.26^{\circ})} = 1223\ MW \tag{7}$$

Equation (7) reveals the concentrated power by adapting power transmission line reactance by series capacitor Units.

3.2. By implementing Series-Capacitor to upsurge the Power on the GCC Electrical-power grid

The SSSC has a tremendous competence to upturn the electrical Power flow and improve dynamic steadiness to supplant usual and uneconomical fixed series capacitors are expended to govern the prerequisite reactive current at midpoint in between Qatar and Kingdom of Saudi Arabia for recompense [4].

Instill/Injected current with series capacitor is computed by exercising in (8), (9).

$$I_m = \frac{Vsin\frac{\delta}{2}}{(1 - K_{series})Z_a sin\frac{\theta}{2}} = 2.1159pu \tag{8}$$

Instilled/injected voltage as given below

$$\hat{V_r} = 1.5237 \times tan\frac{\theta}{2} = 0.8881pu \tag{9}$$

Equation (9) reveals the in a Phase injected/instill voltage

$$V_r = \hat{V_r} \times \frac{V}{\sqrt{3}} = 205.03\ kV \tag{10}$$

Equation (10) reveals the instilled/injected voltage in between QATAR and Kingdom of Saudi Arabia. Herein determined power flow can be sustained with unremitting voltage inserted by the SSSC at midpoint. Firstly, the SIL computed and determined by using Equation (11)

$$\hat{P} = \frac{P}{P_n} = Acos\delta + Bsin\frac{\delta}{2} \tag{11}$$

amount of real power can be dispensed in line with P_n, this is identified as a SIL in contradiction of the real electrical-power delivery.

Where

$$A = \frac{1}{sin\theta} = 1.40115\ \ ;\ \ B = \frac{\hat{V_r}}{2sin\frac{\theta}{2}} = 1.14735 \tag{12}$$

Equation (13) reveals electrical Power flow with unremitting reactive voltage instilled/injected whereas the $\hat{P}$ is determined

when $\frac{d\hat{P}}{d\delta} = 0$

$$\frac{d\hat{P}}{d\delta} = Acos\delta - Bsin\frac{\delta}{2} \tag{13}$$

Equation (14) resultant after substituting $\frac{d\hat{P}}{d\delta} = 0$ value

$$0 = Acos\delta - Bsin\frac{\delta}{2} \tag{14}$$

Unassumingly $cos\delta$ and $sin\delta$ are factorized as resultant in (15)

$$cos\delta = cos\left(\frac{\delta}{2} + \frac{\delta}{2}\right)\ \ ;\ \ sin\delta = sin\left(\frac{\delta}{2} + \frac{\delta}{2}\right) \tag{15}$$

Where

$$\cos(A + B) = CosA.CosB - SinA.SinB \tag{16}$$

Equation (17) resultant after replacing $cos\delta$ and $sin\delta$ values in (16)

$$cos\left(\frac{\delta}{2} + \frac{\delta}{2}\right) = cos\frac{\delta}{2}.cos\frac{\delta}{2} - sin\frac{\delta}{2}.sin\frac{\delta}{2} \tag{17}$$

Equation (17) is the streamlined and resultant in (18)

$$cos\delta = cos^2\frac{\delta}{2} - sin^2\frac{\delta}{2} \tag{18}$$

Equation (18) $cos\delta$ value subtract in (14) and resultant from (19)

$$0 = A\left(cos^2\frac{\delta}{2} - sin^2\frac{\delta}{2}\right) - Bsin\frac{\delta}{2} \tag{19}$$

Whereas

$$cos^2\frac{\delta}{2} = \left(1 - sin^2\frac{\delta}{2}\right) \tag{20}$$

Equation (20) $cos^2\frac{\delta}{2}$ subtracted in (19) and developed new (21)

$$0 = A\left(\left(1 - sin^2\frac{\delta}{2}\right) - sin^2\frac{\delta}{2}\right) - Bsin\frac{\delta}{2} \tag{21}$$

By adding the $x = sin\frac{\delta}{2}$

$$A(1 - x^2 - x^2) - Bx = 0 \quad (22)$$

$$-2Ax^2 + A - Bx = 0 \quad (23)$$

$$-(2Ax^2 - A + Bx) = 0 \quad (24)$$

$$2Ax^2 + Bx - A = 0 \quad (25)$$

Equation resultant from (25) to convert into Quadratic form.

$$\int(x) = \left(x + \frac{b}{2a}\right)^2 + \frac{4ac - b^2}{4a} \quad (26)$$

Equation (26) expressed the Quadratic function as detailed below with factors.

$$\left(x + \frac{B}{4A}\right)^2 + \frac{(4 \times 2A \times -A) - B^2}{4A} = 0 \quad (27)$$

$$\left(x + \frac{B}{4A}\right)^2 - \frac{8A^2 + B^2}{4A} = 0 \quad (28)$$

Equation (27) and (28) Quadratic roles are defined and extracted and resultant in (29)

$$x = -\frac{B}{4A} \pm \frac{\sqrt{8A^2 + B^2}}{4A} \quad (29)$$

Equation (29) subtracted A and B factors values as defined and calculated in (12) and resultant in new (30)

$$x = -\frac{1.401}{4 \times 1.4735} \pm \frac{\sqrt{8 \times 1.4735^2 + 1.401^2}}{4 \times 1.4735} \quad (30)$$

Finally, the "x" values are signified as given below:

$$x = 0.5314 = sin\frac{\delta}{2} =\rightarrow \delta = 64.2° \quad (31)$$

Equation (31) reveals the $x = 0.5314$ and electrical power angle resultant

Where

$$\hat{P}_{max} = Asin\delta + Bcos\theta \quad (32)$$

$$\hat{P}_{max} = Asin(64.2°) + Bcos(45.5°)\hat{P}_{max} = 2.06519 \quad (33)$$

Equation (32) subtract the A, B, $\emptyset$ and δ values and resultant the $\hat{P}_{max}$ value.

$$P_{max} = \hat{P}_{max} \times \frac{400^2}{316.228} = 1044.9\,MW \quad (34)$$

Equation (34) signifies the concentrated electrical power dispensed and distributed in the GCC Electrical-power grid by retaining SSSC to uphold persistent Vr vigorously in the Midpoint of QATAR and Kingdom of Saudi Arabia as illustrated in Figure 4 [5].

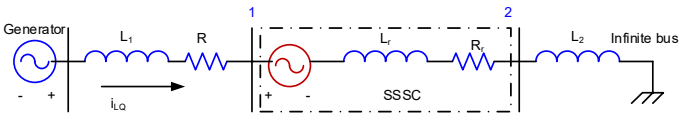


Figure 4: SSSC connected in between Qatar and Kingdom of Saudi Arabia

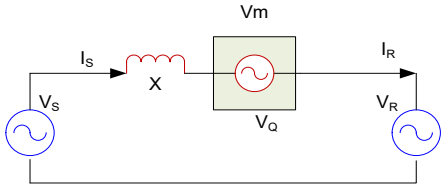


Figure 5: SSSC connected in between Qatar and Kingdom of Saudi Arabia

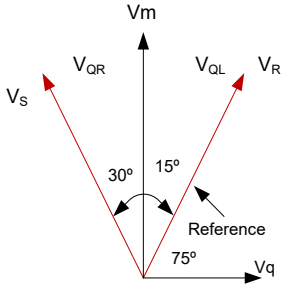


Figure 6. Phasor diagram of voltage profile

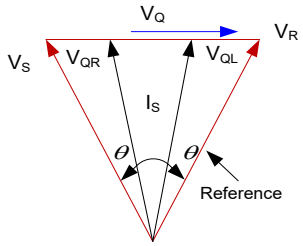


Figure 7. SSSC connected at the midpoint and instilled/injected V_Q toward V_R

Figure 5: signifies that counterbalance device employed in the middle of the of the power transmission line, this is voltage source only dispense reactive electrical power not a real power. Figure 6 displays the influence of location if counterbalance device employment on voltage profile, the voltage on left side of the counterbalance device would be V_{QL} and right side V_{QR}. Figure 7 illustrates that the series counterbalance device employed in the middle the worst voltage-profile occurred at each side of the counterbalance device, whereas the voltage-profile vector aligns with the current vector [5]. Therefore, the midpoint voltage-profile is the same with/without SSSC functions as exhibited in (37) and (38) SSSC power flow with $\delta = 0$[6].

$$P_{max} = 1.147 \times \frac{400^2}{316.2} = 580MW$$

$$\hat{V}_m = \frac{cos\frac{\delta}{2}}{cos\frac{\theta}{2}} \qquad \delta = 0 \quad (35)$$

Equation (35) reveals the midpoint voltage-profile if $\delta = 0$ and resultant midpoint voltage denotes from (36)

$$\hat{V}_m = \frac{1}{cos \times 22.76°} = 1.1\ pu \quad (36)$$

Equation (35) reveals the midpoint voltage if $\delta = 0$ therefore, $V_s and\ V_R$ voltage-profile are computed in (37) and (38). Sending end from the Kingdom of Saudi Arabia and receiving end at QATAR

$$V_s = V_m \angle 0 - j\frac{V_r}{2} = 1.171\angle 0 - j\frac{0.8881}{2} \quad (37)$$

$$V_R = V_m \angle 0 + j\frac{V_r}{2} = 1.171\angle 0 - j\frac{0.8881}{2} \quad (38)$$

3.3. SSSC input Data

Table 1: operating parameters and associated countries

Countries	Power Exchange
To/from Qatar to main transmission line	400kV, 50Hz, 750MW
To/from Kingdom of Saudi Arabia to main power-transmission line	400kV, 50Hz, 600 MW

Table 2: Conventional Reactive Power

Existing Power System	Reactive Power	Distance
VAR Demand at Al Jasra substation Kuwait	850MVAR	Ghunan Substation to Slawa Substation total distance 288Km
Reactive Power demand at Al-Al-Zour Substation	500MVAR	Slawa Substation total distance = 97Km

Table 3: Multivariable Controller's Configuration

Al-Zour Substation	Iq regular: Kp 14dB; Ki: 0.014 dB	Counterbalanced and regulated voltage at 0-2%
Rated SSC: +/- 500 MVAR	Ref V: 1.0 pu (400kV)	(VC voltage change from 2% to 4%)
P (proportional): 0.33, 0.38, 0.40, I (integral) 0.7, 0.9, 1.2	Droop: 0.033pu/100MVA; Kp:14dB; Ki: 3500dB	(VC voltage change from 4-6%)

4. Results and Discussion

4.1. Maximum level Vr voltage-injection

It has been simulated and authenticated that further real power can be up-surged in the power transmission line to touch its design specification by up-surging the SSSC compensation factor from 50% to 65% reactive voltage instilled/injected voltage would be 0.8864pu which is equivalent to 204kV and current Im = 2.1pu and $\boldsymbol{\delta = 63.87°}$ will activate consequently at different functioning condition total 1200MW power will be produced and distributed on the GCC electrical-Electrical-power grid. This is known as a determined compensation by implementing a SSSC and its compensation factor from 50% to 65% to obtained projected power produced and delivered in the electrical power transmission network on the GCC electrical-Electrical-power grid.

4.2. Medium level Vr voltage-injection

As simulated by Matlab/SIMULINK and scientific/ mathematical model V_r instilled/injected voltage 0.5936pu which is uniform and match to 137kV at 50% SSSC reactive voltage compensation whereas the midpoint current I_m = 1.4459pu, and $\boldsymbol{\delta = 71.32°}$ will activate consequently in this functioning circumstance to a total 1079 MW power which can be produced and distributed on the GCC Electrical-power grid as validated by the results. Infact, QATAR and Kingdom of Saudi Arabia system has been devised to produce and distribute concentrated 1200MW electrical-power in the GCC Electrical-power grid during a peak load, therefore Margin is obtainable to upsurge the power up to its projected specification. This is also identified as a medium compensation by a SSSC compensation factor from 20% to 50%. Therefore, the Control function block expended to readjust PI controller response as exhibited in Figure 10 to diverge the firing angle of the Thyrister control [7].

4.3. Minimum level Vr voltage-Injection

In the third challenge of operations, further this model was simulated and authenticated by exercising SSSC Controller. At minimum compensation of the SSSC compensation real power reduce considerably in the power transmission line to touch its projected specification by decreasing the SSSC compensation factor from 5% to 20% reactive voltage instilled/injected voltage would be 0.3710pu which is equal to 86.67kV and current Im = 0.974pu and $\boldsymbol{\delta = 77.57°}$ will activate consequently in this functioning condition total 986MW power will be produced and distributed on the GCC Electrical-power grid. This is identified as a minutest compensation by using a SSSC Controller and its compensation factor from 5 to 20% to achieve projected electrical power delivery of transmission network on the GCC Electrical-power grid. As instilled/injected voltage up-surged based on compensation from 5% to 20% there is considerable power flow and steadiness enhancement but very insignificant influence on voltage-profile as verified simulated consequences in Figure 8 (a) reactive power injection-waveform 8(b) indicates the total power delivered in between Qatar and Kingdom of Saudi Arabia, this is also computed in equation (34). The PI controller configured and demonstrated its operational response in Figure 9. Detailed configuration discussed as given in the appendices [8].

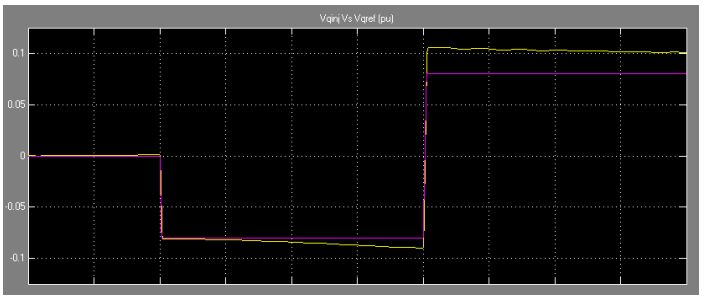


Figure 8 (a). Reactive power injection on the GCC Power Grid

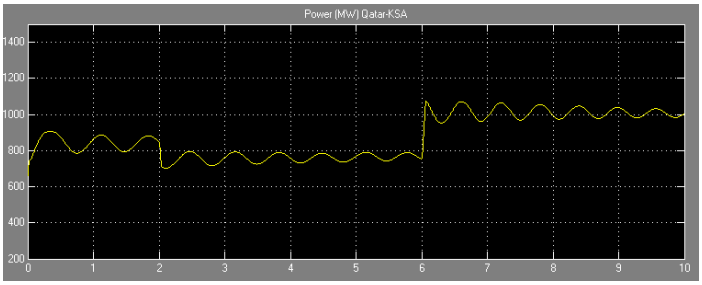


Figure 9 (b). Total active power delivered on the GCC Electrical-power grid

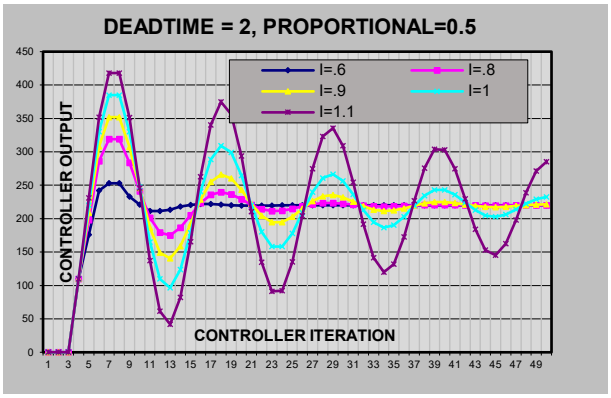


Figure 10. demonstrates the PI controller parameters

5. Conclusion

Fundamental creativity of the SSSC application in the GCC Electrical-power grid transmission line compensation is the subject of extensive significance. In this study, it has been computed and simulated how much maximum power can be produced and distributed in between QATAR and Kingdom of Saudi Arabia without any process vagueness. It has been reasonable how much reactance must be provided by the series capacitance to double the power supply in between Qatar and Kingdom of Saudi Arabia. It also has been also identified and unwavering amount of reactive voltage to be instilled/injected in between Qatar and Kingdom of Saudi Arabia in order to provide concentrated power delivery on the GCC Electrical-Power grid.

If the load angle as decreased below from 65 to 0 amount of electrical-power will be produced by SSSC in between Qatar and Kingdom of Saudi Arabia. Herein, the GCC electrical-power network will exclusively upsurge the power system loadability, decrease the losses and improve sustainability of the electrical-power system functioning. The results also illustrates that SSSC is multipurpose equipment with stupendous active capability to improve power system stability margin on the GCC Electrical-power grid. Based on these results, SSSC is a very strong-candidate to be instigated at GCC Electrical-power grid. It also demonstrates optimistic influence on neighboring countries' power system functions at the GCC Electrical-power grid.

Table 4 Terms and Abbreviations

Name	Abbreviations
GCC	Gulf Cooperative Council
Ir	Reactive current
Za	Line Impedance
Vm	Midpoint voltage
Kshunt	series compensation factor in the network
Pn	Surge Impedance
KSA	Kingdom of Saudi Arabia

Conflict of Interest

No conflict of interest in this paper has been identified.

Acknowledgment

I extend my special gratitude to all my coauthors in this paper in particularly late Prof. Dr Abdel-ETY Edris (Icon of FACTS Technology) at Santa Clara University for their unprecedented support and exceptional contribution to make this study a great success. I am also very grateful to my little Doll Princesses Fatima Masood, who kept me busy full time not giving me space during my research on smart control

References.

[1] Tariq Masood, R.K. Aggarwal, S.A. Qureshi "Novel Control of a SSSC Connecting the Oman and United Arab Emirates in the GCC power Grid Interconnection" published in IJAPE International Journal of Automation and Power Engineering, Volume 2, issue 1 January, 2013.

[2] Tariq Masood, Abdel Aty Edris, Suahil Aftab Qureshi, Muhammad Tajammal, Murtaza Hashmi DP Kothari, Samer Kareem "SSSC tailored to optimization performance in Between United Arab Emirates and Oman on the GCC Power Grid" published on 14-16 February, 2017 in the 13th IET International Conference on AC and DC power Transmission Conference.

[3] Mohammed Rasetgar, Mehdi Saradarzeda "An Improved D-SSSC Voltage and Current Load Balancing Control Strategy Under Unbalanced Load published in 25th Iranian IEEE ICEE 2017.

[4] Mohammed Abdul, Mohd Akram, "Power System Stability Enhancement using Static Synchronous Series Compensator" Published in IEEE (SCOPES)-2016

[5] S. Jamali, A. Kazemi, H. Shateri "Voltage Inversion due to Presence of SSSC on Adjacent Lines and Distance Relay Mal-Operation" published in 2008 IEEE Conference.

[6] D. Rai, S.O. Faried, "Impact of Imbalanced Phase Operation of SSSC on Damping Subsynchronous Resonance published in 2011 IEEE Conference,

[7] Tariq Masood "Improvement of Voltage and power flow control in the GCC Electrical-power grid by using coordinated FACTS Controllers" Ph.D Thesis, published in 2013 at University of Bath, UK.

[8] Tariq Masood, R.K. Aggarwal, Nasser Al-Eamdi "Enhanced Power Flow/Voltage Control in the GCC Grid by SSSC (SUBSYNCHRONOUS SERIES CONTROLLER) Robustness" published by Lambert in 2013 ISBN: 978-3-659-40995-0.

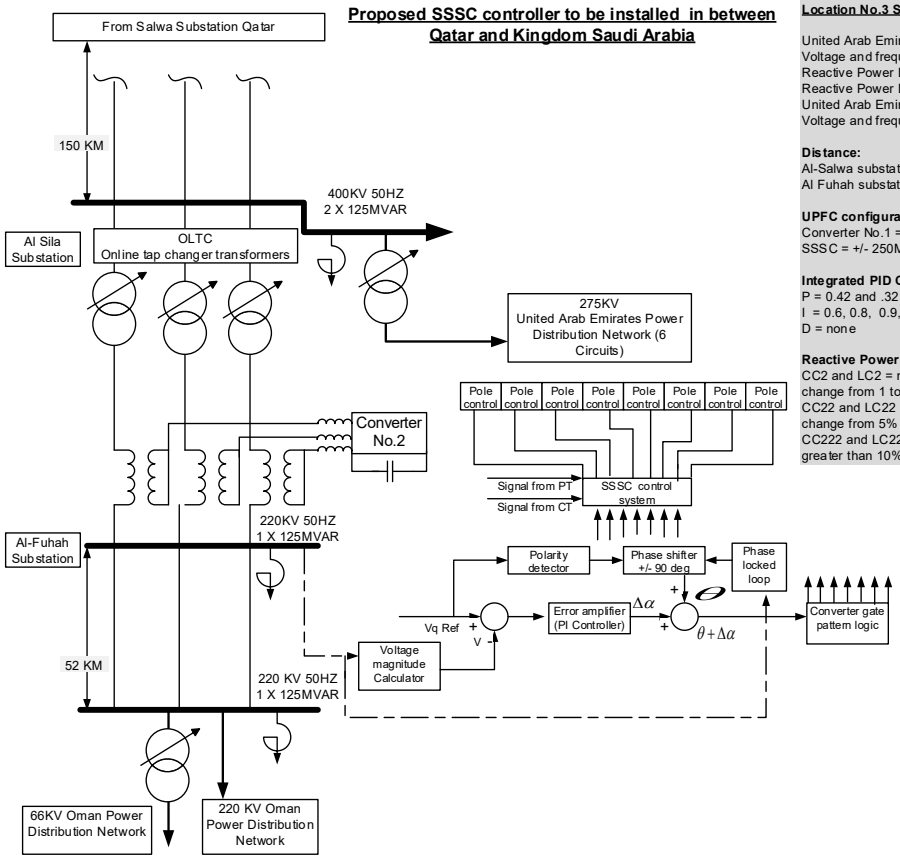
Appendix No.1
PI Controller response computed

Proportional	0.5				
Integral	0.8				
SP	220.0				
Iteration	PV	ERROR	ABSOLUTE ERROR	OUTPUT	INTEGRAL
				0.00	0.00
				0.00	0.00
1	0.0	220.00	220.00	110.00	88.00
2	0.0	220.00	220.00	198.00	176.00
3	0.0	220.00	220.00	286.00	264.00
4	110.0	110.00	110.00	319.00	308.00
5	198.0	22.00	22.00	319.00	316.80
6	286.0	-66.00	66.00	283.80	290.40
7	319.0	-99.00	99.00	240.90	250.80
8	319.0	-99.00	99.00	201.30	211.20
9	283.8	-63.80	63.80	179.30	185.68
10	240.9	-20.90	20.90	175.23	177.32
11	201.3	18.70	18.70	186.67	184.80
12	179.3	40.70	40.70	205.15	201.08

13	175.2	44.77	44.77	223.47	218.99
14	186.7	33.33	33.33	235.65	232.32
15	205.2	14.85	14.85	239.75	238.26
16	223.5	-3.46	3.46	236.53	236.87
17	235.7	-15.65	15.65	229.05	230.61
18	239.7	-19.75	19.75	220.74	222.71
19	236.5	-16.53	16.53	214.45	216.10
20	229.0	-9.05	9.05	211.58	212.48
21	220.7	-0.74	0.74	212.11	212.19
22	214.5	5.55	5.55	214.96	214.41
23	211.6	8.42	8.42	218.62	217.78
24	212.1	7.89	7.89	221.72	220.93
25	215.0	5.04	5.04	223.45	222.95
26	218.6	1.38	1.38	223.64	223.50
27	221.7	-1.72	1.72	222.64	222.81
28	223.4	-3.45	3.45	221.09	221.43
29	223.6	-3.64	3.64	219.61	219.98
30	222.6	-2.64	2.64	218.66	218.92
31	221.1	-1.09	1.09	218.38	218.49
32	219.6	0.39	0.39	218.68	218.64
33	218.7	1.34	1.34	219.31	219.18
34	218.4	1.62	1.62	219.99	219.83
35	218.7	1.32	1.32	220.49	220.36
36	219.3	0.69	0.69	220.70	220.63
37	220.0	0.01	0.01	220.64	220.63
38	220.5	-0.49	0.49	220.39	220.44
39	220.7	-0.70	0.70	220.09	220.16
40	220.6	-0.64	0.64	219.84	219.91
41	220.4	-0.39	0.39	219.71	219.75
42	220.1	-0.09	0.09	219.70	219.71
43	219.8	0.16	0.16	219.79	219.78
44	219.7	0.29	0.29	219.92	219.89
45	219.7	0.30	0.30	220.04	220.01
46	219.8	0.21	0.21	220.11	220.09
47	219.9	0.08	0.08	220.13	220.13
48	220.0	-0.04	0.04	220.11	220.11
49	220.1	-0.11	0.11	220.05	220.06
50	220.1	-0.13	0.13	220.00	220.01

Appendix No. 2

SSSC Controller configuration



Location No.3 SSSC input data:

United Arab Emirates to Oman = 400 MW
Voltage and frequency = 220KV and 50HZ
Reactive Power Demand at UAE Side = 125 MVAR
Reactive Power Demand at Oman Side = 125 MVAR
United Arab Emirates Power demand= 900 MW
Voltage and frequency = 400KV, 50 HZ

Distance:
Al-Salwa substation (Qatar) to Al-Sila substation (UAE) =150 KM
Al Fuhah substation (UAE) to Al-Wasset Substation (Oman) = 50 KM

UPFC configuration:
Converter No.1 = SSSC (Series connection Al-Fuhah Substation UAE)
SSSC = +/- 250MVAR

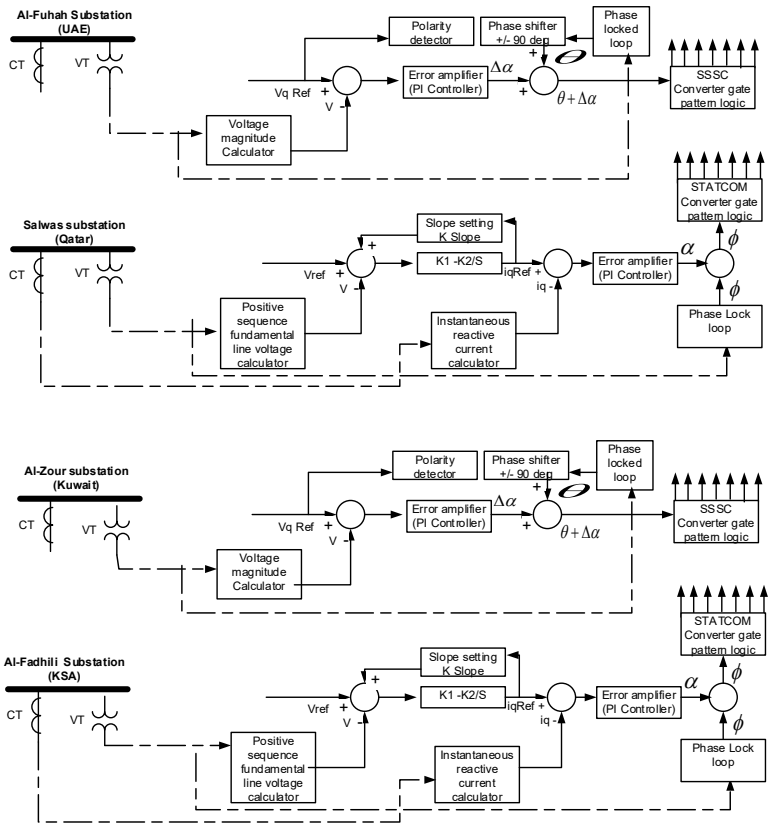
Integrated PID Controller's input data:
P = 0.42 and .32
I = 0.6, 0.8, 0.9, 1.0, 1.1
D = none

Reactive Power Compensation codes are derived in both Mode operations:
CC2 and LC2 = minimum compensation capacitive or inductive (VC voltage change from 1 to 5%)
CC22 and LC22 = medium compensation capacitive or inductive(VC voltage change from 5% to 10%)
CC222 and LC222 = full compensation capacitive or inductive(VC voltage change greater than 10%)

Appendix No. 3

SSSC Centralized and Decentralized Control

Decentralized controller configuration and implementation to control and monitor individual FACTS device operations at GCC Power Grid



Location Controller's input data

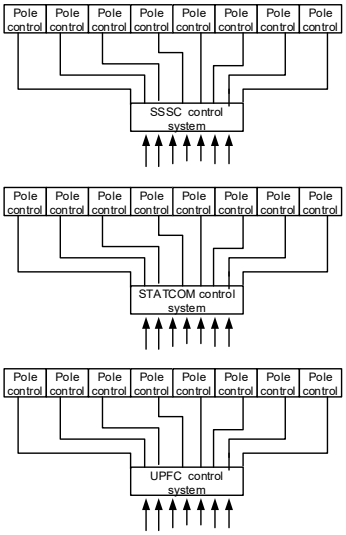
UPFC configuration:
Converter No.1 = STATCOM (shunt connection at Kingdom Saudi Arabia side)
Converter No.2 = SSSC (Series connection Kuwait side)
STATCOM = 250 MVAR
SSSC = 250 MVAR
RATED UPFC = +/- 500 MVAR
Location in between Al-Zour (Kuwait) and Al-Fadhil substation KSA
Droop =0.45/100VA

STATCOM configuration:
STATCOM (shunt connected at Salwa substation in between Bahrain and Doha)
RATED STATCOM = +/- 500 MVAR
Droop =0.37/100VA

SSSC configuration:
Converter No.1 = SSSC (Series connection Al-Fuhah Substation UAE)
SSSC = +/- 250MVAR

Integrated PID Controller's input data:

P = 0.42 and .32
I = 0.6, 0.8, 0.9, 1.0, 1.1



ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 21-27 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Community Detection in Social Network with Outlier Recognition

Htwe Nu Win[*,1], Khin Thidar Lynn[2]

[1]*Web Mining, University of Computer Studies, Mandalay, ZIP Code 05071, Myanmar*

[2]*Faculty of Information Science, University of Computer Studies, Mandalay, ZIP Code 05071, Myanmar*

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 05 February, 2018
Online: 08 March, 2018

Keywords:
Social Network
Community
Outlier
Edge Structure
Overlapped Neighbor Node

ABSTRACT

Exploring communities and outliers in Social Network is based on considering of some nodes have overlapped neighbor node within the same group as well as some nodes have no any link to the other node or have no any overlapped value. The existing approaches are based on the overlapping community detection method were only defined the overlap nodes or group of overlap nodes without thinking of which nodes might have individual communities or which nodes are outliers. Detecting communities can be used the similarity measure based on neighborhood overlapping of nodes and identified nodes so called outliers which cannot be grouped into any of the communities. This paper proposed method to detect communities and outliers from Edge Structure with neighborhood overlap by using nodes similarity. The result implies the best quality with modularity measurement which leads to more accurate communities as well as improved their density after removing outliers in the network structure.

1. Introduction

This paper is an extension of work originally presented in 18thIEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and parallel/Distributed Computing (SNPD 2017) [1] is used undirected and unweighted graph data based on overlapping neighborhood. It is explicitly focused on edge structure to detect outliers and significant communities with nodes similarity. Unfortunately, the number of outliers that implemented is higher than that we expected. Therefore, we continue improve our proposed approach to get the significant result by detecting outliers and community.

Social networks can be considered in graph theory point of view. Thinking of detection community in Social networks plays the important role in recent year. It is defined the presence of groups of nodes that are high tightly links connected with each other than with less links connected to nodes of different groups. The challenges in considering of community detection method are become popular in Social Network. The previous authors had finished thinking about them in various points of corners. In overlapping community detection method were only defined the overlap nodes or group of overlap nodes without thinking of which nodes might be included in its own individual communities . Then, the consideration of community detection does not force each node into a certain group, some independent nodes, which cannot be grouped into any communities, are allowed far outside the detected groups can be measured by the predefined threshold of minimal valid size (mvs) of communities as outliers. However, there are still challenges in considering of some nodes have no any common node within the same group as well as some nodes have no any link to the other node.

It can be used similarity measure based on neighborhood overlapping of nodes to organize communities and to identify outliers which cannot be grouped into any of the communities.

In this paper, we detect communities and outliers from Edge Structure with neighborhood overlap by using nodes similarity. This paper explores the use of neighborhood overlapping by using vertex similarity method for detecting outlier and significant

[*]Corresponding Author:Htwe Nu Win, Web Mining, University of Computer Studies, Mandalay, ZIP Code 05071, Myanmar, htwenuwin99@gmail.com

community. The heart of this approach is to represent the underlying dataset as an undirected graph, where a user refers to each node and friendship between two users represents each edge. Before we measure the similarity among neighborhood overlap, finding seed node by using the degree centrality is necessary which is designed to find nodes that are most "central" to the graph. We operate similarity from the most centrality node and its neighborhood nodes. The values of zero similarity are then used to identify as outliers. To illustrate, consider tiny graph which contain8 nodes and 11 edges as shown in Figure 1. Upon applying communities, rounded with two circles are group, and node h so called outlier is saturated with outside them. It can be seen clearly the significant communities and outliers in this toy example.

This paper proposed the method to detect communities, nodes which are high tightly linked each other as community and outliers which do not have links overlap values with another.

The rest of paper is organized as follows. We briefly surveyed related work in section 2. In section 3, we describe the background methodology of our work. And then, we briefly describe about our propose system in section 4. In section 5, we discuss about the experiment and evaluation of our work. Then, we conclude our work in section 6 and talk about our future idea.

2. Related Work

As described in section 1, this paper is an extension of work originally presented in 18thIEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and parallel/Distributed Computing (SNPD 2017) [1] is used undirected and unweighted graph data based on overlapping neighborhood. It was explicitly focused on edge structure to detect outliers and significant communities with nodes similarity. Unfortunately, the number of outliers that implemented is higher than that we expected. Therefore, we continue improve our proposed approach to get the significant result by detecting outliers and community.

Many approaches of community and outlier detection algorithm have been proved over years. Each trend has efficient and effective in their ways. The proposal of a community discovery algorithm for large networks that iteratively finds communities based on only on local information only without considering the global information which was described in [2], this work is cited it in section. The paper based on considering of defining and evaluation networks community based on ground-truth [3] was discussed how to evaluate and measure the proposed system by comparing the previous and extracted community. In [4], the authors used proximity and random walks, to assess the normality of nodes in bipartite graphs. On the other hand, in [5], the authors proposed the one could use to compute various measures associated with the nodes in the given graph structure, dyads, triads, egonets, communities, as well as the global graph structure. The paper which was proposed an algorithm to detect anomalous nodes in weighted graphs is called OddBall was discussed in [6] that easily reveal nodes with strange behavior. However, it was only considered the side of outlier without community. This work is used some methods in the synthesis Lectures on Data Mining and Knowledge Discovery which is describe in [7] was the series publications on topics pertaining to data mining, web mining, text mining, and knowledge discovery, including tutorials and case studies.

In [8] , the authors proposed the efficient algorithm combined the hierarchical and spectral clustering in non-overlapping community. Then, [9] was represented the new community detection method on network model. In [10], the authors described modularity and community structure in networks based on the problem of detecting and characterizing this community structure was one of the outstanding issues in the study of networked systems. The proposed method for detecting such communities, built around the idea of using centrality indices to find community boundaries was described in [11] . [12] was presented several key features of Gephi in the context of interactive exploration and interpretation of networks. In [13], the authors proposed the method of detecting outliers with community by using minimum valid size (mvs). If the minimum size was 2, it would be chosen the single node and marked it as outlier. In [14], the authors surveyed on Social Community Detection which was also specially focus on to community evaluation since this step becomes important in social data mining. The results of proposed method of applying the metric, modularity, and several popular community quality metrics to two real dynamic networks was described in [15]. Continually, in [16], the authors discussed the attempting of a thorough exposition of the topic, from the definition of the main elements of the problem, to the presentation of most methods developed, with a special focus on techniques designed by statistical physicists, from the discussion of crucial issues like the significance of clustering and how methods should be tested and compared against each other, to the description of applications to real networks. Then, the paper in [17] was presented the graph-based approaches to uncovering anomalies in domains where the anomalies consist of unexpected entity/relationship deviations that resemble non-anomalous behavior. Using synthetic and real-world data, it was evaluated the effectiveness of these algorithms at discovering anomalies in a graph-based representation of data. The original citation of dataset, Zachary Karate Club was described in [18]. The paper was presented by the algorithms OCNS for detecting community overlapping base on node similarity was discussed in [19]. The following sections were described the method of the above references in detail which will be the background methodology for this paper.

3. Basic Definitions and Notations

3.1. Outliers

Nodes which have no overlapped values to its adjacent nodes as well as each node in a graph cannot be grouped into any of the communities is defined outliers. This system is based on edge structure approach in the graph to remove outliers before detecting community.

3.2. Node Degree and Its Neighborhood

In network G, the degree of any node i is the number of nodes adjacent to i. Generally, the more degree that the node has, the more important it will be. Two vertices v and u are called

neighbors, if they are connected by an edge. Let N_i be the neighborhood of vertex i in a graph, i.e., the set of vertices that are directly connected to i via an edge.

3.3. Community

Communities in Social network can be defined as group of nodes which have more links connecting nodes of the same group and comparatively less links connecting nodes of different groups. Communities may be groups of related individuals in social networks. Identifying communities in a network can be provided valuable information about the structural properties of network, the interactions among nodes in the communities, and the role of the nodes in each community. In an undirected graph (V, E), where the total number of node, $|V|=n$ and total number of edges,$|E|=m$ are defined. We can identify set of communities such that $Coms=\{V_1',V_2',\ldots\ldots..,V_{cn}'\}$ where $\bigcup_{i=1}^{cn} V_i' \subseteq V_i$and cn is the total number of communities Coms should satisfy, $V_i' \cap V_j' = \phi$.

Thus, the goal of community detection is to identify sets of nodes with a common (often external/latent) function based only the connectivity structure of the network. Then it can be considered an axiomatic approach and define four intuitive properties that communities would ideally have. Intuitively, a "good" community is cohesive, compact, and internally well connected while being also well separated from the rest of the network. This allows us to characterize which connectivity patterns a given structural definition detects and which ones it misses.

3.4. Degree Centrality

The importance of a node is determined by the number of nodes adjacent to it. The larger the degree of node, the more important the node is. Those high-degree nodes naturally have more impact are considered to be more important. The degree centrality is defined as

$$(v_i) = d_i = \Sigma A_{ij} \tag{1}$$

When one needs to compare two nodes in different networks, a normalized degree centrality should be used,

$$CD'(v_i) = d_i/(n-1) \tag{2}$$

Here, n is the number of nodes in a network. It is the proportion of nodes that are adjacent to node v_i.

Let N_i denote the neighbors of node v_i. Given a link $e(v_i, v_j)$the neighborhood overlap is defined as;

$$Overlap(v_i, v_j) = \frac{number\ of\ shared\ friends\ of\ both\ v_i\ and\ v_j}{number\ of\ friends\ who\ are\ adjacent\ to\ at\ least\ v_i\ and\ v_j} = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} - 2; \tag{3}$$

We have -2 in the denominator just to exclude $v_i\ and\ v_j$ from the set $N_i \cup N_j$. If there are no overlap vertices in any two $N_i\ and\ N_j$ means $|N_i \cap N_j| = \emptyset$, we can identified N_j as outliers of N_i. Assuming like that, our work identify outliers are appeared with among separated communities.

3.5. Vertex Similarity

It can be assumed that communities are groups of vertices similar to each other. We can compute the similarity between each pair of vertices after searching seed nodes. Most existing similarity method are based on the measurement of distance called Euclidean, Manhattan and etc., Although, to considered the similarity between selected node and is neighborhood, Jaccard Similarity is more convenient in this work which we will measure the similarity based on the neighborhood overlap of seed nodes.

3.6. Graph

The definition of Social network can be imaged as graphs, users or things might be nodes and their relationship might be edges. This system will be represented their social network as an undirected and unweighted graph which mean no distinction between the two vertices associated with each edge. The notation of a graph $G = (V,E)$ consists of two sets V and E. The elements of $V = \{v_1, v_2, \ldots, v_N\}$ are the nodes or vertices of the graph G where each vertex vi is associated with the instance x_i from the input data X and the cardinality of $|V|$ is N. The elements of $E = \{e_1, e_2, \ldots, e_M\}$ are links or edges between nodes and the cardinality of $|E|$ is M. An edge connecting the vertices $v_i$and v_j is denoted by e_{ij}.

4. Community Detection Approach

There are two main components will be used to detect communities. They are (i) Finding seed nodes before detecting community and (ii) Detecting community using similarity measure.

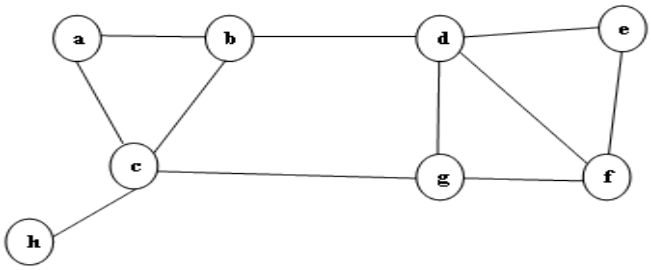
4.1. Finding Seed nodes

This part presents in detail finding for detecting communities and making clear the processes of selecting the initial seed node, associating nodes which incident upon a seed node, and electing new seed nodes. The basic idea is inspired by the well-known degree centrality method. It can be used the centrality of nodes to measure which node is seed within that network. It is the most suitable centrality measure for the default measure which yields the most accurate results and also is easy to compute from which we experimented with our results. It is a simple centrality measure that counts how many neighbors a node has. The importance of a node is determined by the number of nodes adjacent to it. The larger the degree of node, the more important the node is. Those high-degree nodes naturally have more impact are considered to be more important. Degree centrality is defined in section 3.5. For example: by viewing Figure 1, the degree centralities become node c and node d by using equation (2).

4.2. Detecting community using similarity measure

In this component, we used the existing methods which is convenient in our work namely Jaccard Similarity which we will measure the similarity based on the neighborhood overlap of seed

nodes. This system will be assumed that communities are groups of vertices which of their neighbor nodes are overlapped to each other. We can compute the similarity between each pair of vertices after searching seed nodes to detect community. This method is shown in section 3.6. The example of vertex similarity is shown in the following.



Node ={a,b,c,d,e,f,g,h};

Edge ={(a,b),(a,c),(b,c),(b,d),(c,h),(c,g),(d,e),(d,f),(d,g),(e,f),(f,g)};

Figure 1: Tiny Graph

Table1: Finding Node Centrality

Node	Degree	Normalize Degree
a	2	2/7=0.28
b	3	3/7=0.42
c	4	4/7=0.57
d	4	4/7=0.57
e	2	2/7=0.28
f	3	3/7=0.28
g	3	3/7=0.28
h	1	1/7=0.14

The most centralities are node d and node , firstly we choose node c.

$c = \{a, b, h\};$

$$c \text{ and } a = \frac{(c \cap a)}{(c \cup a) - 2} = \frac{1}{4 - 2} = 0.5;$$

$$c \text{ and } b = \frac{(c \cap b)}{(c \cup b) - 2} = \frac{1}{4 - 2} = 0.5;$$

$$c \text{ and } h = \frac{(c \cap h)}{(c \cup h) - 2} = \frac{0}{0 - 2} = 0;$$

We got, there is no overlapped value between c and h. Therefore, we identified node h as outlier, then $node\ a$ and $node\ b$ are the members of community corresponding by their node centrality ($node\ c$). The community is shown in Figure 2.

Then, choose the most centrality value from the rest of the graph. Now, $node\ d$ is the most centrality.

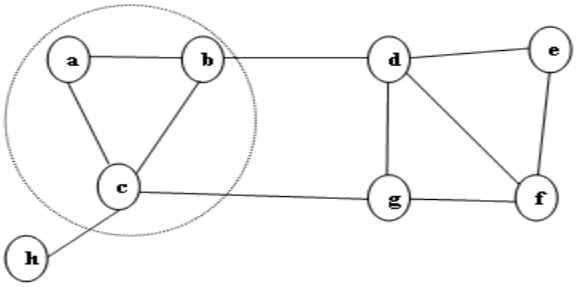


Figure 2: Community of node c Centrality

$d = b, g, e, f;\ b = a, c, d;\ e = d, f;\ f = d, e, g;\ g = c, d, f;$

$$d \text{ and } e = \frac{(d \cap e)}{(d \cup e) - 2} = \frac{1}{5 - 2} = 0.33;$$

$$d \text{ and } f = \frac{(d \cap f)}{(d \cup f) - 2} = \frac{2}{5 - 2} = 0.667;$$

$$d \text{ and } g = \frac{(d \cap g)}{(d \cup g) - 2} = \frac{1}{6 - 2} = 0.25;$$

We got, $node\ e$, $node\ f$ and $node\ g$ are the members of community corresponding by their node centrality ($node\ d$). Figure 3 and Figure 4 are the extracted communities after identifying outlier.

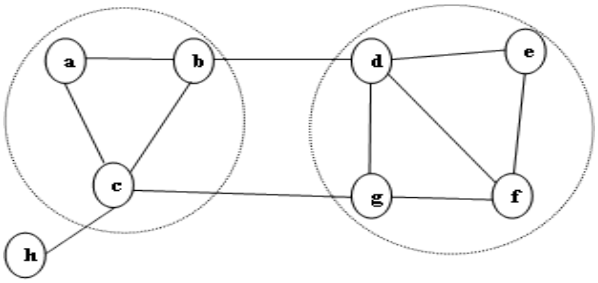


Figure 3: Community of node d Centrality

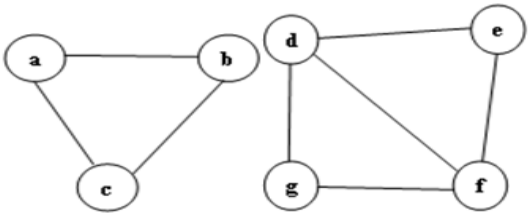


Figure 4: Two Communities after detecting and removing outliers

4.3. Descriptions of Algorithm

For undirected and un-weighted networks dataset, outliers is determined by a node which have no common values then identifying the communities based on the similarity measure in this approach. The following procedures are the steps of detecting outliers and communities:

Step1: Determine seed node by using vertex centrality method.
Step2: Compute the neighborhood overlap of seed nodes by using Jaccard Similarity measure.
Step3: If a node is adjacent to seed node and have overlapped value, determine it as a member of that seed node.
Step4: One node, the member of that seed node within the same community, has another node which is adjacent to its member node and there is overlapped value between them, then identify the linked node as the member of that related community.
Step5: One node which is adjacent to seed node but there is no overlapped values to that seed node is defined as outlier.
Step6: If there is no more adjacent node for that seed node; find another seed node which has the highest centrality value in the left stack of the degree centrality measurement and then repeat the process until there is no more node to be considered.

5. Experiments

5.1. Description of Dataset

In this paper, real undirected networks, Zachary Karate Club Dataset is used. In this Dataset statistics, "nodes" represents the number of friends; "edges" represents the number of friendship in the network. Zachary's karate club network is one of the popular studies in social network analysis and has been used as one of the typical test examples by many researchers to detect community structures in complex network. There are 34 member nodes, 78 edges and splits into two clubs, one is indicated as circles and the other is indicated as squares which are shown in Figure 5.

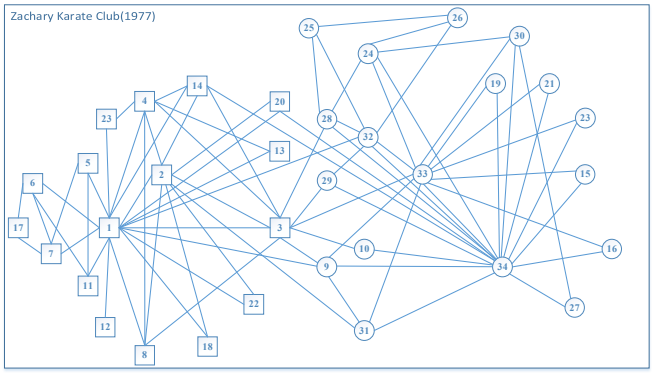


Figure 5: Zarchary Karate Club, Squares and circles indicate two communities observed by Zarchary (1977)

5.2. Evaluation Method

In this section, we show the results of communities and outliers. Before showing our experimental result, we first introduce the evaluation methods that are used in our proposed approach.

When there's no ground-truth for the communities to assess the quality of detected communities, it could be used quality measurement so called quality scoring functions which characterize how community like is the connectivity structure of a given set of nodes. The idea is that given a community scoring function, one can then find sets of nodes with high score and consider these sets as communities. All scoring functions build on the intuition that communities are sets of nodes with many connections between the members and few connections from the members to the rest of the network. And it can be formalized the intuition that "good" communities are both compact and well-connected internally while being relatively well-separated from the rest of the network. In quality measurement, it can be showed the quality of detecting communities in different point of view. There are many possible ways to mathematically formalize this intuition.

Among them it will be used the most useful five kinds of quality functions are as follows:

- Scoring functions based on internal connectivity
- Scoring functions based on external connectivity
- Scoring functions based on community goodness metrics
- Scoring functions based on a network model

5.2.1. Scoring functions based on internal connectivity

Internal density is defined as ratio between the intra-community edges and all the edges in the graph and is one of the simplest measures for community quality which is biased towards coarse-grained communities.
Internal density: $f(S) = \frac{m_S}{n_S(n_S-1)/2}$ is the internal edge density of the node set S where S be the set of nodes, m_S is the number of edges between the members of S and n_s is the number of nodes in S. The greater the value is the better for internal density.

5.2.2. Scoring functions based on external connectivity

The fraction of existing edges (out of all possible edges) leaving the cluster $f(S) = \frac{c_S}{n_S(n-n_S)}$ where n be the number of nodes in the network and c_S be the number of edges on the boundary of S. The smaller value is the better for this external connectivity compare with the internal density.

5.2.3. Scoring functions based on community goodness metrics

A clustering coefficient is a measure of the relationship to which people in a network tend to group together. Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterized by a relatively high density of ties; this likelihood tends to be greater than the average probability of a tie randomly established between two nodes. It is a real number between zero and one that is zero when there is no community, and one for maximal community, which happens when the network consists of disjoint cliques.

Clustering coefficient (CC) $CC = \frac{2N_V}{K_V(K_V-1)}$ is based on the premise that network communities are manifestations of locally inhomogeneous distributions of edges, because pairs of nodes with common neighbors are more likely to be connected with each other.

5.2.4. Scoring function based on a network model

Modularity is defined as having more internal edges and less external edges as is defined as modularity measures the strength

of each partition by considering the degree distribution. One main problem with modularity approach is that it cannot detect well defined small communities when the graphs are extremely large. let us talk about the brief definition of modularity. The Q value is between 0 and 1 and the real network modularity function value is generally between 0.3 and 0.7. Community structure is more obvious with the greater value of modularity.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

5.3. Results

To prove the proposed approach we first test it on the well-known karate friendship studied by Zachary, which has been become a classical studies workbench by many researchers for community detection algorithm testing. There are 34 member nodes, 78 edges and splits into two communities.

As shown in figure, node 34 and node 1 are seed nodes, the most important node among their neighboring is the main part of our proposed approach. We had to start those nodes to detect communities and identify outliers by using similarity measurement based on edge structure. We found that nodes which have no any overlapped value among their neighbors are determined as outliers which are node 10 and node 12. Table 2 is shown about the detected node community memberships better correspond to ground-truth node community memberships.

Table 2: Communities of Ground-Truth and Proposed Approach, their members and number of members in corresponding communities

Karate Dataset	Communities	Member (Node) in Community	No. of Members
Ground-Truth	Community1	34, 9,10, 15, 16, 19, 21, 23, 24, 25, 26,27, 28, 29, 30, 31, 32, 33	18
	Community2	1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22	16
Proposed Approach	Community1	34, 9, 15, 16, 19, 21, 23, 24, 25, 26 27, 28, 29, 30, 31, 32, 33	17
	Community2	1, 2, 3, 4, 5, 6, 7, 8, 11, 13, 14, 17, 18, 20, 22	15

Table 3: Community C, Number of Intra-Nodes (Intra-N), Number of Intra-Edges(Intra-E), Internal Connectivity(IC), External Connectivity(EC) and Average Clustering Coefficient (ACC)

C	Intra-N	Intra-E	IC	EC	ACC
1	17	34	0.25	0.12	0.7033
2	15	32	0.3	0.11	

Some methods based on the idea that nodes can be the member of two or more communities. But some condition, in thinking of which user should be situated on individual community. Our proposed approach intends to split the multiple communities clearly and remove the nodes which are not necessary to group into any communities. In OCNS method had been proved that they can detect the overlapping node definitely. However, they could not be considered which overlapped node is the member of which community exactly. In considering of detection community without studying the overlapping node combines spectral methods with clustering techniques, and uses the concept of modularity in order to develop a working algorithm and the quantitative of individual communities are different from the ground truth community. Moreover, even removing the outliers, the modularity value (Q) of the proposed approach is better than the other system as shown in Table4.

Table 4: Comparison with other methods in Modularity Value(Q).

Algorithms	Q	No. of Communities	Outlier (node)
GN	0.4013	5	-
Detecting network communities: a new systematic and efficient algorithm	0.4	5	-
OCNS	0.4304	2	-
Gephi	0.416	4	-
Proposed Approach	0.534	2	10,12

6. Conclusion

This proposed approach was used neighborhood overlapping with vertex similarity to detect community and outlier based on edge structure. It showed the steps of detecting outliers and communities in detail. Then, it was discussed about the evaluation measurement in different point of view because of different community criteria. The experiments on real Zachary Karate club network show that our algorithm outperforms other community based algorithms in terms of modularity value, number of communities and members in communities. On the other hand, our approach gets the high quality measurement in the assumption of no ground-truth community. In similarity measurement, nodes

which have the overlapped values are detected as communities correspond with its vertex centrality. Nodes which have no any overlapped values in the communities or which need not be necessary to group into the community is defined by outliers. However, in case of node which has no any overlapped value and is connected with multiple communities are still leaving as our future work.

References

[1] Htwe Nu Win and Khin Thidar Lynn, "Community Detection in Facebook with Outlier Recognition", 18thIEEE/ACIS International Conference on Softerwar Engineering, Artificial Intelligence, Networking and Parallel /Distrubuted Computing (SNPD 2017) , June 26-28, 2017, Kanazawa, Ishikawa, Japan, Pages 155-159.

[2] J.Y. Chen, O. R. Zaiane, R. Goebel, "Detecting Communities in Large Networks by Iterative Local Expansion", International Conference on Computational Aspects of Social Networks, 2009.

[3] Jaewon Yang and Jure Leskovec,"Defining and Evaluating Network Communities based on Ground-truth", the Proceedings of 2012 IEEE International Conference on Data Mining (ICDM), 2012.

[4] Jimeng Sun, HuimingQu, DeepayanChakrabarti, and Christos Faloutsos, "Neighborhood formation and anomaly detection in bipartite graph", ICDM, November 27-30 ,2005.

[5] Keith Henderson, Tina Eliassi-Rad, Christos Faloutsos, Leman Akoglu, Lei Li, Koji Maruhashi, B. AdityaPrakash, and Hanghang Tong, "Metricforensics: A multi-level approach for mining volatile graphs", In Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC, 2010, pages 163–172.

[6] Leman Akoglu, Mary McGlohon and Christos Faloutsos , "Anomaly Detection in Large Graphs", November 2009, CMU-CS-09-173.

[7] Lei Tang and Huan Liu, "Community Detection and Mining in Social Media", ISBN:9781608453559, A Publication in the Morgan & Claypool Publishers series, 2010.

[8] Luca Donetti and Miguel A Mu˜noz,"Detecting network communities: a new systematic and efficient algorithm", 2004 IOP Publishing Ltd PII: S1742-5468(04)87880-4 , doi:10.1088/1742-5468/2004/10/P10012, Journal of Statistical Mechanic: Theory and Experiment, an IOP and SISSA Journal, http://stacks.iop.org/JSTAT/2004/P10012.

[9] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", Phys. Rev. E, vol. 69, p. 026113, Feb 2004.

[10] M. E. J. Newman, "Modularity and community structure in networks", Proceedings of the National Academy of Sciences, vol. 103, no. 23, pp.8577–8582, 2006.

[11] M.Girvan, M.E.J.Newman, "Community structure in social and biological networks", Proc.Natl. Acad. Sci. USA 99, 7821-7826, 2002.

[12] Mathieu Bastian, SebastienHeymann and Mathieu Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks", Copyright 2009, Association for the Advancement of Artificial Intelligence (www. aaai.org)

[13] Meng Wang, Chaokun Wang, Jeffrey Xu Yu and Jun Zhang, "Community Detection in Social Networks: An In-depth Benchmarking Study with a ProcedureOriented Framework", 41st International Conference on Very Large Data Bases, August 31st September 4th 2015.

[14] Michel Plantie' and Michel Crampes, "Survey on Social Community Detection", Social Media Retrieval, Springer Publishers, Computer Communications and Networks, 978-1-4471-4554-7, pp.65-85, 2013.

[15] Mingming Chen, Tommy Nguyen and Boleslaw K. Szymanski, "On Measuring the Quality of a Network Community Structure", Proceedings of the IEEE Social Computing Conference, Washington DC, September 8-14, pp. 122-127,2013.

[16] S. Fortunato, "Community Detecction in Graph", Phys. Rep. 486, 75 (2010).

[17] William Eberle and Lawrence B. Holder."Discovering structural anomalies in graph-based data", In ICDM Workshops, 2007, pages 393– 398.

[18] Zachary, W. W.: "An information how model for conflict and fission in small groups", J. Anthropol. Res, 33 452-473. 1977.

[19] Zuo Chen, MengyuanJia , Bing Yang, and Xiaodong Li," Detecting Overlapping Community in Complex Network Based on Node Similarity", Computer Science and Information Systems 12(2):843–855 DOI: 10.2298/CSIS141021029C , 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 25-29 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Virtual Memory Introspection Framework for Cyber Threat Detection in Virtual Environment

Himanshu Upadhyay*, Hardik Gohel, Alexander Pons, Leo Lagos

Applied Research Center, Florida International University, Miami, 33172, United States

ARTICLE INFO

Article history:
Received: 01 November, 2017
Accepted: 13 December, 2017
Online: 18 January, 2018

Keywords :
Cybersecurity,
Virtualization,
Linux threat detection,
Hypervisor,
Feature selection technique,
Memory forensic analysis,
Virtual machine introspection

ABSTRACT

In today's information based world, it is increasingly important to safeguard the data owned by any organization, be it intellectual property or personal information. With ever increasing sophistication of malware, it is imperative to come up with an automated and advanced methods of attack vector recognition and isolation. Existing methods are not dynamic enough to adapt to the behavioral complexity of new malware. Widely used operating systems, especially Linux, have a popular perception of being more secure than other operating systems (e.g. Windows), but this is not necessarily true. The open source nature of the Linux operating system is a double edge sword; malicious actors having full access to the kernel code does not reassure the IT world of Linux's vulnerabilities. Recent widely reported hacking attacks on reputable organizations have mostly been on Linux servers. Most new malwares are able to neutralize existing defenses on the Linux operating system. A radical solution for malware detection is needed – one which cannot be detected and damaged by malicious code. In this paper, we propose a novel framework design that uses virtualization to isolate and monitor Linux environments. The framework uses the well-known Xen hypervisor to host server environments and uses a Virtual Memory Introspection framework to capture process behavior. The behavioral data is analyzed using sophisticated machine learning algorithms to flag potential cyber threats. The framework can be enhanced to have self-healing properties: any compromised hosts are immediately replaced by their uncompromised versions, limiting the exposure to the wider enterprise network.

1. Introduction

Dependency on computer systems has been growing exponentially in recent years. Government offices and business organizations are major targets of malicious actors for stealing highly valuable data. These entities are major Linux adopters nowadays because of its reputed safety and security. They are attempting to protect themselves from cyberattacks with digital defense techniques like encryption, firewalls and heuristic or signature scanning packages. Meanwhile, the number of attacks that involve infiltrating military data centers, targeting power grids, and stealing trade secrets from both private and public organizations continues to increase. The detection, response and reporting of these kinds of intrusions as well as other incidents involving computer systems, are crucial for cybersecurity professionals. This paper proposes a Virtual Memory Introspection based framework for detection, analysis and monitoring of malware behavior using memory forensics, during cyberspace attacks in a virtualized environment. The framework provides advanced instrumentation tools for control and monitoring of malware, fine-grained introspection of operating system Kernel and user process behavior using the well-known LibVMI technology, and analysis of behavioral data using state-of-art machine learning techniques. In study of research literature, one of the major methods of malware detection that has emerged over the years is Linux memory forensics [1] [2]. A number of authors have described novel detection systems using Memory Forensics, or using kernel data structure invariants as a reference frame to identify rootkit intrusions [3] [4]. The goal of this test technology is to facilitate threat assessment of malware, to understand its goals, and degrade impact on the compromised systems [5].

Further, as cyber-attacks continues to expand and the sophistication of the adversaries grows, defenders must adapt quickly in order to survive. There are numerous kinds of malware prevalent today, including Backdoors, Bots, Downloaders and Droppers, Ransomware, Rootkits, Scareware, Spyware, Exploits, Logic Bombs, Trojans, Viruses, Worms, etc. The detection,

*Dr. Himanshu Upadhyay, Email: upadhyay@fiu.edu

response and reporting of these kinds of intrusions as well as other incidents involving computer systems, are crucial for cybersecurity professionals. Present static methods used by anti-viruses are insufficient to stop advanced malware threats, which are capable of disabling the anti-viral software via root-kit mechanisms.

2. Virtualization & Data Analytics Framework

The proposed framework has the following major components as shown in Figure1:

2.1. Virtualization

The virtual machines, which are monitored for malware attacks, are collectively called the system under test (SUT). This system is hosted on a Xen hypervisor based virtualization platform. The behavioral data on the guest VMs is captured through Virtual Memory Introspection (VMI) using the LibVMI framework

2.2. Data Analytics

This platform consists of various traditional machine learning algorithms used to train the models and perform prediction using various test vectors consisting of malware and rootkits on the virtual machines (SUT). In this proposed framework, the authors are building an advanced data analytics platform on the database server, R runtime with in-memory analytics providing the scale and performance. Various traditional machines learning algorithms like Random Forest, Support Vector Machines, Logistics regression etc. will be utilized to perform Data Analytics. This platform analyzes the memory data structure captured on the VMI framework using cutting-edge machine learning algorithms. These algorithms are used to build the model to predict the malware behavior and display the results.

2.3. Test Control Center

The Test Control Center is an application that provides control and administration of the entire framework with an intuitive web based interface. The operator can create virtual machines (SUT), install benign applications and malware, capture behavioral information and use the Data Analytics platform to build the model and test virtual machines for malware using various traditional machine learning algorithms. Test Control Center is a centralized application to manage the test bed and execute various test cases. In the proposed framework, the authors are developing various modules like virtual machine management, network map, test case management, model management, configuration, testbed administration and help.

The three platforms working together will enable an organization to detect malware almost in real time and monitor its behavior in a virtual environment. The user provides a set of well-defined inputs to the virtualization platform based on predetermined test cases, and captures the kernel data structures information and transfers it to the Data Analytics platform. The Data Analytics platform uses traditional machine learning algorithms to build the model and predicts results based on the information extracted from guest virtual machine memory and displays the results on the Test Control Center. The following diagram in Figure 1 provides the overall system design of the framework:

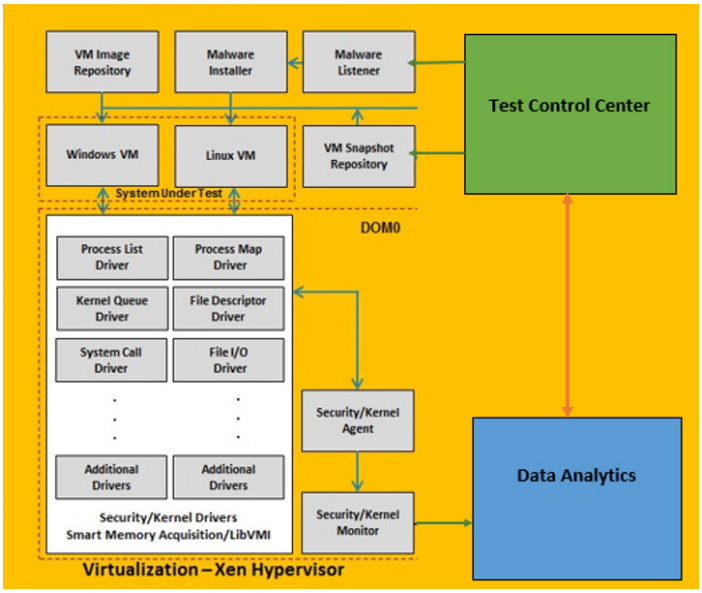


Figure 1: Virtualization & Data Analytics Framework System Diagram

The various components of the virtualization platform are discussed below.

I. VM Image Repository

Images of the different virtual machines will be stored in this location. It provides the user the ability to select the virtual machine images that will create the virtual machines for SUT with different configurations.

II. Malware Listener and Installer

This module listens to the command from the malware management module and communicates with the malware installer on the virtual machine to install the selected malware.

III. VM Snapshot Repository

This module will store the snapshots of normal operation which will later be used to roll back to the clean state after the test execution.

IV. Multiple Linux VMs

These are the guest virtual machines created by the operator based on predefined configurations.

V. Kernel Agent

Security/kernel agent manages all the drivers to extract the kernel data structures from the SUT. They are deployed on DOM 0 which is the secured location on the hypervisor with required privileges.

VI. Kernel Drivers

These the individual drivers which are used for smart memory acquisition of the kernel data structures from the SUT. These drivers use the LibVMI libraries for virtual memory introspection from the Dom0 level to extract kernel data structures. They are managed by Security/Kernel Agent.

VII. Kernel Monitor

This module communicates with data analytics platform and sends the data extracted by the Security/kernel agent to the listener on the data analytics platform to store the data in to the database.

VIII. Kernel Agent Listener

The Agent Listener plays a crucial role in the virtualization framework. This module communicates with the security/kernel monitor in the virtualization framework, and receives the extracted data from the security/kernel agent and stores it into the central database.

3. Virtual Memory Introspection (VMI) Framework

The VMI platform of this framework is built on the "Xen Hypervisor" which enables the operator to create multiple virtual machines with different configuration and perform system testing with the environment. The Linux machines are used as SUT, to be monitored for malware intrusion are hosted on this hypervisor. Xen hypervisor [6] provides two security domains for the hosted virtual machines. The virtual machine running under Dom0 (Domain 0) on the Hypervisor controls the resource allocation for the virtual machines that runs the Linux Guest operating system [7].

The major key objectives of virtualization framework are listed below:

I. Develop innovative ways to detect and monitor malware in a virtualized testbed with smart memory acquisition.

II. Deployment of virtualized environments along with advanced tools developed through this research for control and monitoring of the cyberspace test environment.

III. Cyber-attack emulation through infection and propagation of simulated endpoints.

IV. Virtual machine introspection, data collection, and monitoring of various aspects of the infrastructure through a centralized system.

V. Display results on the Test Control Center to monitor the impact of malware on SUT.

The SUT can be Linux or Windows guest virtual machines. Using the Test Control Center, the operator is able to create guest VMs on the Xen hypervisor from a virtual machine repository, introduce pre-determined malware into them, and capture kernel data structure information and stores them into the central database. We have developed an application called "Kernel Agent" that runs on the Dom0 virtual machine, and uses kernel drivers to perform smart memory acquisition on the guest operating system. This kernel information, which includes process details, memory data structures, and file system information, is written to a central database for data analytics.

Xen hypervisor is used to host, configure, and control the guest Linux virtual machine that will be tested for malware. The introspection is carried out through the "LibVMI" framework in conjunction with Google's "Rekall" profiles. The Rekall profile is a JSON file that contains the address mappings of all Linux Kernel data structures. By integrating LibVMI with Google Rekall, the process of extracting Kernel configuration parameters and data structures at run-time can be automated. This facilitates the smart memory acquisition of process behavioral data from the kernel. We are using Google's "Go" programming language for writing the introspector application, and to push the acquired data to the central database. The Go programming language can integrate with suitable C/C++ libraries, and significantly reduces the time needed for development and testing new build of each module. It provides better control over distribution and parallelization mechanisms of the Security Agent. We are using "Libvirt" library for management of the guest virtual machines through a custom built Introspector. Introspection requests are serviced by the LibVMI library, while the Libvirt library is used to facilitate the creation, starting, stopping, pausing and resuming of the guest Linux virtual machines [8].

The Introspector is comprised of several core subsystems that are necessary for lifecycle management and introspection of virtual machines on a host. The Introspector also maintains two socket connections that listen to requests for introspection or virtual machine administration coming from the Test Control Center.

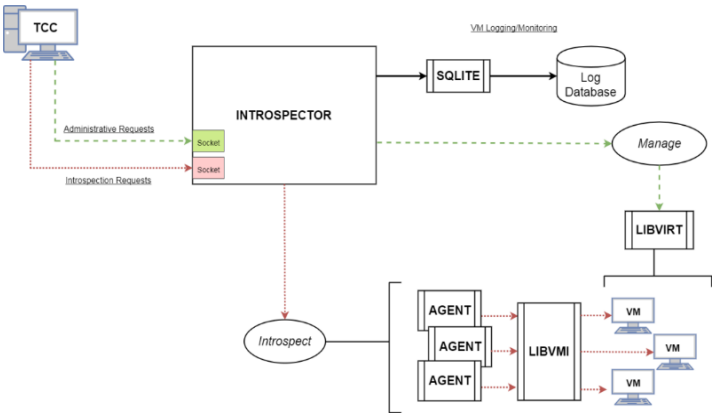


Figure 2: Proposed Virtual Machine Introspection Framework

Introspector: startup

Introspector (introspector.go) builds a Settings object (settings.go) by reading the values in the introspector.conf settings file. The configuration file is used to determine which IP/Port combination to use for handling virtual machine administration requests and introspection requests. The Introspector will also start the StateManager that manages a SQLite database containing the current state of the virtual machines in the system.

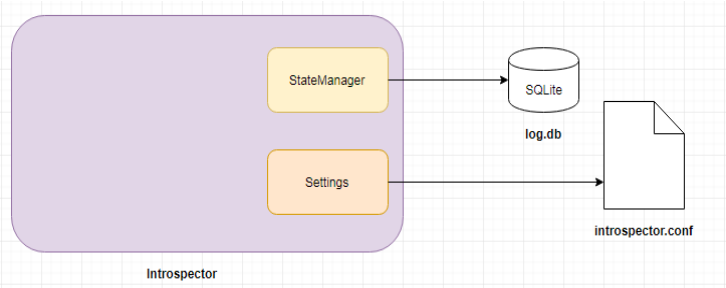


Figure 3: Introspector setup with Database

Next, a thread is spawned that opens a socket connection to listen for incoming requests for virtual machine administration from the Test Control Center while the main Introspector thread waits on another socket for incoming introspection requests. Administration requests are XML requests that have the root node <LibvirtServerMessage>

4. Research on Kernel Data Structures

The analysis conducted on the footprint of a process in kernel space is in near real-time, to distinguish between benign and malicious processes. The Linux Kernel maintains data structures that contain information about every action and resources used by a process. The study thus far has identified 15 features from 118 features which can effectively distinguish between benign and malicious processes [9] [10]. The next steps in the pipeline is to perform virtualization to extract threads, system call , invariant data structures like system call table and interrupt descriptor table, IP addresses, network sockets, URLs, open files, passwords, catches, clipboards and other user generated content, encrypted keys, and configurations of hardware and software [11].

5. Benign and Malware Applications

We have also performed research to extract the kernel data structures from the Linux kernel task structure using LibVMI library. The following Benign processes and Malwares samples have been identified to be installed on the Linux virtual machines. The Linux kernel version under consideration as of the writing of this paper is 3.16.0-23- generic.ko in Ubuntu 14.04 [11] [12].

GUI Applications	CLI Applications
1) Firefox - Web Browser 2) Thunderbird - Email client application 3) gpaint - Paint Application 4) Libre - Office writer 5) Rythmbox - Music player 6) Connectagram - Word unscrambling game 7) Arora - Cross platform web browser 8) Empathy- Internet messaging application 9) Alarm clock 10) Shotwell- Photo manager	1) Alpine-Pico- Is a text editor, uses the pine email client for writing email messages. Run from the terminal as pico, pico.alpine 2) Aaphoto- An image manipulation tool for automatic color correction of photos 3) ACL2- Programming language in which user can model computer systems and a tool to help prove properties of those models. 4) Python- Programming language 5) Gedit Text editing tool 6) Calcoo - scientific calculator 7) Calcurse- text based calendar and todo manager 8) Clamav-daemon- antivirus utility for unix-scanner daemon run from the terminal 9) Gzip- Zipping application

6. Linux Kernel Data Structure Extraction

At the outset we are going to extract the following task structure list from Linux data structures. Based on our research, the following list of features has been identified for extraction to test the virtualization framework in the pilot stage. These are the primary features of the processes that will run in the kernel of the Linux virtual machine at runtime [13] [14].

Sr. No.	Features' Name	Description
1	map_count	Number of memory regions of a process
2	page table lock	Used to manage the page table entries
3	hiwater rss	Number of page frames that a process owns
4	shared_vm	Number of pages in shared file memory mapping of a process
5	exec_vm	Number of pages in exec. memory mapping of process
6	nr_ptes	Number of page tables owned by a process
7	utime	Execution time of a process in user mode (tick count)
8	stime	Execution time of a process in kernel mode (tick count)
9	nvcsw	Volunteer context switches of a process
10	nivcsw	Involuntary context switches
11	total_vm	Size of process's address space in terms of Number of pages
12	min_flt	Minor page faults of a process
13	alloc lock.raw lock.slock	Used to lock memory manager, files and file system etc.
14	hiwater_vm	Max Number of pages appeared in memory region of process
15	fs.count	fs_struct's usage count to indicate the restrictions

7. Data Analytics and Results

The research on data analytics is currently on-going [15] [16]. The Data Analytics platform consists of centralized database servers for analytics and processing. The research will be focused on identifying the classification and anomaly detection machine learning algorithms that includes open source and commercial platforms, libraries with these platforms [17] [18] to detect malware [19] [20] that primarily focus on the Linux data structure extracted by the VMI Framework.

Once the framework is narrowed down and tuned as a viable solution for malware detection, the results will be summarized to demonstrate which machine learning models are powerful for malware detection and monitoring. The research outcome will consist of three components: memory data structures, algorithms with pros/cons and machine learning model performance data.

8. Conclusion and Future Work

The goal of the VMI framework research is to provide solid foundation for security of the Linux operating system as well as the capabilities to identify and monitor cyber threats in virtual environment. The key challenges of this research is to identify various data structures affected by modern malware, and in-depth

examination of various machine learning algorithms with memory forensics to solve key cybersecurity issues.

The focus of this research is to develop a system that can work with the latest Linux operating systems providing memory forensics capability in a virtual environment. In the future, we plan to explore all possible ways of data analytics with big data technologies and deep learning. With sufficient time and effort, the development of this framework can result in an extremely powerful system for early threat detection and defense.

Acknowledgment

This research is supported by Department of Defense (DOD) – Test Resource Management Center (TRMC), USA. We thank our colleagues who provided insight and expertise, directly or indirectly, that greatly assisted the research.

References

[1] M H Ligh, A Case, J Levy, A Walters. "The Art of Memory Forensics", 2014
[2] M Wade, "Memory Forensics: Where to Start" at http://www.forensicmag.com/ article/2011/06/ memory-forensics-where-start, 2011
[3] Baliga, A., Ganapathy, V. and Iftode, L., Detecting kernel-level rootkits using data structure invariants. IEEE Transactions on Dependable and Secure Computing, 8(5), pp.670-684, 2011.
[4] D. Levy, H. A. Gohel, H. Upadhyay, A. Pons and L. E. Lagos, "Design of Virtualization Framework to Detect Cyber Threats in Linux Environment," 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), New York, NY, pp. 316-320, 2017
[5] H. Gohel. "Introduction to Network & Cyber Security", 2015
[6] Xen Project available at https://www.xenproject.org/, 2013
[7] H. Gohel. "Looking Back at the Evolution of the Internet." CSI Communications - Knowledge Digest for IT Community Vol. 38 Issue. 6 pp. 23-26, 2014
[8] N Joshi, D.B.Choksi, "Implementation of process forensic for system calls", International Journal of Advanced Research in Engineering and Technology (IJARET), ISSN 0976 – 6480(Print), ISSN 0976 – 6499(Online) Vol. 5, Issue. 6, pp. 77-82, 2014
[9] Blackbag Team, "MEMORY FORENSICS", at https://www.blackbagtech.com/blog/2016/03/07/ windows-memory-forensics/2016
[10] Farrukh Shahzad, M. Shahzad, Muddassar Farooq, "In-execution dynamic malware analysis and detection by mining information in process control blocks of Linux OS" Information Sciences, Volume 231, pp. 45-63, ISSN 0020-0255,2013
[11] N Joshi, D. B. Choksi, "Implementation of Process Forensic for System Calls", International Journal of Advanced Research in Engineering & Technology (IJARET), Vol. 5, Issue. 6, pp. 77 - 82, ISSN Print: 0976-6480, ISSN Online: 0976-6499, 2014.
[12] H. Upadhyay, H. Gohel "Security Corner: Cyber Threat Analysis with Memory Forensics", CSIC – Knowledge Digest for IT Community, Vol. 40, Issue. 11, ISSN 0970-647X, pp. 17-19, 2017
[13] H. Upadhyay, H. Gohel "Design of Advanced Cyber Threat Analysis Framework for Memory Forensics", International Journal of Innovative Research in Computer and Communication Engineering, ISSN (Online): 2320-9801, ISSN (Print): 2320-9798, Vol. 5, Special Issue 2. pp.132-137, 2017
[14] F. Shahzad, S. Bhatti, M. Shahzad and M. Farooq, "In-Execution Malware Detection Using Task Structures of Linux Processes," 2011 IEEE International Conference on Communications (ICC), Kyoto, pp. 1-6, 2011.
[15] Z. Gu, Z. Deng, D. Xu and X. Jiang, "Process Implanting: A New Active Introspection Framework for Virtualization," IEEE 30th International Symposium on Reliable Distributed Systems, Madrid, pp. 147-156, 2011
[16] Hal Pomeranz, "Detecting Malware with Memory Forensics", at http://www.deer-run.com/~hal/ Detect_Malware_w_Memory_Forensics.pdf, 2015
[17] Hizver, Jennia, and Tzi-cker Chiueh. "Real-time deep virtual machine introspection and its applications." ACM SIGPLAN Notices. Vol. 49. No. 7. ACM, 2014.
[18] Hardik, Gohel. "Data Science - Data, Tools & Technologies." CSI Communications Knowledge Digest for IT Community, Vol. 39 Issue. 3, pp. 8-10, 2015
[19] H Gohel, P Sharma. "Study of Quantum Computing with Significance of Machine Learning." CSI Communications - Knowledge Digest for IT Community, Vol. 38, Issue. 11, pp. 21-23, 2015
[20] E Mariconti, O Lucky, P. Andriotis, "Detecting Android Malware by Building Markov Chains of Behavioural Models", NDDS'17, San Diego, USA, 2017

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 6, 28-39 (2017)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

A novel beamforming based model of coverage and transmission costing in IEEE 802.11 WLAN networks

Mehdi Guessous*, Lahbib Zenkouar

Mohammadia School of Engineers, Electrical Engineering Departement, Mohammed V University In Rabat, 10000, Morocco

ARTICLEINFO

Article history:
Received: 21 August, 2017
Accepted: 19 September, 2017
Online: 30 Spetember, 2017

Keywords:
Beamforming
Co-channel Interferences
Radio Resource Management
Transmit Power Control
Wireless Coverage
Wireless Local Area Network

ABSTRACT

IEEE 802.11 WLAN indoor networks face major inherent and environmental issues such as interference, noise, and obstacles. At the same time, they must provide a maximal service performance in highly changing radio environments and conformance to various applications' requirements. For this purpose, they require a solid design approach that considers both inputs from the radio interface and the upper-layer services at every design step. The modelization of radio area coverage is a key component in this process and must build on feasible work hypotheses. It should be able also to interpret highly varying characteristics of dense indoor environments, technology advances, service design best practices, end-to-end integration with other network parts: Local Area Network (LAN), Wide Area Network (WAN) or Data Center Network (DCN). This work focuses on Radio Resource Management (RRM) as a key tool to achieve a solid design in WLAN indoor environments by planning frequency channel assignment, transmit directions and corresponding power levels. Its scope is limited to tackle co-channel interference but can be easily extended to address cross-channel ones. In this paper, we consider beamforming and costing techniques to augment conventional RRM's Transmit Power Control (TPC) procedures that market-leading vendors has implemented and related research has worked on. We present a novel approach of radio coverage modelization and prove its additions to the cited related-work's models. Our solution model runs three algorithms to evaluate transmission opportunities of Wireless Devices (WD) under the coverage area. It builds on realistic hypotheses and a thorough system operation's understanding to evaluate such an opportunity to transmit, overcomes limitations from compared related-work's models, and integrates a hierarchical costing system to match Service Level Agreement (SLA) expectations. The term "opportunity" in this context relates also to the new transmission's possibilities that related-work misses often or overestimates.

1 Introduction

Designers face problems of two natures when building WLAN networks: issues that are inherent to radio resources' planning and those concerning their integration with other network parts and upper-layer services in response to developing customers' application needs such as mobility, real-time, interactive applications, business intelligence, and on-presence analytics. An accurate design should consider thoroughly both aspects and base its decision on a solid theoretical approach, on-field feedbacks, and best practices.

Design best practices dictate running two site surveys: one at the beginning of the project, before network's deployment, and a second, after network's deployment, to confirm resources' planning results and to calibrate the system's parameters in accordance to the real network's condition. An optional survey may be required if radio environment or its utilization has

*Mehdi Guessous, 7 avenue du General de Gaulle 91120 Palaiseau France, +33(0)614945485 & mehdiguessous@research.emi.ac.ma

https://dx.doi.org/10.25046/aj020604

changed and might affect the overall network capacity. The first survey qualifies mainly the exposure to on-reach foreign autonomous radio systems, and radio characteristics of the surface to cover (walls, floors, and other obstacles). The second site survey confirms the conformance of the actual deployment to design constraints in two ways:

1. by running a passive in-data-path survey that is based on passively reported statistics such as RSSI, SNR, PHY, MAC errors, packet count, etc. at radio interface level,

2. and an active one where in-data-path patterns simulate wireless client's traffic classes to gather actively statistics on Jitter, RTT, or any other measure concerning the rendered service, application or experience.

This approach is not very accurate as it lies on a set of preconfigured unchangeable parameters that result from special case deployments and testing contexts, which may be completely different from the on-field reality. In the market-leading implementations such as Cisco TPC or Aruba ARM, we observe that at a certain point, a reference is made to an "ideal" preconfigured parameter. The third neighbor's RSSI and a predefined hysteresis threshold [1] paremeters are used in the first case, and the "coverage index" [2] parameter in the second case, to decide on transmit power level planning network wide.

Inaccuracies come also from studies when they model the system as a pre-shaped radio coverage area in the form of a hierarchy of disks that delimit transmission ranges from interference or no-talk ones. Some limitations arise also from no-site-survey techniques that base their radio resource management solely on WLAN reported measures. Other limitations come from simplifications such as to consider that a client association and the corresponding CSMA/CA mechanisms rely entirely on the strongest access point's RSSI, because most of the interferences do not occur at MAC level, and radio measurements may differ from a vendor to another. Then it is necessary to ensure that model analysis inputs and outputs do not result in a contradiction.

Motivated by advances in beamforming techniques, especially from an array signal processing perspective, which make it possible to radiate energy in any direction and to easily estimate signal's direction of arrival, this work solution model aims at fixing many of previous limitations by considering the opportunity for an access point to transmit in one specific direction. This opportunity processing accepts data from upper-layer services such as to achieve a complete end-to-end integration with other network parts.

As it will be demonstrated in this work, the number of transmit directions at access point level, could be optimized in conjunction with planning transmit power levels and channel assignment to maximize the overall network capacity. A cost-based approach prices in a timely manner each direction for an eventual transmission. Access points in this scheme discover each other over-the-air and report useful information on transmit directions, power levels and channels, to a central intelligence that coordinates the overall network operation. The ultimate result is:

1. Cancellation of co-channel interferences in a majority of deployment cases.

2. Processing of new transmission opportunities that are missing in conventional work.

3. Enhancement of end-to-end network performance.

In the upcoming section, we present a foundation on unified WIFI architectures and beamforming as they relate to this work. In Sect. 4, we discuss more formally the problem and in Sect. 5, we present our solution model. Sect. 6 discusses our solution results. In the end, we conclude and further our work. This paper is an extension of work originally presented in 2017 International Conference on Information Networking (ICOIN) [3].

2 Related Work

Co-channel interference in WLAN networks is a major problem that has been widely studied. To tackle this issue RRM and related techniques such as Transmit Power Control (TPC), Dynamic Channel Assignment (DCA), constitute a very accurate solution. If we consider TPC that is the focus of this work, the related-work's approaches differ from each other in terms of the adopted interference's model, inputs TPC may work on, and their processing (predictive, heuristic, etc.). We further focus on related-work's adopted interference models.

To our knowledge, all models and derived TPC processing have considered only the transmission power level as a degree of freedom to limit co-channel interference. This work claims that the transmission direction could be an additional degree of freedom to overcome other models' limitations when it comes to modeling irregular coverage areas in indoor WLAN environments.

Authors in this work [4] concentrate, from a lower layer perspective, on the co-channel interferences as a function of the estimated distance between the Access Points (AP) and the Wireless Devices (WD). In this scheme, the WDs estimate the distance based on the RSSI information sent by the APs in RTS control packets. The transmission power level adjusts accordingly to achieve a better network performance and interference cancellation. The issue with this model is that the power level adjustment affects other on reach WDs and APs communications. In addition, this model does not include cases where the WD does not simply receive the RTS packets.

In [5], a WD measures the amount of interference as a function of the local AP's load, its signal's

strength and loads from the distant APs. Here interference is a linear function of the AP's load. This is interesting but as opposed to our work, this model misses the interpretation of interference from an energy radiation perspective. In our work, we still include AP's load consideration in the SLA applied to the processing of transmit opportunities.

Depending on the coverage area pattern: disks or Voronoi zones, the amount of interference in [6], is a function of different AP's areas overlap in the first case and to their borders in the second one. This work model is very accurate to describe how an AP's energy radiation could affect other neighboring APs. However, it is still insufficient to interpret many cases specific to indoor WLAN deployments that relate to on-reach and unreachable AP's. Subsection C of Sect. 4 discusses these cases in detail.

The authors in [7] consider both MAC and PHY layer inputs: data payload length, path loss condition, and frame retry counts, to find out the suitable PHY mode and power level combination. This procedure requires a control packet exchange that is similar to conventional RTC/CTS for this adaptation. This model is easy to implement but it is insufficient to interpret many cases specific to indoor WLAN similarly to the previous work. In addition and similarly to the first cited work, any PHY mode and corresponding power level change apply to all neighboring APs and WDs indistinguishably.

In [8], the authors concentrate on upper layer application performance such as FTP and HTTP, to find out an optimal power scheme or more exactly, an optimal combination between transmit power and supported load for a given distance between nodes. This model considers SLA as in the second and fourth cited works and augments its processing to include "behavioral" protocol aspects. However, the corresponding model misses the lower-level interpretation of AP's radiated energy.

The authors in [9], approach the problem from an inter-protocol coordination point of view: WIFI, BT, etc. which may be very helpful to base an integrated-system wide transmit decision or "etiquette" on the utilization of shared radio resources. In this scheme, all wireless devices broadcast the information on spectrum usage. This information includes transmit power level and frequency channel of use. In addition and based on costing policies, the system allocates resources: frequency, power and time, to wireless devices. This work is a generalization of the first and fourth previously cited works and may incur the same limitations.

In this work, we review the conventional co-channel wireless coverage models and show that the system could process more transmission opportunities and still conform to the SLA. In these works [4] [5], some of the transmission opportunities were missed in areas that were considered falsely as interference ranges or worst as no-talk ones. Our solution model considers inputs from upper-layer services to conform to the SLA on the end-to-end wireless client's experience as in works [7] [8] [9]. In addition, it takes advantage of the technological advances in beamforming and related techniques, to introduce, at a conceptual level in indoor WLAN networks, the direction of transmission as a new radio resource added to frequency channel and power level, to enhance work's like [9] results.

3 Theoretical Background

The next two subsections describe WIFI's unified architecture in addition to the beamforming techniques as they relate to our study. The main objective of this foundation is to justify the feasibility of the hypotheses we base our study on. These hypotheses relate to the possibility to correlate reported radio interface's information from different APs, to coordinate their operation network-wide, to concentrate the radiated energy in any direction, in a timely manner, and to estimate accurately the direction of known and unknown sources' signals.

3.1 WIFI's Unified Architecture

In dense indoor networks, the transmission is difficult to model and depends on radio environment's characteristics such as obstacles, interference, background noise, density of wireless clients, mobility applications, etc.

In this context, standalone APs' architectures do not scale with high number of APs and WDs that require high class QoS treatment. Such architectures are replaced gradually by controller-based or "unified" ones. They are marked as unified for two reasons:

1. a central decision-making and intelligence, that is a Wireless LAN Controller (WLC), manages the overall network's operation.
2. the WLAN integrates with the overall network parts: LAN, WAN, DCN, etc. from especially a QoS perspective.

The market-leading vendors implement such a central decision-making processor mainly in three ways:

1. appliance WLC-based,
2. virtual system WLC-based,
3. distributed AP-based.

In the latter implementation, the APs take over the WLC's role. The first two implementations require a WLC, a virtual or physical appliance, which is reachable by all network APs over the wire. Examples of such WLCs are Cisco 8540 Wireless Controller and Aruba 7280 Mobility Controller.

Cisco unified architecture [10] defines two protocols:

1. CAPWAP, used by APs to build protocol associations to the active WLC.

2. NDP, to send messages over the air (OTA) for APs to exchange proprietary and standard management or control information.

In addition to these protocols, Cisco APs integrate a set of on-chip RRM techniques: CLIENTLINK and CLEANAIR that monitor and measure radio environment characteristics among other features. The WLC gathers this information via the already established CAPWAP tunnels to APs. Further, platforms such as Cisco Prime Infrastructure (CPI) and Mobility Services Engine (MSE) extend the capability of CLEANAIR to perform analytics, locate clients, interferers, and build heat maps. Based on the APs' reported information, the WLC computes the channel assignment and power levels map network wide. Its decision still conforms to the pre-configured policy sets. Each of these sets defines a number of configurable variables and unchangeable settings such as acceptable signal levels, tolerable noise levels, usable power levels, frequencies and channels, etc.

In this work context, the WIFI unified architecture guarantees a coordinated operation between all network elements at a very low level, an accessible distributed intelligence, and a centralized decision-making. It focuses on how to enhance the centralized decision-making at WLC level and especially RRM processing. RRM is the set of tools, algorithms and services that processes inputs from AP's radio and LAN interfaces and outputs a wireless area coverage plan or map that conforms to a predefined utility function.

3.2 Beamforming

The majority of WLAN communicating systems are equipped with omnidirectional antennas or a set of limited number of unidirectional antennas to obtain an equivalent energy radiation's pattern. Ease and cost of such hardware configurations are some of the benefits of such designs but they lack many aspects regarding interference and noise cancellation, signal sources' localization, and power consumption.

Beamforming techniques by definition relate to both the direction of energy radiation and the estimation of signal's arrival direction [11] and from an array signal processing perspective, they offer the possibility to tackle previous issues and design limitations to the next level in three ways:

1. the radiation of energy toward actual receivers allows better performance and less impact on the other system's elements.

2. the estimation of direction of known and unknown sources allows better costing of a given directed transmission.

3. the on-chip beamforming techniques allows for a hardware real-time adaptation to extremely changing radio environments like in our context.

With reference to [11], beamforming concerns both sides of a communicating system: the emitter and the receiver. It allows the radiated energy to be concentrated in one direction and to estimate its corresponding source's location for both legitimate and rogue interferers at receiver level. Basically, beamforming could be achieved mechanically by manually or electronically gearing a directional antenna. From a signal processing perspective, beamforming is achieved by changing certain transmitter's characteristics such as array elements' phase shifting, source signals' amplitude increasing or addition weighting. Additionally, the desired array gain and beam's direction depend on the number and geometry of array elements.

To ease our study we consider that array beamformers achieve equivalent performance to mechanical ones concerning both the resultant beam radiation's angle and signal's gain.

Let us consider an antenna with a beamwidth of $2 * \theta_n$ such as the radiated energy in the zone delimited by this angle is equal to the maximum energy radiated by the antenna minus $n\ dB$. n is calculated in such manner to allow an AP to transmit in all corresponding directions at the same time without interfering with each other. A rough estimation of this number is:

$$N \approx (2 * \pi)/(2 * \theta_n) \tag{1}$$

In the contexts of [12] [13] studies on antenna arrays, the beamwidth is expressed as a function of the distance between array elements d, the number of elements M, and the wave length λ:

$$\Delta\theta_{main} \approx (2 * \lambda)/(d * M) \tag{2}$$

In general, the evolving antenna array's beamformers could be classified as data independent, statistically optimum, adaptative or partially adaptative [11]. In this work, data dependent data beamformers (statistically optimum and adaptative) are of interest. Some of these beamformers are Multiple Sidelobe Canceller (MSC), Reference Signal (RS), Max SNR, Linearly Constrained Minimum Variance (LCMV), Lean Mean Square (LMS), Recursive Least Squares (RLS).

The other part of beamforming relates to the estimation of the Direction of Arrival (DOA). DOA estimation methods may include Bartlett, Capon, MUSIC, Min-Norm, DML, SML, WSF, Root-Music, ESPRIT, IQML, and Root-WSF. SML, WSF and Root-WSF are statistically the most efficient depending on the array geometry [14]. They could be yield into two main categories: spectral-based techniques, which are computationally optimal, and parametric methods, which are more accurate than the previous ones.

Recent algorithms are more accurate in the estimation of signal's arrival of both known and unknown sources. They depend on the array size, antenna's radiation pattern, source signals' correlation,

data statistics, corresponding data generation framework, among others. However, for the rest of this study we could consider that beamforming techniques allow:

1. sufficient number of directions to mimic an omnidirectional operation,
2. geographically any direction is achievable,
3. patterns from different directions at the same AP level do not interfere and are uncorrelated,
4. on-chip real-time transmits and synchronization.

4 Problem Statement

In this section, we formalize mathematically the problem of determining the opportunity to transmit at any point under the wireless coverage area. We first focus on related-work's models. Then in the next sections, we present our solution model and compare its results to preceding studies.

In this problem statement, we focus mainly on how related-work estimates the interference at a given point under the wireless coverage area, applies radio and upper-layer constraints, and processes the corresponding transmit opportunity.

To ease this work we consider the wireless coverage area as a two-dimensional Euclidean plane. The APs and WDs are points of this plane. They transmit in the same channel and cause the majority of co-channel interference in this area. This study focuses on co-channel operation but is easily extendible to tackle cross-channel interference or noise issues in the context described before.

Let us define:

$\{AP_1, AP_2, \ldots, AP_n\}$— set of n access points.

P_j — a point under wireless coverage area.

(x_i, y_i)— coordinates of AP_i or P_i points in an Euclidean two-dimensional plane.

$w_i, w_{min}, w_{max}, w_{opt}$— AP_i's corresponding current, minimum, maximum and optimum transmit power levels.

$d_{i,j}$ — distance between AP_i and P_j points.

$I(), O()$— interference and transmit opportunity functions.

We categorize related-work's models into two categories: Range-based and Zone-based. In the upcoming subsections, we describe each of them and discuss their limitations.

4.1 Range-based Model

In works similar to [4], an AP's wireless coverage corresponds to one of these ranges: a transmission, interference or no-talk range. These ranges correspond to the estimation of the distance between the AP and a receiving point P (AP or WD). Further, they consider that an AP's wireless coverage pattern is omnidirectional in the form of a circle or a disk. In this configuration, the three circles that correspond to each range type and centered at the AP define the whole AP's wireless coverage area as in Figure 1.

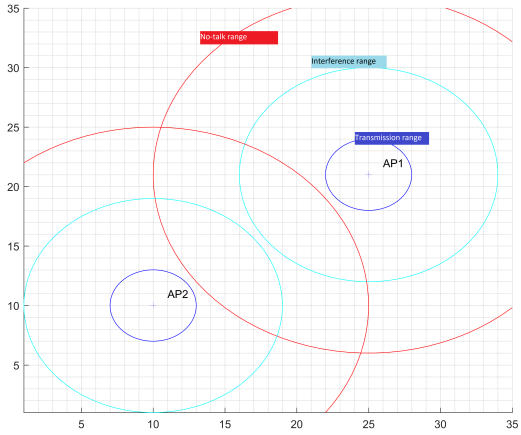


Figure 1: Range-based model's representation of an AP's wireless coverage

Let us further define:

$C_{i,tx}, C_{i,if}, C_{i,nt}$— AP_i attached circles that delimit corresponding transmission, interference and no-talk ranges.

$R_{i,tx}, R_{i,if}, R_{i,nt}$— AP_i corresponding ranges.

If we suppose that:

$$R_{i,tx} \leq R_{i,if} \leq R_{i,nt} \tag{3}$$

Then the interference and transmit opportunity are given as follows:

$$I(P_i) \approx \sum_{k=1}^{3} \alpha_k \sum_{j \neq i} \beta_j * intersection(C_{j,k}, C_{i,1}) \tag{4}$$

$$O(P_i) \approx \frac{1}{\sum_{k=1}^{3} \alpha_k \sum_{j \neq i} \beta_j * intersection(C_j, P_i)} \tag{5}$$

A weighted version of these functions is given further to account for inaccuracies and other constraints. The α and β weights are meant to reflect the non-linearity of interference addition, the power level's effect on interference strength, and the type of transmission's range. Here we consider that the interference at a given point P_i is a function of other AP_j's

signals except from the AP_i to which the point is associated. The opportunity processing is slightly different as it considers all AP's interference to P_i that is not associated to any AP_i yet.

4.2 Zone-based Model

Works such as [5] [6] that are based on this model are different from the previous ones as ranges here are not function only of the corresponding AP transmission's characteristics: channel, power level, etc., but depend also on the neighboring APs. The result of this is that the transmission range shape is no more a solid circle but a convex polygon with straight sides. Each straight side defines a borderline that separate two neighboring APs' transmission ranges. Consequently, it is important to note that a point in a transmission range of one AP could not be in another AP's transmission range. A Zone-based AP's wireless coverage is represented in Figure 2.

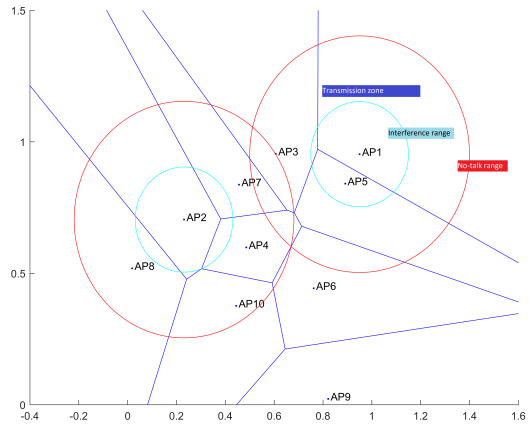


Figure 2: Zone-based model's representation of an AP's wireless coverage

Let us further define:

G_i — AP_i corresponding transmission range polygon or zone.

Then the interference and transmit opportunity are given as follows:

$$I(P_i) \approx \alpha_1 \sum_{j \neq i} \beta_j * intersection(C_{j,1}, C_{i,1})^{-\lambda} + \sum_{k=2}^{3} \alpha_k \sum_{j \neq i} \beta_j * intersection(C_{j,k}, C_{i,1}) \quad (6)$$

$$O(P_i) \approx \frac{1}{\sum_{k=2}^{3} \alpha_k \sum_{j \neq i} \beta_j * intersection(C_{j,k}, G_i)} \quad (7)$$

Similarly to formulas 4 and 5, α and β weights are meant to represent the impact of AP's transmission characteristics. λ is representative of the neighboring APs impact on AP_i borderlines. In this configuration, only transmission range is affected. The processing of other ranges are the same as in the previous model. It is also to note that opportunity in this model correspond to P_i in AP_i transmission's range or zone G_i rather than intersection of the neighboring APs like in the previous model.

4.3 First Observations

The presented models' functions of interference and opportunity do not reflect the accuracy of interference and opportunity measurements but rather the concept behind them and their possibilities to account for different WLAN design cases and aspects. At this stage, we see clearly that both models are limited in these ways:

1. both models are limited to consider that the strength of interference is only inversely proportional to the distance of an AP from interfering neighbors.
2. both models would interpret an increase in a transmission power level as an expanded reach in all directions: uniformly in case of Range-based models but depending on neighboring APs in case of Zone-based ones.
3. a point could not be in two transmission ranges of two different APs at the same time in Zone-based models.
4. transmission range does not depend on the strength of interference in Range-based models.
5. both models would interpret falsely obstacles to the signal propagation, as a weaker signal from an AP in the context of indoor WLAN networks does not mean necessarily that this AP is out of reach.
6. similarly to the previous point a stronger signal does not mean necessarily that an AP is in reach: it signal may be guided under some conditions.

The consequence of these limitations in indoor WLAN networks, is to reduce or overestimate the transmit opportunities that the network may offer. The upcoming section presents our solution to fix these modelization limitations.

5 Solution model

Our solution model is suitable for WLC-controlled indoor networks in the already explained conditions. For the purpose of this study, we suppose that APs are equivalent, support MIMO technologies such as beamforming and DOA, and operate over the same channel. The wireless devices (WD) are equipped with omnidirectional antennas and transmit at a lowered power level in comparison with APs.

The model runs three algorithms:

1. Transmit Direction Discovery (TDD) algorithm that builds a per direction neighbor discovery map, computes the optimal number of transmit directions and the corresponding per direction power levels.
2. Transmit Direction Mapping (TDM) algorithm estimates the impact that may have an AP on the network when it transmits in a specific direction and builds consequently, a per-direction transmit cost-based map that represents this information.
3. Transmit Direction Opportunity (TDO) algorithm processes data from both TDM and the upper-layer network services to optimize the transmission opportunities and build a per-direction transmit opportunity-based map to represent this information.

The next workflow in Figure **??** is an overview of our solution processing.

At initialization, the number of transmission directions is set to the minimum and the power level to the maximum. TDD processing stops after all APs are discovered and exactly one or no neighbor AP is discovered at any AP direction.

Based on results from TDD, TDM categorizes the coverage area points and costs a potential transmission at each of them. The coverage area points belong to one of these three categories: Not-On-Any-Transmit-Direction (NOATD), On-Discovered-Neighbor-Direction (ODND), and Not-ODND-On-Transmit-Direction (NOOTD). After the coverage area point's categorization, the system processes the transmission costs.

TDO algorithm evaluates the opportunity to transmit at a specific coverage area point. It returns the AP that may handle the transmission and the corresponding direction. This processing accepts inputs from TDM and the upper-layer SLA. The SLA processing is tight to the AP's wired network interface as opposed to the radio interface.

5.1 TDD algorithm

TDD algorithm processes the optimal number of transmission directions that APs may support and the corresponding per-direction transmit power levels. To ease this study, we consider these hypotheses and simplifications:

1. only a 2D plane operation is considered.
2. a uniform beamwidth in all directions.
3. APs are able to achieve an emulated omnidirectional operation.
4. two signals from the same AP in two different directions could not interfere.
5. all APs have the same optimal and maximal numbers of transmission directions and corresponding power levels.
6. neighbor discovery is bidirectional, occurs at the opposite direction, and at the same transmit power level.
7. an AP could not discover the same neighbor AP in two separate directions.
8. an AP could not discover itself.
9. at least one neighbor AP is reachable in any AP direction.

Points 2, 3, 4 and 5 are idealistic physical characteristics of AP's beamforming implementation. The 9^{th} point indicates that a co-channel interference condition is detected from a neighboring AP. The 6^{th}, 7^{th} and 8^{th} assumptions are further developed in separate work to cover and interpret cases where:

1. the optimal numbers of transmission directions and power levels are not the same among APs.
2. neighbor discovery is unidirectional or asymmetric.

Let us define:

(d_i, w_i)— the direction and power level associated with AP_i.

$d_{-j}(-x,-y)$— the opposite direction of $d_j(x,y)$ in an euclidean 2D plane.

$L(AP_i, d_j, w_k)$— the segment that represents the transmission range of AP_i in d_j direction at w_k transmit power level.

$a(i,j)$, $b(i,j)$, $c(i,j)$— respectively the segment $L()$ slope, intercept and end point.

$X(AP_i, d_j, w_k)$— the set of discovered AP neighbors by AP_i in d_j direction at w_k power level.

$O(AP_i, AP_j)$— a segment that represents an obstacle between AP_i and AP_j.

We formalize the problem as:

$$\forall i \in \mathbb{N}, \quad \exists j,k \in \mathbb{N} \quad \text{s. t.} \quad X(AP_i, d_j, w_k) \neq \emptyset \tag{8}$$

$$\forall i,j,k \in \mathbb{N}, \quad AP_i \notin X(AP_i, d_j, w_k) \tag{9}$$

$$\forall i,j,j',k,k' \in \mathbb{N}, \quad \text{and} \quad j \neq j', \quad X(AP_i, d_j, w_k) \cap X(AP_i, d'_j, w'_k) = \emptyset \tag{10}$$

$$\forall i,i',j,j',k,k' \in \mathbb{N}, \quad \text{and} \quad i \neq i', \quad AP_{i'} \in X(AP_i, d_j, w_k) \equiv AP_i \in X(AP_{i'}, d_{j'}, w_{k'}) \quad \text{and} \quad d_{j'} = d_{-j}, \quad w_k = w_{k'} \tag{11}$$

$$\forall i,j \in \mathbb{N}, \quad L(AP_i, d_j) = \left\{ a_{i,j} * x + b_{i,j} \quad | \quad x \in \left[x_i, x_i + c_{i,j} \right] \right\} \tag{12}$$

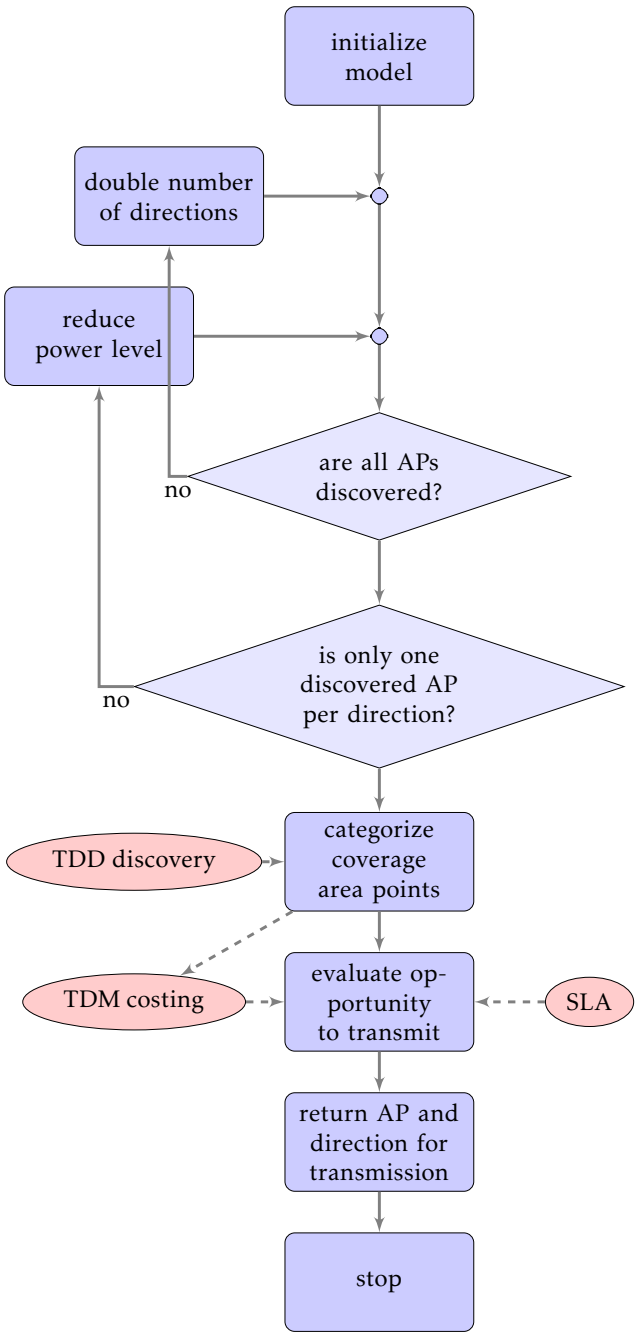


Figure 3: Flowchart of our solution model processing

The APs' distribution on the plan could be any random function or resulting from on-site surveys. However, we consider that the minimum and maximum distances between any of these APs are linear functions respectively of the corresponding minimum and maximum supported power levels. Figure 4 represents a simple network of seven APs with eight transmission directions each. Eight beams in red color are formed roughly around AP_i corresponding transmission directions.

The AP neighbor's per-direction discovery is based on the distance between APs that are on each other theoretical transmission range, on-site surveys, and radio interface sensing. Additionally the solution model allows the discovery of these two design special-case neighbors:

1. "isolated" neighbor AP that is at reach but separated by an obstacle.
2. "guided" neighbor AP that is not at reach but received with an acceptable signal level.

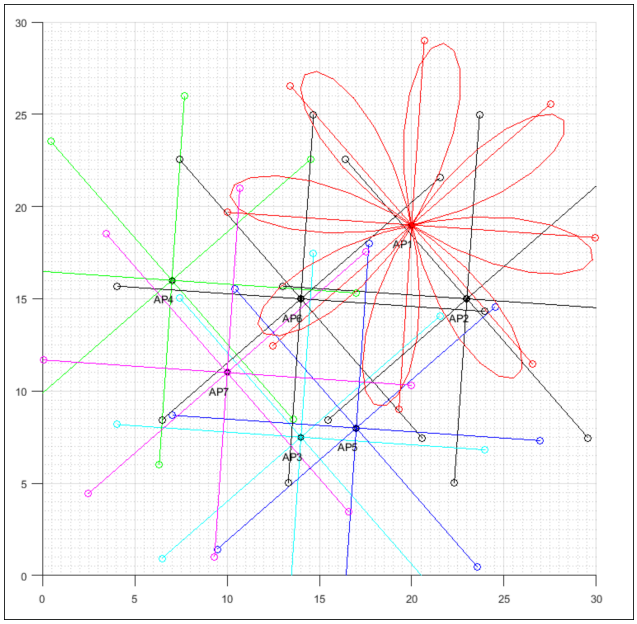


Figure 4: TDD Discovery Map example of 7 APs and 8 transmit directions

With reference to Figure 4 network's representation, the results of neighbor AP's discovery are given in Table. 1.

Table 1: Neighbor AP's discovery results

AP	d_1	d_2	d_3,	d_4	d_5	d_6	d_7	d_8
AP_1	AP_6	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	AP_2	$\emptyset$
AP_2	AP_5	AP_6	AP_1	$\emptyset$	$\emptyset$	$\emptyset$	AP_6	$\emptyset$
AP_3	$\emptyset$	$\emptyset$	AP_7	AP_6	$\emptyset$	AP_5	$\emptyset$	$\emptyset$
AP_4	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	AP_6	$\emptyset$	AP_7	$\emptyset$
AP_5	$\emptyset$	AP_3	$\emptyset$	$\emptyset$	AP_2	$\emptyset$	$\emptyset$	$\emptyset$
AP_6	AP_7	AP_4	$\emptyset$	$\emptyset$	AP_1	AP_2	$\emptyset$	AP_3
AP_7	$\emptyset$	$\emptyset$	AP_4	$\emptyset$	AP_6	$\emptyset$	AP_7	$\emptyset$

The optimization of APs transmission directions' number is a three-step process. For directions with more than one discovered neighbor, we calculate a new number such as to have the least possible number of neighbors per any AP's direction. Otherwise, we adjust the power levels to get some of these APs out of reach from the local AP. If no success, we revert to Range-based-like interference scheme to cost this transmission zone.

A simpler form of the proposed algorithm is given next:

1: $(d^{opt}, w^{opt}) \leftarrow (d^{min}, w^{max})$
2: $X \leftarrow \bigcup_{j=1}^{opt} X(AP_i, d_j^{max}, w_{opt})$
3: **for** $j = 1$ to opt **do**
4: **if** $opt \leq max$ **then**
5: **if card**$(X(AP_i, d_j^{opt}, w_{opt})) > 1$ **then**

$d_{opt} \leftarrow d_{2*opt}$
end if
else if card$(X(AP_i, d_j^{opt}, w_{opt})) > 1$ **then**
repeat
if $\bigcup_{j=1}^{opt} X(AP_i, d_j^{opt}, w_{opt-1}) = X$ **then**
$w_{opt} \leftarrow w_{opt-1}$
end if
until $\forall j$, **card**$(X(AP_i, d_j^{opt}, w_{opt})) \leq 1$
end if
end for

The optimum numbers of directions and power levels are respectively initialized to the minimum and the maximum numbers supported by AP_i. For every AP's direction, we check the number of discovered neighbors. If many neighbors are detected, we multiply the initial number of directions by two until we obtain only one neighbor per direction. If the maximum number of direction is reached then we reduce by one level the power and try to get the same result without altering X. X is the set of all neighboring APs that were detected previously in the maximum supported directions and at the maximum supported power level.

5.2 TDM algorithm

In the previous section, TDD returned the optimal number of transmission directions, the corresponding optimal number of power levels and the sets of AP's per-direction discovered neighbors. In addition to these elements, TDD interprets the presence of two kinds of obstacles: those that are preventing isolated neighbors from seeing each other, and those that are creating new transmission opportunities among guided neighbors. One way for TDD to interpret such kind of obstacles is by comparing the on-the-air discovery's results with on-site surveys' information. Based on the previous elements, TDM algorithm builds a transmission-costing map for APs to reach any point under the overall wireless coverage area.

A simpler form of the proposed algorithm is given next:

for $j = 1$ to opt **do**
if $P_i \in L(AP_i, d_j)$ and $X(AP_i, d_j) \neq \emptyset$ **then**
$Case_1$
else if $P_i \in L(AP_i, d_j)$ and $X(AP_i, d_j) = \emptyset$ **then**
$Case_2$
else
$Case_3$
end if
end for

TDM costing system categorizes WDs as NOATD, ODND, or NOOTD. These WDs could be represented roughly by circles with sufficiently small radii in comparison with APs transmit ranges, to not false processing results. We process carefully NOOTD points, which are not on discovered-neighbor APs' directions, as they are hard to detect.

Let us define:

$C_i^W D$— WD attached circle.

ρ_i — weights associated with each category.

$Cost(P_i)$— function of transmissions cost at P_i

Then for this preliminary work, the cost of a transmission could be expressed as follows:

$$Cost(P_i) \approx \begin{cases} \rho_1 I(P_i), & \text{if } NOATD \\ \rho_2 I(P_i), & \text{if } ODND \\ \rho_3 I(P_i), & \text{if } NOOTD \end{cases} \quad (13)$$

Where,

$$I(P_i) \approx \alpha_1 \sum_j \sum_k \beta_{j,k} * intersection(L_{j,k}, C_i^{WD}) + \sum_{k=2}^{3} \alpha_k \sum_{j \neq i} \beta_j * intersection(C_{j,k}, C_{i,1}) \quad (14)$$

Cost calculation results were plotted on Figure 5, for the same Figure 4 network example set of seven APs and eight transmission directions. On this map, transmission costs range from white color, lowest cost points, to red color that correspond to the highest costly points.

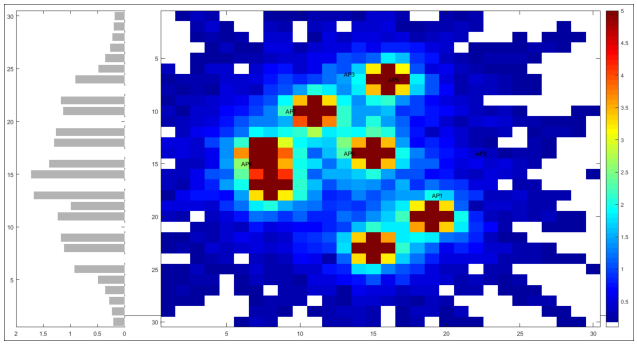


Figure 5: TDM Cost Map example of 7 APs and 8 transmit directions

5.3 TDO algorithm

TDO takes into consideration the worst-case analysis done by TDM when all the APs transmit at the same time in all supported directions at the maximum power level. Then for a given transmission, it prioritizes the directions with the least cost in duality with TDM but in addition, it integrates data from the upper service layers and returned SLA tests both on the passive and active communication's paths.

A more complex form of this opportunity processing relies on a system-wide APs transmit operation's synchronization to cancel costly directions and on-the-air tests to optimize NOOTD costs.

Let us define:

$s_{1,i}, s_{2,i}$— AP_i weights associated with defined SLA passive and active test results.

$O(P_i)$— opportunity of transmission function at P_i.

Then for this preliminary work, the transmission opportunity calculation at any point P_i is given by:

$$O(P_i) = s_{1,i} s_{2,i} \overline{I(P_i)} \tag{15}$$

The transmission opportunity processing results were plotted on Figure 6 for a set of seven APs and eight transmit directions.

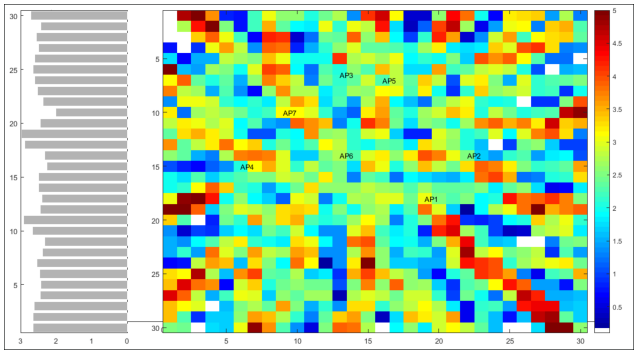


Figure 6: TDO Opportunity Map example of 7 APs and 8 transmit directions

$s_{1,i}$ and $s_{2,i}$ weights are calculated based on the simulated SLA active test results that may include: Packet Loss Rate, Round-Trip Time (RTT), Jitter, and on SLA passive measurements at PHY and MAC levels like: error rates, transmitted packets, etc. that are directly taken from the radio and LAN interfaces. In addition, these weights take into consideration the nature of the end-user application: real-time applications do not require the same service from the network as bulk ones.

5.4 Layered Transmit Opportunity Processing

With regard to this work opportunity processing, a transmission is a four-step procedure: discover, cost, evaluate, and transmit. Discovery is primarily done by APs and could be extended to any WD. The interpretation of discovery results, costing, and opportunity processing, are done at WLC level. Costing in this work has been solely based on interference estimation but could integrate any other radio interface's measurements: Signal to Noise Ratio (SNR), Received Channel Power Indicator (RCPI), and so on. Opportunity evaluation is based both on previous radio interface costing and on measures from upper layer services. These two measures have to be fully uncorrelated to not false results.

5.5 Transmit Synchronization

In this configuration, a transmission is not tight to the access point WD is associated with, but rather to the AP that offers the best transmit opportunity among a set of APs. In fact, the WD is associated with a virtual or pseudo AP that is representative of a set of physical APs with common characteristics with regard to this association. Any physical AP attached to the pseudo-AP could source an effective transmission to the associated WD. In the context of this work, we match the WD's corresponding pseudo-AP to a single physical AP.

6 Evaluation and Conclusion

At this stage, the evaluation is based on Matlab simulations of both our model and models from related-work [4] [5] [8] Range-based, and [6] Zone-based solutions. In the next step, we consider implementing this solution on a Linux-based APs and WDs and compare its performance with simulated related-work and vendor's implementations.

We compare our model performance to both Range-based and Zone-based models in these three transmission conditions:

1. a free space transmission condition,
2. the presence of obstacles, "isolated" neighbor's condition,
3. "guided" neighbor's condition.

We first evaluate the compared transmission opportunity for any set of 30 APs and ten times a randomly selected 100 WDs on a 2D grid. To ease this work we limit the maximum number of APs to 30 and the number of WDs test points to 100.

Figure 7 shows that for 30 APs and 10 random sets of 100 WDs on the grid, the average performance calculations is almost: 77% for our solution, 68% for Zone-based and only 57% for Range-based models. Figure 8 shows the average transmit opportunity for each model per WDs sets.

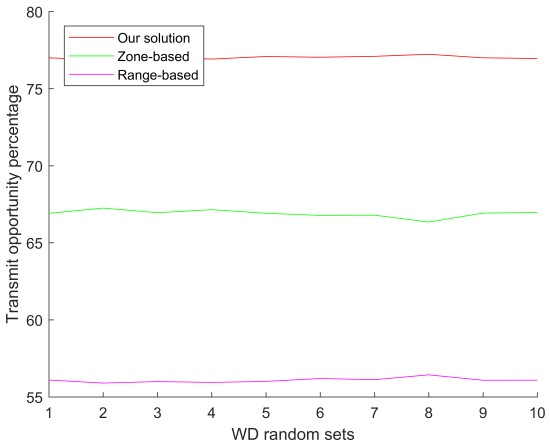


Figure 7: Performance comparison for 30 AP and 10 random sets of 100 WDs

Second, as in Figure 9, we evaluate the APs number impact on the compared performance for any set of 100 WDs. APs number ranges from 1 to 30. In Figure 10, we compare Zone-based to Range-based models.

We notice a general tendency of all models to perform similarly when APs are of limited number. However, for greater APs' numbers, the Zone-based models perform better than Range-based ones as in Figure 10. In Figure 9, our model returns a better transmit opportunity estimation when APs' number increases in comparison to other models.

Third, in Figure 11, we vary the number of the supported directions that is applicable only to our model and observe the compared average performance. For this test purpose, this number is tight to a variable α that takes these values: 12.5%, 25%, 50% and 75%. We only compare our model to Zone-based one.

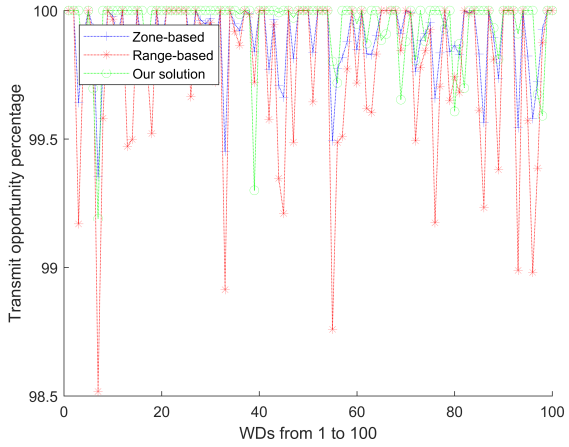


Figure 8: Average performance comparison per WDs in a network of 30 APs and 100 WDs

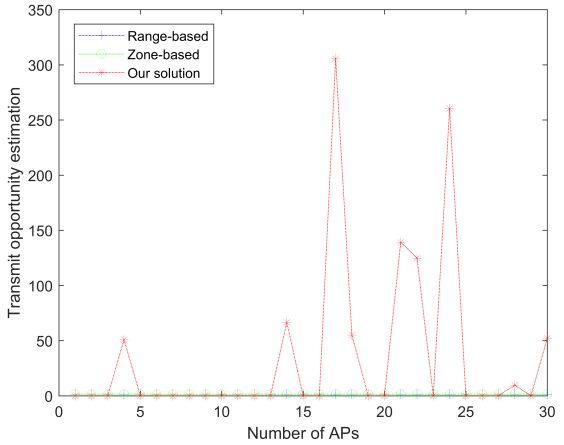


Figure 9: Comparison of our solution, Zone and Range-based models for any set of 100 WDs by APs number

We see clearly that our solution is comparable to Zone-based one when α is almost 75%. Our solution is near optimal when α is 12,5% but this is idealistic. However, achieving a 50% α is realistic with regards to these-days vendors APs system characteristics. A 50% α coefficient may correspond roughly to four transmission directions or beams.

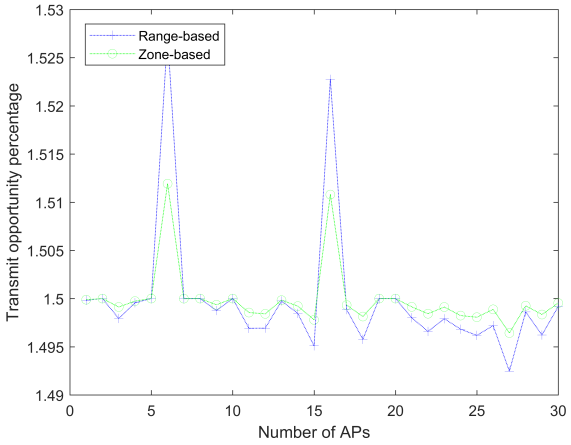


Figure 10: Comparison of Zone and Range-based models for any set of 100 WDs by APs number

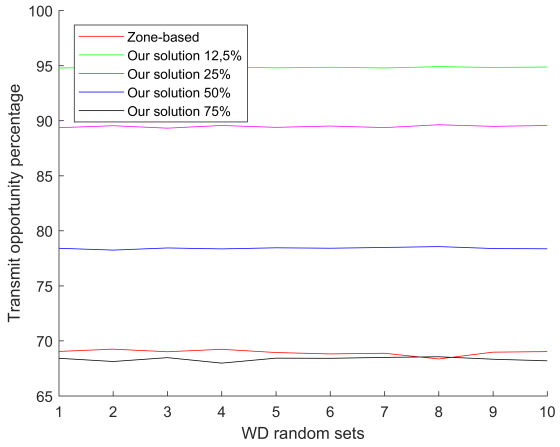


Figure 11: AP directions number impact on performance

As a conclusion, this work transmission processing is built on a new costing and opportunity evaluation model that leverages advances in antenna array processing such as beamforming techniques and WLAN integrated design best practices. In addition to the transmission channels and corresponding transmission power levels, it considers the direction of transmission or beam as an additional degree of freedom of dynamic RRM purposes. A comparison to other models indicates that our model is stable and performs better in very dense WLAN indoor and WLC-based network deployments.

Further work may focus on:

1. more complex neighbor AP discovery scenarios,
2. angle of incidence in TDM costing,
3. APs transmit operation coordination,
4. more complex costing layered approach for end-to-end to communications.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment We would thank colleagues : researchers, engineers, and anonymous reviewers for sharing their precious comments and on-field experience that improved the quality of this paper.

References

[1] "Radio Resource Management White Paper." [Online]. Available: https://www.cisco.com/c/en/us/td/docs/wireless/controller/technotes/8-3/b_RRM_White_Paper.html

[2] "ARM Coverage and Interference Metrics." [Online]. Available: http://www.arubanetworks.com/techdocs/ArubaOS_64x_WebHelp/Content/ArubaFrameStyles/ARM/ARM_Metrics.htm

[3] M. Guessous and L. Zenkouar, "Cognitive directional cost-based transmit power control in ieee 802.11 wlan," in *Information Networking (ICOIN), 2017 International Conference on*. IEEE, 2017, pp. 281–287.

[4] R. Ruslan and T.-C. Wan, "Cognitive radio-based power adjustment for wi-fi," in *TENCON 2009-2009 IEEE Region 10 Conference*. IEEE, 2009, pp. 1–5.

[5] N. Ahmed and S. Keshav, "A successive refinement approach to wireless infrastructure network deployment," in *Wireless Communications and Networking Conference, 2006. WCNC 2006. IEEE*, vol. 1. IEEE, 2006, pp. 511–519.

[6] R. Prateek and P. Om, "Interference-constrained wireless coverage in a protocol model," in *Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems Terromolinos, Spain: ACM*, 2006.

[7] A. Akella, G. Judd, S. Seshan, and P. Steenkiste, "Self-management in chaotic wireless deployments," *Wireless Networks*, vol. 13, no. 6, pp. 737–755, 2007.

[8] D. Qiao, S. Choi, A. Jain, and K. G. Shin, "Adaptive transmit power control in ieee 802.11 a wireless lans," in *Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual*, vol. 1. IEEE, 2003, pp. 433–437.

[9] D. Raychaudhuri and X. Jing, "A spectrum etiquette protocol for efficient coordination of radio devices in unlicensed bands," in *Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on*, vol. 1. IEEE, 2003, pp. 172–176.

[10] "Enterprise Mobility 8.1 Design Guide." [Online]. Available: https://www.cisco.com/c/en/us/td/docs/wireless/controller/8-1/Enterprise-Mobility-8-1-Design-Guide/Enterprise_Mobility_8-1_Deployment_Guide.html

[11] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[12] V. Rabinovich and N. Alexandrov, "Typical array geometries and basic beam steering methods," in *Antenna Arrays and Automotive Applications*. Springer, 2013, pp. 23–54.

[13] T. C. Cheston and J. Frank, "Phased array radar antennas," *Radar Handbook*, pp. 7–1, 1990.

[14] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE signal processing magazine*, vol. 13, no. 4, pp. 67–94, 1996.

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 28-39 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Software and Hardware Enhancement of Convolutional Neural Networks on GPGPUs

An-Ting Cheng*,[1], Chun-Yen Chen[1], Bo-Cheng Lai[1], Che-Huai Lin[2]

[1]*Institute of Electronics Engineering, National Chiao Tung University, Hsinchu, 300, Taiwan*

[2]*Synopsys Taiwan Co., Ltd., Hsinchu, 300, Taiwan*

ARTICLE INFO

Article history:
Received: 30 November, 2017
Accepted: 05 February, 2018
Online: 08 March, 2018

Keywords:
Convolutional Neural Network
GPGPU
Design and Optimization

ABSTRACT

Convolutional Neural Networks (CNNs) have gained attention in recent years for their ability to perform complex machine learning tasks with high accuracy and resilient to noise of inputs. The time-consuming convolution operations of CNNs pose great challenges to both software as well as hardware designers. To achieve superior performance, a design involves careful concerns between exposing the massive computation parallelism and exploiting data reuse in complex data accesses. Existing designs lack comprehensive analysis on design techniques and decisions. The analytical discussion and quantitative proof behind the design criterion, such as choosing proper dimensions to parallelize, are not well studied. This paper performs a series of qualitative and quantitative studies on both the programming techniques and their implications on the GPU architecture. The observations reveal comprehensive understanding on the correlation between the design techniques and the resulting performance. Based on the analyses, we pinpoint the two major performance bottlenecks of CNN on GPGPU: performing computation and loading data from global memory. Software and hardware enhancements are proposed in this paper to alleviate these issues. Experimental results on a cycle-accurate GPGPU simulator have demonstrated up to 4.4x performance enhancement when compared with the reference design.

1. Introduction

Convolutional Neural Networks (CNNs) have gained attention in recent years for their ability to perform complex machine learning tasks. Followed by winning the 2012 ImageNet competition, CNNs have demonstrated superior results in a wide range of fields including image classification, natural language processing and automotive. In addition to high accuracy in object recognition, systems using CNNs are more robust and resilient to noise in the inputs when compared to conventional algorithmic solutions. However, the enormous amount of computing power required by CNNs poses a great challenge to software as well as architecture engineers. The most time-consuming operation in a CNN is the convolution operation, which takes up over 90% of the total runtime. Therefore, the convolution operation becomes one of the most important concerns when implementing CNNs.

GPGPUs have demonstrated superior performance on CNN by exploiting the inherent computation parallelism. Due to the scaling architectures as well as ease-of-programming environment, GPGPUs are among the most widely adopted platforms for CNN. However, it is not a trivial task to have an efficient CNN design on a GPGPU. To achieve superior performance, a design involves careful concerns between exposing the massive computation parallelism and exploiting data reuse in complex data accesses.

This paper is an extension of work originally presented in ICASI 2017 [1]. In this paper, we perform a series of qualitative and quantitative studies on both the programming techniques and their implications on the GPU architecture. The observations reveal comprehensive understanding on the correlation between the design techniques and the resulting performance. There exist several frameworks and libraries that provide solutions to performing convolution on GPGPUs, such as cuDNN [2], Caffe [3], fbfft [4], and cuda-convnet2 [5]. Among the existing solutions, cuda-convnet2 is one of the widely used open-source implementations that enable superior performance on a variety of CNN schemes [6]. It employs design techniques and optimization strategies mainly for NVidia GPGPU architectures. However, while providing a solid implementation, cuda-convnet2 lacks

*Corresponding Author: An-Ting Cheng, National Chiao Tung University, Email: ericcorter78.ee01@g2.nctu.edu.tw

comprehensive analysis on its design techniques and decisions. The analytical discussion and quantitative proof behind the design criterion, such as choosing proper dimensions to parallelize, are not well studied. In addition, current GPGPUs are not designed specifically for convolution. There exist potential architecture enhancements that could significantly enhance the computation efficiency with minor hardware and software cost.

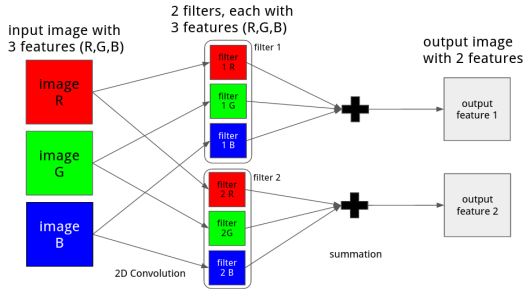


Figure 1: Convolution in CNN

This paper performs a series of qualitative and quantitative studies on both the programming techniques and their implications on the GPU architecture. The studies focus on the widely adopted NVidia GPGPU architecture and CUDA programming environment. The observations reveal comprehensive understanding on the correlation between the design techniques and the resulting performance. Based on the analyses, we pinpoint the two major performance bottlenecks of CNN on GPGPU: performing computation and loading data from global memory. Software and hardware enhancements are proposed in this paper to alleviate these issues. In the computation part, we demonstrate how to avoid excessive local memory accesses, the operations that severely degrade performance, by applying loop unrolling. We then propose two simple yet effective hardware accelerators to speed up the computation of the partial sums. In the data-loading part, we identify that a significant fraction of the time is spent on calculating addresses in the inner loops of CNN. We propose two software techniques to considerably reduce the computation. A low-cost address generator is then introduced to speed up the address calculation. Experimental results on a cycle-accurate GPGPU simulator, GPGPU-sim [7], have demonstrated up to 4.4x performance enhancement when compared with the original cuda-convnet2 design.

The rest of the paper is organized as follows. Section 2 discusses the implementation of the convolution kernel. The techniques of exposing parallelism and data reuse will also be discussed in Section 2. Section 3 and 4 describe the proposed software and hardware improvements to the convolution kernel. Section 5 discusses previous related works and Section 6 presents the conclusions.

2. Implementing Convolution on GPGPU

2.1 Convolution in CNN

Convolution is the basic building block as well as the most time-consuming operation in CNNs. The convolution operation in CNN does more than just convolving two 2D matrices. It takes a set of trainable filters and apply them to the input images, creating one output image for each input image. Each input image consists of one or more multiple feature maps, which means that every pixel in an image contains several features. For example, each pixel in an RGB picture contains three features: red, green and blue. Each filter also has the same number of features that correspond to the input images. When applying the filters to an input image, the filters are convolved across the width and height of the image, and the product of each feature is summed up to produce an output feature map. The number of features in each output image is therefore equal to the number of filters. Figure 1 illustrates an example of the convolution between one image and 2 filters with 3 features, producing one output with 2 features.

The high-level algorithm of the convolution operation in CNN is listed in Figure 2, where conv2 represents computing the 2D convolution between two 2D matrices. The inputs and outputs of the algorithm are all arranged in 4-dimensional arrays. Inputs to the algorithm are the image array images (image_count, height, width, image_features) and the filter array filters (image_features, filter_size, filter_size, filter_count). The output is the array outputs (image_count, height, width, filter_count). Each of the parameters used in the algorithm is described as below.

```
01: for (i = 0; i < image_count; i++) {
02:   for (j = 0; j < filter_count; j++) {
03:     result = zeros(height, width);
04:     for (k = 0; k < image_features; k++)
05:       result += conv2(images(k, :, :, i),
06:                       filters(k, :, :, j));
07:     outputs(j, :, :, i) = result;
08:   }
09: }
```

Figure 2: Pseudocode of convolution operations

image count. This parameter is the size of input mini-batch. A mini-batch contains multiple independent input images to be processed. Each image in the mini-batch will be processed by the same set of filters to produce one output image.

width, height. These two parameters are the width and height of the input images. All input images in the mini-batch have the same size. In this paper, the size of output images is the same as the size of the input images.

image_features. This parameter indicates the number of features maps in an input image. This is also the number of feature maps in a filter.

filter_size. In the context of CNNs, filters are always square-shaped. Therefore, we use only one parameter, which is filter_size, to represent both the width and height of a filter. As a result, the number of pixels in a filter is (filter_size * filter_size).

filter_count. This parameter is the total number of filters. Because each filter produces an output feature map, filter_count is also the number of output feature maps.

2.2 Exploiting Parallelism and Data Reuse in Convolution

This paper uses cuda-convnet2 [5] as the reference implementation of CNN on GPGPUs. Cuda-convnet2 is one of the widely used open-source implementations that enable superior

performance on a variety of CNN schemes [6]. It employs design techniques and optimization strategies mainly for NVidia GPGPU architectures. This section will discuss how to implement the convolution algorithm efficiently with CUDA [8].

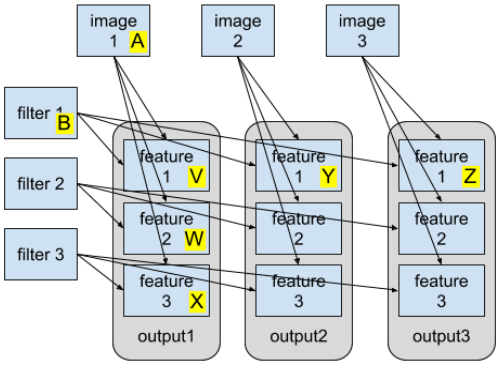


Figure 3: The relation between the input data (images and filters) and the output images.

CUDA is a design environment for massively parallel applications. CUDA applications exploit parallelism provided by the GPGPU by breaking down the task into blocks containing the same number of threads. Functionally speaking, blocks are independent to each other. Although some of them may be assigned to the same computing tile (a.k.a. SM (Streaming Multiprocessor in NVIDIA GPU), the computing model dictates that one block cannot communicate with another. Threads in a block are assigned to the same computing tile so that they can communicate and share data. This computing model leads to two important design decisions: 1) which dimensions to parallelize; and 2) what data to share and reuse between threads within a block.

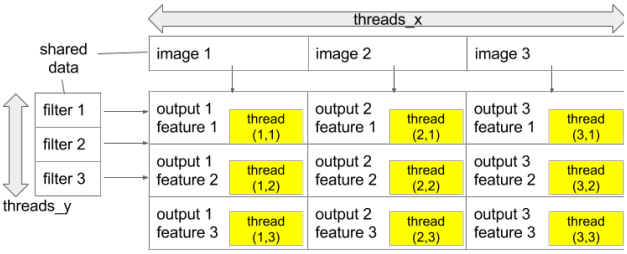


Figure 4: Configuration of threads in a block

Parallelizing one dimension means that the task is divided into smaller pieces by splitting at that dimension. For example, if we choose to parallelize the image_count dimension, computation for different images are divided into smaller tasks. In this case, each task is responsible for a small number of images. There are six dimensions in the inputs: image_count, image_features, width, height, filter_size and filter_count. A designer needs to decide which of these dimensions should be parallelized.

One limitation of the CUDA programming model is that different blocks cannot communicate with each other. Therefore, we can only parallelize dimensions that divide the problem into smaller independent tasks. In other words, outputs of each task should be stored separately without being combined into larger results. Based on this criterion, we will separately examine each of the dimensions to determine if it can be parallelized.

The dimension image_count represents the number of images. Because each image produces its own output independent of other images, we can parallelize the image_count dimension. The width and height dimensions can also be parallelized because computation of each output pixel is independent. Next, each filter produces its own output feature map, so the filter_count dimension can be parallelized. Now we are left with two dimensions to examine. The features dimension represents the number of feature maps in input images as well as filters. Because the results from different feature maps are summed to form a single output feature map (see Figure 1), we do not parallelize the feature dimension. Finally, we also do not parallelize filter_size because the products of filter pixels and image pixels are summed to form a single output pixel. To summarize, the dimensions that we are going to parallelize are image_count, width, height and filter_count.

The next step is to decide what data is reused and shared by different threads within a block. Figure 3 depicts the relation between the input data (images and filters) and the output images. Some input/output data set are labeled with capital letters for ease of explanation. The arrows in the figure indicates the input-output relation of the data. For example, the arrow pointing from A (image 1) to V (feature 1 of output 1) means that computation of V depends on A. From the figure we can see that different features in the same output image share the same input image (e.g. V, W, X all depend on A). Also, the same feature in different output images share the same filter (e.g. V, Y, Z all depend on B). These observations reveal some data-reuse opportunities.

To benefit from reusing both images and filters, a block should load pixels of multiple images and filters from the global memory. The threads also need to efficiently reuse the loaded data. This can be done by arranging the threads into a 2D configuration as show in Figure 4. The total number of threads in a block is threads_x * threads_y, and each thread is responsible of computing a single output pixel. All threads have an x and y index, where the x index determines which image the thread uses and the y index determines the filter. By doing this, the kernel only needs to load threads_x image pixels and threads_y filter pixels to compute threads_x * threads_y outputs. Each loaded image pixel is reused by threads_y threads, and each loaded filter pixel is reused by threads_x threads. As a result, this design reduces the memory access required to load the images to 1 / threads_y and that of the filters to 1 / threads_x.

```
01: __shared__ float
    images_pixel[features][filter_size*filter_size][threads_x];
02: __shared__ float
    filter_pixel [features][filter_size*filter_size][threads_y];
03: <collaboratively load images and filters>
04: float result = 0;
05: for (int i = 0; i < filter_size*filter_size; i++)
06:     for (int f = 0; f < features; f++)
07:         result +=
            image_pixel[f][i][threadIdx.x]*
filter_pixel[f][i][threadIdx.y];
08: <write result back to global memory>
```

Figure 5: High-level structure of the CUDA kernel

The high-level structure of the CUDA kernel is listed in Figure 5. Shared memory arrays are allocated to store the image and filter pixels to be reused. At the beginning, all threads work together to load the image and filter pixels required by the block. Then, each thread computes the output pixel it is responsible for. Finally, the computed results are written back to global memory.

```
01: __shared__ float
    images_pixel [features][cached_pixels][threads_x];
03: __shared__ float
    filter_pixel [features][cached_pixels][threads_y];
05:
06: <collaboratively load images and filters>
07:
08: float result = 0;
09: for (int p = 0; p < filter_size*filter_size; p += cached_pixels){
10:     for (int i = 0; i < cached_pixels; i++)
11:         for (int f = 0; f < features; f++)
12:             result += image_pixel[f][i][threadIdx.x] *
                filter_pixel[f][i][threadIdx.y];
13: }
14:
15: <write result back to global memory>
```

Figure 6: Modified kernel that loads data in chunks

This kernel is functional, but there are some details that can be improved. The first one is that this kernel loads the entire filter into shared memory, resulting in higher shared memory usage. This can be improved by loading the filter pixels in fixed-size chunks instead of loading as a whole. Due to the commutativity and associativity of summation, the final result is equal to the sum of the partial result of each chunk. Now we can modify the program by adding a parameter cached_pixels to set the size of the chunk. The modifications are listed in Figure 6. Note that both images and filters are loaded in chunks, or tiles. This technique is normally referred as tiling. As listed in (1), the shared memory usage after tiling ($SharedMem_{tile}$) can be greatly reduced from the original cost ($SharedMem_{original}$).

$$SharedMem_{tile} = x = \frac{\text{cached_pixels}}{\text{filter_size}^2} \times SharedMem_{original} \quad (1)$$

Currently, each thread in the kernel computes only one output pixel. We can generalize the kernel to compute multiple output pixels per thread. An output pixel is computed from one image and one filter, so we will add two extra kernel parameters images_per_thread and filters_per_thread to make the kernel compute multiple output pixels. These two parameters decide how many images and filters should each thread use. Because each pair of image and filter generate one output pixel, the number of outputs of each thread is images_per_thread * filters_per_thread. The modification to the program is listed in Figure 7. By adjusting images_per_thread and filters_per_thread, we can make each thread compute multiple outputs and therefore reduce the total number of blocks for the same problem size. For example, if the original kernel (equivalent to the modified kernel with images_per_thread = filters_per_thread = 1) has N blocks in total, the new kernel with images_per_thread = filters_per_thread = 2 only has N/4 blocks.

This kernel is used as the baseline program on which we propose enhancements. In the next section, we will describe how we choose the input sizes to use in our experiments.

2.3 Choosing the Input Size for Experiments

GPGPU-sim can obtain detailed execution behavior of CUDA programs on a GPU architecture similar to GTX-480. However, performance simulation in GPGPU-sim is very slow compared to a physical GPU. For example, a program that takes 100ms on an NVidia GTX480 can take up to 3 hours when running on GPGPU-sim. With such long simulation periods, changes to the program or the simulator itself cannot be quickly tested. Therefore, we think it is beneficial to reduce the size of the input dataset for shorter simulation time. But reducing the input data size, if done improperly, might produce inaccurate simulation results that deviates from the behavior of the original data. In this section, we will discuss a way to choose the size of the reduced data.

```
01: __shared__ float
    images_pixel [images_per_thread][features]
[cached_pixels][threads_x];
02: __shared__ float
    filter_pixel [filters_per_thread][features] [cached_pixels][threads_y];
03: <collaboratively load images and filters>
04: float result[images_per_thread][filters_per_thread] = {};
05: for (int p = 0; p < filter_size*filter_size; p += cached_pixels) {
06:     for (int i = 0; i < cached_pixels; i++)
07:         for (int f = 0; f < features; f++)
08:             for (int ii=0; ii<images_per_thread; ii++)
09:                 for (int if=0; if<filters_per_thread; if++)
10:                     result += image_pixel[ii][f][i][threadIdx.x] *
                        filter_pixel[if][f][i][threadIdx.y];
11: }
12: <write result back to global memory>
```

Figure 7: Modified kernel with multiple output in each thread

Computation in CUDA is broken down into independent blocks. Each block is assigned to a streaming multiprocessor (SM) so that every thread in the block can be run concurrently via fine-grain context switching of warps (groups of 32 threads). Also, each SM is capable of executing multiple blocks at the same time. On GTX480, the maximum number of blocks that a SM can run concurrently is limited by the following limiting factors:

1. An SM has 32768 registers.
2. An SM has 48kB of shared memory.
3. There should be no more than 8 blocks assigned to the same SM at the same time.

Each thread in a block has its own set of registers, so the number of registers used by each block is block *size*×*register per thread*. The limitation of blocks per SM due to the limiting factor 1 mentioned above is

$$\left\lfloor \frac{32768}{registers\ per\ block} \right\rfloor \quad (2)$$

Another limitation is the size of shared memory. Because an SM only has 48kB of shared memory, the limiting number of blocks per SM due to limiting factor 2 is

$$\left\lfloor \frac{48\text{kB}}{shared\ memory\ per\ block} \right\rfloor \quad (3)$$

At last, due to scheduler hardware limitations (limiting factor 3), a SM can only run at most 8 blocks concurrently. Concluding

(2), (3) and the hardware limitation, the actual maximum number of blocks per SM can be obtained by (4).

$$\min\left(\left\lfloor\frac{32768}{registers\ per\ block}\right\rfloor, \left\lfloor\frac{48\text{kB}}{shared\ memory\ per\ block}\right\rfloor\right) \tag{4}$$

In GTX480, there are 15 SMs. If each SM can run eight blocks in parallel, then the GPU can run 120 blocks concurrently. While using very large datasets, the total number of blocks will be much larger than 120 so that most of the time, all SMs on the GPU is doing some work. However, if the reduced dataset has less than 120 blocks, then some of the SMs on the GPU will be idle all the time, making occupancy lower than it should have been in larger datasets. Also, if the number of blocks is slightly more than 120 blocks, the GPU will execute the first 120 blocks in its full capability, and then execute the remaining blocks using only some of the SMs while leaving other SMs idle. This also makes measurements inaccurate in the same way.

Therefore, it is preferable to adjust the reduced data size so that blocks fill in all SMs during the entire simulation period. In other words, the total number of blocks divided by the maximum number of concurrent blocks should be a whole number or slightly less than a whole number (ex. 1.99). By reducing the data size like this, measurements will be closer to that of larger data sets.

Input parameters		Kernel parameters	
image_count	128	images_per_thread	4
image_width	5	filters_per_thread	4
image_height	6	threads_x	16
image_features	3	threads_y	4
filter_count	32	cached_pixels	4
filter_width, filter_height	32		

Table 1: Parameters used in the experiments

The actual input data size and kernel parameters we use in the experiment are listed in Table 1. According to these parameters, we can compute the total number of blocks using (5).

$$number\ of\ blocks = \left\lceil\frac{image_count}{thread_x \times images_per_thread}\right\rceil \times \left\lceil\frac{image_width \times image_height \times filter_count}{thread_y \times filters_per_thread}\right\rceil \tag{5}$$

After compiling this program, the compiler outputs the following information:

1. Each thread uses 28 registers.
2. Each block uses 3900 bytes of shared memory.

Now we compute the number of blocks per SM. The limitation caused by registers is

$$\left\lfloor\frac{registers\ per\ SM}{registers\ per\ thread \times block\ size}\right\rfloor = \left\lfloor\frac{32768}{28 \times (16 \times 4)}\right\rfloor = 18\ blocks\ per\ SM \tag{6}$$

The limitation caused by shared memory is

$$\left\lfloor\frac{\text{shared memory per SM}}{shared\ memory\ per\ block}\right\rfloor = \left\lfloor\frac{48\text{kB}}{3900}\right\rfloor = 12\ blocks\ per\ SM \tag{7}$$

Because both (6) and (7) exceed the hardware limitation of eight blocks per SM, maximum numbers of blocks per SM in this case is 8. Multiplying with the total numbers of SMs on GTX480, we get 8×15=120 blocks that can be executed on the GPGPU in parallel. Since there are exactly 120 blocks to be executed, all the blocks can be run in parallel without any SM stalling.

2.4 Summary of Design Concerns

In this section, we discussed an implementation of convolution in CUDA step-by-step. This convolution kernel employs a 2D block configuration to enable sharing of onboard data between threads through the shared memory, reducing accesses to the global memory. The kernels are also parameterized to enable adjusting the amount of work done by each thread. We also explained how we choose a relatively small input size that utilize all SMs. This method of choosing input sizes reduces the error of the experiment results caused by idle SMs. In all the following experiments, we will use the input sizes listed in Table 1.

According to simulation using GPGPU-sim, we identified that the two bottlenecks of the convolution kernel are computation of partial sums and loading of image and filter data. The following two sections, Section 3 and Section 4, will elaborate the proposed software and hardware improvements to speed up these two bottlenecks.

3. Accelerating Computation Part

The result of profiling the baseline implementation is shown in Figure 8. According to the profiling result, the most time-consuming part in the entire kernel is the computation part, which takes up over 97% of the overall runtime. In this section, we will focus on reducing the computation cost by various techniques including software and hardware modifications.

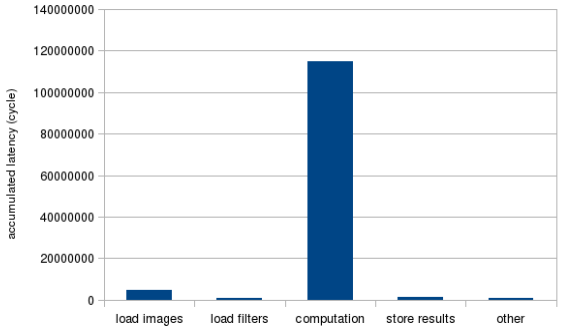


Figure 8: Initial Performance Breakdown

3.1 Avoid Local Memory Access

In order to improve the performance of the computation part, we need to understand what is responsible for its relatively long latency. Instruction-level breakdown of the profiling result listed in Figure 9 reveals that the latencies of instructions accessing local

memory are orders of magnitude greater than other instructions. In this section, we will discuss why there is local memory accessing in the code and how can we avoid it.

latency	instruction
…	…
900003	add.u64 %rd55, %rd54, %rd51;
825928	mul.lo.u64 %rd56, %rd55, 4;
909892	add.u64 %rd57, %rd13, %rd56;
92553	ld.shared.f32 %f47, [%rd57+0];
25725679	ld.local.f32 %f48, [%rd53+0];
1059732	mad.f32 %f49, %f43, %f47,%f48;
84348	st.local.f32 [%rd53+0], %f49;
26358148	ld.local.f32 %f50, [%rd53+4];
1059915	mad.f32 %f51, %f44, %f47, %f50;
84135	st.local.f32 [%rd53+4], %f51;
26520275	ld.local.f32 %f52, [%rd53+8];
1061173	mad.f32 %f53, %f45, %f47, %f52;
84381	st.local.f32 [%rd53+8], %f53;
26074621	ld.local.f32 %f54, [%rd53+12];
1062764	mad.f32 %f55, %f46, %f47, %f54;
80641	st.local.f32 [%rd53+12], %f55;
168702	add.u32 %r98, %r98, 1;
650839	add.u64 %rd53, %rd53, 16;
...	…

Figure 9: Local memory access latency

3.1.1 Local Variables in CUDA

Before going into discussion, we will first briefly introduce how local variables (also known as automatic variables) are handled in NVidia GPGPUs. In the CUDA programming language, local variables are normally placed in the stack frame of the current function call. But compilers are also allowed to put them in registers if the architecture permits. Execution stack of CUDA threads are placed in a special memory space called local memory. Local memory resides in the global memory but is partitioned and allocated to each thread. Each thread can only see its own copy of local memory. Because global memory is much slower than registers, it is often preferable to put local variables in registers instead of local memory.

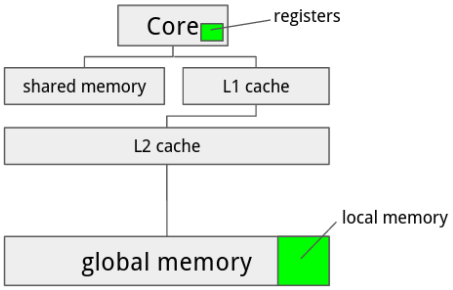


Figure 10: Registers and Local Memory

However, there are two limitations in CUDA regarding the use of registers. One limitation is that Fermi GPGPUs only have 32768 register in each core, so the total number of registers used by all threads in a block cannot exceed 32768. This limitation is less of a problem in convolution because the number of threads per block is relatively small (less than 100), and the number of register required for each thread is around 60. The other limitation is that registers have no addresses. This implies that an array can be put in registers only if no indexing is performed on them.

3.1.2 Loop Unrolling

In the convolution kernel, the partial sums computed by each thread are stored in a local array and accumulated over all pixels as shown in Figure 11. The loop counter f and g are used with the subscript operator to access the array element for each image and filter. In this case, the compiler needs to put the array in local memory because registers cannot be indexed. This will cause the GPU to access global memory in every iteration of the inner loop.

```
01: float result[filters_per_thread][images_per_thread];
02: for (int pixel = 0; pixel < filter_pixels; pixel += cached_pixels)
03: {
04:  <load filter pixels>
05:  <load image pixels>
06:     for (int i =0; i < cached_pixels*image_features; i++)
07:     {
08:        for (int f=0; f<filters_per_thread; f++)
09:        {
10:           for (int g=0; g<images_per_thread; g++)
11:           {
12:              result[f][g] += image_pixel[i][g] * filter_pixel[i][f];
13:           }
14:        }
15:     }
16: }
```

Figure 11: Local Array in the Convolution Kernel

It can be seen in the figure that the overall latency is dominated by local memory access (highlighted in boldface). The bottleneck can be totally avoided if the array elements are put in registers instead of local memory. However, the compiler cannot put the array in register because the array needs to be dynamically indexed.

One way to get around this limitation is to apply loop unrolling, which effectively eliminates all dynamic indexing on the array by expanding the loop and replacing the indices with constants. As long as the array is not dynamically indexed, the compiler can allocate registers for the array elements.

```
01: for (int i = 0; i < cached_pixels * image_features; i++) {
02:  #pragma unroll
03:    for (int f = 0; f < filters_per_thread; f++)
04:    {
05: #pragma unroll
06:       for (int g = 0; g < images_per_thread; g++)
07:       {
08:          result[f][g] += image_pixel[i][g] * filter_pixel[i][f];
09:       }
10:    }
11: }
```

Figure 12: Applying Loop Unrolling

Loop unrolling in CUDA can be enabled by setting up the preprocessing hint during compilation. A directive #pragma unroll is provided to let the programmer issue unrolling hints so that the compiler knows which loops should be unrolled. Using the unroll directive, we can apply loop unrolling to the original program (shown in Figure 12). By using loop unrolling on the two inner-most loops, the `result` array is expanded into multiple independent registers. As a result, the inner loop no longer requires local memory access. As shown in the profiling result in

Figure 13, latency of local memory accesses is completely eliminated after unrolling the loop.

latency	instruction
189822	ld.shared.f32 %f59, [%rd13+0];
1681	ld.shared.f32 %f60, [%rd16+0];
17345	mov.f32 %f61, %f2;
25172	mad.f32 %f62, %f59, %f60, %f61;
8650	mov.f32 %f63, %f62;
1681	ld.shared.f32 %f64, [%rd13+4];
18890	mov.f32 %f65, %f4;
24716	mad.f32 %f66, %f64, %f60, %f65;
6933	mov.f32 %f67, %f66;
1699	ld.shared.f32 %f68, [%rd13+8];
17569	mov.f32 %f69, %f6;
24335	mad.f32 %f70, %f68, %f60, %f69;
5724	mov.f32 %f71, %f70;
1684	ld.shared.f32 %f72, [%rd13+12];
17055	mov.f32 %f73, %f8;
24019	mad.f32 %f74, %f72, %f60, %f73;
192675	mov.f32 %f75, %f74;
1707	ld.shared.f32 %f76, [%rd16+4];
16959	mov.f32 %f77, %f10;
...	…

Figure 13: Profiling result after applying loop unrolling

Before applying loop unrolling, it takes 524k cycles for the program to finish. The letter k is a postfix indicating 1,000. After applying loop unrolling, the number of cycles is reduced to 148k, resulting in a 71% improvement on the overall performance. The latency breakdown in Figure 14 shows that the computation part is dramatically improved by loop unrolling.

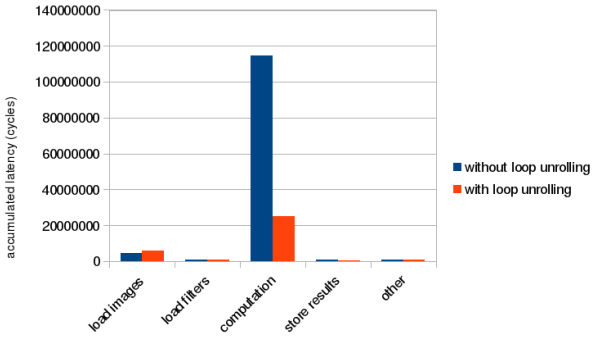


Figure 14: Loop unrolling performance improvement

3.2 Adding Inner Product Engine

The computation part (shown in Figure 15) is still the bottleneck even after applying loop unrolling, taking up about 69% of the overall execution time. Looking at the loop as a whole, what it does is computing the product between each image and filter pixel and sum them together. The partial sums are then accumulated in the array `result`.

```
01: for (int i = 0; i < cached_pixels * image_features; i++) {
02:   #pragma unroll
03:     for (int f = 0; f < filters_per_thread; f++)
04:     {
05:     #pragma unroll
06:       for (int g = 0; g < images_per_thread; g++)
07:       {
08:         result[f][g] += image_pixel[i][g]* filter_pixel[i][f];
09:       }
10:   }
11: }
```

Figure 15: Computation loop

If we exchange the order of the outmost loop with the two inner loops in Figure 16, the inner loops (Line 05 to Line 11) becomes an inner product operation between the two arrays. This modification does not change the behavior of the original function because there is no dependency between each loop iteration. Since the two outer loops are unrolled, the inner product becomes the only bottleneck.

```
01:  #pragma unroll
02:  for (int f=0; f < filters_per_thread; f++) {
03:    #pragma unroll
04:    for (int g=0; g < images_per_thread; g++) {
05:      for (int i=0;  i<cached_pixels* image_features; i++)
06:      {
07:        result[f][g] += image_pixel[i][g] * filter_pixel[i][f];
08:      }
09:    }
10: }
```

Figure 16: Inner product in the computation loop

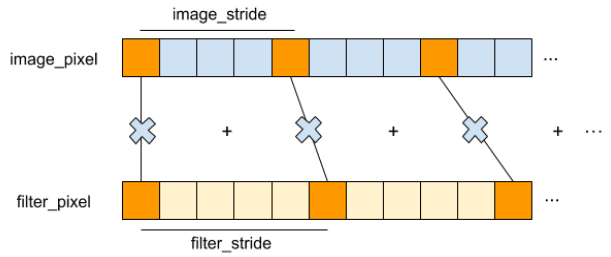


Figure 17: Inner Product

To reduce the bottleneck, we propose adding a hardware inner product accelerator to the cores, one unit for each thread. The accelerator loads pairs of image and filter pixels from shared memory, compute the product of each pair, and then sum all the products together. The program passes the starting addresses and strides of both arrays and the number of elements to the accelerator. These arguments are then stored in the internal registers of the accelerator and are reused until their values are changed again. The final result is also stored in the unit and can be retrieved in the program.

The hardware architecture of the inner product accelerator unit is shown in Figure 18. For each iteration, it loads two elements from shared memory and accumulate the product of them to a register. It requires one multiplier, one adder and a register to store the partial results. In the actual implementation, the multiplier and adder can be fused into a fused multiply-add (FMA) circuit. We assume that it requires 2 cycles to load the two elements from shared memory, and the latency of the FMA is also assumed to be 2 cycles. A 2-stage pipeline can then be used to repeatedly load elements and compute FMA. In our work, we model the latency of the inner product unit as 2 * N, where N is the number of elements to compute.

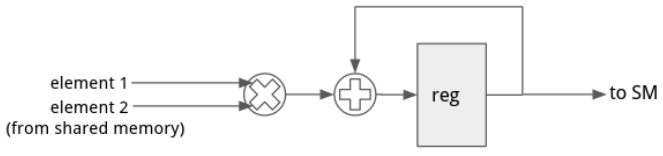


Figure 18: Inner Product Unit

Modeling the accelerator requires modifying both GPGPU-sim and the application. We modeled the inner product accelerator in GPGPU-sim by intercepting memory access to specific addresses. We modified the load (ld) and store (st) handlers to check if the source or target address matches any of the designated addresses. List of the addresses we use and their functions are listed in Table 2. If the address matches, the original memory access is skipped and the corresponding function of the accelerator is performed. For example, as soon as the thread writes 16 to address 0xffffffc8, GPGPU-sim sets the stride register of the accelerator to 16.

Table 2: Addresses used by the inner product engine

address	direction	function
0xffffffc0	write	address of image array
0xffffffc8	write	stride of image array
0xffffffd0	write	address of filter array
0xffffffd8	write	stride of filter array
0xfffffff0	write	number of elements
0xfffffff8	read	start computing and retrieve the result; stall the thread for N cycles

To use this inner product accelerator, the application needs to use the addresses to manipulate the registers inside the accelerator. First, the application should pass the starting address of the image and filter array to 0xffffffc0 and 0xffffffd0 respectively. The image stride is threads_x * images_per_thread, and the filter stride is threads_y * filters_per_thread. These two values are constant as the dimensions of the arrays are known in compile time, so it is only necessary to write them once at the beginning of the kernel. The fifth parameter, number of elements, is also known in compile time and only needs to be write once. Finally, after all parameters are set, the program needs to read from address 0xfffffff8 to retrieve the result of the inner product. The modified program is listed in Figure 19.

```
01: // beginning of the kernel
02: *(unsigned long long*)(0xffffffc8)  =  sizeof(shm_images[0]);
03: *(unsigned long long*)(0xffffffd8)  =  sizeof(shm_filters[0]);
04: *(unsigned long long*)(0xfffffff0)  =  cached_pixels * image_features;
05: // inside the computation part
06: #pragma unroll
07: for (int f = 0; f < filters_per_thread; f++) {
08:    *(float**)(0xffffffd0) = &shm_filters[0][threadIdx.y * filters_per_thread
       + f];
09: #pragma unroll
10:    for (int g = 0; g < images_per_thread; g++) {
11:       float partial_sum;
12:          *(float**)(0xffffffc0) = &shm_images[0][threadIdx.x *
             images_per_thread + g];
13:       partial_sum = *(float*)0xfffffff8;
14:       result[f][g] += partial_sum;
15:    }
16: }
```

Figure 19: Program modified to use the inner product engine

The profiling result of the modified program is listed as follows in Figure 20. As shown in the figure, the computation part is improved by 88% and results in 66% improvement on the overall performance. The bottleneck of the program is no longer the computation part.

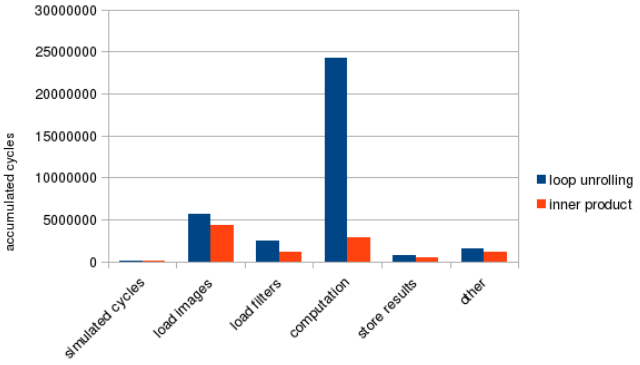


Figure 20: Loop unrolling performance improvement

3.3 Outer Product Engine

In this section, we will discuss an alternative way to accelerate at the computing part. The two inner loops (shown in line 03~09 in Figure 21) can also be viewed as computing the products of each pair of elements in image_pixel[i] and filter_pixel[i]. This operation is called the outer product, which can be represented as multiplying a column matrix with a row matrix as shown in the figure. Inputs to the outer product are two arrays image_pixels[i] and filter_pixel[i], each with length images_per_thread and filters_per_thread. Each element in the first array is multiplied with each element in the second array, producing a total of images_per_thread * filters_per_thread numbers. The outer product is performed cached_pixels * image_features times, and the output of each time is accumulated to produce the partial sums.

```
01: for (int i = 0;  i < cached_pixels * image_features;  i++) {
02:    for (int f = 0; f < filters_per_thread; f++) {
03:       for (int g = 0; g < images_per_thread; g++)
04:       {
05:          result[f][g] += image_pixel[i][g] * filter_pixel[i][f];
06:       }
07:    }
08: }
```

Figure 21: Outer product in the computation loop

We propose an accelerator to speed up the computation of the outer product. The accelerator has at least images_per_thread * filters_per_thread internal registers to store the computed partial sums. It loads two arrays, compute the product of each pair of elements, and accumulate the results to the internal registers. Values of the registers can be read or reset to zero by the program.

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \begin{bmatrix} w & x & y & z \end{bmatrix} = \begin{bmatrix} aw & ax & ay & az \\ bw & bx & by & bz \\ cw & cx & cy & cz \\ dw & dx & dy & dz \end{bmatrix}$$

Figure 22: Outer product

The hardware architecture of the outer product accelerator unit is shown in Figure 23. For each iteration, it loads four elements from image array and one element from filter array

multiply them to a register, and the adder adds previous partial sum to the original register. It consumes four multiplier, one adder and number of image array * filter array registers to save the partial results. Suppose the multiplier requires 1 cycle to produce the result, storing result to register and loading pixel from shared memory also consumes 1 cycle. Then, a 3-stage pipeline can be applied to repeatedly load elements, compute outer product and store results. In our implementation, we model the latency of the outer product unit which assume that the accelerator can load one pixel from the shared memory each cycle. It will need to preload some pixels from the image array before using a 3-stage pipeline, and the number of preloaded pixels depends on image array length. Therefore, we can assume that the total latency of the accelerator is image array length + (2*N)-2 ,where N is the number of elements to compute.

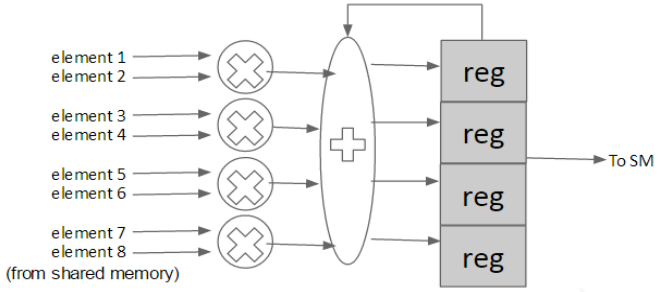


Figure 23: Outer Product Unit

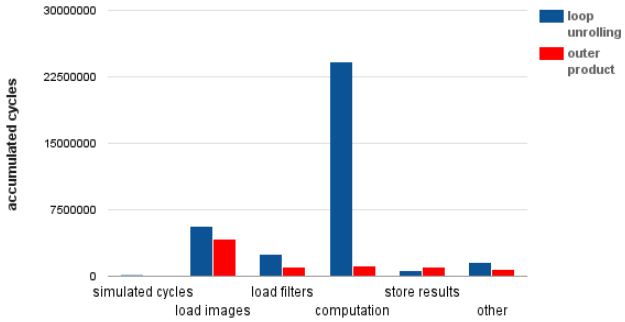


Figure 24: Outer product performance improvement

4. Accelerating Data-Loading

After applying the improvements described in Section IV, the computation part is improved a lot. As illustrated in *Figure 25*, the data-loading part, including loading of images and filters, becomes the bottleneck of the program.

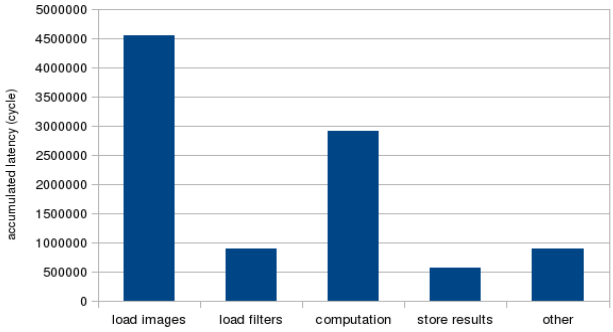


Figure 25: Performance breakdown after improvements

The data-loading part is responsible for loading image and filter pixels from global memory to shared memory. It is composed of deeply nested control structures of loops and conditionals. The control structures themselves also take time to execute, especially for the inner loops. Any subtle overhead inside inner loops can build up and become major bottlenecks. In this section, we propose two software approaches to reduce the overhead inside inner loops. We also propose a hardware accelerator to speed up address calculation in the image-loading part.

4.1 Strength Reduction

One of the major bottlenecks in inner loops is the computation of array indices for each iteration. In the data-loading loops, array indices in inner loops can contain complex arithmetic expressions that translates into larger number of instructions. Because the arithmetic instructions are executed in the inner loops, latencies of them can quickly build up and become a major bottleneck.

Take the program in Figure 26 as an example. To compute the array index for images, it needs to compute two multiplications and two additions for each iteration (line 4). If we work out the total number of operations, we will find that the program needs to carry out image_features * images_per_thread multiplications and additions in total.

```
01: for (f = 0; f < image_features; f++) {
02:    for (i = 0; i < images_per_thread; i++) {
03:       image_pixel[f][i] = images[ base + f * stride + i * threadIdx.x];
04:    }
05: }
```

Figure 26: Arithmetic operation in the inner loop

We propose a method based on strength reduction to improve the performance of this program. This method takes advantage of the fact that some arithmetic operations can be reduced to successive simpler operations. For example, multiplication can be done with repeated addition of the multiplier. Instead of computing the multiplication in each iteration, we can use a separate counter variable idx to accumulate the index throughout the entire loop and update it according to the following rule.

The arithmetic expression for computing the array index can be broken down into 3 parts:

a. loop invariants (terms that does not change throughout the nested loops)
b. multiples of the outer loop counter f
c. multiples of the outer inner loop counter i

Terms belonging to Part a is constant with respect to both the inner and outer loops. Therefore, the term `base` is used to initialize idx before entering the loops. Part b contains the term f * stride, whose value increases by stride whenever f is incremented. Therefore, stride is added to idx at the end of the outer loop. Part c contains the term i * threadIdx.x. The value of this term goes from 0 to (images_per_thread-1) * threadIdx.x in the inner loop and returns to zero again. Therefore, we will first save the value of idx before entering the inner loop, increment idx in each iteration, and restore idx after leaving the inner loop. The resulting program is listed in Figure 27.

```
01: for (int f = 0; f < image_features; f++) {
02:    for (int i = 0; i < images_per_thread; i++) {
03:       if (image_index + i < image_count) {
04:          /* load image `image_index + i` */
05:       }
06:    }
07: }
```

Figure 27: Applying strength reduction

In this transformed program, it only needs to compute one addition for each iteration in the inner loop and one addition for each iteration in the outer loop. In total, we get image_features * (images_per_thread+1) additions. The total number of operations is cut in half compared to the original program, so the modified program should run faster in the data-loading part. Profiling results in Figure 28 supports this prediction by showing that the performance of the image-loading and filter-loading parts are improved significantly. The impact of strength reduction is 25.9% on the overall performance.

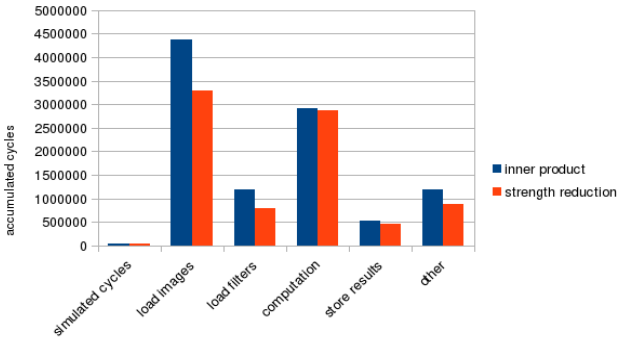


Figure 28: Strength reduction performance improvement

4.2 Extract Conditionals

In addition to strength reduction, we also identified another useful optimization strategy in the data-loading loop. The idea is to prevent unnecessary condition checks in inner loops by checking whether the conditional is necessary in the loop before entering the loop. To illustrate the case, consider the following code snippet taken from the convolution kernel:

```
01: int idx = base;
02: for (f = 0; f < image_features; f++) {
03:    int idx_save = idx;
04:    for (i = 0; i < images_per_thread; i++) {
05:       image_pixel[f][i] = images[idx];
06:       idx += threadIdx.x;
07:    }
08:    idx = idx_save;
09:    idx += stride;
10: }
```

Figure 29: Boundary checking in the data-loading loop

In the inner-most loop, it checks whether the index of image to load (image_index + i) is within bounds and only load the image if so. Because the total number of image is not always a multiple of images_per_thread, the boundary condition check is necessary here to ensure that the index to load is within bounds. However, checking for the boundary condition every time in an inner loop degrades performance. To eliminate redundant checks in inner loops, the loop is duplicated and modified into two variants: one with boundary checking, and the other without them. Without the checking overhead, the one without boundary checking will run faster. The problem we are left with is how to choose between these two variants.

Because the loop variable i is always less than or equal to images_per_thread - 1, image_index + i will always be less than or equal to image_index + images_per_thread - 1. If given image_index + images_per_thread - 1 < image_count, we will automatically get image_index + i < image_count. In other words, if image_index + images_per_thread <= image_count, there is no need to check for the boundary condition. As a result, we will choose the faster loop without boundary checking if image_index + images_per_thread <= image_count, or the slower loop otherwise. The resulting code is listed in Figure 30.

```
01: if (image_index + images_per_thread <= image_count) {
02:   for (int f = 0; f < image_features; f++) {
03:     for (int i = 0; i < images_per_thread; i++)
04:     { // no boundary check
05:       // load image `image_index + i`
06:     }
07:   }
08: } else {
09:   for (int f = 0; f < image_features; f++) {
10:     for (int i = 0; i < images_per_thread; i++)
11:     {
12:       if (image_index + i < image_count)
13:       { // boundary check
14:         // load image `image_index + i`
15:       }
16:     }
17: }
18: }
```

Figure 30: Reduced boundary check code

Profiling result before and after applying this technique is shown in Figure 31. This technique improves the overall performance by 2.6% compared to the previous version using only strength reduction.

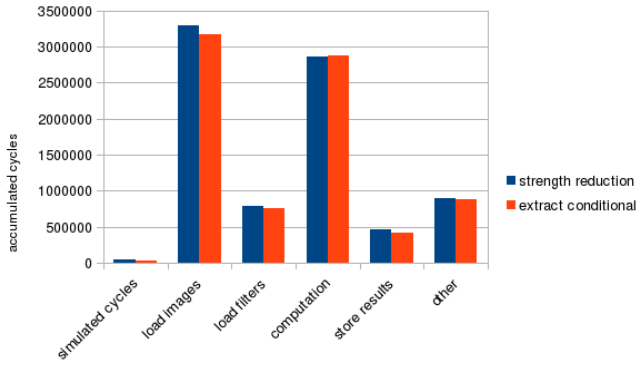


Figure 31: Extract conditionals performance improvement

4.3 Index Conversion Accelerator

In the data-loading part of the convolution kernel, filter and image pixels are loaded to shared memory one after another. Before loading each pixel from global memory, the program must compute the index of the pixel in the input arrays. In each iteration of the loop, the program first loads several filter pixels. While loading the filter, there is no need to compute the index because the loop counter itself represents the index of the filter pixel to load. However, for each filter pixel, it is still necessary to load the corresponding image pixel. Computing the index of image pixels is more involved. Sometimes the filter pixel is placed outside the

image and doesn't overlap with any pixel in the input. When this happens, the loaded image pixel should be set to zero because zero-padding is used outside the edges of the image. If the filter pixel is placed within the image, the filter pixel index should be converted into the image pixel index, which is then used to load the image pixel from global memory.

```
01:for (int p = 0; p < cached_pixels; p += threads_y) {
02:   int pixel_index = pixel + p + threadIdx.y;
03:   int x = image_pixel_x – filter_size / 2 + pixel_index % filter_size;
04:   int y = image_pixel_y - filter_size / 2 + pixel_index / filter_size;
05:   if (y >= 0 && y < image_height && x >= 0 && x < image_width)
06:   {
07:      int image_pixel_index = (y * image_width + x) * image_count;
08:      <load image pixel from image_pixel_index>
09:   } else {
10:     <set image pixel to 0>
11:   }
12: }
```

Figure 32: Index conversion in the data-loading loop

Computing the image pixel index and checking for the boundary condition takes up about 1/3 of the image-loading time. Therefore, we propose adding an accelerator to speed up this two tasks at the same time. Before describing what this accelerator should do, we will first look at the code to convert filter pixel index to image pixel index.

From the code listed above, we can see that there are three steps involved in converting filter pixel index to image filter index. The first step is break down filter pixel index to its x and y component and offset the coordinates by the location of the kernel. Then, boundary check is performed on the (x, y) point to ensure that there is a corresponding pixel in the input. Finally, the (x, y) is converted to the image pixel index. These steps are translated to tens of instructions and slows down the program.

```
01: for (int p = 0; p < cached_pixels; p += threads_y) {
02: <invoke index converter using inline assembly>
03:  // the result is stored in image_pixel_index
04:     if (image_pixel_index >= 0) {
05:     <load image pixel from image_pixel_index>
06:     } else {
07:     <set image pixel to 0>
08:     }
09: }
```

Figure 33: Modified program using the index conversion accelerator

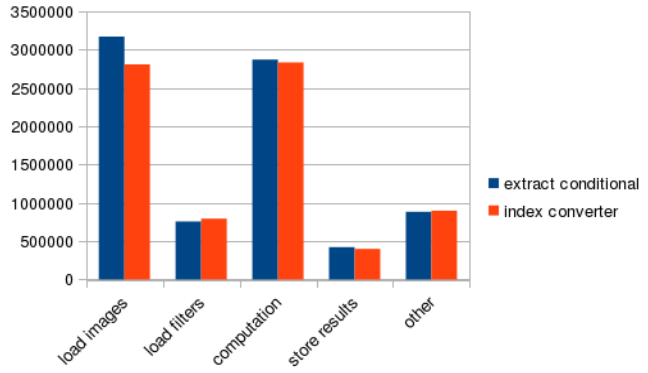


Figure 34: Index converter performance improvement

We propose adding a hardware accelerator to do the index conversion. The accelerator we propose will do the three steps altogether in one instruction. If the (x, y) point is within bounds, it returns the index of the image pixel. Otherwise, it returns -1. We implement the accelerator in GPGPU-sim as an instruction and use inline assembly in the convolution kernel to invoke the instruction. The modified program is listed in Figure 33. The performance improvement of using the index converter is listed in Figure 34. The overall performance is improved by 5%.

5. Related Work

Accelerating convolutional neural networks is a very popular research topic. Accelerators have been developed in different hardware technologies. Eyeriss [9][15] developed by Yu-Hsin Chen et al. is an ASIC CNN accelerator that can run AlexNet at 35fps with only 278mW of power consumption. There are other ASIC accelerators proposed to exploit the redundancy of CNN networks [16][17][18]. Cheng Zhang [10] implemented a CNN accelerator on FPGA and achieved 61.62 GFLOPS under 100MHz clock frequency. He also proposed an analytical design scheme using the roofline model.

The GPGPU is also a widely-used platform for CNN. NVidia developed a software library named cuDNN [2] that uses GPGPU to speed up convolution. When integrated with the Caffe framework, it can improve the performance by up to 36%. However, the cuDNN library is proprietary and cannot be studied by the community. The fbfft library [4] also uses GPGPU to speed up CNN, but it employs a different algorithm (FFT) to compute convolution. Cuda-convnet2 is an efficient implementation of CNN for NVidia GPGPU. It is by far the fastest open-source CNN implementation. However, it lacks analysis on the techniques it uses to improve the performance.

6. Conclusion

This paper describes an implementation of the convolution operation on NVidia GPGPU and analyze the techniques in the implementation. We propose software and hardware enhancements to the program to speed up the computation of partial sums and loading the input data, which are the two major bottlenecks. The experiments have shown that the proposed modifications have achieved 4.4x speedup compared with the baseline implementation.

References

[1] Lin, C.-H.; Cheng, A.-T.; Lai, B.-C.; "A Software Technique to Enhance Register Utilization of Convolutional Neural Networks on GPGPUs," IEEE International Conference on Applied System Innovation, May 2017. http://ieeexplore.ieee.org/document/7988499/

[2] Chetlur, S., Woolley, C., Vandermersch, P., Cohen, J., Tran, J., Catanzaro, B., & Shelhamer, E. (2014). cudnn: Efficient primitives for deep learning. arXiv preprint arXiv:1410.0759. https://arxiv.org/abs/1410.0759

[3] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 675-678). ACM.

[4] Vasilache, N., Johnson, J., Mathieu, M., Chintala, S., Piantino, S., & LeCun, Y. (2014). Fast convolutional nets with fbfft: A GPU performance evaluation. arXiv preprint arXiv:1412.7580. https://arxiv.org/abs/1412.7580

[5] Alex Krizhevsky. cuda-convnet2. https://code.google.com/p/cuda-convnet2/, 2014. [Online; accessed 23-January-2015].

[6] Lavin, A. (2015). maxDNN: an efficient convolution kernel for deep learning with maxwell gpus. arXiv preprint arXiv:1501.06633. https://arxiv.org/abs/1501.06633

[7] Bakhoda, A., Yuan, G. L., Fung, W. W., Wong, H., & Aamodt, T. M. (2009, April). Analyzing CUDA workloads using a detailed GPU simulator. In Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on (pp. 163-174). IEEE. http://ieeexplore.ieee.org/abstract/document/4919648/

[8] Nickolls, J., Buck, I., Garland, M., & Skadron, K. (2008). Scalable parallel programming with CUDA. Queue, 6(2), 40-53.

[9] Chen, Y. H., Krishna, T., Emer, J., & Sze, V. (2016, January). 14.5 Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. In 2016 IEEE International Solid-State Circuits Conference (ISSCC) (pp. 262-263). IEEE. http://ieeexplore.ieee.org/document/7738524/

[10] Zhang, C., Li, P., Sun, G., Guan, Y., Xiao, B., & Cong, J. (2015, February). Optimizing fpga-based accelerator design for deep convolutional neural networks. In Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (pp. 161-170). ACM. https://dl.acm.org/citation.cfm?id=2689060

[11] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105). https://dl.acm.org/citation.cfm?id=2999257

[12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://arxiv.org/abs/1409.1556

[13] Chen, T., Du, Z., Sun, N., Wang, J., Wu, C., Chen, Y., & Temam, O. (2014, February). Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. In ACM Sigplan Notices (Vol. 49, No. 4, pp. 269-284). ACM. https://dl.acm.org/citation.cfm?id=2541967

[14] Cavigelli, L., Magno, M., & Benini, L. (2015, June). Accelerating real-time embedded scene labeling with convolutional networks. In Proceedings of the 52nd Annual Design Automation Conference (p. 108). ACM. http://ieeexplore.ieee.org/document/7167293/

[15] Y.-H Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," IEEE Journal of Solid-State Circuits, 2017.

[16] D. Kim, J. Ahn, and S. Yoo, "A novel zero weight/activation-aware hardware architecture of convolutional neural network," in Design, Automation, and Test in Europe (DATE), 2017.

[17] A. Parashar, M. Rhu, A. Mukkara, A. Puglielli, R. Venkatesan, B. Khailany, J. Emer, S. W. Keckler, and W. J. Dally, "SCNN: An accelerator for compressed-sparse convolutional neural networks," in 44th Annual International Symposium on Computer Architecture (ISCA), 2017, pp. 27–40.

[18] S. Zhang, Z. Du, L. Zhang, H. Lan, S. Liu, L. Li, Q. Guo, T. Chen, and Y. Chen, "Cambricon-X: An accelerator for sparse neural networks," in 49th IEEE/ACM International Symposium on Microarchitecture (MICRO), 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 30-37 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

A Test Code Generation Method for Coding Standard Input/Output with Exception Handling in Java Programming Learning Assistant System

Ei Ei Mon[1], Nobuo Funabiki[*,1], Ryota Kusaka[1], Khin Khin Zaw[1], Wen-Chung Kao[2]

[1]*Okayama University, Department of Electrical and Communication Engineering, Okayama, Japan*

[2]*National Taiwan Normal University, Department of Electrical Engineering, Taipei, Taiwan*

ARTICLEINFO

Article history:
Received: 30 October, 2017
Accepted: 13 December, 2017
Online: 30 January, 2018

Keywords:
Java programming
JPLAS
Test code
Test case
Automatic generation
JUnit

ABSTRACT

To advance Java programming educations, we have developed the Java Programming Learning Assistant System (JPLAS) that provides the code writing problem. This problem asks a student to write a source code to satisfy the specification of a given assignment, where the correctness is verified by running test code on JUnit. For a novice student, a code of implementing the standard input/output with the exception handling should be mastered at the early stage as the first step programming for human interfaces. However, for a teacher, it is not easy to write the test code for it. In this paper, we propose a test code generation method to generate the test code using the reference source code for the assignment. In the evaluation of this proposal, all the students completed the codes using the generated test codes for exception handling, although the use of exception handling functions was sometimes insufficient or incorrect.

1 Introduction

Recently, the objected oriented programming language *Java* has been widely used in various practical application systems in societies and industries due to the high reliability, portability, and scalability. Java was selected as the most popular programming language in 2015 [2]. Therefore, there have been strong demands from industries for Java programming educations. Correspondingly, a plenty of universities and professional schools are currently offering Java programming courses to meet this challenge. A typical Java programming course consists of grammar instructions in the class and programming exercises in computer operations.

To advance Java programming educations, we have developed the Web-based *Java Programming Learning Assistant System (JPLAS)* [3]-[7]. JPLAS inspires students by offering sophisticated learning environments via quick responses to their answers for self-studies. At the same time, it supports teachers by reducing loads of evaluating codes. JPLAS has several types of problems to cover a variety of students at different learning levels. Among them, the *code writing problem* [4] asks a student to write a source code to satisfy the specification of a given assignment.

The *code writing problem* is implemented based on the *test-driven development (TDD) method* [8], using an open source framework *JUnit* [9]. *JUnit* automatically tests the codes on the server to verify their correctness using the *test code* when they are submitted by students. Thus, students can repeat the cycle of writing, testing, modifying, and resubmitting codes by themselves, until they can complete the correct codes for the assignments.

To register a new assignment for the code writing problem in JPLAS, a teacher has to prepare a *problem statement* describing the code specification, a *reference source code*, and a *test code* using a Web browser. It is noted that the reference source code is essential to verify the correctness of the problem statement and the test code. Then, a student should write a source code for the assignment while referring the statement and the test code, so that the source code can be tested by using the given test code on *JUnit*.

However, teachers at schools are not accustomed to writing a test code that can run on *JUnit*. A teacher may spend much time in struggling to write a test code, and may

[*]Nobuo Funabiki, Department of Electrical and Communication Engineering, Okayama University, Okayama, Japan, funabiki@okayama-u.ac.jp
This paper is an extension of work originally presented in 31st IEEE International Conference on Advanced Information Networking and Applications (AINA-2017) [1]

https://dx.doi.org/10.25046/aj030105

register an incomplete test code that does not verify some requirements described in the problem statement correctly. This incomplete test code must be avoided because it may produce inappropriate feedback to a student and undermine confidence to JPLAS. On the other hand, a commercial tool for generating a test code is usually expensive, and may not cover a test code that verifies the standard input/output with exception handling in a source code. The code of implementing the standard input/output with exception handling should be mastered by novice students at the early stage of Java programming educations as the first step programming for human interfaces.

In this paper, we propose a *test code generation method* for the code writing problem in JPLAS that generates a test code using a reference source code to test the standard input/output with exception handling. This method can generate a test code through the following steps: 1) a *test code template* is provided by our proposal, 2) a set of standard inputs to be tested are made by a teacher, 3) by running the reference code with each input, the corresponding expected standard output is extracted correctly, and 4) this pair of the standard input and the standard output are embedded into the test code template. By repeating steps 3) and 4) for every test standard input, the test code can be completed. To run the source code using the test code on *JUnit*, it introduces the classes to handle the standard input/output functions as the memory access functions in [10].

To evaluate the proposed method, first, we applied it to 97 source codes in Java programming textbooks or Web sites that contain the standard input/output. It has been proved that the generated test codes could correctly verify the source codes except for one code using a random generator. Then, we generated the test codes for three problems and asked five students who are currently studying Java programming to write the source codes using them. It was found that they completed the codes that can pass the test codes, whereas the use of exception handling functions was sometimes insufficient or incorrect.

The rest of this paper is organized as follows: Sections 2 and 3 introduce the TDD method and JPLAS respectively. Section 4 presents the test code generation method. Section 5 shows the evaluation result. Sections 6 shows related works. Finally, Section 7 concludes this paper with some future works.

2 Test-driven Development Method

In this section, we introduce the test-driven development method along with its features.

2.1 Outline of TDD Method

In the TDD method, the test code should be written before or while the source code is implemented, so that it can verify whether the current source code satisfies the required specifications during its development process. The basic cycle in the TDD method is as follows:

1) to write the test code to test each required specification,
2) to write the source code, and
3) to repeat modifications of the source code until it passes each test using the test code.

2.2 JUnit

In JPLAS, we adopt *JUnit* as an open-source Java framework to support the TDD method. *JUnit* can assist the unit test of a Java code unit or a *class*. Because *JUnit* has been designed with the Java-user friendly style, its use including the test code programming is less challenging for Java programmers. In *JUnit*, a test is performed by using a given method whose name starts from *assert*. This paper adopts the *assertThat* method to compare the execution result of the source code with its expected value.

2.3 Test Code

A test code should be written using libraries in *JUnit*. Here, by using the following **source code 1** for *MyMath* class, we explain how to write a test code. *MyMath* class returns the summation of two integer arguments.

source code 1

```
1  public class Math {
2    public int plus(int a, int b) {
3      return( a + b );
4    }
5  }
```

Then, the following **test code 1** can test the *plus* method in the *MyMath* class.

test code 1

```
1  import static org.junit.Assert.*;
2  import org.junit.Test;
3  public class MathTest {
4    @Test
5    public void testPlus() {
6      Math ma = new Math();
7      int result = ma.plus(1, 4);
8      assertThat(5, is(result));
9    }
10 }
```

The names in the test code should be related to those in the source code so that their correspondence becomes clear:

- The class name is given by the *test class name + Test*.
- The method name is given by the *test + test method name*.

The test code imports *JUnit* packages containing test methods at lines 1 and 2, and declares *MathTest* at line 3. *@Test* at line 4 indicates that the succeeding method represents the test method. Then, it describes the test method.

The test code performs the following functions:

1) to generate an instance for the *MyMath* class,
2) to call the method in the instance in 1) using the given arguments,
3) to compare the result with its expected value for the arguments in 2) using the *assertThat* method, where the first argument represents the expected value and the second one does the output data from the method in the source code under test.

2.4 Features in TDD Method

In the TDD method, the following features can be observed:

1. The test code can represent the specifications of the source code, because it must describe the function tested in the source code.
2. The test process for a source code becomes efficient, because each function can be tested individually.
3. The refactoring process of a source code becomes effective, because the modified code can be tested instantly.

Therefore, to study the TDD method and writing a test code is useful even for students, where the test code is equivalent to the source code specification. Besides, students should experience the software test that has become important in software companies.

3 Java Programming Learning Assistant System

In this section, we review the outline of our Java programming learning system *JPLAS*.

3.1 Server Platform

JPLAS is implemented as a Web application using *JSP/-Java*. For the server platform, it adopts the operating system *Linux*, the Web server *Apache*, the application server *Tomcat*, and the database system *MySQL*, as shown in Figure 1. For the browser, it assumes the use of *Firefox* with *HTML*, *CSS*, and *JavaScript*.

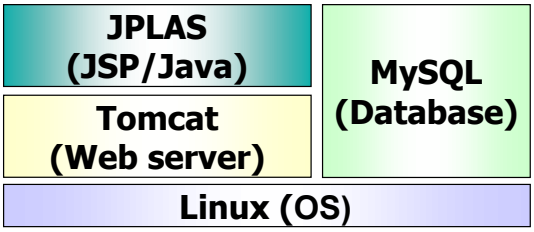


Figure 1: JPLAS server platform.

3.2 Teacher Service Functions

JPLAS has user functions both for teachers and students. *Teacher service functions* include the registration of courses, the registration and management of assignments, and the verification of source codes that are submitted by students. To register a new assignment, a teacher needs to input an assignment title, a problem statement, a reference (model) source code, and a test code. After the registration, they are disclosed to the students except for the source code. Note that the test code must be able to test the model code correctly. Using the correspondence between a source code and a test code in Section 2.3, it is possible to automatically generate a template for the test code from the source code. Then, a teacher merely needs to specify concrete values for the arguments in each test method to complete the test code.

To evaluate the difficulty of assignments and the comprehension of students, a teacher can refer to the number of submissions for code testing from each student. If a teacher finds an assignment with plenty of submissions, it can be considered as quite difficult for the students, and should be changed to an easier one. If a teacher finds a student who submitted codes in many times whereas other students did in a few times, this student may require additional assistance from the teacher.

3.3 Student Service Functions

Student service functions include the view of the assignments and the submission of source codes for the assignments. A student should write a source code for an assignment by referring the problem statement and the test code. It is requested to use the class/method names, the types, and the argument setting specified in the test code. JPLAS implements a Web-based source code editor called *CodePress* [11] so that a student can write codes on a Web browser. All submitted source codes will be stored in the database on the server as a reference for students.

4 Proposal of Test Code Generation Method

In this section, we propose the test code generation method for coding the standard input/output with exception handling.

4.1 Scope of Source Code under Test

At the early stage of the Java programming education, the responsibility of a student is to master how to write a source code that contains the standard input/output with exception handling. Thus, a teacher in a Java programming course should prepare a considerable number of assignments for writing source codes containing them, where many Java programming textbooks offer such assignments for novice students.

The source code in this paper must contain the functions for the standard input/output and the exception handling. Then, if the proper data is given to the code from the standard input, it must handle it correctly and outputs the message specified in the assignment to the standard output. On the other hand, if the improper data is given, it must handle it using the exception handling command without abortion and outputs the corresponding message.

4.2 Requirements in Test Code

Subsequently, the test code must satisfy the following requirements:

1. The input data from the standard input (keyboard) must be described in the test code to test the standard input in the source code.
2. The output data to the standard output (console) must be received by the test code to test the standard output in the source code.

3. The input data must be elaborated in the test code for the standard input.

4. The input data in the test code should cover any possible one for the standard input, including the proper and improper ones.

5. The expected output data for each input data must be narrated in the test code correctly.

4.3 Solutions for Requirements

Test code generation method adopts following functions and commands to solve the above mentioned requirements by referring the test code implementation in [10]:

- To describe the standard input data to the source code, the *Inputln* method in *StandardInputSnatcher* class is adopted in the test code. It is noted that *StandardInputSnatcher* class is extended from *InputStream* class.

- To receive the standard output data from the source code, the *readLine* method in *StandardOutputSnatcher* class is adopted in the test code. It is noted that *StandardOutputSnatcher* class is extended from *PrintStream* class.

- Any possible standard input data is prepared by a teacher beforehand. It is used in the argument of *Inputln*.

- To obtain the expected standard output data from the code for each input data, the reference source code is executed with this input data.

- Each pair of the standard input and output data is embedded into the test code.

4.4 Conditions of Source Code

Currently, to avoid the complexity, the proposed method confines the applicable source code that satisfies the following conditions:

1. it has the *main* method only.

2. it contains the standard input function.

3. it contains the standard output function for handling the proper input.

4. it contains the standard output function for handling the exception.

It is noted that a source code containing multiple standard input/output functions can be handled by increasing the number of *Inputln* or *assertThat* in the test code accordingly. Besides, if a code does not have the *main* method, it can be handled by describing the proper statements to execute the method for the standard input/output in the test code.

An example source code in this scope is as follows:

source code 2

```
1  import java.util.Scanner;
2  public class Sample {
3    public static void main(String args[]){
4      int number;
5      Scanner scan = new Scanner(System.in);
6      try{
7        System.out.print("Enter an integer");
8        String actual = scan.nextLine();
9        number = Integer.parseInt(actual);
10       System.out.println(number +": is input
         number");
11     } catch(NumberFormatException e) {
12       System.out.print("
         NumberFormatException occurs!");
13     }
14   }
15 }
```

source 2 accepts an integer data from a console and outputs a message with this data on a display. In this source code, 1) it has only the *main* method at line 3, 2) *scan* object of *Scanner* class is defined at line 5 as the standard input function, 3) *System.out.println* is called at line 10 as the standard output function for handling the proper input, and 4) *System.out.println* is called at line 12 as the standard output function for handling the exception.

4.5 Test Code Template

Then, the proposed method provides the *test code template* containing the required functions for the above mentioned source code. The following code describes the core part of the test code template starting from *@Test*. In advance, several *import* statements to use related libraries, and the instance generations for the *StandardInputSnatcher* and *StandardOutputSnatcher* classes are necessary. Besides, the definitions of these classes are also required to complete the test code template.

In this template, *in.Inputln* at line 29 gives the standard input data to the source code, where *in* is an instance of *StandardInputSnatcher* class. The statements at lines 30-37 run the source code and read the standard output data for this input data, where *out* is an instance of *StandardOutputSnatcher* class. *expected* at line 38 represents the expected output data of the source code. The blanks ″ ″ at lines 29 and 38 should be filled by the standard input and output data. *assertThat* at line 39 compares the expected data with the output data of the code. The whole statements at lines 25-40 should be prepared for each input data.

test code template

```
1  import static org.hamcrest.CoreMatchers.is;
2  import static org.junit.Assert.assertThat;
3  import static org.junit.Assert.*;
4  import java.io.InputStream;
5  import org.junit.Before;
6  import org.junit.Test;
7  import Snatcher.StandardOutputSnatcher;
8  import java.io.BufferedReader;
9  import java.io.ByteArrayOutputStream;
10 import java.io.IOException;
11 import java.io.InputStream;
12 import java.io.PrintStream;
13 import java.io.StringReader;
14
15 public class TemplateTest {
16   private StandardInputSnatcher in = new
       StandardInputSnatcher();
```

```
17      private StandardOutputSnatcher out = new
            StandardOutputSnatcher();
18
19      @Before
20      public void setUp() {
21        System.setIn(in);
22        System.setOut(out);
23      }
24
25      @Test
26      public void test1() throws Exception {
27        StringBuffer bf = new StringBuffer();
28        String actual,line,expected;
29        in.Inputln(""); // standard input
30        Sample.main(new String[0]);
31        System.out.flush();
32        while((line = out.readLine()) != null) {
33          if (bf.length() > 0)
34          bf.append("\n");
35          bf.append(line);
36        }
37        actual = bf.toString();
38        expected = ""; // expected standard output
39        assertThat(actual,is(expected));
40      }
41    }
```

4.6 Test Code Generation Procedure

The test code generation procedure using the *test code template* in the proposed method is as follows:

1) A teacher prepares the reference source code for the assignment.

2) He/she prepares a set of possible standard input data to the source code.

3) He/she runs the source code by using each standard input data and observes the corresponding standard output data.

4) He/she embeds the standard input data into ″ ″ at line 29 and the observed standard output data into ″ ″ at line 38 in the test code template.

As the possible standard input data in step 2), the following five data types should be considered. Then, the teacher needs to select one value for each data type, which is used in step 3).

- positive integer: 5
- negative integer: -14
- zero integer: 0
- floating-point number: 0.5
- one-byte character: abc
- two-byte character: A B C

4.7 Generated Test Code Example

This subsection introduces an example of the test code generated by applying the proposed method to **source code 2**. The file name for the generated test code is given as *SampleTest.java*. The following **test code 2** shows a part of the test code.

test code2

```
1   .....
2
3     @Test
4     public void test1() throws Exception {
5       StringBuffer bf = new StringBuffer();
6       String actual,line,expected;
7       in.Inputln("5"); // proper standard input data
8       Sample.main(new String[0]);
9       System.out.flush();
10      while((line = out.readLine()) != null) {
11        if (bf.length() > 0)
12        bf.append("\n");
13        bf.append(line);
14      }
15      actual = bf.toString();
16      expected = "Enter an integer" +
17       "5: is input number";
18      assertThat(actual,is(expected));
19    }
20
21    @Test
22    public void test2() throws Exception {
23      StringBuffer bf = new StringBuffer();
24      String actual,line,expected;
25      in.Inputln("abc"); // improper standard input
          data
26      Sample.main(new String[0]);
27      System.out.flush();
28      while((line = out.readLine()) != null) {
29        if (bf.length() > 0)
30        bf.append("\n");
31        bf.append(line);
32      }
33      actual = bf.toString();
34      expected = "Enter an integer" + "
          NumberFormatException occurs!";
35      assertThat(actual,is(expected));
36    }
37  .....
```

5 Evaluation

In this section, we evaluate the effectiveness of the proposed *test code generation method* in terms of generating test codes from existing source codes and writing source codes using the test codes by students.

5.1 Test Code Generation Results

First, we evaluate the method in generating test codes from source codes. For this purpose, 97 source codes were collected from Java programming textbooks or Web sites [12]-[16], and the test codes were generated by applying the proposed method. It is noted that some codes in [15] were modified to using the standard input/output through the console instead of using the dialog box. Then, the correctness of each test code was examined by testing the original source code. It was found that our method generated the test codes that can pass original codes correctly except for one source code, which outputs a random number generated in the code. Thus, the effectiveness of the proposed method was confirmed.

The following **source code 3** shows an example source code in [12] where the method successfully generates the test code shown in **test code 3**. It is noted that *try - catch* is used here instead of *throws* in the original source code.

source code 3

```
1  import java.io.*;
2  class Sample3 {
3    public static void main(String[] args) throws
       IOException {
4      try {
5        System.out.println("Enter two integers
         ");
6        BufferedReader br =
7          new BufferedReader(new
         InputStreamReader(System.in));
8        String str1 = br.readLine();
9        String str2 = br.readLine();
10       int num1 = Integer.parseInt(str1);
11       int num2 = Integer.parseInt(str2);
12       System.out.println("The sum is " + (
         num1+num2) + ".");
13     } catch(NumberFormatException e) {
14       System.out.print("
         NumberFormatException occurs!");
15     }
16   }
17 }
```

test code3

```
1  .....
2
3    @Test
4    public void test1() throws Exception {
5      StringBuffer bf = new StringBuffer();
6      String actual,line,expected;
7      in.Inputln("2"); in.Inputln("7");// proper
         standard input data
8      Sample.main(new String[0]);
9      System.out.flush();
10     while((line = out.readLine()) != null) {
11       if (bf.length() > 0)
12       bf.append("\n");
13       bf.append(line);
14     }
15     actual = bf.toString();
16     expected = "Enter two integers" +"The
         sum is 9.";
17     assertThat(actual,is(expected));
18   }
19
20   @Test
21   public void test2() throws Exception {
22     StringBuffer bf = new StringBuffer();
23     String actual,line,expected;
24     in.Inputln("0.5"); in.Inputln("-3");//
         improper standard input data
25     Sample.main(new String[0]);
26     System.out.flush();
27     while((line = out.readLine()) != null) {
28       if (bf.length() > 0)
29       bf.append("\n");
30       bf.append(line);
31     }
32     actual = bf.toString();
33     expected = "Enter two integers" + "
         NumberFormatException occurs!";
34     assertThat(actual,is(expected));
35   }
36 .....
```

5.2 Source Code Writing Results

Next, we evaluate the proposed method in writing source codes with generated test codes by five students who are currently studying Java programming and have same technical levels. For this purpose, we prepared the following three problems, where all the students completed the source codes that pass the test codes for any problem.

5.2.1 Problem #1

In problem #1, the code accepts an integer data from a console, and outputs a message with this data to a console, where **source 2** is the reference source code and **test 2** is the test code. The source code from a student is expected to use *NumberFormatException* to check the input data format. Then, three students use this class for the exception handling, and one uses *Exception*. However, one student does not use it where he implements the data format checking function.

5.2.2 Problem #2

In problem #2, the code accepts an integer index from a console, and outputs the indexed data from the data array. The student code is expected to use *ArrayIndexOutofBoundException* to check the range of the index. Then, only one student uses this class. The other students implement the index checking function in the codes. Two students use *IOException*, and two students do not use any class for the exception handling. No student use *NumberFormatException* to check the input data format, although the class was requested in problem #1. Unfortunately, many students cannot integrate the knowledge that has been studied sequentially.

5.2.3 Problem #3

In problem #3, the code accepts a file path from a console, and outputs the string at the first line in the file. The student code is expected to use *FileNotFoundException* or *IOException* to check the file path. Then, three students use *FileNotFoundException*, one uses *Exception*, and one uses *IOException*.

5.2.4 Summary of Student Applications

This simple experiment of our proposal shows that the students can generally complete source codes using standard input/output with exception handling that can pass the generated test codes. However, their use of the class for the exception handling is sometimes insufficient or incorrect. It has been observed that these students are not experts, which causes the difference in their source codes, although they have enough programming skills. To let them understand the correct use, it is necessary to improve the proposed method.

6 Related Works

In this section, we introduce some related works to this paper.

In [17], Fu presented a static exception-flow analysis that computes chains of semantically-related exception-flow links and reports entire exception propagation paths. These chains can be used, 1) to show the error handling architecture of a system, 2) to assess the vulnerability of a single component and the whole system, 3) to support the better testing of an error recovery code, and 4) to facilitate the tracing of the root cause of a logged problem.

In [18], Rashkovits showed that most of college students understand the concept of Java exception handling at the basic level, and the majority of them have difficulty in understanding advanced properties such as use of multiple exceptions, flow of control in the context of exceptions, handling exceptions further up the calling chain, catching and handling hierarchically related exceptions, and overriding methods that throw exceptions. They also provided a tutorial of exception handling, and quoted that exception handling is perceived as a relatively difficult task by novice programmers. In future works, we will consider to adopt their contributions.

In [19], Júnior presented a practical approach to preserve the *exception policy* in a system by automatically checking *exception handling design rules*. They are checked through executions of *JUnit* test cases with *dynamic mock objects* that are generated by the supporting tool. Four versions of *Mobile Media in SPL* were used to evaluate whether the policy was preserved or not. The results show that the approach can effectively detect violations on the policy of software product lines.

In [20], Nakshatri presented an empirical study of exception handling patterns in Java projects. It forces developers to think in sophisticated ways to handle the exceptions. In this study, empirical data was extracted from projects by analyzing data in *GitHub* and *SourceForge* repositories. The results were compared with recommendations for best practices in exception handling presented by Bloch [21]. It has been observed that most programmers ignore checking exceptions, and higher classes in the exception class hierarchy are more frequently used.

In [22], Brunet presented the concept of *design test* that automatically checks whether the code conforms to the specific *design rule* by using a test-like program. To support it, *DesignWizard (DW)* had been developed with a fully-fledged API that allows writing design tests for Java codes using *JUnit*. The proposal was applied to three software products in their group and student projects in the undergraduate course. The results showed that this approach was suitable to check conformance between the design rules and the code implementation. Moreover, it has been observed that both designers and programmers appreciated the design tests as executable documents that can be easily kept up to date.

In [23], Akahane presented a Web-based *automatic scoring system* for Java programming assignments to reduce loads of teachers in verifying a huge number of codes and in giving feedbacks to students. The system receives Java application programs submitted by students, and immediately returns the results of *JUnit* tests where the *Java Reflect API* is adopted for testing private classes and methods that have been commonly found in introductory courses. The regular expression is used to compare the output texts of each student program and those of the reference program. Through use in an actual course in their university, it was confirmed that this system was very helpful for students to improve programming skills by correcting mistakes in their programs and repeating their submissions.

In [24], Kitaya presented a Web-based scoring system of programming assignments to students, which is similar to JPLAS. Their test consists of compiler check, *JUnit* test, and result test. The result test verifies the correctness of a student code composed of only the *main* method that reads/writes data from/to the standard input/output devices, by comparing the results of this code and of the reference code. However, the method has several disadvantages from our proposal: 1) it is only applicable to a code composed of the *main* method with the standard input/output, 2) it uses other programs to use the redirection for handling the standard input/output, and 3) it needs several input files to check the correctness for different input data. On the other hand, our method is applicable to a code containing other than the *main* method, it needs only *JUnit* with a test code, and all the input data can be described in a single test code.

7 Conclusion

In this paper, we proposed the *test code generation method* for the code writing problem in JPLAS that requires implementing a Java source code containing the *standard input/output* with *exception handling*. To access the standard input/output from the test code on *JUnit*, the *test code template* is first prepared with the *input/output snatcher classes*. Then, the test code is completed by embedding the input and output extracted by running the *reference source code* into the template. This proposal is helpful in reducing the teacher load in writing the test code for the programming assignment that requires the standard input/output with exception handling, which is common for novice students. The effectiveness is evaluated through applying the method to 97 source codes in Java programming text books or Web sites, and asking five students to write source codes using the generated test codes for three problems. In future works, we will extend the proposed method to handle other input/output functions, other methods than the *main* method, and improve the readability of the generated test code to make it easier for novice students.

References

[1] N. Funabiki, R. Kusaka, N. Ishihara, and W.-C. Kao, "A proposal of test code generation tool for Java programming learning assistant system," Proc. IEEE Int. Conf. Adv. Inform. Netw. Appl., pp. 51-56, March 2017.

[2] S. Cass, The 2015 top ten programming languages, `http://spectrum.ieee.org/computing/software/the-2015-top-ten-programminglanguages/?utm_so.`

[3] N. Funabiki, Tana, K. K. Zaw, N. Ishihara, and W.-C. Kao, "A graph-based blank element selection algorithm for fill-in-blank problems in Java programming learning assistant system", IAENG Int. J. Comput. Science, vol. 44, no. 2, pp. 247-260, May 2017.

[4] N. Funabiki, Y. Matsushima, T. Nakanishi, K. Watanabe, and N. Amano, "A Java programming learning assistant system using test-driven development method", IAENG Int. J. Comput. Science, vol.40, no.1, pp. 38-46, Feb. 2013.

[5] K. K. Zaw, N. Funabiki, and W.-C. Kao, "A proposal of value trace problem for algorithm code reading in Java programming learning assistant system", Inf. Eng. Express, vol. 1, no. 3, pp. 9-18, Sep. 2015.

[6] N. Ishihara, N. Funabiki, and W.-C. Kao, "A proposal of statement fill-in-blank problem using program dependence graph in Java programming learning assistant system", Inf. Eng. Express, vol. 1, no. 3, pp. 19-28, Sep. 2015.

[7] N. Ishihara, N. Funabiki, M. Kuribayashi, and W.-C. Kao, "A software architecture for Java programming learning assistant system", Int. J. Comput. Soft. Eng., vol. 2, no. 1, Sep. 2017.

[8] K. Beck, Test-driven development: by example, Addison-Wesley, 2002.

[9] JUnit, http://www.junit.org/.

[10] Diary of kencoba, http://d.hatena.ne.jp/kencoba/20120831/1346398388.

[11] CodePress, http://codepress.sourceforge.net.

[12] M. Takahashi, Easy Java, 5th Ed., Soft Bank Creative, 2013.

[13] H. Yuuki, Java programming lessen, 3rd Ed., Soft Bank Creative, 2012.

[14] Y. D. Liang, Introduction to Java programming, 9th Ed., Pearson Education, 2014.

[15] Java programming seminar, http://java.it-manual.com/start/about.html.

[16] Kita Soft Koubo, http://kitako.tokyo/lib/JavaExercise.aspx.

[17] C. Fu and B. G. Ryder, "Exception-chain analysis: revealing exception handling architecture in Java server applications", Proc. Int. Conf. Soft. Eng., pp. 230-239, May 2007.

[18] R. Rashkovits and I. Lavy, "Students' understanding of advanced properties of Java exceptions", J. Inform. Tech. Edu., vol. 11, pp. 327-352, 2012.

[19] R. J. S. Júnior and R. Coelho, "Preserving the exception handling design rules in software product line context: a practical approach", Proc. Latin-American Symp. Depend. Comp. Work., pp. 9-16, 2011.

[20] S. Nakshatri, M. Hegde, and S. Thandra "Analysis of exception handling patterns in Java projects: an empirical study", Proc. IEEE/ACM Work. Conf. Mining Soft. Rep., pp. 500-503, May 2016.

[21] J. Bloch, Effective Java, 2nd Ed., Prentice Hall PTR, Upper Saddle River, NJ, USA, 2008.

[22] J. Brunet, D. Guerrero, and J. Figueredo, "Design tests: an approach to programmatically check your code against design rules", Proc. Int. Conf. Soft. Eng., pp. 255-258, May 2009.

[23] Y. Akahane, H. Kitaya, and U. Inoue, "Design and evaluation of automated Scoring Java programming assignments" Proc. Int. Conf. Soft. Eng., Art. Intel., Net. Para./Dist. Comp., pp.1-6, 2015.

[24] H. Kitaya and U. Inoue, "An online automated scoring system for Java programming assignments", Int. J. Inform. Edu. Tech., vol. 6, no. 4, pp. 275-279, April 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 38-48 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Framework for the Formal Specification and Verification of Security Guidelines

Zeineb Zhioua[*,1], Rabea Ameur-Boulifa[2], Yves Roudier[3]

[1]*EURECOM, Digital Security, France*

[2]*LTCI, Télécom ParisTech, Université Paris-Saclay, France*

[3]*I3S - CNRS - Université de Nice Sophia Antipolis, France*

ARTICLEINFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords:
Security Guidelines
Formal specification
Model Checking
Information Flow Analysis
Program Dependence Graph

ABSTRACT

Ensuring the compliance of developed software with general and application-specific security requirements is a challenging task due to the lack of automatic and formal means to lead this verification. In this paper, we present our approach that aims at integrating the formal specification and verification of security guidelines in early stages of the development lifecycle by combining both the model checking analysis together with information flow analysis. We present our framework that is based on an extension of LTS (labelled transition Systems) by data dependence information to cover the end-to-end specification and verification of security guidelines.

1 Introduction

This paper is an extension of work originally presented in Pacific Rim International Symposium on Dependable Computing (PRDC 2017) [1]. About 64% of the 2500+ vulnerabilities in the National Vulnerability Database NVD were due to programming mistakes [2], and the majority of software vulnerabilities are caused by coding errors. Flaws and errors can be introduced during the different phases of the software development lifecycle, from design to development. The missed programming errors can turn into security vulnerabilities at run-time, and can be exploited by intruders who may cause serious damage to the software critical assets and resources. The undetected flaws can cause a cost increase, comprising maintenance and flaw correction fees. Using code analysis tools would avoid such issues and help produce safe and secure software. The last decades have witnessed the development of many analysis techniques that aim at detecting security vulnerabilities in the early stages of development lifecycle; however, most attention was devoted to control-flow, somehow ignoring the data dependencies source of vulnerabilities that can remain undetected.

On the one hand, it is important for developers to be aware of domain-specific requirements as failure to pass the verification and validation phase, and subsequent application corrections and maintenance may be costly, time-consuming, and affect the company's reputation. On the other hand, security guidelines should be expressed in a way that allows their understanding and easy implementation for developers who may not be security experts to develop and deliver secure software.

Security guidelines or security best practices serve as recommendations to developers to reduce the application exposure to security issues, and to ensure that the developed system will behave as expected in hostile environments.

The problems described above entail the need to add on top of the development process different types of verification such as the compliance with the general and application-specific security requirements. Applying formal methods in the different phases of the software engineering process can help further understanding of the system, and detect design flaws rather early in the development.

In our papers [3,4] we conducted a survey on the static code analysis methods with the objective of identifying the approach that best fits our needs in terms of information flow properties detection and

[*]Corresponding Author: Zeineb Zhioua, zeineb.zhioua@eurecom.fr

https://dx.doi.org/10.25046/aj030106

validation, as well as on the system abstraction models that constitute a strong basis in order to carry out the analysis.

The main problem we are tackling in this paper is how to automatically verify the systematic application and compliance of (being) developed software with security requirements expressed in natural language. This problem requires the transformation of the guidelines written informally into a precise formalism by security expert(s), and this is a very tedious task. Nowadays, formal methods, in particular formal verification, are increasingly being used to enforce security and safety of programs.

We propose a framework that first provides security experts with the means to express security guidelines in a more formal way than plain text. Then, our framework verifies the adherence to the guidelines over an abstraction of the program, and provides understandable and clear feedback to the developer to indicate the exact program location where the error occurred. The innovation of our framework relies on the combination between model checking and data dependencies together with the analysis of the system behavior without actually executing it. We focus on the current version of our framework on the Java programming language.

Below we provide a sample code (Figure 1) that presents an implicit violation of the guideline from the *OWASP Cryptographic Storage Cheat Sheet* [4]:"Store unencrypted keys away from the encrypted data" [1] explaining the encountered risks when the encryption key is stored in the same location as the encrypted data. If we want to verify if the code below meets this guideline or not, then we have first to annotate the sources and the sinks, and run the analysis. We need first to highlight several elements in the codes below; the data key *k* and *encrypted_cc* are stored respectively in file **keys.txt** and **encrypted_cards.txt**. One may conclude that the guideline is met, as key *k* and *encrypted_cc* are stored in separate files. However, the two files are located in the same file system, which constitutes a violation of the guideline. Let us look at the details of the code. The developer encrypts the secret data credit card number, and stores the cipher text into a file. At line 115, the developer creates a byte array *y* used as parameter for the instantiation of a SecretKeySpec named *k* (line 116). At line 119, the key *k* is stored in a file, through the invocation of method *save_to_file* (Figure 1). Once created, key *k* is provided as parameter to the method *save_to_file(String data, String file)* (Figure 1) The developer then encrypts the secret variable creditCardNumber using method *private static byte[] encrypt(Key k, String text)* which uses key *k* as parameter. The encrypted data is then stored using method *save_to_file(String data, String file)* (Figure 1). One can conclude that the guideline is met, as key *k* and *encypted_cc* are stored in separate files. But if we take a closer look, we would notice that the two files are located in the same file system. Hence, the code violated the guideline.

In this paper, we go through the approach that we propose in order to help solving the difficulty of capturing implicit and subtle dependencies that can be source of security guidelines violations.

```
106  public static void main(String[] args)
107      throws NoSuchAlgorithmException,
108      NoSuchProviderException,
109      FileNotFoundException {
110  int c = 123456;
111  Payment p = new Payment();
112  p.setCreditCardNumber(c);
113
114  String x = "0xe04fd020ea3a6910a2d808002b30309d";
115  byte[] y = hexStringToByteArray(x);
116  SecretKeySpec k = new SecretKeySpec(y, "AES");
117
118  // save
119  save_to_file(k,"C:/[redacted]"
120      + "//[redacted]//src//secGuidelines//keys.txt");
121
122  // encrypted data
123  byte[] encrypted_cc = encrypt(k, Integer.toString
124      (p.getCreditCardNumber()));
125
126  // save
127  save_to_file(encrypted_cc,"C:/[redacted]"
128   + "//[redacted]//src//secGuidelines//encrypted_cards.txt");
129  }
```

Figure 1: Sample code for the encryption of credit card number

```
146  public static void save_to_file(String data, String file) {
147  try (PrintWriter out = new PrintWriter(file)) {
148      out.print(data +"\r\n");
149  } catch (FileNotFoundException e) {
150      // TODO Auto-generated catch block
151      e.printStackTrace();
152      System.out.println("file error");
153  }
154  }
```

Figure 2: Source code of save_to_file method

The flow of this paper is as follows: Section 2 introduces the security guidelines and discusses the motivation behind this work. In Section 3, we explain in detail the approach we carried out. We highlight several issues related with the guidelines presentation to developers in Section 4. In Section 5, we introduce the notion of information flow analysis that we make use of in our framework for the detection of implicit dependencies. In Section 6, we present our model construction methodology. We provide in Section 7 a comprehensive application of our formalism to specify and to verify the selected guideline. In section 8, we outline the related work, followed by a discussion in Section 9. And finally, Section 10 concludes the paper and discusses future work.

[1] https://www.owasp.org/index.php/Cryptographic_Storage_Cheat_Sheet#Rule_-_Store_unencrypted_keys_away_from_the_encrypted_data

2 Security Guidelines

Organizations and companies define non-functional security requirements to be applied by software developers, and those requirements are generally abstract and high-level. Security requirements such as confidentiality and integrity are abstract, and their application requires defining explicit guidelines to be followed in order to meet the requirements. Security guidelines describe bad as well as good programming practices that can provide guidance and support to the developer in ensuring the quality of his developed software with respect to the security aspect, and hence, to reduce the program exposure to vulnerabilities when delivered and running on the customer platform (on premise or in the cloud). Bad programming practices define the negative code patterns to be avoided, and that can lead to exploitable vulnerabilities, while good programming practices represent the recommended code patterns to be applied on the code.

Official sources, such as OWASP [5], Oracle [6], CERT [7], NSA [8], NIST [9] propose rules and examples of good/bad programming practices. The presentation of the security guidelines differ from one source to another. For instance, CERT Oracle Coding Standard for Java [7] provides for each guideline a textual description, followed by a compliant sample code, and another sample code violating the guideline. OWASP [5] provides for most guidelines a detailed description, and examples of compliant and non-compliant solutions.

Motivation The OWASP Foundation[5] for instance introduces a set of guidelines and rules to be followed in order to protect data at rest. However, the guidelines are presented in an informal style, and their interpretation and implementation require security expertise, as stressed in [3]. In the *OWASP Storage Cheat Sheet* [5], OWASP introduces the guideline "Store unencrypted keys away from the encrypted data" [5] explaining the encountered risks when the encryption key is stored in the same location as encrypted data. This guideline recommends not to store encrypted together with the encryption key, as this operation can result in a compromise for both the sensitive data and the encryption keys. However, encryption keys can be declared as byte arrays with insignificant names, which makes their identification as secret and sensitive data very difficult.

Correctly applying this guideline would provide a strong protection mechanism against this attack scenario: an attacker can get access to the encryption server or client, and can retrieve the encrypted data with the encryption key. Fetching those two elements allows the deciphering of the encrypted sensitive information. This reminds the well known HeartBleed[2] [10] attack that occurred couple of years ago (April 2014), and that allowed an attacker to read the memory, steal users credentials directly from the systems protected by the vulnerable version of OpenSSL. This example emphasizes the critical attacks that can be performed if the guideline is not respected. OWASP provides a set of security guidelines that should be met by developers, but does not provide the means to ensure their correct implementation. We aim at covering this gap through the formal specification of security guidelines and their formal verification using formal proofs.

In the sample code that we provided in Section 1, we showed how we could implicitly violate the guideline. Detecting this violation is not trivial, as it includes subtle dependencies that should be analyzed with due consideration. This is the main objective of this paper, and we could achieve this through the approach that we depict in details in the next Section.

3 Approach

In this section, we go through the details of our approach that aims at filling the gap between the informal description of security guidelines presented in natural language, and their automatic verification on the code level to provide precise and comprehensive feedback to the developer.

We started first by doing attempts to extract security properties from the code level, but we found out that we needed to have a reference against which we can compare the extracted program parts. This brought the idea of performing a deep survey on the guidelines that we gathered from different sources, as explained in Section 4.

The positive (resp. negative) security guidelines serve to express the desired (resp. undesired) program behavior. However, we operate on the code level, meaning that we do not monitor the program execution. We need then to approximate the program behavior but still from a static point of view. This requires that we transform the program into a formal model that allows us to exploit its properties, and approximate its behavior. In addition, the program model that we construct should be able to represent the whole flow of information in order to be able to reason about how data propagates, and capture possible information leakage. This induces the need to choose a formalism to represent the information flows that can be checked over this program model. Having covered the aforementioned aspects, we proceed to verify whether the program meets the specified guidelines or not. As we aim to decrease considerably the heavy load on the developer, we propose an automatic formal verification of the guidelines. The purpose is to verify that security guidelines are met, and not to prove that the program is correct. To the best of our knowledge, no prior work has filled in this gap between the informal description of security guidelines and their automatic formal verification on the code level.

[2] http://heartbleed.com/

The big picture of the proposed approach is shown in Figure 3, highlighting the relevant steps towards fulfilling the transformation of security guidelines written in natural language into exploitable formulas that can be automatically verified over the program to analyze.

The crucial part of the work is the explicit mapping between abstract security guidelines formal specification, and concrete statements on the code.

In order to make the process more concrete, a *separation of duties* needs to be made. We make the distinction between the *security expert* and the *developer*. The former carries out the formal specification of the security guidelines and their translation from natural language to formulas and patterns. He establishes also the mapping between the abstract labels and possible Java language instructions. The latter invokes the framework that makes use of this specification to make the mapping between abstract labels and the program logic, and then to verify the compliance of his developed software with the security requirement.

The idea we propose, as depicted in Figure 3, is the following:

3.1 Formal Specification of Security Guidelines

Starting from the guidelines presented in an informal manner, we make the strong assumption that the **security expert** formally specifies the security guidelines by extracting the key elements, and builds the formulas or patterns based on formalism. The established formulas or patterns can be supported by standard model checking tools. We present in Section 7 how the guideline that we consider in this paper can be modeled in a formalism, and can be formally verified using a model checking tool.

3.2 Program Model Construction

Choosing a program representation depends on the intended application. In our case, the program should be abstracted in a way that preserves its properties, such as the explicit and the implicit dependencies, hence allowing the performance of deep information flow analysis. In our framework, we have chosen the Program Dependence Graph (PDG) (see section 6) as the representation model, for its ability to represent both control and data dependencies. The generated PDG is then augmented with details and information extracted from the formulas and patterns of the security guidelines. We performed Information Flow Analysis over the constructed PDG in order to augment it with further security-related details. This analysis aims at capturing the different dependencies that may occur between the different PDG nodes, hence, augmenting the generated PDG with relevant details, such as annotations mapping the PDG nodes to abstract labels of the security guidelines. Then, we generate form the augmented PDG, a Labelled Transition System tha is accepted by model checking tools.

3.3 Verification

As previously mentioned, security guidelines will be modeled in the form of sequence of atomic propositions or statements representing the behavior of the system. The security guidelines will then be verified over the Labelled Transition System that we generate from the PDG augmented with implicit and subtle dependencies. The verification phase can have the following outcomes:

- The security guideline is valid over all the feasible paths
- The security guideline is violated

The first case can be advanced further, meaning that the verification can provide more details to the developer (or the tester) about circumstances under which the security guideline is valid. In the second case, recommendations to make the necessary corrections on the program can then be proposed . The concrete mapping between the abstract propositions in the formal security guidelines and the program model is managed in the Security Knowledge Base (Section 9).

4 Security Guidelines Analysis

Security guidelines are usually presented in an informal and unstructured way. Their presentation differs from one source to another, which can be misleading to developers. We did the effort of analyzing the guidelines from different sources, and we raised different problems that we have discussed in [4]. Upon this survey, we noticed a lack of precision and a total absence of automation. From developers perspective, the understanding and interpretation of guidelines is not a trivial task, as there is no formalization that exposes the necessary program instructions for each guideline, or that explains how to apply them correctly in their software. In order to overcome this weakness, we come up with a centralized database that gathers the possible mappings between guidelines and Java instructions in the Security Knowledge Base. In the OWASP Secure Coding Practices guide [5], a set of security guidelines are presented in a checklist format arranged into classes, like Database Security, Communication Security, etc. The listed programming practices are general, in a sense that they are not tied to a specific programming language. Another programming practices guide we can consider for instance is the CERT Oracle Coding Standard for Java[7] ; for each guideline, the authors provide a detailed textual explanation. For most, there are also provided examples of compliant and non-compliant sample codes in addition to the description. We want to pinpoint another key element that gathered our attention; there is a huge effort invested in order to build and maintain the catalogs, but no attempt was undertaken to instrument their automatic verification on the code level.

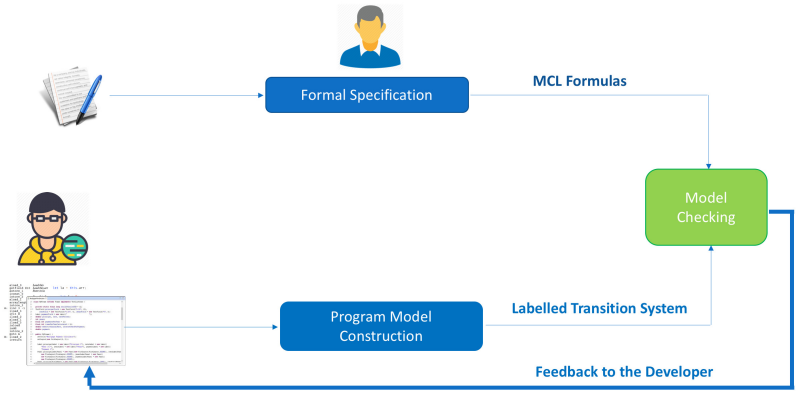


Figure 3: Framework for the formal specification and verification of Security Guidelines

5 Information Flow Analysis

In our framework, we make use of the Information Flow Analysis for many purposes, mainly the detection of implicit and subtle dependencies that can be source of covert channels and sensitive information leakage. From a security perspective, this might have serious damages on the security of the infrastructure as well as on users sensitive data. One might argue about using known security mechanisms such as access control to control the information propagation in a program. This aspect is of a paramount importance when dealing with information security. However, from a historic point of view, access control mechanisms, are used to verify the access rights at the point of access, and then, to allow or deny the access to the asset over which the mechanism is set. Access control mechanisms, just like encryption, can't provide assurance about where and how the data will propagate, where it will be stored, or where it will be sent or processed. This entails the need for controlling information flow using static code analysis. This same idea is emphasized by Andrei Sabelfeld and Andrew C. Myers [11], who deem necessary to analyze how the information flows through the program. According to the authors, a system is deemed to be secure regarding the property confidentiality, if the system as a whole ensures this property.

The main objectives of the information flow control [12] are to preserve the confidentiality and integrity of data; the former objective consists in guaranteeing that confidential data don't leak to public variables. As for the second objective, it consists in verifying that critical data is independent from public variables/output. Information flow control analyzes the software with the objective of verifying its compliance and conformance to some security policies. Different approaches have been proposed for Information Flow Control, where we can distinguish between *language-based* and *type-based* information flow control. The former has the advantage of exploiting the program source code and the programming language specificity, but falls short in covering different aspects such as physical side channels that is covered by other approaches[13] and execution environment properties. Language-based security mechanisms have been treated in the literature, including the bytecode verifiers and sandbox model. Those mechanisms enforce security through the Java language, but only the bytecode verifiers make use of the static code analysis. Type-based information flow control, on the other hand, basically makes use of the typing rules that capture illegal flows of information throughout a program, however, they are neither flow-insensitive, context-sensitive nor object-sensitive, which leads to imprecision, which in turn leads to false alarms.

6 Program Model Construction

The starting key element for this step is the standard PDG that we generate from the Java program bytecode using the JOANA tool [14]. In this PDG, control and (explicit/implicit) data dependencies are captured, which constitutes a strong basis to perform a precise analysis. Since our main objective is to automatically verify the adherence of programs to formalized security guidelines, we need to model check the guidelines MCL formulas over the program model. However, the PDG is not formal, and doesn't consist a basis for the formal verification through model checking. Thus, we need to construct from the augmented PDG a model that is accepted by a model checking tool, and that can be verified automatically through model checking techniques. We depict in Figure 4 the program model construction flow that we have adopted to generate the Labelled Transition System (LTS) from the PDG, that we generate from the program sources.

- Augmented Program Dependence Graph: this component first builds the program dependence graph (PDG) from the Java bytecode (.class) using the JOANA IFC tool [14]. We have chosen the Program Dependence Graph (PDG) as the abstraction model for its ability to represent both control and (explicit/implicit) data dependencies. The generated PDG is then annotated

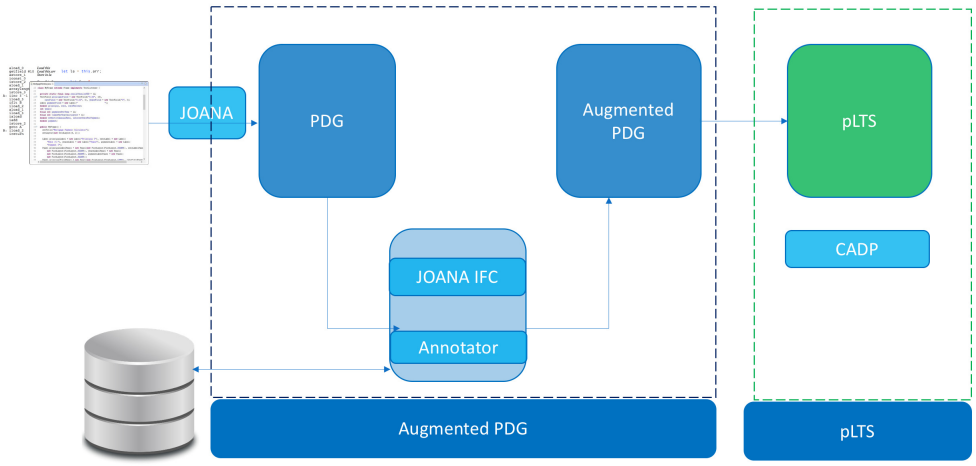


Figure 4: Methodology for the model construction: From program sources to the Augmented Program Dependence Graph, to the Labelled Transition System

by the **PDG Annotator** with specific annotations (labels in the MCL formulas). The **PDG Annotator** retrieves the nodes details (method signature) from the PDG, and fetches from the Security Knowledge Base the matching label if it exists. We run the information flow analysis using the JOANA IFC, that is formally proven [14] in order to capture the explicit and the implicit dependencies that may occur between the program variables. The operation results in a new PDG that we name the **Augmented PDG**. We show in Figure 5 the Augmented PDG of the sample code (Figure 1) that we consider in this paper.

- LTS Construction: this component translates automatically the **Augmented PDG** into a parameterized Labelled Transition System (pLTS) that is accepted by model checking tools. The annotations on the PDG nodes are transformed into labels on the transitions in the pLTS.

- Java Classes Parser: This component that we have developed [3] takes as input the URL of the Java class official documentation, and parses the HTML code (Javadoc) in order to extract all the relevant details: the class name, the inheritance, the description, the attributes, the constructor(s), the methods signatures, their return type and their parameters. This component populates the Security Knowledge Base with the extracted information.

6.1 Program Dependence Graphs

PDG (Program Dependence Graph) is a language-independent representation of program. This model was first proposed by Ferrante et al. [15] as a program representation taking into consideration both control and data relationships in a program. Formally, the PDG is a directed graph whose nodes correspond to program statements and whose edges model dependencies in the program. Those dependencies can be classified as either control or data dependencies. The nodes are predicates (variable declarations, assignments, control predicates) and edges are data and control dependencie representation; both types are computed using respectively control-flow and data-flow analysis.

PDGs have the ability to represent the information flow in a program, and have different properties, such as being *flow-sensitive*, *context-sensitive* and *object-sensitive* [13]. Being flow-sensitive is the ability of considering the order of statements in the program. The context-sensitivity is perceived from the fact that if the same method is invoked multiple times, then each call site will be represented a separate node in the graph, and will be analyzed separately. In other words, the methods calling context is considered, and this increases precision. The object sensitivity, on the other hand, is the ability to extend the analysis to the attributes level for Object Oriented Programs; object, which is an instance of a class, is not considered as an atomic entity, hence the analysis will be extended to the attributes level.
PDG abstracts away irrelevant details, such as independent and non-interacting program statements, that represent the unfeasible paths.

In our framework, we make use of the JOANA tool, that stands for Java Object sensitive Analysis [16] for the construction of the PDG. JOANA [14] is a framework that statically analyzes the byte code of Java programs; the tool first generates from the program byte code a PDG, which constitutes an over-approximation of the information flow in the program to analyze. The PDG contains apart from the nodes representing the statements and the variable declarations in the

[3] https://github.com/zeineb/Java-classes-parser

program, contains also edges referring to control and data dependencies between nodes. The dependencies represent explicit dependencies as well as transitive and implicit dependencies.

JOANA's strength relies on the ability to track how information propagates through a program, and captures both the explicit and implicit information flows.

6.2 Augmented Program Dependence Graph

We extend the definition of program dependence graphs to accommodate propositions over the set of program variables. Once the PDG is built, we compute all the propositions that are defined over the program. These propositions are parametrized by the variables defined within the program. Each node is annotated with a set of propositions.

There are two categories of information which can all be used as annotations. The first category is obtained by identifying the set of standard instructions: variables assignments, method calls, etc. at a given node. Example of methods are *encypt*, *hash*, *log*, *normalize*, *sanitize*. Second category is dedicated to relationship between variables, the dependencies in terms of explicit and implicit data dependencies between variables in a program.

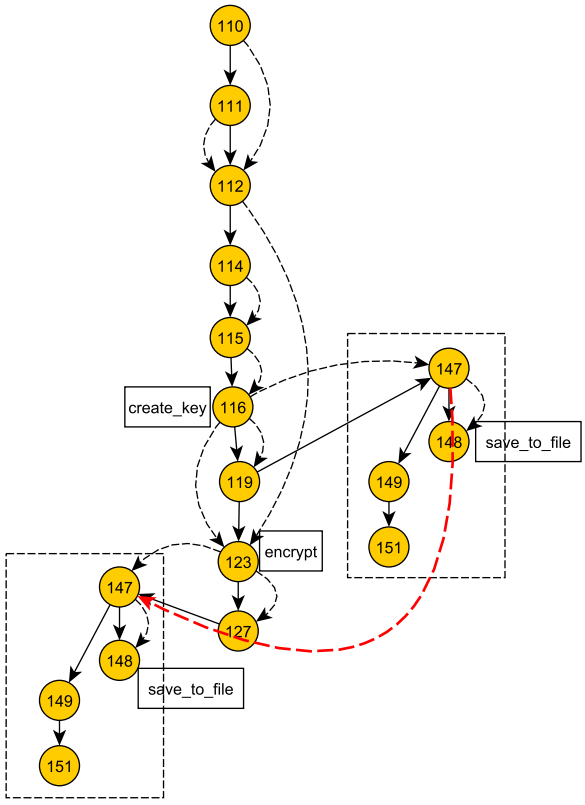


Figure 5: Augmented Program Dependence Graph for the sample code. Strong edges represent the control flows, the dashed edges refer to explicit and implicit data flows. Nodes are labeled with their corresponding instructions line numbers

Note that referring to our security knowledge base, different annotations on the PDG were pre-computed, like for instance the *save*, *userInput* and the *encyption_key*.

Automatic annotations. First, we have created our own annotations based on the atomic propositions of the security guidelines formulas. We made modifications on the source code of JOANA, and added the annotations *hash*, *userInput*, *Password*, *encrypt*, etc. in addition to the predefined annotations *SOURCE* and *SINK*. As shown in Figure 5, different program nodes are annotated with abstract labels, such as node 65 annotated as *Password*, node 73 annotated as *hash* and node 80 annotated as *log* or *store*. The log annotation was pre-computed after the PDG is built, meaning that the security knowledge base was accessed to fetch the concrete possible mappings between known APIs, methods, methods parameters mapped to the abstract labels the formulas are built upon. The mapping between the method invocation *logger.log* and the label *log* is already established. Same for the *hash* label. However, for the *Password*, the automatic annotation requires a semantic analysis to be performed over the code in order to determine the variable names matching password. The semantic analysis is not in the scope of this paper.

Annotations validation by the developer. Once the automatic detection of the atomic propositions on the PDG is performed, the intervention of developer is required to validate the added annotations. There might also be the case where the developer creates a method implementing the hash functionality, then the detection of the label *hash* on the program model will fail. In the sample code, the logging, which is one simple possible storing operation, was invoked. The developer, in our example, annotated the node 65 as *Password*, and the node 80 as *store* (in addition to the *log* automatic annotation).

6.3 Labelled Transition System

A parameterized Labelled Transition System (pLTS) is a labelled transition system with variables; a pLTS can have guards and assignment of variables on transitions. Variables can be manipulated, defined, or accessed in states, actions, guards, and assignments. JML [17], Z [18], B [19] allow to describe the states of the system through mathematics-based objects (machines, sets, etc.), and they describe pre- and post-conditions on the transitions between the states. Those languages deal with sequential programs and do not handle value passing for most.

Definition 1 (pLTS) *A parametrized LTS is a tuple* $pLTS \triangleq (S, s_0, L, \rightarrow)$ *where:*

- S *is a set of states.*
- $s_0 \in S$ *is the initial state.*

- *L is the set of labels of the form* $\langle \alpha, e_b, (x_j := e_j)^{j \in J} \rangle$, *where* α *is a parametrized action,* e_b *is a guard, and the variables* x_j *are assigned the expressions* e_j. *Variables in are assigned by the action, other variables can be assigned by the additional assignments.*
- $\rightarrow \subseteq S \times L \times S$ *is the transition relation.*

Informally, we interpret the behavior of a program as a set of reachable states and actions (instructions) that trigger a change of state. The states express the possible values of the program counter, they indicate whether a state is an entry point of a method (initial state), a sequence state (representing standard sequential instruction, including branching), a call to another method, a reply point to a method call, or a state which is of the method terminates. Each transition describes the execution of a given instruction, so the labels represent the instruction names.

The LTS labels can mainly be of three types: actions, data and dependencies.

- Actions: they refer mainly to all program instructions, representing standard sequential instructions, including branching and method invocations.
- Value passing: as performed analysis involves data, generated LTSs are parametrized, i.e, transitions are labelled by actions containing data values.
- Dependencies: in addition to program instructions, we added transitions that bring (implicit and explicit) data dependencies between two statements with the objective of tracking data flows. Indeed, transitions on LTS show the dependencies between the variables in the code. We label this kind of transition by *depend var1 var2* where *var1* and *var2* are two dependent variables.

7 Verification

With the objective of achieving our main goal consisting in helping a programmer verify that his program satisfies given security guidelines, we translate the augmented PDG into a formal description, which is precise in meaning and amenable to formal analysis. As usual, in the setting of distributed and concurrent applications, we provide behavioral semantics of analyzed programs in terms of a set of interacting finite state machines, called LTS [20]. An LTS is a structure consisting of states with transitions, labeled with actions between them. The states model the system states; the labeled transitions model the actions that a system can perform. Considered LTS are specific; their actions have a rich structure, for they take care of value passing actions and of assignment of state variables. They encode in a natural way the standard instructions of PDGs (as shown in Figure 5). Besides the classical behavior of a PDG, we encode in our LTS the result of tracking of explicit and implicit dependencies between program instructions. These dependencies are encoded by transitions labeled with the action *depend input_data output_data* (see Figure 6), allowing one to prove information flow properties.

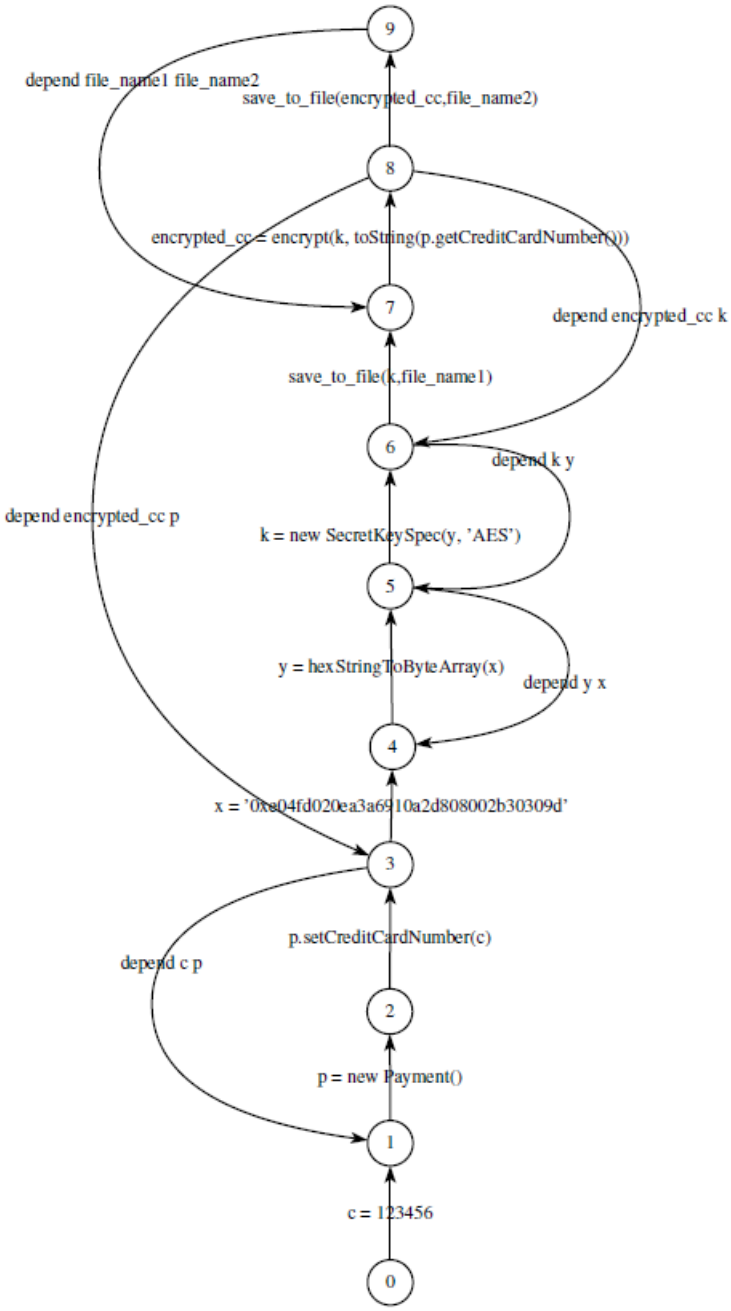


Figure 6: Labelled Transition System for the sample code given in Figure 1

Once the behavioral models are generated, we use model checking technique to automatically verify correctness of guidelines against the model.

For expressing the properties, we adopt MCL logic [21]. MCL (Model Checking Language) is an extension of of the alternation-free regular μ-calculus with facilities for manipulating data in a manner consistent with their usage in the system definition. The

MCL formula are logical formula built over regular expressions using boolean operators, modalities operators (necessity operator denoted by [ ] and the possibility operator denoted by ⟨ ⟩) and maximal fixed point operator (denoted by μ). For instance, the guideline *"Store unencrypted keys away from the encrypted data"* will be encoded directly by the following formula MCL:

```
[true*.{create_key ?key:String}.true*.
({save !key ?loc1:String}.true*.
{encrypt ?data:String !key}.true*.
{save !data ?loc2:String}.true*.
{depend !loc1 !loc2}
|
{encrypt ?data:String !key}.true*.
{save !key ?loc1:String}.true*.
{save !data ?loc2:String}.true*
.{depend !loc1 !loc2})] false
```

This formula presents five actions: the action {*create_key ?key:String*} denoting encryption key *key* (of type *String*) is created, the actions {*save !key ?loc1:String*}, {*save !data ?loc2:String*}, {*encrypt ?data:String !key*} denoting respectively the storage of the corresponding *key* in location *loc1*, the storage of the corresponding *data* in location *loc2*, the encryption of *data* using *key*, and the particular action *true* denoting any arbitrary action. Note that actions involving data variables are enclosed in braces ({ }). Another particular action that we make use of in this formula is {*depend !loc1 !loc2*}, denoting the implicit dependency between the file locations *loc1* and *loc2*; we captured this implicit dependency through advanced information flow analysis on the code.

This formula means that for all execution traces, undesirable behavior never occurs (false). The unexpected behavior is expressed by this sequence of actions: if encryption key *k* is saved in *loc1*, and *k* is used to encrypt *data* that is afterwords stored in *loc2*, then if *loc1* and *loc2* are dependent, the guideline is violated. The second undesirable behavior, expressed in the second sequence of the formula, means that if encryption of data using *k* occurs before the storage of *k* in *loc1*, and if *loc1* and *loc2* are dependent, then the guideline is violated.

We made use of the checker EVALUATOR of the CADP toolbox [22] to verify the property. From a behavioral point of view, the verification result is true, indicating that the guideline is verified. However, from a security point of view the answer should be false, as the variable *xx* containing the password in plain text, was leaked to the logging operation through and the implicit flow between this variable and the logging operation. To guarantee the reliability of the analysis, one needs to check secret/sensitive variables and depending variables as in the presented formula.

No surprise the answer is false. In addition to a *false*, the model checker produces a trace illustrating the violation from the initial state, as in Figure 7.

8 Related Work

Prior work in the area of information-flow security [23] has been developed during the last decades. A line of work [24] [25] adopts the Extended Static Checking, a specific technique for finding source code errors at compile-time. Eau Claire [21] framework operates as follows; it translates C program into Guard Commands (Guarded Command Language), that are afterwards translated into verification conditions for each function of the program. The generated verification conditions serve as input to automatic theorem prover. Adopting the prototype Eau Claire is very much-time consuming, requiring annotations entered by the developer, hence it is hard to integrate in the development phase.

De Francesco et al. [26] combine abstract interpretation and model checking to check secure information flow in concurrent systems. The authors make use of the abstract interpretation to build a finite representation of the program behavior: a labelled transition system . The security properties are specified in temporal logics, and are model checked over the built LTS. Their approach consists in verifying the non-interference property, meaning that the initial values of high level (secret) variables do not influence the final values of low level (public) variables. This approach checks for the non-interference only on two program states, which might miss possible information leakage within the process itself. In addition, the adopted formalism does not support value passing, and does not reason about data propagation.

The Verification Support Environment [27] is a tool for the formal specification and verification of complex systems. The approach adopted by the authors is similar to model-driven engineering, in the sense that the formal specification results in code generation from the model.

SecureDIS [28] makes use of model checking together with theorem-proving to verify and generate the proofs. The authors adopt the Event-B method, an extension of the B-Method, to specify the system and the security policies. The authors do not make it clear how the policies parameters are mapped to the system assets, and they do not extend the policy verification and enforcement on the program level. The work targets one specific system type (Data Integration System), and is more focused on access control enforcement policies, specifying the subject, the permissions and the object of the policy. However, access control mechanisms are not sufficient for the confidentiality property, as they can't provide assurance about where and how the data will propagate, where it will be stored, or where it will be sent or processed. The authors target system designers rather than developers or testers, and consider a specific category of policies focused on data leakage only.

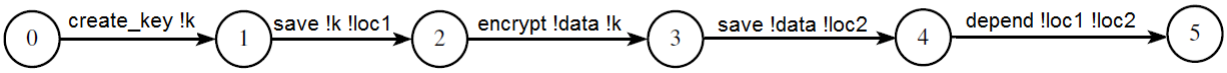


Figure 7: Path violating the guideline

GraphMatch [29] is a code analysis tool/prototype for security policy violation detection. GraphMatch considers examples of security properties covering both positive and negative ones, that meet good and bad programming practices. GraphMatch is more focused on control-flow security properties and mainly on the order and sequence of instructions, based on the mapping with security patterns. However, it doesn't seem to consider implicit information flows that can be the source of back-doors and secret variables leakage.

PIDGIN [30] introduces an approach similar to our work. The authors propose the use of PDGs to help developers verify security guidelines throughout the exploration of information flows in their developed software and also the specification and verification of adherence to those policies. Privacy policies are encoded in LEGALEASE language that allows to specify constraints on how user data can be handled, through the clauses ACCEPT and DENY [**?**]. The specification and verification of security properties rely on a custom PDG query language that serves to express the policies and to explore the PDG and verify satisfiability of the policies. The parameters of the queries are labels of PDG, which supposes that the developer is fully aware of the complex structure of PDGs, identify the sensitive information and the possible sinks they might leak to. For example, the authors propose a policy specifying that the guessing game program should not choose a random value that is deliberately different from the user's guess provided as input.

9 Discussion

We tackled in this paper the problem of verifying the adherence of the developed software to security guidelines that are presented in formal language. We raised the main issue regarding the interpretation, implementation and verification of the guidelines that are written in natural language, which might be subject to misinterpretation by developers. We worked towards stripping away ambiguities in [19] capturing implicit information flows that can be source of information leakage. We provided a centralized repository (Security Knowledge Base) gathering the security guidelines and patterns that we have discussed in detail in [19]. Apart from the patterns, we have centralized the labels that are used to build the formulas, mapped to the traditional information annotations (*SOURCE*, *SINK* and *DECLASS*) together with their security level (*HIGH* / *LOW*).

Security Knowledge Base *Security Knowledge Base* is a centralized repository gathering the labels of the formulas mapped to APIs, instructions, libraries or programs. This helps the automatic detection of labels on the system model. We built **Security Knowledge Base** using a Java classes parser [4] that operates as follows: for the different Java classes used in the program to analyze, we launch the parsing of this given class (html code, javadoc), and we extract all the relevant details, such as the description, the attributes, the constructors, the methods signatures and their parameters. Then, we made the effort of performing a semi-automatic semantic analysis to detect key elements, such as the keyword *secure*, *key*, *print*, *input*, etc. This operation is of a paramount importance, as it allows us to map the key words used to build the formulas, to the possible Java language instructions (methods invocations, constructors invocations, specific data types declarations, etc). For example, the Java API *KeyGenerator.generateKey()* is mapped to the label *isKey*. This label is also mapped to the traditional information flow annotation **high** level **source**.

As part of the Security Knowledge Base, we have also considered the vulnerabilities the program could eventually be exposed to if the guideline is violated. The Security Knowledge Base is rich yet extensible repository, that can be extended if new security concepts are introduced. For instance, the same guideline might be expressed through different MCL formulas and using different terms that are semantically equivalent. Let us take the example of the guideline *"Store unencrypted keys away from the encrypted data"* that we have previously formulated using those keywords: *create_key*, *save* and *encrypt*. Among the keywords contained in the dictionary that we provide to the security expert, the *create_key* is semantically equivalent to *isKey*, the word *save* is equivalent to *store*, and so on and so forth. Hence, the used keywords can be replaced with their equivalent as long as they do not alter the semantics of the guideline.

10 Conclusion

In this paper, we presented on a high level the approach that we propose with the objective of filling the gap for the verification of the security guidelines. We pinpoint different issues with the security guidelines that are present in different sources, but no verification means is provided to the developers to make sure that their being developed software adheres to those guidelines. Security guidelines are meant mainly for developers, but the way they are

[4] https://github.com/zeineb/Java-classes-parser

presented presents ambiguities, and this might lead to misinterpretation. Formalizing the guidelines would help strip away ambiguities, and prepare the ground for the formal verification. We stressed on the need for performing model checking as verification approach. This allows to have an automatic verification, hence to reduce the intervention of a human operator, whether the developer or the security expert leads this verification.

References

1. Zeineb Zhioua and Stuart Short and Yves Roudier, "Towards the Verification and Validation of Software Security Properties Using Static Code Analysis", in International Journal of Computer Science: Theory and Application,
2. Jon Heffley, Pascal Meunier, "Can Source Code Auditing Software Identify Common Vulnerabilities and Be Used to Evaluate Software Security?", in Proceedings of the 37th Hawaii International Conference on System Sciences-2004, https://doi.org/10.1109/HICSS.2004.1265654
3. Zeineb Zhioua and Stuart Short and Yves Roudier, "Static Code Analysis for Software Security Verification: Problems and Approaches", in 2014 IEEE 38th Annual International Computers, Software and Applications Conference Workshops,
4. Zeineb Zhioua and Yves Roudier and Stuart Short and Rabea Ameur-Boulifa, "Security Guidelines: Requirements Engineering for Verifying Code Quality", in ESPRE 2016, 3rd International Workshop on Evolving Security and Privacy Requirements Engineering
5. OWASP, Cryptographic Storage Cheat Sheet
6. Oracle, Secure Coding Guidelines for Java SE
7. CERT, SEI CERT Oracle Coding Standard for Java
8. NSA, Juliet Test Suite
9. National Institute of Standards and Technologies and Elaine Barker, "Recommendation for Key Management", 2016, http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-57pt1r4.pdf
10. Durumeric, Zakir and Kasten, James and Adrian, David and Halderman, J Alex and Bailey, Michael and Li, Frank and Weaver, Nicolas and Amann, Johanna and Beekman, Jethro and Payer, Mathias and others, "The matter of heartbleed", in Proceedings of the 2014 Conference on Internet Measurement Conference
11. Sabelfeld, Andrei and Sands, David, "Declassification: Dimensions and Principles"
12. Denning, Dorothy E. and Denning, Peter J, "Certification of Programs for Secure Information Flow", https://doi.org/10.1145/359636.359712
13. Hammer, Christian and Krinke, Jens and Snelting, Gregor, "Information flow control for java based on path conditions in dependence graphs" in IEEE International Symposium on Secure Software Engineering
14. Jrgen Graf and Martin Hecker and Martin Mohr and Gregor Snelting, "Checking Applications using Security APIs with JOANA", in 8th International Workshop on Analysis of Security APIs
15. Ferrante, Jeanne and Ottenstein, Karl J. and Warren, Joe D., "The Program Dependence Graph and Its Use in Optimization", in ACM Trans. Program. Lang. Syst., July 1987, https://doi.org/10.1145/24039.24041
16. Jurgen Graf and Martin Hecker and Martin Mohr, "Using JOANA for Information Flow Control in Java Programs - A Practical Guide", in Proceedings of the 6th Working Conference on Programming Languages (ATPS'13)
17. Leavens, Gary T and Baker, Albert L and Ruby, Clyde, "JML: a Java modeling language" in Formal Underpinnings of Java Workshop (at OOPSLA98)
18. Potter, Ben and Till, David and Sinclair, Jane, "An Introduction to Formal Specification and Z"
19. Lano, Kevin, "The B language and method: a guide to practical formal development"
20. A. Arnold, "Finite transition systems. Semantics of communicating systems"
21. Sabelfeld, Andrei and Myers, Andrew C, "Language-based information-flow security", in IEEE Journal on selected areas in communications
22. Mateescu, Radu and Thivolle, Damien, "A Model Checking Language for Concurrent Value-Passing Systems" in "Proceedings of the 15th International Symposium on Formal Methods" , https://doi.org/10.1007/978-3-540-68237-0_12
23. Garavel, Hubert and Lang, Frederic and Mateescu, Radu and Serwe, Wendelin "CADP 2010: A Toolbox for the Construction and Analysis of Distributed Processes" in "Tools and Algorithms for the Construction and Analysis of Systems: 17th International Conference, TACAS 2011", https://doi.org10.1007/978-3-642-19835-9_33
24. Flanagan, Cormac and Leino, K. Rustan M. and Lillibridge, Mark and Nelson, Greg and Saxe, James B. and Stata, Raymie, "PLDI 2002: Extended Static Checking for Java", http://doi.acm.org/10.1145/2502508.2502520
25. Chess, Brian V "Improving computer security using extended static checking"
26. De Francesco, Nicolette and Santone, Antonella and Tesei, Luca "Abstract Interpretation and Model Checking for Checking Secure Information Flow in Concurrent Systems"
27. Serge Autexier and Dieter Hutter and Bruno Langenstein and Heiko Mantel and Georg Rock and Axel Schairer and Werner Stephan and Roland Vogt and Andreas Wolpers, "VSE formal methods meet industrial needs", https://doi.org/10.1007/s100099900022
28. Akeel, Fatimah and Salehi Fathabadi, Asieh and Paci, Federica and Gravell, Andrew and Wills, Gary, "Formal Modelling of Data Integration Systems Security Policies", https://doi.org/10.1007/s41019-016-0016-y
29. John Wilander and et al, "Pattern Matching Security Properties of Code using Dependence Graphs"
30. Andrew, Johnson and Lucas, Waye and Scott, Moore, "Exploring and Enforcing Security Guarantees via Program Dependence Graphs", https://doi.org/10.1145/2737924.2737957

"Formal specification of security guidelines for program certification"

"Abstract Interpretation and Model Checking for Checking Secure Information Flow in Concurrent Systems" in Proceedings of Fundamenta Informaticae, 2003

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 40-48 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Technology

ASTESJ

ISSN: 2415-6698

A New profiling and pipelining approach for HEVC Decoder on ZedBoard Platform

Habib Smei[*1, 2], Kamel Smiri [1, 3], Abderrazak Jemai [1, 4]

[1]Université de Tunis El Manar, Faculté des Sciences de Tunis, Laboratoire LIP2, 2092, Tunis, Tunisie

[2]Direction Générale des Etudes Technologiques, Institut Supérieur des Etudes Technologiques de Rades, Rades, Tunisie

[3]Université de Manouba, Institut Supérieur des Arts Multimédias Manouba, Campus Universitaire Manouba, 2010, Tunisie

[4]Université de Carthage, INSAT, B.P. 676, 1080 Tunis, Cedex, Tunisie

ARTICLE INFO

Article history:
Received: 01 August, 2017
Accepted: 22 September, 2017
Online: 02 October, 2017

Keywords:
Co-Design flow
Profiling
Performance estimation
Embedded Systems
Pipeline
HEVC decoder
Zynq
Zedboard

ABSTRACT

New multimedia applications such as mobile video, high-quality Internet video or digital television requires high-performance encoding of video signals to meet technical constraints such as runtime, bandwidth or latency. Video coding standard h.265 HEVC (High Efficiency Video Coding) was developed by JCT-VC to replace the MPEG-2, MPEG-4 and h.264 codecs and to respond to these new functional constraints. Currently, there are several implementations of this standard. Some implementations are based on software acceleration techniques; Others, on techniques of purely hardware acceleration and some others combine the two techniques. In software implementations, several techniques are used in order to decrease the video coding and decoding time. We quote data parallelism, tasks parallelism and combined solutions. In the other hand, In order to fulfill the computational demands of the new standard, HEVC includes several coding tools that allow dividing each picture into several partitions that can be processed in parallel, without degrading neither the quality nor the bitrate.

In this paper, we adapt one of these approaches, the Tile coding tool to propose a pipeline execution approach of the HEVC / h265 decoder application in its version HM Test model. This approach is based on a fine profiling by using code injection techniques supported by standard profiling tools such as Gprof and Valgrind. Profiling allowed us to divide functions into four groups according to three criteria: the first criterion is based on the minimization of communication between the different functions groups in order to have minimal intergroup communication and maximum intragroup communication. The second criterion is the load balancing between processors. The third criterion is the parallelism between functions. Experiments carried out in this paper are based on the Zedboard platform, which integrates a chip Zynq xilinx with a dual core ARM A9. We start with a purely sequential version to reach a version that use the pipeline techniques applied to the functional blocks that can run in parallel on the two processors of the experimental Platform. Results show that a gain of 30% is achieved compared to the sequential implementation.

1. Introduction

This paper is an extension of the work originally presented in [1].

In recent years, the number of applications processing digital video is steadily growing. Streaming videos, videoconferencing, web cameras, mobile video conversations are examples of digital video that require good video quality.

In addition, statistics show that by 2020 [2], Internet video streaming and downloads will increase to over 80% of all consumer Internet traffic. This growing demand for digital video processing encourages digital video coding market operators to design and develop new solutions that can meet this growing need.

*Corresponding Author: Habib Smei, Université de Tunis El Manar, Faculté des Sciences de Tunis, Laboratoire LIP2, 2092, Tunis, Tunisie
Email: habibsmei@gmail.com

Indeed, we are witnessing an increase in companies and research groups working to develop standards of video codec with better quality.

The JCT-VC (Joint Collaborative Team on Video Coding) group, formed by VCEG and MPEG, was created at the end of 2009 to create a new standard that meets the new requirements of video processing [3]. Following the work of this group, a new standard called H.265 / HEVC (High Efficiency Video Coding) was created with the aim of compressing a half-rate video from the previous H.264 / AVC standard to one Quality. This improvement in terms of efficiency is necessary to manage beyond high definition resolutions like HD, Quad-HD and Ultra-HD.

Many research groups and companies participate in standardization meetings and contribute to the growth of the standard with new algorithms. As a result, many implementations of this standard have been developed. One of these implementations is the test model software [4], which is a complete encoder and a decoder for the new algorithms.

To design video processing devices that implement standard video encoders, designers investigate several solutions. All solutions are based on improvement techniques based on software or hardware optimization approaches or a combination of both approaches.

The work undertaken in this paper is part of the LIP2 co-design flow.

In a co-design process, designers typically begin to run embedded application code on a host machine. Then, a thin profiling step is done to describe each function. This description can affect the execution time, memory size, number of calls, relationship between functions and other parameters that may be useful for the designer to make the best design decisions.

In this paper, we present a new approach to implement a pipeline solution of the HEVC H.265 application decoder based on the Zedboard platform [5]. The results show that a gain of 30% is achieved compared to the sequential implementation.

The rest of this article is organized as follows: Section 2 presents an overview of the HEVC Standard and HEVC h.265 decoders. In section 3, we present some works in relation with HEVC parallelization.

Section 4 gives a detailed description and profiling of the version of the test model of the decoder and analyses the results obtained. Section 4 shows the execution of the decoder pipeline on the Zedboard platform. Finally, Some conclusion remarks and future directions are given in Section 6.

2. Overview of the HEVC Standard

Due to the complexity of multiprocessor systems, the probability of failure is all the more important that it requires special consideration. Indeed, during a failure, a part of the application state disappears and the application may pass in an inconsistent state that prevents it from continuing normal execution.

The HEVC [5,6] is the acronym of "High Efficiency Video Coding". This is the latest video-coding format used by JCT-VC. This standard is an improvement of the H.264 / AVC standard. It was created on January 2013 [7], when a first version was finalized. It was developed jointly by the ISO / IEC Moving Picture Experts Group (MPEG) & ITU-T Video Coding Experts Group (VCEG).

The main objective of this standard is to significantly improve video compression compared to its predecessor MPEG-4 AVC / H.264 by reducing bitrate requirements by as much as 50% compared to H.264/ AVC, with equivalent quality.

The HEVC supports all common image definitions. It also provides support for higher frame rates, up to 100, 120 or 150 frames per second.

Video coding standards have evolved mainly through the development of the well known ITU-T and ISO / IEC standards. ITU-T produced H.261 [8] and H.263 [9], ISO / IEC produced MPEG-1 [10] and MPEG-4 Visual [11], and the two organizations together produce H. 262 / MPEG -2 Video [12] and H.264 / MPEG-4 Advanced Video Coding (AVC) [13,14] standards.

The two standards that have been produced jointly have had a particularly strong impact and have found their way into a wide variety of products that become increasingly common these days. In the course of this evolution, efforts have been multiplied to increase the compression capacity and to improve other characteristics such as the robustness against data loss. These efforts take into account the capability of practical IT resources to be used in products at the time of the expected deployment of each standard.

The HEVC standard is intended to complement various objectives and to meet even stronger needs to encode videos with an efficiency more important than H.264 / MPEG -4. Indeed, there is a growing diversity of services such as ultra high definition television (UHDTV) and video with a higher dynamic range.

On the other hand, the traffic generated by video applications targeting peripherals and mobile tablets and transmission requirements for video-on-demand services impose serious challenges in current networks. An increased desire for quality and superior resolutions also occurs in mobile applications.

The HEVC standard defines the process of encoding and decoding the video. As an input, the encoder will process an uncompressed video. It performs the prediction, transformation, quantification and entropy coding processes to produce a bitstream conforming to the H.265 standard.

The decoding process is divided into four stages. The first stage is the entropy decoding for which relevant data such as reference frame indices; intra-prediction mode and coding mode are extracted. These data will be used in the following stages. The second is called reconstruction step, which contains the inverse quantization (IQ), inverse transform (IT) and a prediction process, which can be either intra-prediction or motion compensation (inter-prediction). In the third stage, a de-blocking filter DF is applied to the reconstructed frame. Finally a new filter called Sample Adaptive Offset (SAO) is applied in the fourth stage. This filter adds offset values, which are obtained by indexing a lookup table to certain sample values [14].

In HEVC, the coding structure is based on a quaternary tree representation allowing partitioning into multiple block sizes that

easily adapt to the content of the frames. Three units are defined to treat each frame. First, the Coding Unit (CU) that defined the frame‘s based size on pixels. It allows sizes ranging from 8x8 to 64x64 pixels to be adapted according to the application. Second, the Prediction Unit (PU) that defines the partitioning size according to predicts type (inter and intra). Finally, the Transform Unit (TU) defines the quantification and transforms size to be applied to a prediction unit. Four levels of decomposition are possible for this TU, which take sizes ranging from 4x4 to 32x32 pixels.

Three types of coding structures are defined; ALL intra or intra-only (AI), Low Delay (LD) and Random Access (RA). In the first structure, all pictures are encoded as intra, which yields very high quality, and no delay; this mode is mainly aimed for studio usage.

In the low delay one, the first is an intra frame while the others are encoded as generalized P or B pictures (GPB). This structure is conceived for interactive real-time communication such as video conferencing or real-time uses where no delay for waiting for future frames is permitted.

Finally, the random access structure is similar to hierarchical structure and intra pictures are inserted periodically, at the rate of about one per second. It is designed to enable relatively frequent random access points in the coded video data. This coding order has an impact on latency, since it requires frame reordering: for this reason the decoder might have to wait to have decoded several frames before sending them to output. This is the most efficient mode for compression but also requires the most computational power.

Each of these three modes has a low complexity variant where some of the tools are disabled or switched into a faster version. As an example low complexity uses CAVLC instead of CABAC

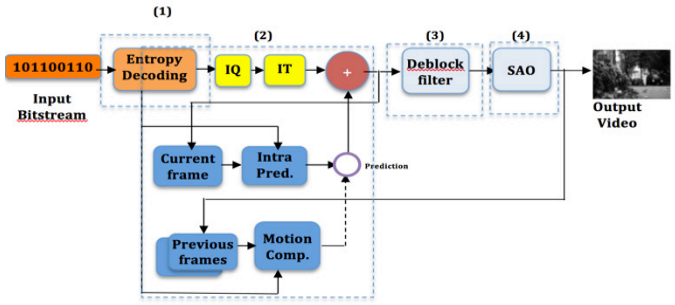


Figure 1. Functional structure of the HEVC decoder

3. Related works

HEVC includes several parallel specifications for many distributed multi-core systems. This allows division of each picture into several partitions that can be processed in parallel.

For these specifications, coarse grain communications levels are used. We cite the Group Of Pictures (GOP) level, the Slice Level, the Tile level and the wavefront parallel processing (WPP) level.

All these communication granularity has been justified by the small data dependency between processes acting on separate partition blocks.

For an embedded multiprocessor System on Chip (SoC) implementation, the communication granularity specified is not appropriate in all cases given the limited on-chip resources (memory, CPU frequency, bandwidth, …). For this, more fine grain communications granularities (CTU, TU, PU) are specified in HEVC standard based on a fine partition of a frame.

In [15], Authors carried out the analysis of parallel processing tools to understand the effectiveness of achieving the purpose for which they are targeted. In [16], an optimized method for MV-HEVC is proposed. It uses multi-threading and SIMD instructions implementation on ARM processors. The proposed method of MV-HEVC showed improvement in terms of processing speed on advanced RISC multi-core processors mobile platforms (ARM). In [17], Authors used the Wavefront Parallel Processing (WPP) coding and implemented it on multi- and many-core processors. The implemented approach is named Overlapped Wavefront (OWF), an extension of WPP tool that allows processing multiple picture partitions as well as multiple pictures in parallel with minimal compression losses. Results of her experimentations show that exploiting more parallelism by increasing the number of cores can improve the energy efficiency measured in terms of Joules per frame substantially. [18] discussed various methods in which the throughput of the video codec has been improved including at the low level in the CABAC entropy coder, at the high level with tiles, and at the encoder with parallel merge/skip tool. In [19], many optimizations are proposed to achieve HEVC real-time decoding on a mobile processor. These code optimizations include the adoption of single-instruction multiple-data (SIMD). An important speedup is accomplished with multiple threads assigned each one to a picture to be decoded. However, in the proposed solution, the SAO filter has not been implemented and optimized. Further, for the AI configuration, no additional speedup is achieved since no CPU resources are available anymore to decode a picture. In [20], a hybrid parallel decoding strategy for HEVC is presented. It combines task and data parallelism without any constraint or coding tools. The proposed approach aims to balance execution time of different stages, especially with SSE (Streaming SIMD Extensions) optimization. Another parallelism approach based on entropy slices is presented in [21]. This approach is not based on many slices because it can reduce the coding efficiency. It assigns one thread per LCU block to parallelize the HEVC decoder. Moreover, these threads are synchronized using the Ring-Line Strategy to maintain the wave front dependencies. This solution is great for high resolutions; however, it’s not the same for others.

4. Profiling of Test Model HEVC Decoder

In the specification and modelling phase of a co-design flow, the complete embedded system is written in a high level language such as C, C++, Java, Matlab and then the software is functionally verified. It will then be simulated on a host machine in order to understand its behaviour and measure the runtime performance of the program and eventually return feedback and performance statistics to the designer

This specification is an input to the profiling step, which consists in understanding the behaviour of the system in the various execution cases in order to deduce a clear representation of the various functions that represent it.

The behaviour of the system can be understood and mastered by following a parametric analysis of its execution in the different configurations and situations that the system is subject to.

This analysis can only be undertaken after having completed a detailed description of different functions of the system. The description can include the execution time of each function, its sizes, call graphs, dependencies between them and other information that can help designers to make the best decisions and technical choices. This process of identification and description is called profiling.

Once the profiling is done, designers can test the system with several configurations and estimate its performance parameters.

Several works dealing with performance estimation of embedded systems have been realized. In the case of the HEVC codec, performance estimation aims to test the ability of codecs to process video for real-time applications and to support new resolutions (HD, UHD, Full HD, 4k, 8k) and to measure footprint and the space occupied by the hardware components that make up the device.

In the work carried out by [22], a complete profiling of the encoder was carried out. The aim of this work is to present the different functions of the encoder, their execution times and the types of operations carried out in order to deduce the functions that are candidates for a hardware migration. The results are presented in terms of types of assembly level instructions in each encoder function. In [23], the authors propose a hybrid parallel decoding strategy for HEVC, which combines task level parallelism and data level parallelism. In [24], the authors use a performance estimation analysis to prove a power model based on bit derivations that estimates the energy required to decode a given HEVC coded bitstream. In [25], authors proposed a method to improve H.265/HEVC encoding performance for 8K UHDTV moving pictures by detecting amount or complexity of object motions.

4.1. Functions of the Test Model

HEVC HM Test Model is an open source project under BSD license. It is intended for the implementation of an efficient C++ HEVC decoder. The version used in this work was released from [4] and it is compliant with most of the HEVC standard. The code of the application is composed of C++ classes, and the main classes and functions are listed in the table 1.

The graph below (Figure 2) shows the most important classes in the decoding process and their associated functions. We note that the most significant classes are TComDataCU, TComTrQuant and TComSlice. The TComDataCU class represents the declaration of the CU data structure. The TComTrQuant class includes inverse transformation and inverse quantization, and the TComSlice class includes the decoder decompression process.

The optimization of the codes of the functions making up these classes can have a positive impact on the execution time and the footprint of the decoder.

As DecodeCTU, DecompressCtu called recursive function xDecompressCU which, in turn, decompresses each CU with the adequate prediction mode. When all CTUs in a frame are processed, the loop ends and the decoder performs both DF and SAO filters to correct artefacts.

Table 1: Main classes and principal functions of test model decoder

Main Classes	Principal functions	Size of code (Ko)
TDecGop	FilterPicture decompressSlice	8
TAppDecCfg	parseCfg	9
TAppDecTop	Decode x WriteOutput	21
TDecEntropy	decodePredInfo decodePUWise decodeInterDirPU	28
TdecTop	Decode Executeloopfilter	32
TComLoopFilter	xDeblockCU LoopFilterPic	35
TDecCU	decompressCtu decodeCtu xdecodeCU xcompressCU xReconInter xReconIntraQT	44
TComSlice	decompressSlice	126
TComTrQuant	partialBufferflyInverse32 partialBufferflyInverse16 partialBufferflyInverse8 xDeQuant	135
TComDataCU	initCu copySubCu	140

As shown in the diagram in Fig. 3, the process of decoding a frame is done by calling the DecompressCTU function of the TDecSlice class. This function executes a read loop of all CTUs in the current frame. Each CTU is decoded through the DecodeCtu function of the TDecCu class. The XdecodeCu function of the same class, recursively performs the decoding of all CUs in the same CTU. When decoding a CTU, the TDecSlice makes a second call to the DecompressCTU function of the TDecCu class. Like DecodeCTU, DecompressCtu call the recursive function xDecompressCU, which, in turn, decompresses each CU with the proper prediction mode. When all CTUs in a frame are processed, the loop ends and the decoder perform DF and SAO filters to correct the artefacts.

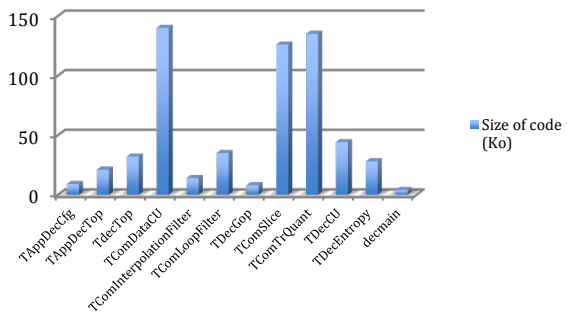


Figure 2. Main functions of Test Model Decoder

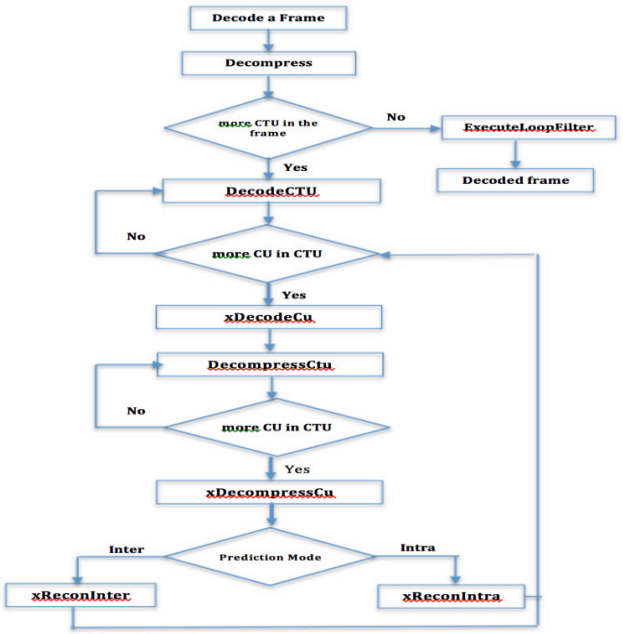


Figure 3. Frame decode Diagram of HM version

4.2. Call Graph functions of HM HEVC Decoder

Our goal in this step is to check the behaviour of the decoder described above, and to evaluate the performance of its various components in detail.

To achieve the profiling operation, we made use of profiling techniques based on code injection and also the use of tools such as gprof [26, 27] and Valgrind [28]. Gprof provides a flat profiling to give the execution time of each function of the decoder and also the number of times this function is called. Gprof can give us as a result, the call graph. This indicates for each function code, the number of times it was called, either by other functions or by itself. This information shows the relationships between the different functions and can be used to optimize certain code paths. Roughly, the results of profiling the performance generated by GProf revealed a runtime complexity of balance in terms of computing time in the functions Motion Compensation (36%) and ExecuteLoofilters of (16.7%).

Valgrind is a GPL licensed programming tool that can be used for profiling under the Linux operating system. It includes several tools, one of which is Callgrind. We used Callgrind to obtain the exact number of operations that were performed while decoding an entire video sequence. The results are separated in terms of classes, and in every class in terms of functions.

The resulting graph shows for each function, the execution time as a percentage of the total execution time (Valingring) and the functions to which they appealed. Normally, the graph shows that the most time-consuming functions are Motion Compensation and ExecuteLoofilters.

The aim of the profiling step is to understand the behaviour of the decoder described above and to evaluate the performance of its various components in detail.

To do so, we used profiling tools such as Gprof and Valgrind. These tools allow us to present the call graphs of the different functions and their execution times, and thus the state of the stack used. In order to have a complete profiling which includes the execution sequences in the different execution scenarios and configurations, a manual injection of code has been realized into significant functions.

The Gprof tool can give as a result, the call graph. This indicates for each function, the number of times it was called. This information shows the relationships between the different functions and can be used to optimize some code paths.

The profiling results generated by GProf revealed a computational complexity in terms of execution time in Motion Compensation (36%) and ExecuteLoofilters (16.7%) functions.

Valgrind is a GPL licensed tool. It includes several tools, one of which is Callgrind. We used Callgrind to get the exact number of operations performed when decoding a complete video sequence. The results are separated in terms of classes and in each class in terms of functions.

The resulting graph shows, for each function, the execution time as a percentage of the total execution time and the functions to which they appealed.

The graph shows that the most time-consuming functions are Motion Compensation and ExecuteLoofilters.

4.3. Profiling and functions Execution trace

As explained above, in order to determine the sequence of execution of the various functions of the decoder and the number of calls, a combination of code injection (manual code injection) and sampling profiling (using tools such as valgrind , Gprof, ...) has been made. The code injection in each function of the decoder has allowed us to locate called and calling functions. This also helps us to know the execution sequential order of different functions.

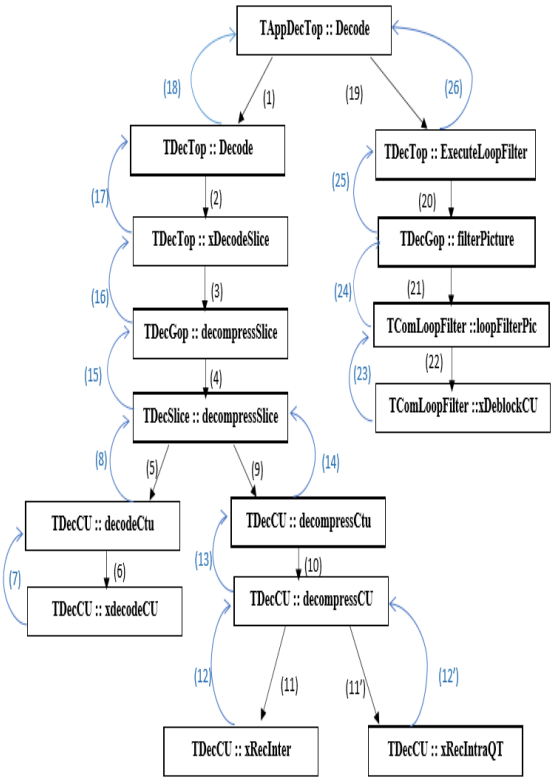


Figure 4. Functions execution Sequences

Values in parentheses indicate the execution order of decoder's functions.

The distribution of functions into groups is mainly based on three criteria: The first is to minimize communication between different groups of functions in order to have minimal inter-group communication and maximum intra-group communication. The second criterion is load balancing between the processors. Indeed, the target platform (Zedboard) is equipped with the two ARM V9 processors. To ensure optimal utilization of two processors, we must balance the charge of activity between the two cores. The third criterion is the execution parallelism. Indeed, we must ensure maximum parallel execution to minimize the average processing time of images.

Taking into account all the criteria, we came to the distribution of assigning the group functions 1, 2 and 4 to processor 0 (proc0) and functions of the group 3 to processor 1 (Proc1).

The calculations we have developed suggested the division mentioned above with the following performance data:

■ CPU.0:

Group 1: (1) (2) (3) (4) (15) (16) (17) (18) = 24.46%
Group 2: (5) (6) (7) (8) = 6.85%
Group 4: (19) (20) (21) (22) (23) (24) (25) (26) = 16.70%
Total activity time = 48.01%

■ CPU.1

Group 3: (9)(10)(11)(11') (12)(12')(13)(14)= 51.95%

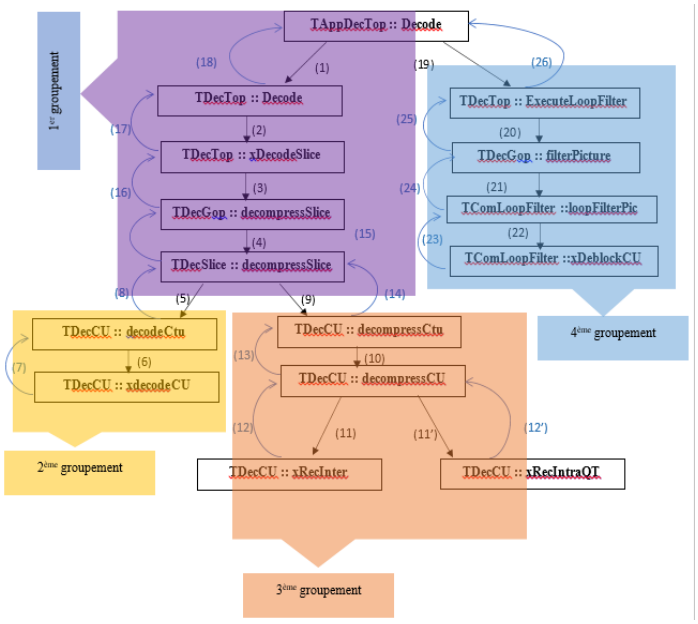


Figure 5. Grouping decoder functions

According to profiling procedure, we can divide the code on four main parts (Figure 4). The first contains the configurations functions. The second group represents the entropy decoding functions (decodeCTU and xdecodeCU). The third group is the reconstruction. Finally, the HEVC filters are grouped. We analyse their decoder's execution time in order to know its behaviour at runtime.

5. Experimental results

5.1. Test sequences

As an effort to carry out a good evaluation of the standard, the JCT-VC developed a document with some reference sequences and the codec configuration, which should be used with each one [29]. The sequences are divided into 6 groups (Classes) based on their temporal dynamics, frame rate, bit depth, resolution, and texture characteristics

A subset of six video sequences was selected from this list. These six video sequences were selected from three classes, such that two sequences from each class, namely Class A, Class C, Class F. Detailed descriptions of the sequences are given in Table II.

Table 2: Bitstreams used in the experimentations

Resolution (Sequence class)	Name	Frame rate	Frame number
2560x1600 (Class A)	*PeopleOnStreet*	30	150
	Traffic	30	300
832x480 (Class C)	*BQMall*	60	600
	PartyScene	30	500
1024x768 (Class F)	ChinaSpeed	50	500
	Slideediting	30	300

5.2. Zedboard Platform

Zedboard platform [30] (table III) is an evaluation platform based on a Zynq-7000 family [31]. It contains on the same chip two components. A dual-core ARM Cortex MPCore based on a high-performance processing system (PS) that can be used under Linux operating system or in a standalone mode and an advanced programmable logic (PL) from the Xilinx 7th family that can be used to hold hardware accelerators in multiple areas.

The two parts (PS and PL) interact between them by using different interfaces and other signals through over 3,000 connections [32]. Available four 32/64-bit high-performance (HP) Advanced eXtensible Interfaces (AXI) and a 64-bit AXI Accelerator Coherency.

Table 3: Zedboard technical specifications

Component	Characteristics
Processeur ZYNQ-7020 AP SOC XC7Z020-7CLG484CES	2 ARM Cortex A9 cores at 667 MHz
Memory	512 MB DDR3, 256 MB Quad-SPI Flash et SD Card
Communication	10/100/1000 Ethernet, USB OTG et USB UART
Extension	FMC (Low Pin Count) et 5 Pmod headers (2*6)
Display	HDMI output, VGA output et 128*32 OLED
Input / Output	8 switches, 7 push butons et 8 leds
Current and Voltage	3.0 A (Max) et 12V DC input
Certification	CE and RoHS certifier

In our experimentations, we have configured and compiled a custom Linux kernel downloaded from the official Linux kernel from Xilinx [31].

5.3. Pipeline execution approach

Since ZedBoard is based on Zynq chip that has dual cores, it is possible to distribute functions between the two cores.

Because of strict sequentially that we have seen in the functions execution of the HEVC decoder (HM Test Model version), the two processors will necessarily be in mutual exclusion of activity. To create parallel activity, we can transform the sequential video processing by a "pipelined" treatment.

If we refer to the grouping of functions explained above, we can consider that video sequence processing goes through four stages:

Stage 0: Download to a frame buffer (FB) the frame to be processed. This step corresponds to transition (1) in the graph (Fig. 5)

Stage 1: Decoding of the frame, corresponding to transitions (2, 3, 4, 5, 6, 7, 8). This stage corresponds to the Entropy decode process.

Stage 2: Decompression, corresponding to transitions (9, 10, 11-11 ', 12-12', 13, 14). This stage corresponds to the reconstruction process (that contains Inverse Transform, Inverse Quantification, Inter- prediction and Intra-prediction).

Stage 3: Filtering, which corresponds to transitions (15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26). This stage corresponds to the Loop Filter process

A parametric analysis of the application has enabled to deduce that the overall execution time of Stages 0, 1 and 3 is approximately equivalent to the execution time of Stage 2. Therefore, we can affect stages 0, 1 and 3 in the first PROC0 processor and Stage 2 to the second processor PROC1.

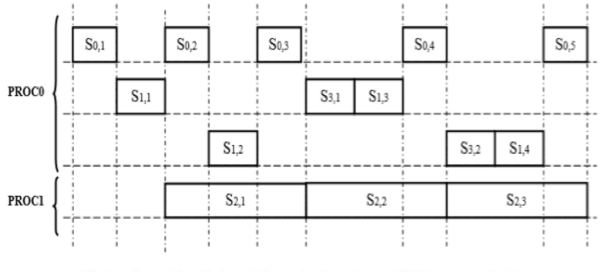


Figure 6. Pipeline execution based on dual cores Processor

Si,j represents the execution sequence of stage i applied to the frame number j. The first three rows reflect the sequencing of stage Si,j on the same processor PROC0:

Process P0, P1 and P2 support respectively the first, second and the third row.

The fourth row reflects the execution sequence of stage2 of the processor PROC1 applied to successive frames. A process P3 supports it.

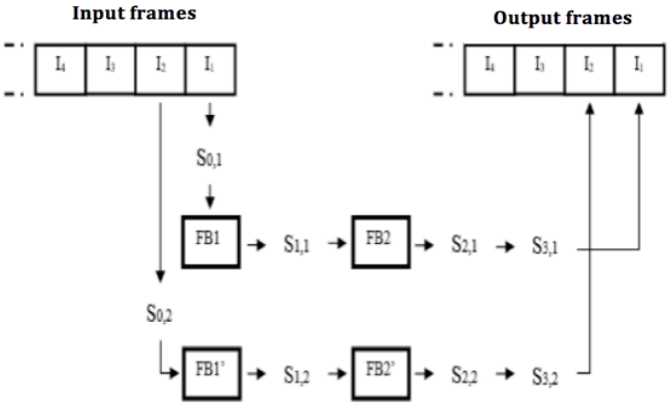


Figure 7. Sequence Process on two processors (cores)

Results obtained for the RA, AI and LD configurations are presentaed in tables IV, V and VI..

Table 4: Results for AI configuration

Class	Sequence	Frames Number	QP	Sequential Time (ms)	Pipelined Time (ms)
A	People On Street	150	22	594.28	386.28
			32	454.59	295.49
	Traffic	150	22	610.68	396.94
			32	432.75	281.29
Average execution time				523.075	340
Gain					35%
C	Party Scene	500	22	307.41	218.26
			32	219.16	155.61
	BQ Mall	600	22	259.30	184.10
			32	191.85	136.21
Average execution time				244.43	173.54
Gain					29%
F	China Speed	500	22	382.39	287.68
			32	295.93	207.15
	Slide Editing	300	22	279.75	199.83
			32	242.71	169.9
Average execution time				300.19	210.14
Gain					28%

Table 5: Results for RA configuration

Class	Sequence	Frames Number	QP	Sequential Time (ms)	Pipelined Time (ms)
A	People On Street	150	22	409.51	274.37
			32	256.97	172.17
	Traffic	150	22	293.26	196.48
			32	179.19	120.06
Average execution time				284.77	190.77
Gain					33%
C	Party Scene	500	22	153.41	107.38
			32	89.64	62.75
	BQ Mall	600	22	131.76	93.55
			32	84.08	59.70
Average execution time				114.72	80.84
Gain					29.5%
F	China Speed	500	22	216.84	156.12
			32	140.75	101.34
	Slide Editing	300	22	56.93	38.14
			32	50.25	33.67
Average execution time				116.19	82.31
Gain					29.15%

Table 6: Results for LD configuration

Class	Sequence	Frames Number	QP	Sequential Time (ms)	Pipelined Time (ms)
A	People On Street	150	22	446.65	294.79
			32	266.68	176.01
	Traffic	150	22	310.86	220.71
			32	164.25	116.62
Average execution time				297.11	202.03
Gain					32%
C	Party Scene	500	22	176.65	125.42
			32	91.99	60.71
	BQ Mall	600	22	142.03	99.42
			32	85.04	56.13
Average execution time				123.9275	85.42
Gain					31%
F	China Speed	500	22	217.94	155.84
			32	143.83	99.93
	Slide Editing	300	22	45.43	35.35
			32	41.39	27.85
Average execution time				112.14	74.01
Gain					28.89%

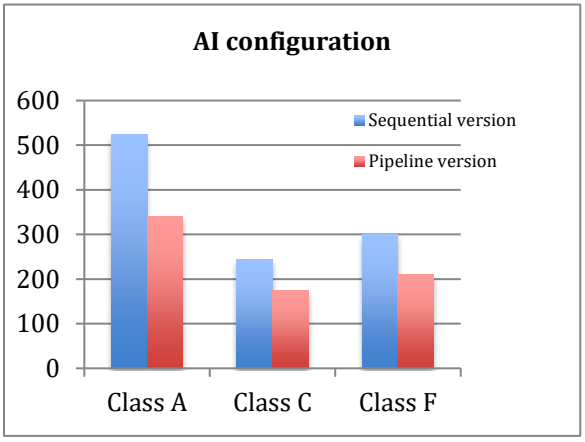


Figure 8. Execution time gain (AI configuration)

The experimental tests use reference sequences for the three configurations (AI, RA and LD) and with different quantification parameters (QPs). For the different configurations, results show a considerable gain compared to the sequential version (around 30%). We also note that class A frames have more processing time than the other classes (C and F) with a slightly more significant gain. This is logical since the resolution in this class is more important than for the other classes.

6. Conclusion

The objective of this article was to study the behaviour of the HEVC decoder represented by the reference application Test Model (HM). This study permitted us to discover the different functions of the decoder, their size, the call graphs, the percentage of CPU utilization and the number of instructions for each function.

Moreover, we have identified the most consuming functions in terms of CPU execution time and memory size, and then we have represented the execution trace of the functions by the use of an injection code technics. All these profiling results allowed us to divide the functions into groups. Once the regrouping is done, we proposed a parallelism approach based on the pipeline principle in order to run the application on the two cores of the Zedboard Platform. The experimental results found show an acceleration of about 30% compared to the sequential version.

Other technics can be added to improve this approach such as the increase of the number of processors or the implementation of hardware accelerators in the PL side.

References

[1] H. Smei, K. Smiri, A. Jemai, “Pipelining the HEVC Decoder on ZedBoard Plateform”, in 11th International Design & Test Symposium, IDT 2016, Hammamet, Tunisia, December 18-20, 2016. IEEE 2016, ISBN 978-1-5090-4900-4(IDT 2016).

[2] Cisco, Visual Networking Index (VNI): Forecast and Methodology, 2015-2020.

[3] JCT-VC - Joint Collaborative Team on Video Coding - [Online]:http://www.itu.int/en/ITU-T/studygroups/2013-2016/16/Pages/video/jctvc.aspx

[4] BSD Licence HEVC decoder (HM), Reference web site : [Online]:https://hevc.hhi.fraunhofer.de, code available Online at https://github.com/bbc/vc2-reference. Accessed, September 2016

[5] Bingjie Han , Ronggang Wang, Zhenyu Wang, Shengfu Dong, Wenmin Wang, Wen Gao, “HEVC decoder acceleration on multi-core x86 platform” IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.

[6] D. Marpe and T. Wiegand, J. Sullivan, Microsoft Corporation, IEEE Communications Magazine. August 2006.

[7] Article about HEVC - [Online]:http://en.wikipedia.org/wiki/High_Efficiency_Video_Coding

[8] Video Codec for Audiovisual Services at px64 kbit/s, ITU-T Rec. H.261, version 1: Nov. 1990, version 2: Mar. 1993.

[9] Video Coding for Low Bit Rate Communication, ITU-T Rec. H.263, Nov. 1995 (and subsequent editions).

[10] Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s—Part 2: Video, ISO/IEC 11172-2 (MPEG-1), ISO/IEC JTC 1, 1993.

[11] Coding of Audio-Visual Objects—Part 2: Visual, ISO/IEC 14496-2 (MPEG-4 Visual version 1), ISO/IEC JTC 1, Apr. 1999 (and subsequent editions).

[12] Generic Coding of Moving Pictures and Associated Audio Information—Part 2: Video, ITU-T Rec. H.262 and ISO/IEC 13818-2 (MPEG 2 Video), ITU-T and ISO/IEC JTC 1, Nov. 1994.

[13] Advanced Video Coding for Generic Audio-Visual Services, ITU-T Rec. H.264 and ISO/IEC 14496-10 (AVC), ITU-T and ISO/IEC JTC 1, May 2003 (and subsequent editions).

[14] H.Krichene Zrida, A.Jemai, A.C.Ammari, M.Abid, “High Level H.264/AVC Video Encoder Parallelization for Multiprocessor Implementation”, Date’09, International Conference on Design, automation &Test. Nice, France.

[15] P. GB, P. NS and R. Adireddy. “Analysis of HEVC/H265 Parallel Coding Tools”. PathPartner Technology consulting Pvt. Ltd. White Paper. (2014).

[16] W. Liu, J. Li and Y. B. Cho. “A novel architecture for parallel multi-view HEVC decoder on mobile device”. Journal on Image and Video Processing (2017).

[17] C. Ching, M. Alvarez-Mesa, J. L. Ben Juurlink and T. Schierl. “Parallel HEVC Decoding on Multi- and Many-core Architectures. A Power and Performance Analysis”. J Sign Process Syst (2013).

[18] M. Zhou, V. Sze and M. Budagavi. “Parallel Tools in HEVC for High-Throughput Processing”. Applications of Digital Image Processing XXXV, Vol. 8499, 849910, (2014).

[19] B. Bross and al. “HEVC Real-time Decoding”, SPIE Proceedings Vol. 8856: Applications of Digital Image Processing XXXVI, (2013).

[20] B. Han and al. “HEVC Decoder acceleration multi-core X86 plateform”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2014).

[21] M. Alvarez-Mesa and al. “Parallel video decoding in the emerging HEVC standard”, 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2012).

[22] F. Saab, I.H. Elhajj, A. Kayssi and A. Chehab, “ Profiling of HEVC encoder", Article in Electronics Letters, July 2014.

[23] B.Han, R.Wang, Z.Wang, S.Dong, W.Wang and W.Gao, "HEVC Decoder Acceleration On Multi-Core X86 Platform", IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.

[24] C.Herglotz, D.Springer, M.Reichenbach, B.Stabernack, and A.Kaup, "Modeling the Energy Consumption of the HEVC Decoding Process", DOI 10.1109/TCSVT.2016.2598705, IEEE, 2016.

[25] R.Harada, Y.Matsuo And J.Katto, "Improvement Of H.265/HEVC Encoding For 8K UHDTV By Detecting Motion Complexity", IEEE International Conference on Consumer Electronics (ICCE), 2016.

[26] Gprof web Site. https://sourceware.org/binutils/docs/gprof/

[27] Gprof tutorial. [Online]:https://www.gadgetdaily.xyz/apply-gradual-transition-effects-to-page-elements/

[28] Valgrind Web site. [Online]:http://valgrind.org/.

[29] F. Bossen, "Common HM test conditions and software reference configurations," in 12th Meeting: Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC 1/SC 29/WG 11, Geneva, 2013, document JCTVC-L1100.

[30] Zedboard Plateform web site. www.zedboard.org. Accessed, September 2016

[31] Xilinx, Inc. (2016). Zynq-7000 All Programmable SoC Technical Reference Manual. http:// www.xilinx.com/support/documentation/user_ guides/ug585-Zynq-7000-TRM.pdf.

[32] Linux kernel from Xilinx. [Online]:https://github.com/Xilinx/linux-xlnx

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 40-48 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

An Advanced Algorithm Combining SVM and ANN Classifiers to Categorize Tumor with Position from Brain MRI Images

Rasel Ahmmed[*,1], Md. Asadur Rahman[2], Md. Foisal Hossain[3]

[1]*Department of Electronics and Communication Engineering, East West University, Dhaka, Bangladesh*

[2]*Department of Biomedical Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh*

[3]*Department of Electronics and Communication Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh*

ARTICLE INFO

Article history:
Received: 30 November, 2017
Accepted: 05 February, 2018
Online: 08 March, 2018

Keywords:
Magnetic resonance imaging (MRI)
Temper based K-means & modified fuzzy C-means clustering (TKFCM)
Artificial Neural Network (ANN)
Support Vector Machine (SVM)
Tumor Size; and Tumor Position

ABSTRACT

Brain tumor is such an abnormality of brain tissue that causes brain hemorrhage. Therefore, apposite detections of brain tumor, its size, and position are the foremost condition for the remedy. To obtain better performance in brain tumor and its stages detection as well as its position in MRI images, this research work proposes an advanced hybrid algorithm combining statistical procedures and machine learning based system Support Vector Machine (SVM) and Artificial Neural Network (ANN). This proposal is initiated with the enhancement of the brain MRI images which are obtained from oncology department of University of Maryland Medical Center. An improved version of conventional K-means with Fuzzy C-means algorithm and temper based K-means & modified Fuzzy C-means (TKFCM) clustering are used to segment the MRI images. The value of K in the proposed method is more than the conventional K-means. Automatically updated membership of FCM eradicates the contouring problem in detection of tumor region. The set of statistical features obtained from the segmented images are used to detect and isolate tumor from normal brain MRI images by SVM. There is a second set of region based features extracted from segmented images those are used to classify the tumors into benign and four stages of the malignant tumor by ANN. Besides, the classified tumor images provide a feature like orientation that ensures exact tumor position in brain lobe. The classifying accuracy of the proposed method is up to 97.37% with Bit Error Rate (BER) of 0.0294 within 2 minutes which proves the proposal better than the others.

1. Introduction

The brain controls all psychological and physiological activities of human body. These functional activities can be disrupted or damaged due to the abnormal cell division or growing tumor in our brain that causes miscellaneous problems to the malfunction of our body. A human brain is divided into several major areas and these major areas (see Fig. 1) are related to different functional part of our body. Especially, these major areas are known as frontal lobe (marked with 1), central lobe (marked by 2, 3, & 4), parietal lobe (marked by 5 & 6), occipital lobe (marked with 7), and temporal lobe (marked with 8). Frontal lobe functions to control our thinking, emotion, innovation, and other cognitive works. Central lobe is a part of the frontal lobe and this part controls our movement related functions. The temporal lobe is responsible for listening and it helps to avoid the uncertainty principle. Occipital lobe helps us to see or observe anything. Another major part that controls the speech processing through reading compression area, sensory speech area and motor speech area of Broca is parietal lobe. These major areas can be affected by tumors which are definitely threat to our normal living. Therefore, the proper detection is the first priority for the remedy. This paper presents an efficient method for the detection of the tumor size as well as the position with classified tumor stages from MRI images and this work is an extension of our conference paper [1] that was presented in ECCE-2017.

[*]Rasel Ahmmed, East West University, Email: rsa@ewubd.edu

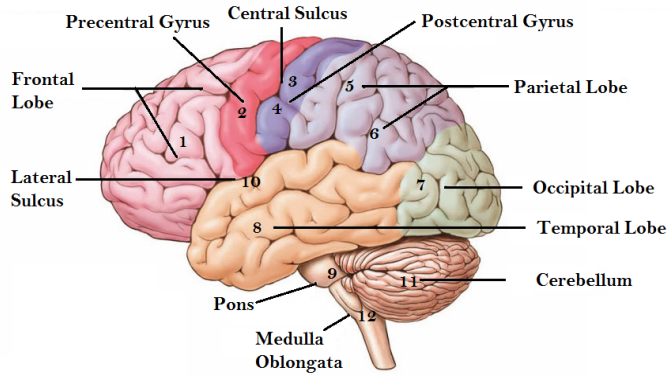


Figure 1: Major functional areas of human brain

In the recent era, there is a radical change in the field of medical science because it has experienced a great invention to the brain imaging. Previously it was predicted by the experts about brain hemorrhage on the basis of symptoms of a patient. But in recent decades the imaging techniques ease the way to detect any abnormality in the brain. There are several techniques of brain imaging like Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Ultrasound Imaging, Positron Emission Tomography (PET), and so on. Among the aforementioned imaging techniques, MRI is a gold accepted modality by which the brain can properly be imaged avoiding any harm.

In case of imaging by MRI, water density in soft tissue is considered and we know that water density in brain tissue is comparatively high than other tissue like bone. Due to the inhomogeneity of brain structure, their contrast values differ randomly in MRI and that is why the proper detection of tumor or tumor size becomes often difficult for the general observer [2]. Recently, proficient brain detection has been a remarkable challenge for medical engineering science. In this case, MRI gets more attention because of its grayscale images. Though MRI provides good contrast value than the other techniques, a suitable segmentation of brain MRI image is ostensible for abnormality detecting from the brain. Since the brain structure is quite complicated, MRI image segmentation needs the sophisticated method and precise results [3]. The segmentation of an MRI image describes notable image regions to attain region(s) of interest (ROI's) like as tumors, edema, legions, necrotic tissues, etc. from brain MRI images [4].

In medical diagnosis, most of the doctors rely on MRI images due to its robustness and that is why the accuracy of the prediction algorithms is very important because the result is crucial for patient treatment. Region growing algorithm [3] is used to categorize brain tumor through statistical feature based brain MRI image classification. The results are attained through some predefined criteria such as intensity information and/or edges. Consequently, the connected regions of an image are being extracted. The primary limitation of this method is finding seed points through manual interaction. The principal component analysis (PCA) algorithm with K-means discussed in [4] can be used to define the tumor class on the basis of some correlated pixel of the MRI images. The increased number of features and samples cause more time to consume and increase inaccuracy in results of PCA based K-means algorithm [5].

The classification with K-means clustering is claimed for CT-Liver image. Along with experimental application in brain MRI explained that proper segmentation can be possible with exact thresholding [6]. Again the *K*-value does not exceed greater than 3 if there is any gray level intensity more than that. The research work presented in [7], has introduced a masking algorithm for the classification with the aid of any classifier algorithm. It is effective in finding automatic seed point and neighbors but the dimension of the mask is to change manually for different brain MRI images [7]. In [8]-[10], the Fuzzy C-means algorithm (FCM) was proposed for segmentation. After that, an expert system was introduced with predefined membership and clustered centroid to trace a landmark tissue comparing with a prior model. On the other hand, FCM is described in [8] has limitation due to its noise sensitivity and inadequacy in the detection of abnormality in brain MRI images like a tumor, edema, and cyst. One of the most acceptable and used techniques for brain MRI image classification is Artificial Neural Network (ANN). An ANN technique is discussed in [11] with convincing results. Nonetheless, the procedure of ANN actually requires a perfect pixel classifier, high dimensional training data, and long time to attain the results those are susceptible conditions for the acute patients.

Most of the methods mentioned above are good in some specific point of view like better accuracy but time-consuming or low accuracy with less time consumption. To enhance the overall performance there be requisite of hybridization of those methods in a way that can be able to overcome these limitations.

An advanced algorithm combining SVM and ANN for tumor classification is introduced in this research work. Brain MRI images with normal and abnormal behavior are firstly enhanced through some filter and preprocessing steps. Thereafter, for the detection and classification of the tumor in the brain, proposed segmentation processes namely temper based K-means and modified Fuzzy C-means (TKFCM) clustering algorithm is used. In this technique, the K-values vary from 1 to 8 those are limited to only 1 to 3 in conventional *K*-means and the automatically updated membership function eradicates limitation of FCM. Then, two kinds of features are extracted from these segmented images. One is used to classify the tumor with SVM as it is easy to classify two kinds of dataset in this method and another is used to classify tumor with ANN into five categories along with four malignant stages. Again, the extracted features provide the classification through ANN and the orientation of tumor define the exact position of the tumor in the lobe (i.e. right, left &center) of the brain.

This paper is structured to present that in section 2 conventional K-means and Fuzzy C-means algorithm is presented, the proposed algorithm is described step by step in section 3, the results and discussions are in section 4 and finally, total work is concluded with a few words in section 5.

2. State of the Arts

2.1. K-means clustering

The conventional K-means is discussed with proper explanation in [6] and based on that idea, in this section this method is represented with a slight modification concerning the proposed work. Suppose, a data set $\{x_1, \ldots, x_N\}$ contains N number of observations where x is D-dimensional Euclidian variable. The basic intention of K-means is to divide the data set into N numbers of clusters, where the value of K is given. Naturally, this can be considered that a cluster including a group of data points and their inner side distances are small compared to the outer sided distances

of the cluster. This concept can be formulated by familiarizing a set of D-dimensional vectors c_k, where k=1, …, K, in which c_k is an example related to the k^{th} cluster. Therefore, it is considered that c_k is representing the center of the cluster. Now it is consequence of the previous technique to find an assignment of those data points of clusters and a set of vectors $\{c_k\}$, in such an approach that the summation of the squares of each data point distance to its closest vector c_k could be minimum.

Now, it is suitable to define some symbolization to designate the assignment of data points to the clusters. For each data point x_n, a set of binary indicator variables $b_{nk} \in \{0,1\}$ can be introduced, where, k= 1, …, K represents the K clusters. The data point x_n is assigned to cluster k then b_{nk}=1 for $j \neq k$. This is recognized as the 1-of-k coding scheme. Consequently, definition of an objective function for distortion measurement [6] can be written as,

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} b_{nk} \left\| x_n - c_k \right\|^2 \quad (1)$$

The relation (1) represents the summation of the squares of each data point distance to its allotted vector c_k. The objective is to determine the values of $\{b_{nk}\}$ and $\{c_k\}$ so that the system can minimize J. It is usually determined through an iterative procedure in which each iteration comprises two successive steps related to successive optimizations regarding of the values of $\{b_{nk}\}$ and c_k. At first, some preliminary values of c_k are chosen and then the first phase J is being minimized with respect to the $\{b_{nk}\}$, maintaining the values of c_k, fixed. In second phase, J is to minimize in regard to the c_k, maintaining the values of $\{b_{nk}\}$, fixed. This two-stage optimization process is repeated until convergence. These two stages of updating $\{b_{nk}\}$ and c_k correspond respectively to the E (expectation) and M (maximization) steps of the EM algorithm in [12], and to emphasize this and EM is used the terms E step and M step in the context of the K-means algorithm.

Here, $\{b_{nk}\}$ is considered as first determination because J in (1) is a linear function of $\{b_{nk}\}$. This optimization is generally evaluated to offer a closed form solution. The values of n are independent and that is why this can be optimized for each n, separately. By selecting $\{b_{nk}\}$ as 1 gives the minimum value of $\left\| x_n - c_k \right\|^2$ for whatever the value of k. In other words, it can be merely allotted the n^{th} data point to the neighboring cluster center. More strictly, this can be stated as [12],

$$b_{nk} = \begin{cases} 1 & if\ k = \arg\ \min_a \left\| x_n - c_a \right\|^2,\ a = 1,...,k \\ 0 & otherwise \end{cases} \quad (2)$$

Consequently in this situation, it is to consider that the optimization procedure of the c_k with the values of $\{b_{nk}\}$ is occurred immovable. Here, J is the objective function which is actually quadratic function of c_k, and it is commonly minimized through setting its derivative regarding the values of c_k to be zero and this consideration gives the following mathematical relation given in (3).

$$2\sum_{n=1}^{N} b_{nk} (x_n - c_k) = 0 \quad (3)$$

From (3), the values of c_k can be easily evaluated as,

$$c_k = \frac{\sum_n b_{nk} x_n}{\sum_n b_{nk}} \quad (4)$$

The denominator of (4) is equal to the number of points allotted to the cluster k and subsequently this outcome has a simple explanation, explicitly set c_k equal to the mean of all the data points, x_n those are being assigned to the cluster k. Hence, this technique is recognized as the K-means algorithm.

2.2. Fuzzy c-means

The Fuzzy c-means (FCM) algorithm for image clustering with FORTAN code was first introduced in [14]. In this paper, we have presented FCM algorithm with an improvement of earlier clustering methods which actually followed by the explanation given in [13]. According the approach of this paper, suppose R is the set of real number where R^{P} and R^+ are the set of p tuples of real number and set of nonnegative real number, respectively. Here, W_{cn} is a matrix of order $c \times n$ which is called feature space where feature element, $x \in R^P$ & feature vector $x=(x_1, x_2, \ldots, x_p)$ is consists of p real numbers.

Delineation 1: If X is a subset of R^P and every function $u : X \to [0,1]$ is considered to be assigned to each $x \in X$, its grade of membership should be in the Fuzzy set u. The function u is termed a Fuzzy subset of X. It can be noted that there could be infinite Fuzzy sets related to the set X. It is anticipated to make "partition" X by the means of Fuzzy sets. Normally, it is executed by defining a number of Fuzzy sets on X such that for each $x \in X$. The summation of the Fuzzy memberships of x in the previously considered Fuzzy subsets is one.

Delineation 2: It is given that, a finite set $X \subseteq R^P$, $X = (x_1, x_2, ..., x_p)$, and an integer c $(2 \le c \le n)$ originate a Fuzzy c partition of X that can be represented by a matrix $U \in W_{cn}$ whose entries satisfy the following conditions:

i) The number of row i of U or $U_i = (u_{i1}, u_{i2}, ..., u_{in})$ exhibits the i^{th} membership function of X.

ii) The number of column j of U or $U_i = (u_{1j}, u_{2j}, ..., u_{cj})$ revelations the values of the c membership functions of the j^{th} data in X.

iii) The term, u_{ik} will be construed as u_i (x_k) which actually represents the value of the membership function of the i^{th} Fuzzy subset for the k^{th} data.

iv) The summation of the membership values for each x_k will be always one.

v) No Fuzzy subset will be empty.

vi) No Fuzzy subset will contain all elements of X.

M_{fc} denotes the set of the partitions of X in case of Fuzzy c. Here, the distinctive subset $M_c \subseteq M_{fc}$ of X in every u_{ik} is 0 or 1. In addition, the subset is the discrete set of non-Fuzzy c partitions of X. The solution space, M is for the conventional clustering algorithms. The Fuzzy c-means algorithm followed by this proposed work uses the iterative optimization method in order to approximating an objective function minimization which measures similarity on $R^P \times R^P$.

Delineation 3: Suppose that, $U \in W_{fc}$ is a Fuzzy c partition of X, and consider that v is the c tuple $(v_1, v_2, \ldots, v_c)$, $v_i \in R^P$. Therefore, $J_m : M_{fcd} \in R^P \rightarrow R^+$ is described by the following relation:

$$J_m = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{jk})^m (d_{ik})^2 \; ; \quad v = (v_1, v_2, ..., v_c) \in R^{cP} \tag{5}$$

Additionally, $v_i \in R^P$ is considered to be the cluster center or prototype of class i, $1 \leq i \leq c$, and consequently,

$$d_{ik}^2 = \|x_k - v_i\|^2 \tag{6}$$

Here, $\|.\|^2$ represents any inner product norm metric that defines the Euclidian distance [6], and $m \in [1, \infty]$. This distance calculates the distance from cluster centroid to each object. If we take the Euclidean distance with the distance matrix at null iteration, we get following relationship:

$$D = \begin{bmatrix} \alpha_1 & \beta_1 & \chi_1 & \delta_1 \\ \alpha_2 & \beta_2 & \chi_2 & \delta_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{7}$$

Each column in D symbolizes the object. The first row of D relates to the distance of the each object to the 1st centroid and similarly, the second row of D is the distance of the each object to the 2nd centroid. As for example, distance from cluster center c= (x, y) to the first centroid $c_1 = (\alpha_1, \alpha_2)$ is $(x- \alpha_1)^2 + (y- \alpha_2)^2$, and its distance to the second centroid $c_2 = (\beta_1, \beta_2)$ is $(x- \beta_1)^2 + (y- \beta_2)^2$, and so on.

The FCM algorithm by the iterative optimization method produces a J_m Fuzzy C-partition of the data set $X=\{x_1, \ldots, x_n\}$. The basic steps to implement the algorithm can be represented by the following consecutive mathematical phases (followed the explanation given in [13]).

1) Fix the cluster numbers c, $2 \leq c \leq n$ where, n is the number of data items. Besides, it is to fix, m $(1 < m < \infty)$. Choose any inner product induced norm metric $\|.\|$ e.g., $A \in W_{pp}$.

2) Initialize the Fuzzy c partition $U^{(0)}$,

3) At step b, b=0, 1, 2, ...

4) Compute the c cluster for the centers $\{v_i^{(b)}\}$ with $U^{(b)}$ and the formula for the i^{th} cluster center is usually consider as the following relation:

$$v_{il} = \frac{\sum_{k=1}^{n} (u_{ik})^m x_{kl}}{\sum_{k=1}^{n} (u_{ik})^m}, \qquad (l = 1, 2, \ldots, p) \tag{8}$$

5) Bring up-to-date, $U^{(b)}$: determine the memberships in $U^{(b+1)}$ as the following steps a) & b). For k = 1, 2, 3, …, n.

a) Evaluate, I_k and $\tilde{I}_k$:

$$I_k = \{i \mid 1 \leq i \leq c,\ d_{ik} = \| x_k - v_i \| = 0\},$$
$$\tilde{I}_k = \{1, 2, ..., c\} - I_k, \tag{9}$$

b) For data item k, determine the new membership values so that,

i) if $I_k = 1$

$$u_{ik} = \left[\sum_{j=1}^{C} \left\{ \frac{d(x_k, v_i)}{d(x_j, v_k)} \right\}^{\frac{-2}{m-1}} \right] \tag{10}$$

ii) else, $u_{ik} = 0$ for all $i \in I_k$ and $\sum_{i \in I_k} u_{ik} = 1$, next k.

6) Compare $U^{(b)}$ and $U^{(b+1)}$ in a convenient matrix norm; if $\left\| U^{(b)} - U^{(b+1)} \right\| \leq \varepsilon$, (where $\varepsilon = \{0 \; to \; 1\}$) stop; otherwise, set b=b+1, and go to step 4.

The use of FCM algorithm necessitates the determination of several parameters, *i.e.*, $c \in m$, the inner product norm $\|.\|$, and a matrix norm. In addition to that, the set $U^{(0)}$ of initial cluster centers should have to be defined for sure. Although no necessary rules for choosing a good value of m are available in the literature. In most of the cases, the value of m is typically reported as the useful range of values as $1 \leq m \leq 5$. The objective of the algorithm mentioned above is to reduce the computational burden imposed by iterative looping between (9) and (10) when c, p, and n are large.

3. Methodology

3.1. MRI image collection

The data of brain MRI images are collected from internet public repository. The images of normal and tumorous brain with Lower Grade Glioma or Glioblastoma Multiforme are collected from the sources [15]-[17]. The number of the used data for each MRI image for classification of normal and tumor brain through SVM is of 39 images. There are 37 images for the classification of benign and malignant tumor stages.

3.2. Image processing

The enhanced images are achieved from the raw MRI images through some steps described as follow.

- **Image conversion and orientation setting**: The images MRI images are converted from *.mha* format and *.dicom* format into *.jpg* by using MATLAB conversion tools application. On the consequence of the conversion, the sizes and directions are reset. This step is conducted in order to have the same size and direction for all the MRI images. This process is performed automatically by using MATLAB with 256×256 pixels for the betterment of image usage.
- **Image enhancement:** The transformed images from the previous step, at first they are converted to L*a*b* images for the comparatively better view and quality. That is why the values of the luminosity of the images can be spanned with a range from 0 to 100 which should be scaled to [0 1] range (appropriate for MATLAB intensity images of class double) before applying the three contrast enhancement techniques like adjusted, adaptive thresholded, and histogram imaging. In this method, for smoothing the images hybridization of both weiner2 (image, [40, 40]) and median2 filter is assured and acquired good results.

3.3. Proposed TKFCM algorithm for tumor detection

In this segmentation process, K-means algorithm is used to segment MRI images on the basis of gray level. This gray level is selected depending on the temper of the image. Then the modified Fuzzy c-means algorithm which depends on the updated membership is applied to segment the temper based K-means segmented image. The membership of modified Fuzzy c-means is updated with the cluster distances from centroid defined by the features of the tumor MRI image. The TKFCM algorithm is the combination of the K-means algorithm and Fuzzy c-means algorithm with some important modifications. The temper is added in the proposed approach along with the conventional K-means algorithm which is identified by the temper or gray level intensity in the brain MRI images. Besides, the Fuzzy c-means membership and Euclidian distance are also modified by the image features.

Here, the coarse image $B(x_i, y_i)$ which is marked and describing the desired tempers for the *K*-means could be found through convolution of gray level based temper and image given as,

$$B(x_i, y_j) = \sum_{i=n+1}^{M+n} \sum_{j=n+1}^{N+n} P(x_i, y_j) \oplus T_{MN(resize)} \quad (11)$$

Temper based window is selected by T_{MN} that is calculated as,

$$T_{MN} = \sum_{i=n+1}^{M-n} \sum_{j=n+1}^{N-n} P(x_i, y_j) \quad (12)$$

In (12), there is presented a temper based matrix of image with a number of gray level intensities, *G* and number of bins, *S* those are used to detect the temper of the images P(x_i, y_j). Where *n* is defined as *n*= (*window Size*-1)/2. With exact value of the temper, row and column, the desired temper is obtained.

Separately temper based *K*-means and modified Fuzzy *c*-means clustering algorithm for segmentation can be written in equation as below:

$$J_k = \sum_{i=1}^{C} \sum_{j=1}^{K} B(x_i, y_j) \left\| x_i - c_j \right\|^2 \quad (13)$$

$$J_m = \sum_{j=1}^{N} \sum_{i=1}^{K} (U_{ij})^m (d_{ij})^2 \quad (14)$$

Here, *M* and *N* are the row and column of the binary coarse image matrix $B(x_i, y_j)$. The number of data points in clusters, centroid of the cluster, and number of clusters are defined by C, *N*, and *K*, respectively.

Then, from (13) and (14) the contour through which the exact location of tumor portion in any image can be find is

$$J_{km} = \oint_c \left(J_k, J_m \right) \quad (15)$$

The relation given in (15) shows a contour integral of temper based *K*-means image and updated membership based FCM image, where *c* is the contour value. The whole method that has been proposed for tumor detection from brain MRI image using Temper based K-means and modified fuzzy C-means is described by the flow chart given in Figure 2.

3.4. Feature extraction

The system extracts the first and second order statistical features as in [18]. The first order statistical features like contrast, correlation, entropy, energy, and Homogeneity are used to detect exact tumor and its position in the brain MRI image through SVM. On the other hand, second order region based statistical features provide area, eccentricity, and the perimeter is required for distinguishing the malignant and benign tumor. These second order feature values are used as the input of the ANN and provide desired tumor categories. The feature extraction procedure is mentioned step by step in Figure 2.

3.5. Methodology of combining SVM and ANN

In this proposed work, the combination of SVM and ANN is used to classify the tumor and its stages. The hybridization of the method is used firstly to classify normal and tumor dataset using Linear SVM Kernel and then classify the tumor data into different stages through ANN. In SVM there will be a hyperplane between the set of data points as the decision boundary. In this case, there are two classified data of normal and abnormal (Tumor) brain images and the hyper planes of SVMs are used to separate these classified data as normal data and tumor data. The input towards the ANN is the information of feature extraction; the first and second order statistic features. Then, the ANN will generate the output results as of benign and malignant I-IV tumors. The process of using ANN to attain the proposed goal of this research work is explained by the flow diagram presented in Figure 3.

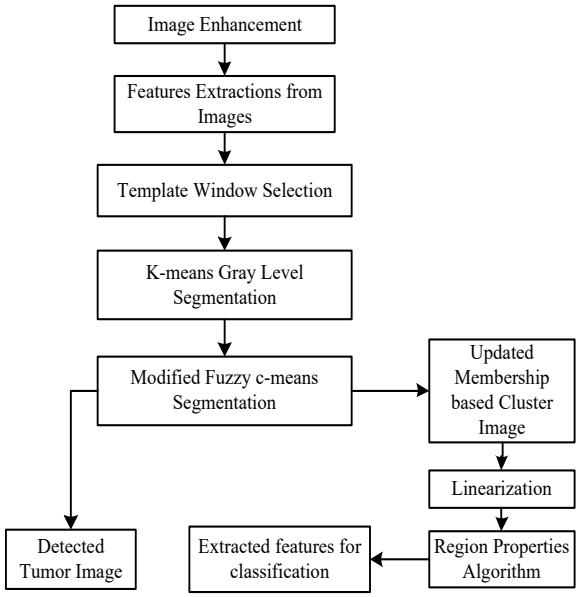


Figure 2: Methodology steps for proposed TKFCM algorithm in tumor detection

4. Results and Discussions

To implement the proposed methodology, first of all, a database of 46 multifaceted brain tumor images is created. It is aforementioned that, the images are collected from the sources of [15]-[17]. To make the quality of the images acceptable for the proposed methodology, some necessary steps were to do and the database was set for the network. To do so for the previous consequence, a thresholding method is applied with threshold level 0.8 and the morphological operation was performed. In addition to

that median and hybrid filters were used to remove the primary noises from the MRI images. The filtered MRI images are presented in Figure 4 (a) & 4(d).

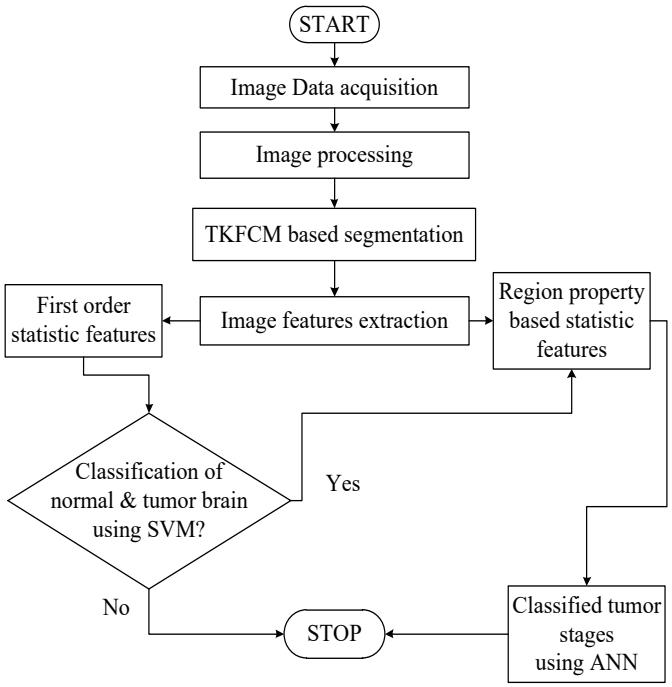


Figure 3: The methodology with their step by step description to implement the proposed hybridization of SVM & ANN algorithms in the filed tumor classification.

The preprocessing steps include filtration, increase of luminosity, and adjust of contrast. After the previous processing, all the images are segmented by the proposed TKFCM based segmentation process. The effects of this segmentation process on images are shown in Figure 4(b) & 4(e) and 4(c) & (f). Temper based K-means (TK) segmentation method shows 8-gray level intensity-based images in Figure 4(b) & 4(e). These figures describe that the input image is segmented by the combined effects of TK-segmentation and the modified approach of FCM algorithm.

The updated membership function with proper Euclidian distance for modified FCM represents the detected tumors in Figure 4(c) & 4(f). These are some examples of the optimum result that can differ this modified and hybrid TKFCM method from the conventional methods. The extracted tumors according to the proposed technique are marked with red color in Figure 4(c) & 4(f). Therefore, these segmented areas can be used to determine the region property based statistical features from the images.

In Figure 5, we presented ten images (given in Figure 5 (a) & (d)) as the input of classification by the proposed TKFCM scheme. These ten images are taken from the created database and used for the classification of the tumor and its area. The enhanced input images for TKFCM (shown in Figure 5(a) & 5(d)) are preprocessed by the steps described by the visual explanation in Figure 4. In Figure 5 (b) & 5(e), the detected brain tumor images marked as the red color those are the outcomes of the previous images by the application of the TKFCM method. Classified brain tumors through linearization of TKFCM are presented in Figure 5 (c) & 5(f). From which the region based features are obtained by the level thresholding, updated membership function, and region properties algorithm.

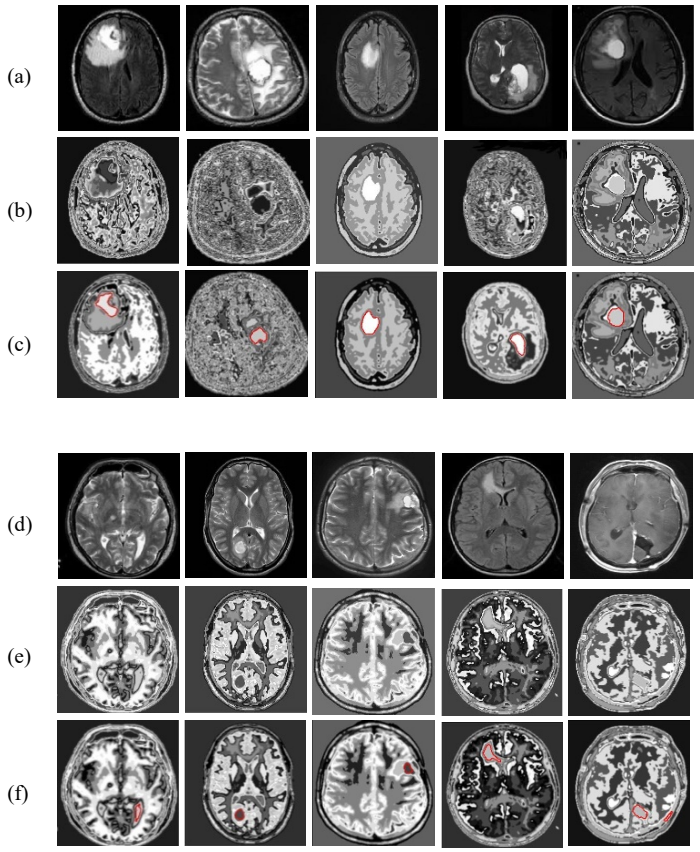


Figure 4: (a) & (d): Filtered images for TKFCM, 4(b & e): Temper based segmented images, 4(c & f): TKFCM based detected tumor images.

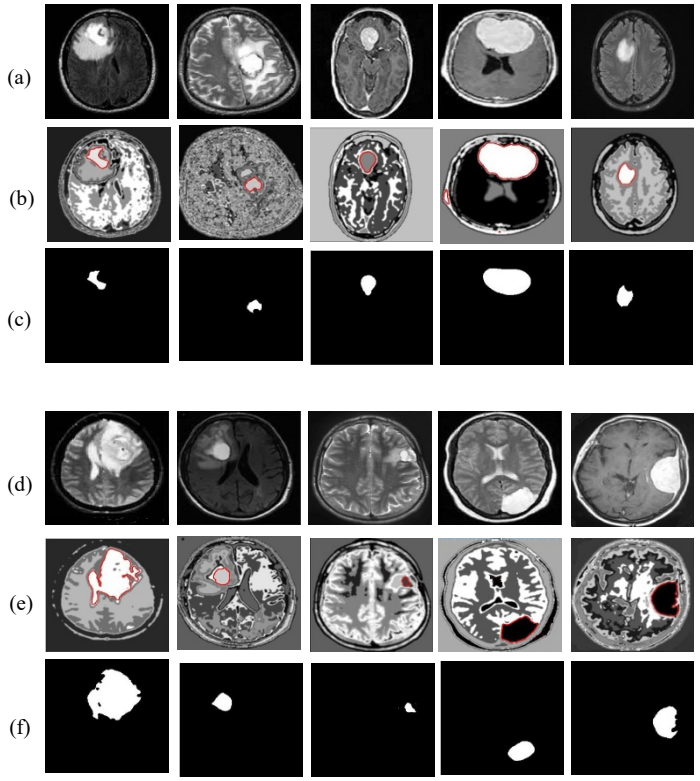


Figure 5: Enhanced Input images for TKFCM in 5(a & d), TKFCM based red marked tumor images in 5(b & e), and classified brain tumor through linearization of TKFCM in 5(c & f).

From the segmented image, two kinds of features are extracted: one is statistical features and other is region property based features. The first kind of statistical features *i.e.*, contrast, correlation, entropy, homogeneity, energy are used to classify whether there is any normal brain or tumorous brain using SVM. Therefore, SVM classifies the images into two categories- tumorous and normal. For this significance, two major kinds of MRI image data of normal and abnormal brain are fed to SVM network and the corresponding results of classification property of the SVM are presented in Figure 6(a). There are 46 images used to classify whether the system classified 37 tumor images, 8 normal images and rest 1 is misclassified. As a result, the accuracy of SVM method is 97.44% which is very convincing to move to the next step c.

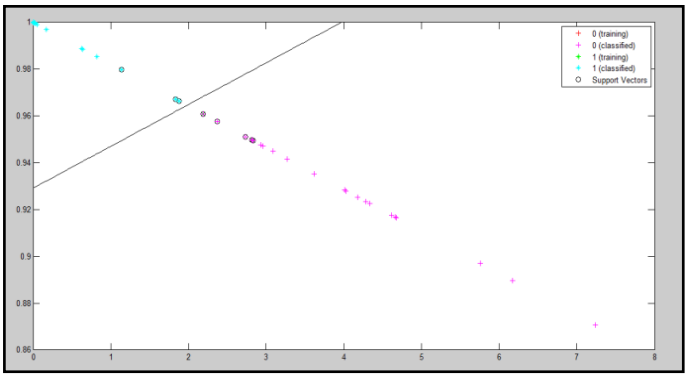
According to the previous approach the tumorous brains have to be classified into 5 defined categories. Due to implementing this classifying network, an ANN-based network was designed as the configuration given in Figure 6(b). The network consists of four input vector layers, one hidden layer with 129 neurons, and five output vector layers. The input feature vectors are evaluated as the second kind of features of MRI images which are acknowledged in this article by the region property based features.

The performances of training, validation, and testing of the proposed network are shown in Figure 6(c). The specifications of the achieved performances were 60 iterations with 0.05 increment order. Additionally, the minimum error was considered up to 0.5e-02, the gradient minima were approximately 1e-10, and the maximum validation check failure was taken 6. From the Figure 6(c), it is found that the desired performances are achieved between 8 and 9 iterations which indicate the less required time for the network compared to the 60 iterations. Best validation performance is 0.17479 at iteration 2 mentioned in Figure 6(c).

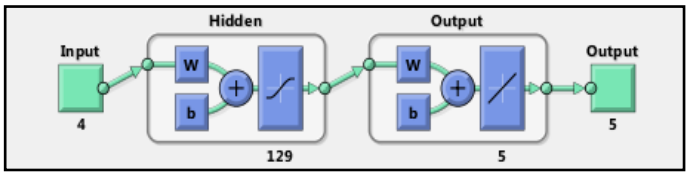
The second kind of features *i.e.*, area, eccentricity, perimeter, orientation are used to differentiate the tumor and different malignant stages of I to IV. The obtained features are denoted as input vectors of the proposed network. On this contrary, we acknowledged in this article that we have five categories of brain tumor those are I-IV malignant groups along with benign. Therefore, the results are given by the confusion matrix in Figure 6(d) are the output vectors or classified 5 groups. The network achieved 97.3% classifying accuracy given in Figure 6(d). This result is very convincing where it is found that there are 9 benign, 17 malignant I, 6 malignant II, 3 malignant III, and 1 malignant IV stages. The network misclassified 1 tumorous brain MRI images.

In Table 1, there are some feature values for classifying tumor categories as well as detecting exact tumor position in our brain lobe. The region-based features are used as input vector of ANN to classify the malignant and benign tumors. Beside one of the features mentioned as 'orientation' of the tumor defines the exact lobe position of the tumor in brain MRI images. These results of orientation by the proposed method provide the information about the position of the tumor either in left or right or center part of the brain.

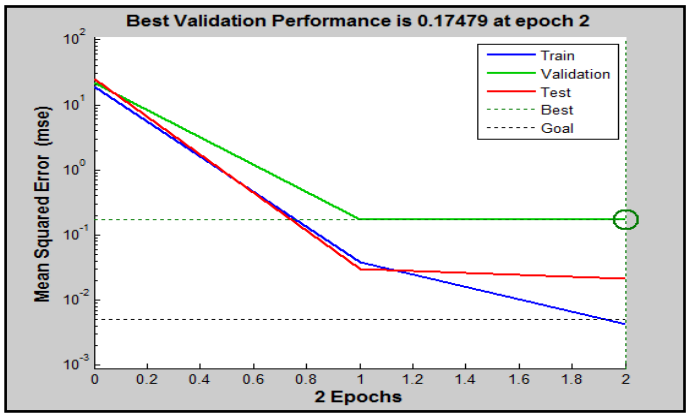
Since the classification accuracy is not the only performance metric, some other performances of the classifier are necessary to be evaluated. Based on the value of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) of the proposed network, some other performances like sensitivity, specificity, beat error rate (BER) can be calculated. In this step,



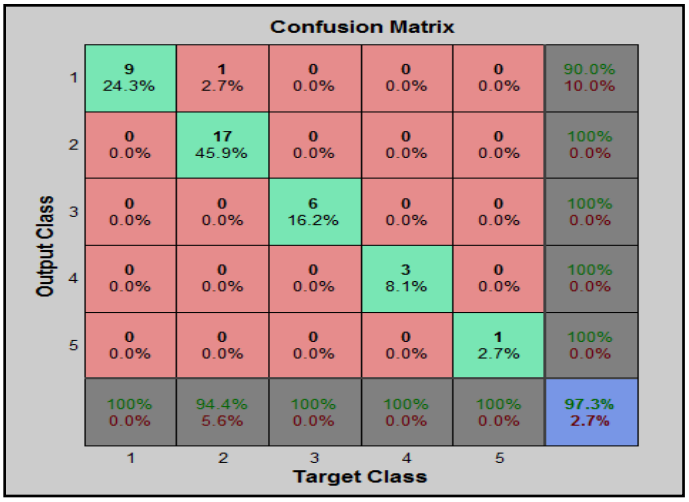
(a)



(b)



(c)



(d)

Figure 6: The relevant simulated results of proposed hybridized SVM and ANN algorithm for classifying tumor categories. 6(a): 2D plot of classified normal and tumor brain data for proposed SVM classifier, 6(b): The network architecture of proposed ANN method, 6(c): The performance curves for the proposed ANN method, and 6(d): The confusion matrix for benign and malignant stages I-IV using proposed ANN classifier for second order region based features.

Table 1: Some examples of classified tumor and identification tumor position using ANN approach

Area in mm^2	Eccentricity in mm	Perimeter in mm	Orientation in degree	Tumor type	Tumor position in lobe
9.9167	0.9754	60.275424	-84.1794	Malignant II	Left
6.472	0.9213	33.55836	-100.0502	Malignant I	Left
8.2561	0.8402	40.736784	-54.547	Malignant I	Left
9.5952	0.9414	36.863376	-57.7349	Malignant I	Left
8.6719	0.7084	31.913904	0.404	Malignant II	Centered
10.2451	0.6974	41.258448	-32.196	Malignant II	Left
21.5625	0.793	82.645728	0.6467	Malignant III	Centered
9.7787	0.6381	37.249872	-88.0731	Malignant II	Left
8.1541	0.8617	32.143056	69.189	Malignant I	Right
7.0295	0.7793	25.955424	14.0189	Malignant I	Right
7.9814	0.9886	18.599592	85.0421	Malignant I	Right
22.4633	0.6164	135.3689	60.2044	Malignant III	Right
8.0379	0.4214	28.77864	-59.6895	Malignant II	Left
8.7599	0.8547	36.848064	1.3184	Malignant I	Right
24.4937	0.8675	135.78576	-77.2517	Malignant III	Left
7.7914	0.7926	36.24984	-49.5862	Malignant I	Left
13.3835	0.8289	49.570752	12.2018	Malignant I	Right
15.8158	0.8813	75.392592	44.3799	Malignant II	Right
11.0597	0.625	89.080464	-55.4867	Malignant II	Left
36.9817	0.4032	141.77	-82.9364	Malignant IV	Left
15.7074	0.671	58.995552	89.6919	Malignant III	Right
12.3489	0.8206	58.698816	-82.3899	Malignant II	Left

Table 2: Comparison between proposed algorithm and conventional methods

Algorithms	Sensitivity (%)	Specificity (%)	Accuracy (%)	BER	Computational Time
Thresholding	85	80	81.3	0.175	~3 min
Region Growing	88.46	75	86.47	0.182	~6 min
ANN	95.42	100	95.07	0.022	~8 min
FCM	86.95	85.7	86.4	0.136	~5 min
SVM	96.2	66.67	90.44	0.0234	~4 min
K-means	75	92.85	83.7	0.160	~160-170 sec
TKFCM	88.9	100	91.89	0.055	~100 sec
Fuzzy Logic Method	96.3	100	96.667	0.018	~ 120 sec
Proposed SVM+ANN Method	**98**	**100**	**97.37**	**0.0294**	**~ 2 min**

*Image size=256x256, Software=MATLAB2014a, Processor= Core2duo, RAM=2GB, windows=7

these important estimations are performed by the following relations (16)-(19) and used for comparison purposed with other conventional methods [19].

$$Sensitivity = \frac{TP}{(TP+FN)} \times 100 \quad (16)$$

$$Specificity = \frac{TN}{(TN+FP)} \times 100 \quad (17)$$

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FN+FP)} \times 100 \quad (18)$$

$$BER = \frac{1}{2}\left(\frac{(FN)}{(TP+FN)} + \frac{(FP)}{(FP+TN)}\right) \quad (19)$$

Comparisons of the proposed method with predictable techniques whose algorithms are coded to cope up with the same dataset are mentioned in Table 2. In this proposed technique, 37 brain tumor images are taken for the simulation. The sensitivity of 98%, specificity of 100%, the accuracy of 97.37%, and BER 0.0294, which are obtained by combining ANN and SVM. The results show that the proposed method is better than the conventional methods like thresholding, region growing, SVM, ANN, FCM, TKFCM, and K-means. The proposed algorithm that is actually combinations of several suitable techniques proves more effective than any other single method. Though the conventional ANN comprehends the value of accuracy of 95.07%, the computational time is so high. But the proposed technique shows less computational time with better accuracy than conventional ANN. So, it can be useful for both detecting tumors and be classifying tumor stages of brain MRI images for experts.

5. Conclusion

This work classified the images as normal and tumorous. In addition, the size of the tumor and its position in brain lobe were also identified by the proposed method from MRI image. To implement such an efficient and intelligent algorithm, a number of statistical and machine learning based algorithms like temper based *K*-means and modified Fuzzy *c*-means clustering, SVM, and ANN are combined. As a result, it is found that the previous limitations like less classifying accuracy, computational time requirement, unsatisfactory of BER, sensitivity, and specificity of conventional ideas has been overcome by this proposed algorithm. Other existing methods have some tradeoff among the performances. This method provides sensitivity 98%, specificity 100%, classifying accuracy 97.37%, BER=0.0294, and required less than 2 minutes to give the result. These results are too convincing to identify the brain tumor and its size, as well. In addition, this algorithm has provided the tumor position accurately. Considering all the performances, it can be concluded that the proposed algorithm is better than others like region growing, thresholding, and FCM compared to every parameter. On the other hand, in point of accuracy and

computational time the proposed method is better than ANN, SVM, and TKFCM. Eventually, we hope that this method can be very helpful for diagnosis of brain tumor.

Conflict of Interest

So far the knowledge, the authors declare no conflict of interest regarding this article.

Acknowledgment

Authors would like to thank Prof. Dr. Md. Abdur Rafiq, Dean, Faculty of Electrical and Electronic Engineering, Khulna University of Engineering & Technology (KUET), Khulna, Bangladesh, for his beautiful lectures on neural network- fuzzy logic and guidelines to implement it in this research work.

References

[1] R. Ahmmed, A. S. Swakshar, M. F. Hossain, and M. A. Rafiq, "Classification of tumors and it stages in brain MRI using support vector machine and artificial neural network," International Conference on Electrical, Computer and Communication Engineering (ECCE-2017), 229-234, February 16-18, Cox's Bazar, Bangladesh. http://doi.org/ 10.1109/ECACE.2017.7912909.

[2] R. Rana, H. S. Bhadauria, and A. Singh, "Study of various methods for brain tumor segmentation from MRI images," International Journal of Emerging Technology and Advanced Engineering, **3(**2), 338-342, 2013.

[3] J. Joshi and A. Padhke, "Feature Extraction and Texture Classification in MRI," International Journal of Computer & Communication Technology (IJCCT), **2**, 130-136, 2010.

[4] A. Ahirwar, "Study of techniques used for medical image segmentation and computation of statistical test for region classification of brain MRI," International Journal on Information Technology and Computer Science, **5**, 44-53, April 2013. doi:10.5815/ijitcs.2013.05.06

[5] I. E. Kaya, A. Ç. Pehlivanli, E. G. Sekizkardes and T. Ibrikci, "PCA based clustering for brain tumor segmentation of T1W MRI Images" Computer Methods and Program in Biomedicines in Elsivier, **140**, 19-28, March, 2017. http://doi.org/ 10.1016/j.cmpb.2016.11.011

[6] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," Annual review on Biomedical Engineering, **2**, 315-337, 2000.

[7] W. Narkbuakaew, H. Nagahashi, K. Aoki, and Y. Kubota, "Integration of modified K-means clustering and morphological operations for multi-organ segmentation in CT liver-images," Recent Advances in Biomedical & Chemical Engineering and Materials Science, 34-39, March 2014.

[8] G. Deng, "A generalized unsharp masking algorithm," IEEE Transactions on Image Processing, **20(**5), 1249-1261, May 2011. http://doi.org/ 10.1109/TIP.2010.2092441

[9] S. R. Kannana, S. Ramathilagam, R. Devi, and E. Hines, "Strong fuzzy c-means in medical image data analysis," Journal of Systems and Software, **85**, 2425–2438, December 2011. http://doi.org/10.1155/2013/972970

[10] V. Harati, R. Khayati, and A. R. Farzan, "Fully automated tumor segmentation based on improved fuzzy connectedness algorithm in brain MR images," Computers in Biology and Medicine, **41**, 483–492, April 2011. http://doi.org/10.1186/s12938-016-0165-2

[11] H. R. Shally, and K. Chitharanjan, "Tumor volume calculation of brain from MRI slices," International Journal of Computer Science & Engineering Technology (IJCSET), **4**(8), 1126-1132, 2013.

[12] X. Descombes, F. Kruggel, G. Wollny, and H. J. Gertz, "An object-based approach for detecting small brain lesions: Application to Virchow-robin spaces," IEEE Transaction on Medical Imaging, **23**(2), 246–255, 2004. http://doi.org/10.1109/TMI.2003.823061

[13] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering amgorithm," Computer & Geosciences, **10(**2-3), 191-203, 1984. https://doi.org/10.1016/0098-3004(84)90020-7

[14] M. Gong, Y. Liang, J. Shi, W. Ma and J. Ma, "Fuzzy c-means clustering with local information and kernel metric for image segmentation," IEEE Transactions On Image Processing, **22**(2), 573-584, February 2013. http://doi.org/10.1109/TIP.2012.2219547

[15] (2017) PE- Brain tumor website. [Online]. Available: http:// www.mayfieldclinic.com /

[16] University of Maryland Medical Center (2017) homepage. Primary brain tumor [Online]. Available: http://umm.edu/health/medical/reports/articles/brain-tumors-primary

[17] Oncology (2017) homepage. Tumor type [Online]. Available: http://www.oncolink.org/types/article.cfm?c=52&aid=247&id=9534

[18] M. Lugina, N. D. Retno, and R. Rita, "Brain tumor detection and classification in magnetic resonance imaging (MRI) using region growing, fuzzy symmetric measure, and artificial neural network back propagation", International Journal on ICT, **1**, 20-28, December 2015.

[19] J. Selvakumar, A. Lakshmi, and T. Arivoli, "Brain tumor segmentation and its area calculation in brain MR images using K-mean clustering and fuzzy c-mean algorithm," International Conference on Advances in Engineering, Science and Management, Tamil Nadu, 186-190, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 49-55 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Flexible Aperture Tuning Solution for Cellular Main Antenna in Metallic Back Cover Mobile Phone

Yew Choon Mark Tan*, Guan Hong Ng, Yew Siow Roger Tay

AAC Technologies Pte. Ltd., Singapore

ARTICLE INFO

Article history:
Received: 06 September, 2017
Accepted: 11 October, 2017
Online: 19 October, 2017

Keywords:
Metallic Back Cover
Tunable Antenna
Aperture Tuning

ABSTRACT

Metal housing has been used extensively on portable communication devices such as on mobile phones and tablets. The choice of metal housing ranges from metallic rim to metallic back cover. This metal housing tends to improve the outlook appearance of the mobile devices, and add mechanical strength towards the mobile devices. However, from the aspect of the communication antenna, the metal housing often posts great challenges towards the flexibility in antenna design and reduction in antenna performance. This paper presents an approach to overcome the challenges by integrating the metal housing of the mobile phone as part of the antenna, along with the introduction of tunable antenna concept to provide different forms of Aperture Tuning to the Cellular Main Antenna, to satisfy its wide frequency band coverages for the 2nd, 3rd and 4th Generation (2G, 3G and 4G) mobile network.

1. Introduction

Over the years, there had been a popular trend for mobile phones to be equipped with metal housing. The metal housing often comes in different forms, which examples are as shown in Figure 1. One form of metal housing is the Closed Metal Ring as shown in Figure 1(a), with a metal ring running around the edges of the mobile phone. Another type of metal housing, quite similar to the Closed Metal Ring, is the Broken Metal Ring (with open slot) as shown in Figure 1(b). In the Broken Metal Ring, there are slots cutting on the metal ring, dividing the metal ring into two or more sections. In recent years, there is another type of metal housing that has been becoming popular. This metal housing comes in the form of a Metallic Back Cover as shown in Figure 1(c). The use of metal housing on mobile phone tends to improve the outlook appearance of the mobile phone. Apart from that, the robustness of the metal housing tends to add mechanical strength to the mobile phone. However, the use of metal housing on mobile phone often post disadvantages towards the mobile phone antenna. These disadvantages include introducing unwanted coupling towards the antenna, creating a shielding effect over the antenna hence reducing the sensitivity of the antenna etc. Additionally, maintaining the aesthetics of the metal housing on the mobile phone comes with an antenna design flexibility cost.



(a) (b) (c)

Figure 1. Different forms of metal housing: (a) Closed Metal Ring; (b) Broken Metal Ring (with open slot); (c) Metallic Back Cover;

This paper serves as an extension of work originally presented in 2017 11th European Conference on Antennas and Propagation (EuCAP2017) [1]. This paper presents a flexible Aperture Tuning solution [2 – 4] for the cellular Main Antenna in a Metallic Back Cover mobile phone, to provide coverage for multi frequency bands in the frequency range of 700MHz to 960MHz (LTE700, GSM850, GSM900), 1710MHz to 2170MHz (DCS1800, PCS1900, WCDMA2100) and 2300MHz to 2690MHz (LTE2300, LTE2600). The flexible Aperture Tuning solution as proposed in this paper aims to overcome the disadvantages post by the metal housing as mentioned above. In the earlier paper [1], the flexible Aperture Tuning solution for the cellular Main Antenna has been briefly introduced, along with the Free-Space passive measurement data of the Main Antenna's Reflection Coefficient S11 and Total Efficiency. This paper will further elaborates on the concept of the different Aperture Tuning solutions that are deployed at different parts of the Main Antenna for tuning the antenna's Low (700 – 960MHz) Band, Mid (1710 – 2170MHz)

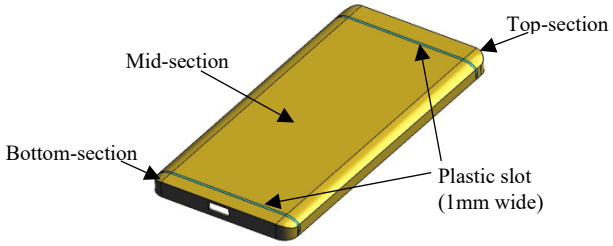
*Corresponding Author: Yew Choon Mark TAN, 85 Science Park Drive, #02-08 The Cavendish, Singapore 118259, (+65) 91767574 & marktan@aactechnologies.com

Band and High (2300 – 2690MHz) Band. Besides the Free-Space passive measurement data, this paper also presents the active measurement data (Total Radiation Power TRP) taken with the mobile phone mockup in the Free-Space and Phantom Head & Hands conditions.

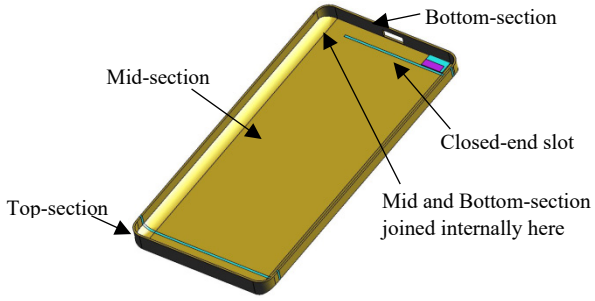
2. Metallic Back Cover Mobile Phone Mockup

2.1. Metallic Back Cover

The metal housing for the mobile phone presented in this paper takes the form of a Metallic Back Cover, as mentioned in Figure 1(c). Figure 2 shows pictures illustrating the exterior and interior appearance of the Metallic Back Cover. From the exterior appearance shown in Figure 2(a), the Metallic Back Cover seems to be divided into three sections, namely Top-section, Mid-section and Bottom-section. In between each section, there is a plastic slot (at 1mm wide) running across the Metallic Back Cover. However, from the interior appearance shown in Figure 2(b), it can be seen that the Mid-section and Bottom-section of the Metallic Back Cover are actually joined together internally. This makes the slot at the Bottom-section more like a Closed-end slot instead of a through slot as seen at the Top-section. The overall dimension of the Metallic Back Cover is 144mm by 70mm by 6.5mm, which is suitable for mobile phone with 5.5inch display.



(a) Exterior appearance



(b) Interior appearance

Figure 2. Metallic Back Cover

In this proposed solution, the Main Antenna is located at the bottom section of the mobile phone. The entire Bottom-section metallic cover and partial of the Mid-section metallic cover are integrated as part of the Main Antenna design. Figure 3 shows pictures mainly highlighting the interior appearance at the bottom section of the Metallic Back Cover. There is a Capacitive Feed Patch located near the bottom left corner of the Metallic Back Cover. This Capacitive Feed Patch is a thin layer of conductive rectangular patch, separated from the inner surface of the Metallic Back Cover by a thin layer of plastic substrate with thickness of 0.5mm. The Capacitive Feed Patch is positioned on the Bottom-section side of the Metallic Back Cover in a way such that the edge of the Capacitive Feed Patch will align well with the edge of the Closed-end slot, as can be seen in Figure 3. Figure 3 also highlights the installation of a Tuner Board on the inner surface of the Metallic Back Cover. The Tuner Board is a tiny piece of Printed Circuit Board (PCB) that is used to house a Tunable Capacitor TC1. The Tuner Board is fully grounded onto the Mid-section of the Metallic Back Cover, positioned just by the edge of the Closed-end slot. The output of the Tunable Capacitor TC1 is bridged across the slot and get into direct contact with the Bottom-section of the Metallic Back Cover.

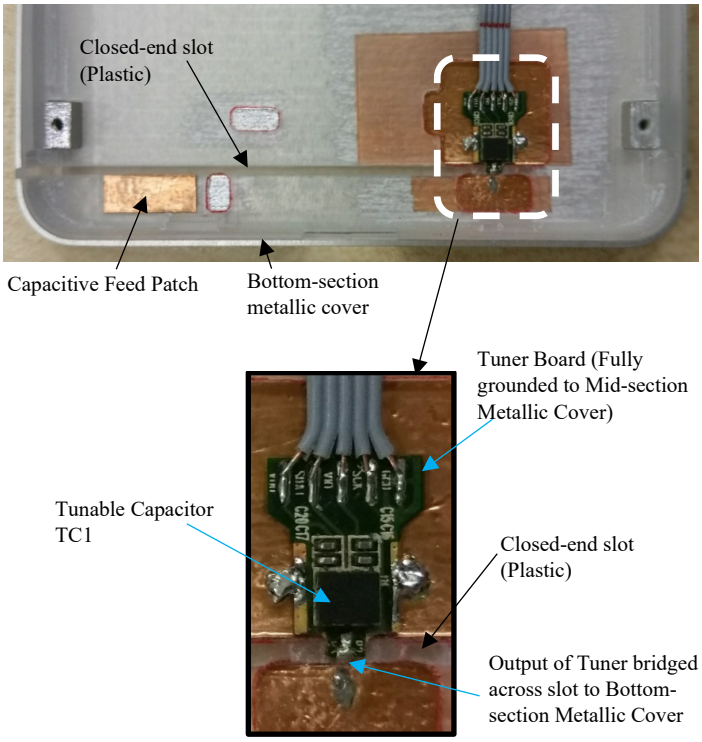


Figure 3. Capacitive Feed Patch and Tuner Board on Metallic Back Cover

2.2. Main Antenna Printed Circuit Board (PCB)

Figure 4 shows partial of the Main Antenna Printed Circuit Board (PCB) used in the mobile phone mockup, mainly highlighting the environment in the Main Antenna region at the bottom section of the mobile phone. The PCB is equipped with a Type-C Micro-USB and a speaker which are components that are commonly found at the bottom section of mobile phone. On the PCB, there is a Feed-Pin that is traced to the source of the Main Antenna. The Feed-Pin can be in the form of a pogo-pin or spring contact to establish a direct connection with the Capacitive Feed Patch at the location (red-dot) as highlighted in Figure 4. Located next to the Feed-Pin is Gnd-Pin1, which is traced to a Tunable LC Resonant Circuitry, with reference to the common ground. This Tunable LC Resonant Circuitry is made up of a shunt inductor in parallel with a shunt Tunable Capacitor TC2, to provide Aperture Tuning to the Main Antenna. The Gnd-Pin1 can be in the form of a pogo-pin or spring contact to make direct contact with the Bottom-section metallic cover at the location (blue-dot) as indicated in Figure 4. Located above Gnd-Pin1 on the PCB is Gnd-Pin2 which is traced to a standalone Tunable Capacitor TC3 connecting to the reference ground. This standalone Tunable Capacitor is meant for providing Aperture Tuning to the Main Antenna. The Gnd-Pin2 makes direct connection with the Mid-

section of metallic cover at the location (purple-dot) as indicated in Figure 4. To the left of Gnd-Pin2, located near the speaker is Gnd-Pin3 which is connected directly to the reference ground. Gnd-Pin3 on the PCB is meant to make direct connection with the Mid-section of metallic cover at the location (green-dot) next to the Tuner Board as shown in Figure 4. This arrangement of the grounding connections between the respective ground pins (Gnd-Pin1, Gnd-Pin2 and Gnd-Pin3) on PCB and the Metallic Back Cover has integrated the Metallic Back Cover as part of the Main Antenna design. The three Tunable Capacitors (TC1, TC2 and TC3) as mentioned above on the Metallic Back Cover and Main Antenna PCB are Micro-Electro-Mechanical Systems (MEMS) based variable capacitors. Each capacitor has 16 different states of capacitance value ranging from 0.3pF to 2.97pF programmed in its register. The proposed tunable solution for the Main Antenna in this paper is not limited to the technology or type of the Tunable Capacitor.

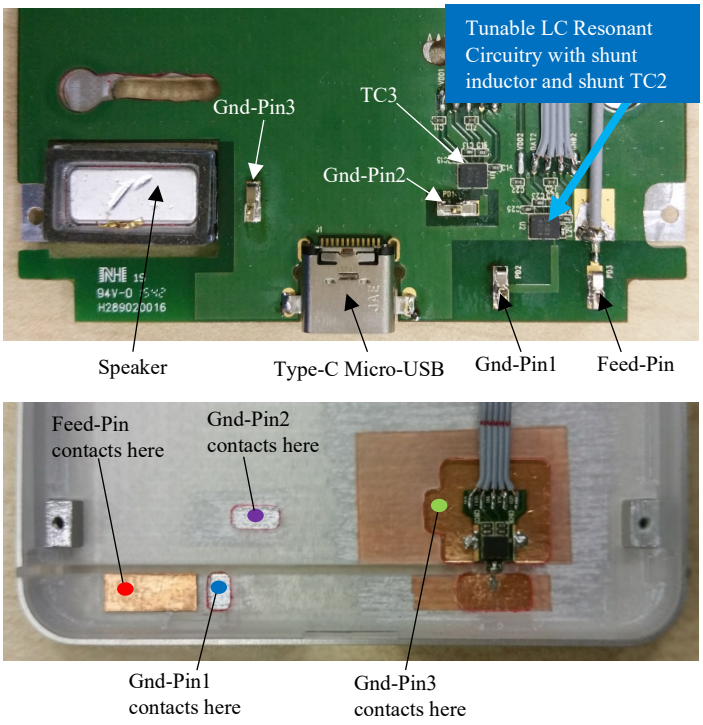


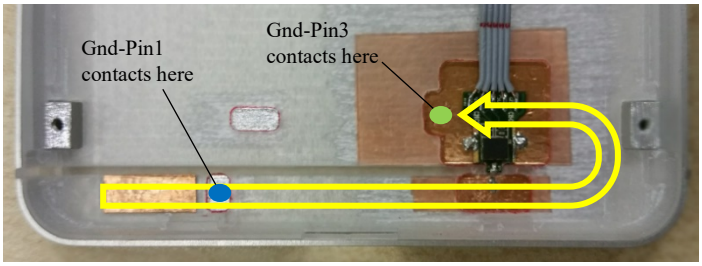
Figure 4. PCB layout for the Main Antenna and integrating the Metallic Back Cover as part of the Main Antenna design

3. Flexible Aperture Tuning Solution for Main Antenna

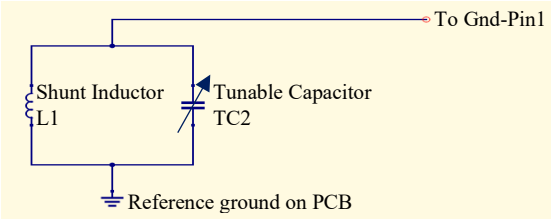
The Main Antenna of the mobile phone has operating frequency bands that are divided into the Low Band, the Mid Band and the High Band. The Low Band covers frequency range from 700MHz to 960MHz, the Mid Band covers frequency range from 1710MHz to 2170MHz and the High Band covers frequency range from 2300MHz to 2690MHz. These ranges of frequency bands are sufficient to provide coverage for the 2G, 3G and 4G network. The Aperture Tuning of the Main Antenna as proposed in this paper will be broken down to sections with respect to the following bands: Low Band, Mid Band and High Band.

3.1. Low Band Tuning (700MHz to 960MHz)

The proposed Main Antenna solution is fed using the Capacitive-Fed Coupling Technique [5 – 6] to the Bottom-section metallic cover via the Capacitive Feed Patch as mentioned above in Figure 3. Energy from the Capacitive Feed Patch is coupled onto the Bottom-section metallic cover. Current on the Bottom-section metallic cover will then flows in the direction as indicated by the yellow-coloured arrow (as shown in Figure 5(a)) to Gnd-Pin3 on the Main Antenna PCB. This forms a Loop Antenna for the Low Band of the Main Antenna. Aperture Tuning is applied to this Loop Antenna at Gnd-Pin1, via the Tunable LC Resonant Circuitry [2 – 4] (as mentioned in Figure 4) which equivalent circuit is as shown in Figure 5(b). The resonance at the Low Band of the Main Antenna can be made to shift within the frequency range of 700MHz to 960MHz by tuning the capacitance value of the Tunable Capacitor TC2 on the PCB. In tuning the Low Band of the Main Antenna, the capacitance value of the Tunable Capacitor TC1 on the Metallic Back Cover and the Tunable Capacitor TC3 at Gnd-Pin2 on the PCB are kept constant at 2.97pF and 0.66pF respectively.



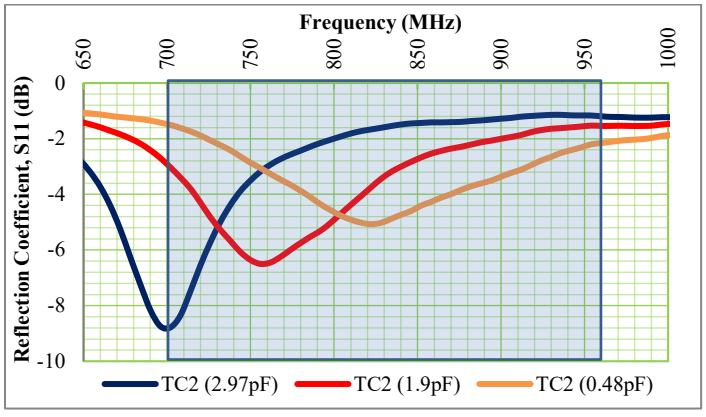
(a) Current flow for Low Band Tuning



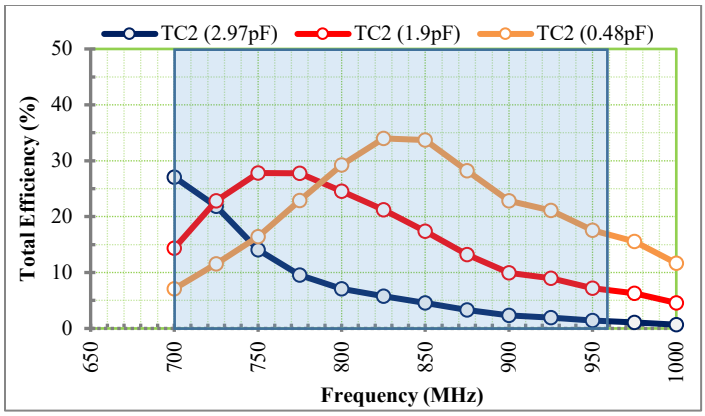
(b) Equivalent circuit of Tunable LC Resonant Circuit

Figure 5. Low Band Tuning of Main Antenna

Figure 6 shows plots of the passive measurement result for the Low Band Tuning of the Main Antenna, in Free-Space condition. Figure 6 (a) is the Reflection Coefficient S11 of the Main Antenna measured using a Network Analyzer. Figure 6(b) shows plots of the Total Efficiency of the Main Antenna in the Low Band region, measured in an Anechoic Over-the-Air (OTA) Chamber. From the two plots shown in Figure 6, it can be seen that three different



(a) Reflection Coefficient S11 in Low Band region



(b) Total Efficiency in Low Band region

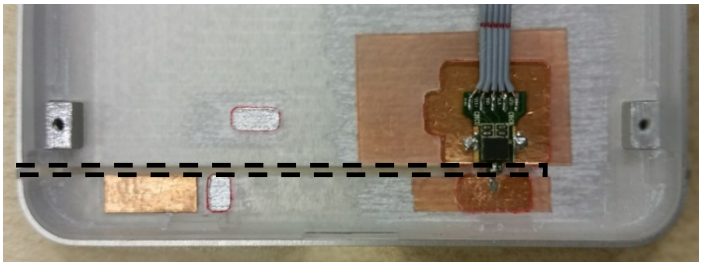
Figure 6. Passive measurement results in Low Band region

capacitance value of TC2 (at 0.48pF, 1.9pF and 2.97pF) are used to shift the Low Band resonance across the Low Band region of 700MHz to 960MHz. This shift in the Total Efficiency plot (as seen in Figure 6(b)) helps to maintain the Total Efficiency in the Low Band region at above 20%.

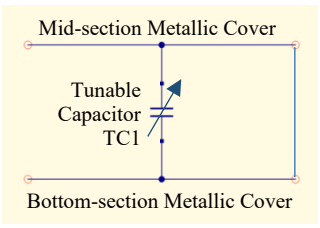
3.2. Mid Band Tuning (1710MHz to 2170MHz)

The Mid Band of the Main Antenna is resulting from a Slot Antenna form on the Metallic Back Cover by the Closed-end slot, as highlighted with the black-dotted line in Figure 7(a). The Tunable Capacitor TC1, which output is bridged across the Closed-end slot on the Metallic Back cover, is being used to provide Aperture Tuning to this Slot Antenna. The configuration of having the Tunable Capacitor TC1 across the Closed-end slot can be best described by the equivalent circuit as shown in Figure 7(b). Hence, by tuning the capacitance value of Tunable Capacitor TC1, the wavelength of the Slot Antenna will change accordingly, shifting the Mid Band resonance within the frequency range of 1710MHz to 2170MHz. In tuning the Mid Band of the Main Antenna, the capacitance value of both the Tunable Capacitor TC2 and TC3 on the Main Antenna PCB are kept constant at 2.97pF.

Figure 8 shows plots of the passive measurement result for the Mid Band Tuning of the Main Antenna, in Free-Space condition. Figure 8(a) is the measured Reflection Coefficient S11 of the Main Antenna and Figure 8(b) shows plots of the measured Total Efficiency of the Main Antenna in the Mid Band region. From the two plots shown in Figure 8, it can be seen that the Mid Band resonance can be made to shift across the Mid Band region of 1710MHz to 2170MHz by tuning the capacitance value of TC1 at 1.55pF, 1.9pF, 2.44pF and 2.97pF respectively. This shift in the Total Efficiency plot (as seen in Figure 8(b)) helps to maintain the Total Efficiency in the Mid Band region at above 20%.

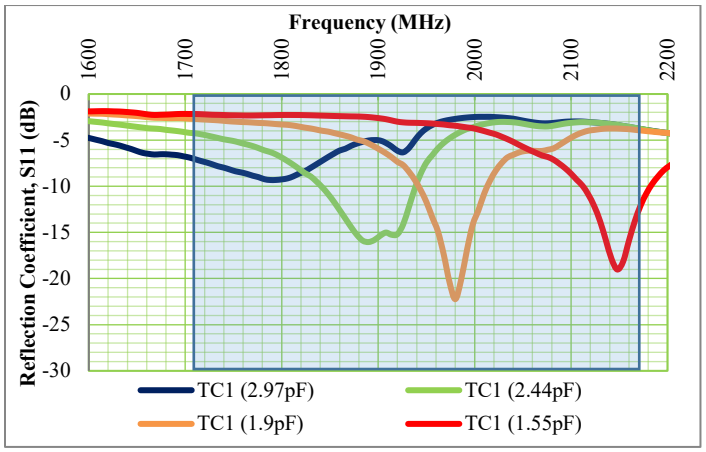


(a) Slot Antenna for Mid Band Tuning

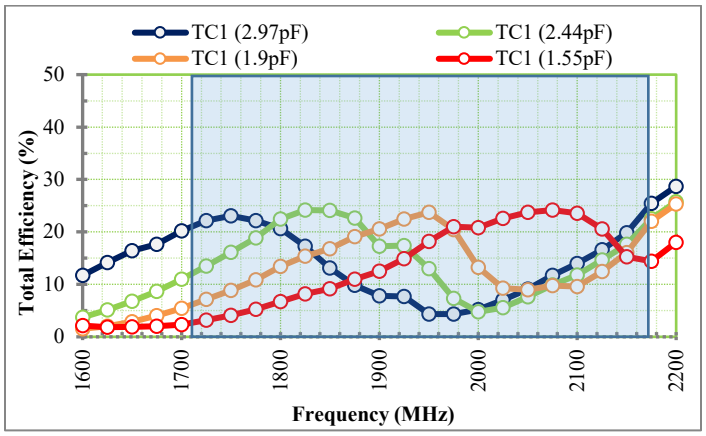


(b) Equivalent circuit of Tuner Board on Metallic Back Cover

Figure 7. Mid Band Tuning of Main Antenna



(a) Reflection Coefficient S11 in Mid Band region



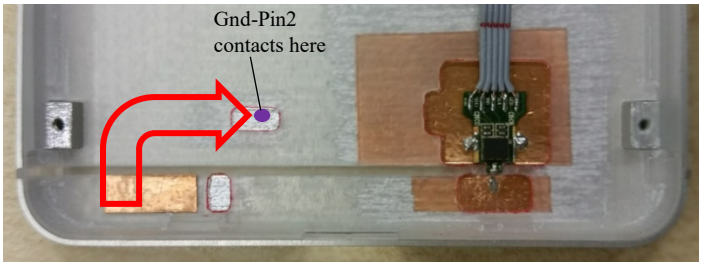
(b) Total Efficiency in Mid Band region

Figure 8. Passive measurement results in Mid Band region

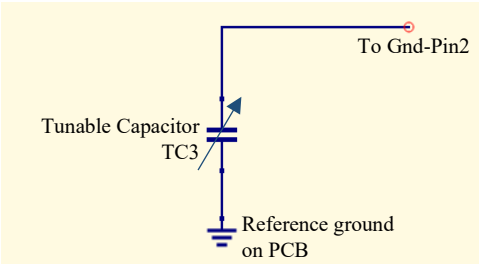
3.3. High Band Tuning (2300MHz to 2690MHz)

In High Band Tuning of the Main Antenna, the energy from the Capacitive Feed Patch on the Bottom-section metallic cover is coupled across the Closed-end slot to the Mid-section metallic cover. Current on the Mid-section metallic cover will then flows in the direction as indicated by the red-coloured arrow to Gnd-Pin2 on the Main Antenna PCB, as shown in Figure 9 (a). This actually forms a small Loop Antenna for the High Band of the Main Antenna. Aperture Tuning is applied to this small Loop Antenna at Gnd-Pin2, via the Tunable Capacitor TC3, which equivalent circuit is as shown in Figure 9 (b). The resonance at the High Band of the Main Antenna can be made to shift within the frequency range of 2300MHz to 2690MHz by tuning the capacitance value of the Tunable Capacitor TC3 on the PCB. In tuning the High Band of the Main Antenna, the capacitance value of the Tunable

Capacitor TC1 on the Metallic Back Cover and the Tunable Capacitor TC2 in the Tunable LC Resonant Circuit on the PCB are both kept constant at 2.97pF.

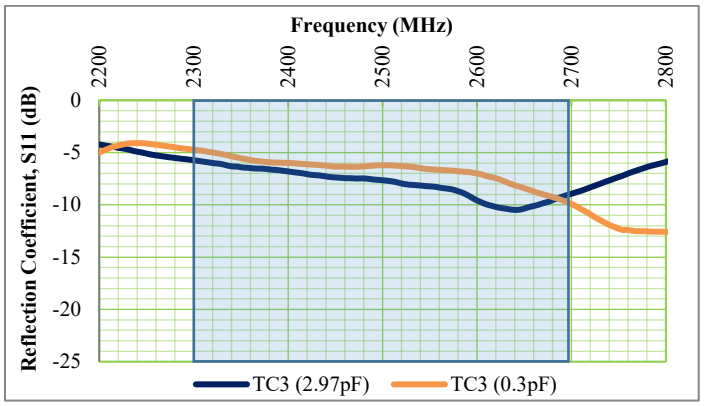


(a) Current flow for High Band Tuning

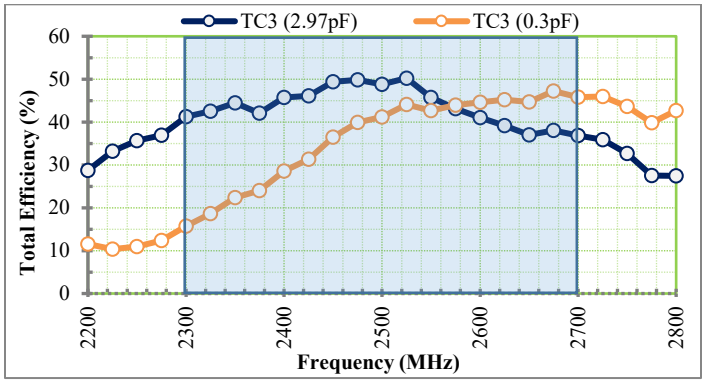


(b) Equivalent circuit of Aperture Tuning at Gnd-Pin2

Figure 9. High Band Tuning of Main Antenna



(a) Reflection Coefficient S11 in High Band region



(b) Total Efficiency in High Band region

Figure 10. Passive measurement results in High Band region

Figure 10 shows plots of the passive measurement result for the High Band Tuning of the Main Antenna, in Free-Space condition. Figure 10 (a) is the measured Reflection Coefficient S11 of the Main Antenna and Figure 10 (b) shows plots of the measured Total Efficiency of the Main Antenna in the High Band region. From the two plots shown in Figure 10, it can be seen that two-different capacitance value of TC3 (0.3pF and 2.97pF) are used to shift the High Band resonance across the High Band region of 2300MHz to 2690MHz. This shift in the Total Efficiency plot (as seen in Figure 10(b)) helps to maintain the Total Efficiency in the High Band region at above 40%.

4. Mobile Phone Mockup in Active Phone Condition

In this section, the Metallic Back Cover mobile phone mockup as reported above is configured to be an active phone. Figure 11 shows a picture illustrating the Metallic Back Cover mobile phone mockup in its active phone configuration. The mockup is equipped with an in-house designed Main PCB which is used to house a Long-Term Evolution (LTE) module for conducting active Total Radiated Power (TRP) measurement in an Anechoic OTA Chamber. The Main PCB is equipped with power contact points to draw power from a mobile phone battery, to power up the LTE Module as well as the Tunable Capacitors (TC1, TC2 and TC3 mentioned above in section 2 and 3) during the active measurement.

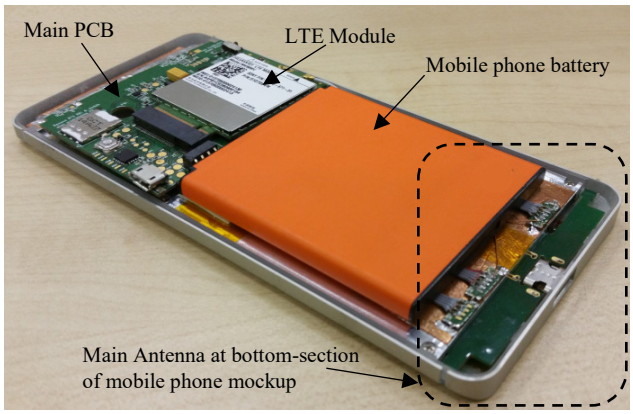


Figure 11. Mobile phone mockup in active condition

4.1. Measurement in Free-Space

The active TRP measurement in the Free-Space condition was conducted in an Anechoic OTA Chamber, for four different 3G bands and five different 4G LTE bands. These bands are WCDMA Band 1, 2, 5 and 8 in the 3G network, and FDD LTE Band 1, 4, 7, 8 and 13 in the 4G LTE network. Table 1 tabulates the active TRP measurement results of the Main Antenna, with the metallic cover mobile phone mockup in Free-Space condition. The first column in Table 1 lists the respective nine frequency bands in the 3G and 4G network. The second column displays the mid channel for each band respectively, and the third column lists the frequency (in MHz) corresponding to respective mid channel in the second column. As can be seen from the third column, this measurement covers the various frequencies which fall within the Low (700MHz to 960MHz), Mid (1710MHz to 2170MHz) and High (2300MHz to 2690MHz) Band regions of the Main Antenna. The fourth column in Table 1 lists the measured Conducted Power (in dBm) of the LTE Module, transmitting at the respective frequencies listed in the third column. And lastly, the measured Free-Space TRP (in dBm) of the Main Antenna on the Metallic Back Cover mobile phone mockup is as listed in the fifth column. From Table

1, the Conducted Power of the LTE Module is ranging from 20 to 23dBm, and the Free-Space TRP ranges from 14 to 20dBm.

Table 1. Active TRP measurement results in Free-Space

Band	Channel	Frequency (MHz)	Conducted Power (dBm)	Free-Space TRP (dBm)
WCDMA Band5	4182	836.6	22.3	17.8
WCDMA Band8	2787	897.6	22.2	16.7
WCDMA Band2	9400	1880.0	22.9	15.1
WCDMA Band1	9750	1950.0	22.4	14.1
FDD LTE Band13	23230	782.0	22.0	16.9
FDD LTE Band8	21625	897.5	21.2	16.6
FDD LTE Band4	20175	1732.5	21.8	15.7
FDD LTE Band1	18300	1950.0	21.7	13.8
FDD LTE Band7	21100	2535.0	20.2	19.7

4.2. Measurement with Phantom Head & Hands

To study the user effect towards the Main Antenna solution on the metallic cover mobile phone mockup, the active TRP measurement was further extended to the Phantom Head & Hands configuration. Figure 12 shows an example of the setup of the Phantom Head + Right Hand (BHHR) in the talk mode configuration. In this setup, the metallic cover mobile phone mockup will be gripped by the Hand, and placed beside the Phantom Head, as according to the CTIA Test Plan [7]. Two Phantom configurations had been considered in this measurement: Phantom Head + Left Hand (BHHL) and Phantom Head + Right Hand (BHHR). This active TRP measurement of the Main Antenna, with the mobile phone mockup on the Phantom Head & Hands was conducted in an Anechoic OTA Chamber, for four different 3G bands (WCDMA Band 1, 2, 5 and 8). Table 2 tabulates the active TRP measurement results of the Main Antenna, with the mobile phone mockup on the Phantom Head & Hands (BHHL and BHHR) talk mode configuration. The first three columns in Table 2 list the frequency bands, its mid channel and the corresponding frequency in the 3G network respectively. The fourth and fifth column in Table 2 list the measured TRP (in dBm) of the Main Antenna, in the BHHL and BHHR configuration respectively. From the BHHL TRP and BHHR TRP results listed in Table2, it can be clearly seen that the left hand has a greater influence on the Main Antenna performance as compared to the right hand. This is likely due to the hand is covering the opening of the Closed-end slot on the Metallic Back Cover in the BHHL configuration. The variation in TRP between BHHL and BHHR is between 1 to 3.3dBm, with the worst case scenarios taking place at WCDMA Band 2 (1880MHz). A comparison between the Free-Space TRP (in Table 1) and BHHL TRP (in Table 2) shows a reduction in the antenna performance in the latter case with variation of between 5.9 to 9.6dBm, and the worst case scenarios taking place at WCDMA Band 8 (897.6MHz). The comparison between the Free-Space TRP and BHHR TRP also shows a reduction in the antenna performance in the latter case, with variation of between 4.6 to 8.4dBm, and the worst case scenarios taking place at WCDMA Band 5 (836.6MHz).

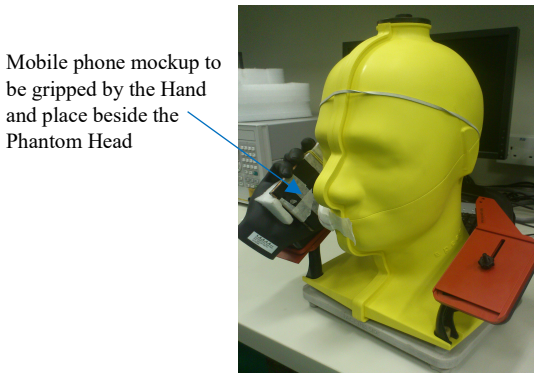


Figure 12. Phantom Head + Right Hand (BHHR) in talk mode setup

Table 2. Active TRP measurement results in BHHL and BHHR

Band	Channel	Frequency (MHz)	BHHL TRP (dBm)	BHHR TRP (dBm)
WCDMA Band5	4182	836.6	8.3	9.3
WCDMA Band8	2787	897.6	7.1	9.6
WCDMA Band2	9400	1880.0	7.0	10.3
WCDMA Band1	9750	1950.0	8.2	9.5

5. Conclusion

It has always been challenging in providing solution for antennas on mobile phones with metal housing. The metal housing often introduces unnecessary coupling, shielding effect towards the antennas as well as reducing the flexibility in designing the antenna. In this paper, a tunable solution has been proposed for the cellular Main Antenna of a mobile phone with metal housing in the form of metallic back cover. The proposed solution demonstrates the integration of the metal housing as part of the antenna design, along with the use of different Aperture Tuning circuitries at different parts of the Main Antenna, to allow flexible tuning of the antenna's Low, Mid and High Band. In an attempt to study the user effect towards the Main Antenna on the metallic cover mobile phone, the measurement results suggested that the Left Hand influence on the antenna is greater as compared to the Right Hand. In the comparison between Free-Space and BHHL/BHHR condition, the antenna performance sees a reduction of less than 10dBm.

References

[1] Mark Y.C. Tan, Guan Hong Ng, Roger Tay, "A tunable LTE main antenna solution for metallic cover mobile phone", 2017 11th European Conference on Antennas and Propagation (EuCAP), 19 – 24 March 2017. https://doi.org/10.23919/EuCAP.2017.7928219

[2] P. A. Tornatta, R. Gaddi, "Aperture tuned antennas for 3G-4G applications using MEMS digital variable capacitor", 2013 IEEE MTT-S International Microwave Symposium Digest (IMS), 2 – 7 June 2013. https://doi.org/10.1109/MWSYM.2013.6697602

[3] Sameer Akhtar Shah, Liu Qianyun, "A novel approach to design an antenna aperture tuning circuit", 2015 Asia-Pacific Microwave Conference (APMC), 6 – 9 Dec 2015. https://doi.org/10.1109/APMC.2015.7413123

[4] Mark Y.C. Tan, Guan Hong Ng, Roger Tay, "Tuning of LTE Main Antenna Using Flexible Tunable LC Resonant Circuitry", 2017 11th European Conference on Antennas and Propagation (EuCAP), 19 – 24 March 2017. https://doi.org/10.23919/EuCAP.2017.7928223

[5] Wei-Yu Li, Chun-Yih Wu, Kin-Lu Wong and Ming-Fang Tu, "Internal small-size PIFA for LTE/GSM/UMTS operation in the mobile phone", 2010 IEEE Antennas and Propagation Society International Symposium (APSURSI), July 2010. https://doi.org/10.1109/APS.2010.5562325

[6] Norakamon Wongsin, Thanakarn Suangun, Chatree Mahatthanajatuphat, Prayoot Akkaraekthalin, "A multiband fractal ring antenna fed by capacitive coupling", 2011 8th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 17 – 19 May 2011. https://doi.org/10.1109/ECTICON.2011.5947808

[7] "Test plan for wireless device Over-the-Air performance", CTIA, version 3.6.2, May 2017. https://www.ctia.org/docs/default-source/certification/ctia-test-plan-for-wireless-device-over-the-air-performance-ver-3-6-2.pdf?sfvrsn=2

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 49-65 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Improving System Reliability Assessment of Safety-Critical Systems using Machine Learning Optimization Techniques

Ibrahim Alagöz[*1], Thomas Hoiss[2], Reinhard German[1]

[1]*Department of Computer Science 7, FAU Erlangen-Nuremberg, 91058, Germany*

[2]*Automotive Safety Technologies GmbH, 85080, Gaimersheim, Germany*

ARTICLEINFO

Article history:
Received: 14 November, 2017
Accepted: 10 January, 2018
Online: 30 January, 2018

Keywords:
Safety-Critical System
Black Box Regression Testing
Linear Classifier
Selection
Prioritization

ABSTRACT

Quality assurance of modern-day safety-critical systems is continually facing new challenges with the increase in both the level of functionality they provide and their degree of interaction with their environment. We propose a novel selection method for black-box regression testing on the basis of machine learning techniques for increasing testing efficiency. Risk-aware selection decisions are performed on the basis of reliability estimations calculated during an online training session. In this way, significant reductions in testing time can be achieved in industrial projects without uncontrolled reduction in the quality of the regression test for assessing the actual system reliability.

1 Introduction

Reliability assessment of safety-critical systems is becoming an almost insurmountable challenge. In the near future, the engineering of new applications for vehicles such as driving assistance functions or even autonomous driving systems will inevitably incur significantly increased engineering sophistication and longer test cycles. Thus, in the automotive domain, functional safety continues to be ensured on the basis of the international ISO 26262 standard. As both the levels of functionality such systems provide and their degree of interaction with their environment increases, an adequate increase in system safety assessment capabilities is required.

This paper is an extension of work originally presented at the 10th IEEE International Conference on Software Testing, Verification and Validation (ICST 2017) [1] and describes a methodology for efficiently assessing system safety. The focus of the paper is on regression testing of safety-critical systems consisting of black-box components. This scenario is common for automotive electronic systems, where testing time is expensive and should be reduced without an uncontrolled reduction in reliability.

The work reported here correspondingly seeks to increase testing efficiency by reducing the number of selected test cases in a regression test cycle. When a selection decision is made, the following two types of errors are possible:

- a test case is selected but would pass (type-I-error, false-positive case) and
- a test case is not selected but would fail (type-II-error, false-negative case).

Accordingly, we model a classifier $\hat{H}$ for solving the following optimization problem.

$$\begin{aligned} &\min\ p_{FP} = P(\hat{H} = H_1|H_0) \\ &\text{subject to } p_{FN} = P(\hat{H} = H_0|H_1) \leq p_{FN,MAX} \end{aligned} \tag{1}$$

A good standard of test efficiency calls for the avoidance of false-positives. This requires minimization of the probability of mistakenly assuming the rival hypothesis (H_1 : test case fails) even though the null hypothesis (H_0 : test case passes) is correct. Conversely, false-negatives mean that system failures remain undetected; the occurrence of this type of error must therefore be avoided with very stringent requirements. Thus, a predefined limit $p_{FN,MAX}$ for the probability of a false-negative is defined.

[1] proposed a concept for the selection of test cases based on a stochastic model. However, this paper proposes a holistic optimization framework for the safety assessment of safety-critical systems based on machine learning optimization techniques. We

[*]Corresponding Author: Ibrahim Alagöz, Hornstr. 1 85051 Ingolstadt, ibrahim.alagoez@gmail.com

https://dx.doi.org/10.25046/aj030107

suggest an incrementally and actively learning linear classifier whose parameters are estimated on the basis of Bayesian inference rules. As a result, our novel approach for modeling a linear classifier outperforms other machine learning approaches in terms of sensitivity.

Furthermore, this paper deals with the following fundamentally important research question: The machine learning approach is trained with data (test evaluations) obtained during a concurrently running regression test. How much training data is enough? When does regression test selection actually start?

We extend the proposed selection method [1] by introducing suitable test case features that are used in the machine learning approach for increasing performance (see [2]). Therefore, each feature introduced increases the complexity of the optimization problem (cf. Eq. 1) as a new dimension for optimization is introduced. Thus [3] and [4] suggest that high dimensional optimization problems can be solved in reasonable timeframes by using evolutionary algorithms instead of a (grid)search-based approach as given in [1]. Accordingly, we propose an evolutionary optimization approach for increasing testing efficiency.

Further extensions, such as the introduction of a prioritization strategy for test cases in order to select higher-priority test cases, will also be presented within this paper. In our novel approach, a linear classifier is trained in an online session; the ordering of the training data on the basis of a prioritization strategy therefore has the potential to improve our classifiers' performance.

We also provide an industrial case study to show the advantages of the suggested selection method. The study uses data from several regression test cycles of an ECU of a German car manufacturer, showing how test effort can be reduced significantly whereas the rates of both false-negatives and false-positives can be kept at very low values. In this example, we can quadruple test efficiency by keeping the false-negative probability at 1%.

We first discuss related work in Sec. 2 and explain basic definitions in Sec. 3. Accordingly, we motivate our research topic in Sec. 4 by giving some background information on regression tests and referring to the challenges. In Sec. 5, we give a brief overview of known machine learning methods' performance in solving safety-critical binary classification tasks. The concept of our novel machine learning approach is presented in Sec. 6. Sec. 7 discusses optimization strategies, and Sec. 8 focuses on the importance of the learning phase for the success of our approach. An industrial case study with real data is then given in Sec. 9. Finally, Sec. 10 presents the paper's conclusions.

2 Related Work

The automotive industry is currently engaged in a laborious quality assessment process around new engineered driver assistance and active and passive safety functions, while functional safety is ensured according to the international ISO 26262 standard [5]. Reliability assessment of systems is therefore, possible through both model-checking and testing.

Model-checking is used for verifying conditions on system properties. Thus [6] states that system requirements can also be validated by model-checking techniques. The idea is to check the degree to which system properties are met and to deduce logical conclusions on the basis of the satisfaction of system requirements. Model-checking has therefore gained wide acceptance in the field of hardware and protocol verification communities [7]. Motivated by the fact that numerical model-checking approaches cannot be directly applied to black-box components as a usable formal model is not available, we focus on model-checking driven black-box testing [6] and statistical model-checking techniques [8]. However, there exist some approaches for interactively learning finite state systems of black-box components (see [9] and [10]), which are proposed as *black box checking* in [11]. Learning a model is an expensive task, as the interactively learned model has to be adapted due to inaccuracy reasons. Nevertheless, some assumptions about the system to be checked, such as the number of internal states, are necessary; furthermore, conformance testing for ensuring the accuracy of the learned model has to be iteratively performed [9].

Therefore, [8] outlines the advantages of statistical model-checking as being simple, efficient and uniformly applicable to white- and even to black-box systems. [6] motivates on-the-fly generation of test cases for checking system properties; here, a test case is generated for simulating a system for a finite number of executions. All these executions are used as individual attempts to discharge a statistical hypothesis test and finally for checking the satisfaction of a dedicated system property.

Model-checking driven testing, or even simply testing a system in order to validate its requirements, is an expensive task, especially where safety-critical systems are concerned. However, the focus is on regression testing, which means that the entire system under test has already been tested once but has to be tested again due to system modifications that have been carried out. The purpose of regression testing is to provide confidence that unchanged parts within the system are not affected by these modifications [12]. White-box selection techniques have been comprehensively researched [13, 14]. However, we are here considering black-box components, and hence selecting test cases that only check modified system blocks gets difficult.

Since the implementation of black-box systems and moreover, the information on performed system modifications is not available [12], reasonably conducting a regression test becomes impossible.

Accordingly, regression testing of safety-critical black-box systems ends up in simply executing all existing test cases; this is a *retest-all* approach [12].

This is also motivated by the fact that in the au-

tomotive industry, up to 80% of system failures [1] that are detected during a regression test have not occurred previously. The reason behind this fact is that often many unintended bugs are introduced during a bug-fixing process. So between two system releases many new unknown errors are often introduced.

For reducing the overall test effort, we apply a test case selection method [1] based on hypothesis tests. Those test cases that are assumed to fail on their executions are accordingly selected. However, type errors while performing hypothesis tests are possible, as, for instance, in statistical model-checking.

We extend the proposed selection method into a holistic machine learning-driven optimization framework that utilizes suitable test case features for increasing testing efficiency (see [2]). Machine learning methods are often trained in so-called *batch* modes. Nevertheless, many applications in the field of autonomous robotics or driving are trained on the basis of continuously arriving training data [15]. Thus, incremental learning facilitates learning from streaming data and hence is exposed to continuous model adaptation [15]. Especially handling non-stationary data assumes key importance in applications like voice and face recognition due to dynamically evolving patterns. Accordingly, many adaptive clustering models have been proposed, including incremental K-means and evolutionary spectral clustering techniques [16].

Furthermore, labeling input data is often awkward and expensive [17] and hence accurately training models can be difficult. Therefore, semi-supervised learning techniques are developed for learning from both labeled and unlabeled data [17]. Motivated by these techniques, we propose a similar approach for effectively learning from labeled data. Hence, we cluster binary labeled data in more than two clusters for improving a classifier's learning capability due to the optimization of an objective function. Our optimization framework thus utilizes evolutionary optimization algorithms for handling the optimization complexity. Minimization of labeling cost on the basis of active learning strategies [18, 19] will also be dealt with in this paper.

3 Basic Definitions

We define the test suite $T = \{t_i \mid 1 \leq i \leq M\}$ consisting of a total of M test cases. $T_{Exec} \in T$ and $T_{\overline{Exec}} \in T$ are subsets of T that contain test cases that are executed and deselected in a current regression test respectively. Based on the test case executions ($\forall t_i \in T_{Exec}$), a system's reliability is actually learned, and thus the machine learning algorithm is trained.

The focus in supervised learning is on understanding the relationship between feature and data (here test case evaluation) [4]. Therefore, a test case needs to code a feature vector so that the indication of the coded features for a system failure can be learned in a supervised fashion. Such an indication is not just a highly probable forecast of an expected system failure, it is rather a particular risk-associated recognition.

First of all, a feature can be any individual measurable property of a test case. The data type of a feature is mostly numeric, but strings are also possible. However, such features need to be informative, discriminative and independent of one another if they are to be relevant and non-redundant The definition of suitable features increases the classifier performance [20]. In our application, a feature can be varied, such as a

- subjective ranking of a test case based on expert knowledge. Such rankings can hint at the error susceptibility of verified parts of the system;
- verified function's safety integrity level, known as the *ASIL* in automotive applications [5];
- name of a function whose reliability is assured;
- reference to any hardware component of a circuit board that is being tested in a hardware-in-the-loop (HiL) test environment;
- number of totally involved electronic control units during the testing of a networked functionality; Such a number can hint at the complexity of the networked functionality and hence at its error susceptibility.

We define the entire set of features $\Phi = \{\phi_f \mid 1 \leq f \leq F\}$ of test cases that might be relevant for understanding the behavior of test cases. Thus, features may be e.g. $\phi_1 = \{'QM','A','B','C','D'\}$ ($ASIL$) or $\phi_2 = \{f_1, f_2, f_3\}$ (function name). Hence, a test case can verify a function f_3 that has an *ASIL* A.

The following passages discuss the selection of suitable features, which is an important strategy for improving a classifier's performance.

- Sometimes less is more - If the defined set Φ is too large it can cause huge training effort, high dimensionality of the optimization problem and overfitting. Thus we define a selection mask $\boldsymbol{b}_s = \begin{bmatrix}1 & 0 & 0 \cdots 1\end{bmatrix}$ of length F for selecting relevant features Φ_s. If the $f-th$ matrix entry of $\boldsymbol{b}_s$ is greater than or equal to 1, then the corresponding feature $\phi_f \in \Phi$ is selected and added to Φ_s, otherwise not.
- The set of main features $\Phi_m \subseteq \Phi_s$ is coded as follows: If the $f-th$ matrix entry of $\boldsymbol{b}_s$ is equal to 2, then the corresponding feature $\phi_f \in \Phi_s$ is at the same time a main feature $\phi_f \in \Phi_m$, otherwise not. The main features are used to establish the overall training data set: The training data is adapted to each test case, and thus it is $T_{t_i} = \{t_j \mid t_j \in T_{Exec} \wedge t_i \overset{\Phi_m}{\equiv} t_j\}$. Hence, we define that two test cases t_i and t_j are equivalent $t_i \overset{\Phi}{\equiv} t_j$ if their features $\forall \phi_f \in \Phi$ have identical values.

Additionally, a cross-product transformation of features $\forall \phi_f \in \Phi_s \setminus \Phi_m$ is performed. Thus we define $\Psi = \phi_{f_1} \times \phi_{f_2} \times ... \times \phi_{f_h} \times ... \times \phi_{f_H}$ with $\phi_{f_h} \in$

$\Phi_s \setminus \Phi_m, 1 \leq h \leq H = |\Phi_s \setminus \Phi_m|$ consisting of features ψ_l that represent individual combinational settings for features $\phi_f \in \Phi_s \setminus \Phi_m$. In simple terms, the cross-product of our sample features is $\phi_1 \times \phi_2 = \{('QM', f_1), ('QM', f_2), ('QM', f_3), ('A', f_1), ...\}$ Finally, a Boolean function $check : T \times \Psi \rightarrow \mathbb{B}$ with $check(t_i, \psi_l) = \begin{cases} 0, & \text{if } t_i\text{'s features are given by } \psi_l \\ 1, & otherwise \end{cases}$ is defined.

In addition, the function $state$: $T \times R \rightarrow S$ is defined; it returns the state of a dedicated test case in a concrete regression test. The state has to be either 'Pass' or 'Fail', except for cases where the test case has not been executed so that its state is undefined. Therefore, $S = \{'Pass', 'Fail', 'Undefined'\}$ defines the set of possible states. Furthermore, the set $R = \{r_k \mid 0 \leq k \leq K\}$ includes r_0 which is the current regression test and older regression tests starting from the last regression r_1 to the first considered regression r_K. Lastly, we define the tuple $history(t_i) = \{state(t_i, r_1), ..., state(t_i, r_K)\}$ containing t_i's previous test results.

4 Motivation

In practice, finding suitable features is a difficult task. Since we focus here on black-box systems, system-internal information is not available that might be useful for understanding the system behavior. As a reason, we can only define the above listed features, which might be too high-level for classifying system failures. To illustrate this fact, Fig. 1 a) shows a typical situation: The behavior of test cases in relation to arbitrarily defined features ϕ_1 and ϕ_2 is given. Passed and failed test cases are presented by green squares and red diamonds respectively, and white circles stand for test cases yet to be executed.

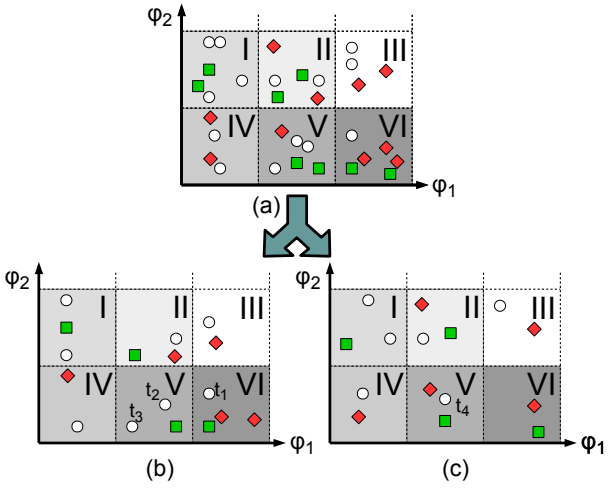


Figure 1: An artificial regression test with test cases.

We can see that the green squares and the red diamonds are widely scattered. Thus, defining a hyperplane in order to set two acceptance regions for passing and failing test cases is no easy matter. In order to solve this complex task, we develop a novel approach that is basically motivated by the following thought experiment: All test cases that are represented in Fig. 1 a) are now either assigned to 1 b) or 1 c) according to a certain mapping. The individual mappings of test cases will be discussed later, in Sec. 6. In the next step, a cross-product transformation is performed in order to group test cases into sub-regions (we refer these later as sub-clusters). In our example, we create six sub-regions. Table 1 lists the empirical failure probabilities of each sub-region.

In order to keep our thought experiment very simple, we will neglect statistical computations for now and focus only on the main idea of our novel approach. The introduction of Bayesian networks and hence the derivation of weights for linear classifiers will be discussed later, in Sec. 6. We assume for now that the calculated failure probabilities of test cases in Fig. 1 b) and Fig. 1 c) are correlated. So our example remains very simple, we also require that the failure probabilities of the corresponding sub-regions are equal. This assumption reduces the complexity of the following classification task enormously. We will classify the following test cases t_1, t_2, t_3 and t_4 in accordance with whether a selection is necessary or not.

Table 1: Failure probability of each sub-region.

$P(H_1)$	I	II	III	IV	V	VI
Fig. 1 a)	0	0.5	1	1	1/3	3/5
Fig. 1 b)	0	0.5	1	1	0	2/3
Fig. 1 c)	0	0.5	1	1	0.5	0.5

Only if t_1 passes will the failure probability of sub-region VI in Fig. 1 b) be equal to the failure probability of sub-region VI in Fig. 1 c). According to this fact, t_1 is assumed to pass, and hence it is deselected. Furthermore, t_2 will be selected as a fail of a test case inside sub-region V is expected. However, t_2 passes, and, based on the same consideration, t_3 is also selected and finally fails. Since now a failure probability of 1/3 is expected in sub-region V, t_4 is assumed to pass, and therefore it does not need to be selected. Table 2 summarizes all decisions executed.

Table 2: Test case states and algorithm decisions.

Test Case	State	Decision	Type of Decision
t_1	Pass	Deselected	True-Negative
t_2	Pass	Selected	False-Positive
t_3	Fail	Selected	True-Positive
t_4	Pass	Deselected	True-Negative

So our novel approach for solving binary classification tasks is based on calculated empirical probabilities and empirically evaluated correlations among those probabilities. The behavior of test cases can be

precisely estimated on the basis of the calculated correlations. In practice, the failure probabilities of same sub-regions in Fig. 1 b) and Fig. 1 c) is often not exactly equal, but these failure probabilities are correlated. So the main task is to find *good* sub-regions for maximizing the empirically evaluated correlations and thus for precisely estimating the behavior of test cases. A more detailed explanation of our novel approach will follow in Sec. 6.

5 Performance of Known Machine Learning Methods

We have already indicated, by showing the example regression test in Sec. 4 (see Fig. 1), that according to the distribution of the input data, many machine learning methods cannot be reasonably applied for solving the constrained optimization problem (cf. Eq. 1). We will now demonstrate briefly that training linear classifiers in the classical sense by minimizing a loss function cannot perform well for solving safety-critical binary classification tasks. The situation is that only a small percentage of the data is actually labeled with one ('Fail'). Furthermore, failed and passed test cases are widely scattered in the feature space, which means detecting failing test cases becomes impossible. Additionally, the performance of deep neuronal networks is validated in the following.

The evaluation results (precision/recall) of these machine learning methods are given in Table 3. Each machine learning method is trained in the *batch* mode. The training data consists of all obtained test evaluations of a special regression test that will also be analyzed in our industrial case study in Sec. 9. For evaluating the machine learning methods, we used the training data first for training and later for testing (training data = test data). Even so, the sensitivity of both machine learning methods is zero, and thus we propose a novel approach for determining a linear classifier's parameters in Sec. 6.

Table 3: Performance of known machine learning methods in solving safety-critical binary classification tasks.

Machine Learning	Precision	Recall / Sensitivity
Linear Classifier (Trained by Minimizing a Loss-Function)	0	0
Deep Neuronal Network	0	0

6 Concept

The concept of our novel approach is shown in Fig. 2. We start with specifying a feature set Φ, and taking its subset Φ_s and finally constitute a cross-product feature transformation to obtain the set Ψ. Based on Ψ and by applying the *check*-function on T_{Exec}, test cases can be grouped. If we look back to the example where we grouped test cases in Fig. 1 a), then we will see that there is a relationship between test cases' features and their assignments to sub-regions.

Correspondingly, we introduce the definitions of clusters and sub-clusters of test cases. In the first step, test cases $\forall t_k \in T_{t_i}$ are assigned to clusters based on their history-tuples. A cluster is basically a partition of T_{t_i} and consists of test cases that have the same history-tuples. Accordingly, the number of distinct history-tuples N determines the total number of clusters. This is the step that has already been shown in Sec. 4, where test cases inside Fig. 1 a) were individually mapped into Figs. 1 b) and 1 c). In this way, already executed test cases depicted in Fig. 1 b) belong to one cluster and, analogously, those executed test cases that are depicted in Fig. 1 c) belong to another cluster.

In the next step, each cluster C_n is subdivided into L sub-clusters. A test case $t_k \in C_n$ is an element of the l-th sub-cluster $C_{n,l}$ if $check(t_k, \psi_l)$ is true. We originally introduced the terminology of sub-regions in Sec. 4. However, we focus in what follows on discrete valued features, which means that grouping test cases into sub-clusters is more appropriate. By introducing the function $eval : T_{Exec} \rightarrow \{0, 1\}$ that is defined as follows

$$eval(t_i) = \begin{cases} 0, & \text{for } state\{t_i, r_0\} = \text{'Pass'} \quad (2) \\ 1, & \text{for } state\{t_i, r_0\} = \text{'Fail'} \quad (3) \end{cases}$$

the calculation of failure probabilities can be given in Eq. 4.

$$p_{n,l} = \frac{1}{|C_{n,l}|} \sum_{\forall t_i \in C_{n,l}} eval(t_i) \quad (4)$$

The given selection decisions in our example in Sec. 4 (cf. Fig. 1) were taken based on calculated failure probabilities. Additionally, the correlations between the failure probabilities were considered. Accordingly, we need a stochastic model for estimating the classifier's sensitivity and specificity. We propose a univariate and also a multivariate stochastic model. The short-comings of the univariate stochastic model for solving the optimization problem (cf. Eq. 1) will be discussed later to motivate the introduction of a multivariate stochastic model. First of all, the next step introduces a multidimensional Gaussian distribution that constitutes a distribution for the failure probabilities of test cases. Based on this distribution, two distinct Bayesian Belief networks for both stochastic models will be introduced.

In the following, we interpret $p_{n,l}, 1 \leq l \leq L$ as realizations of a random variable X_n. X_n is Gaussian distributed based on the following assumption:

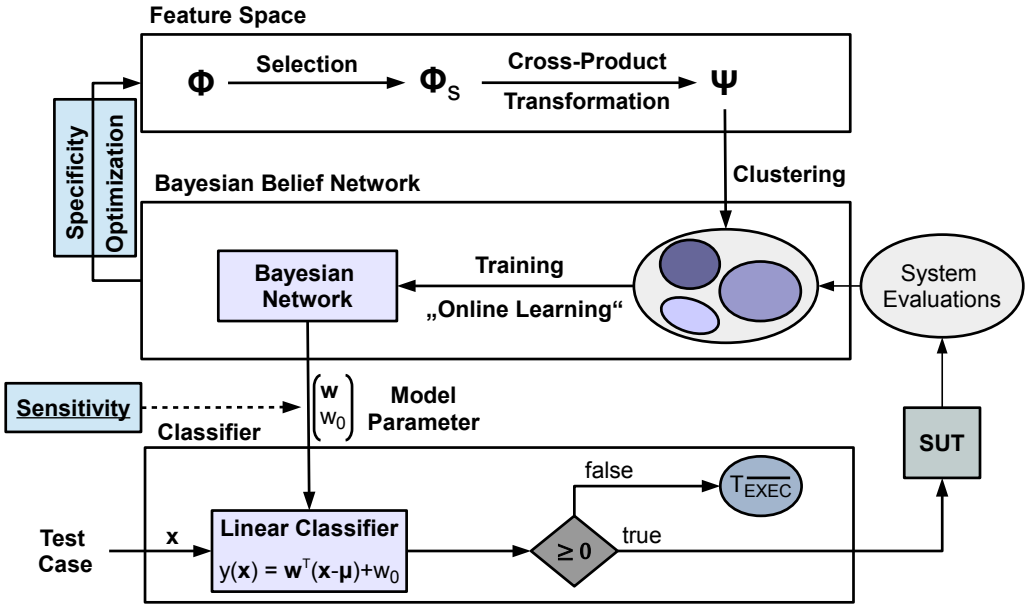


Figure 2: Determining the weights of a linear classifier for maximizing its specificity under the constraint of a specific sensitivity.

Since each test case evaluation is a binary experiment with two possible outcomes ('0' or '1'), it can be regarded as a realization of a binary random variable. As the sum of independent random variables results into a Gaussian random variable according to the central-limit theorem [21], considering test case evaluations as independent random experiments justifies X_n's assumed distribution. However, test cases are executed on the same system and there may be some dependencies between test case evaluations that cannot be directly validated by such means as performing code inspections. As a result, we assume a mix of dependent and independent test case evaluations, and hence the Gaussian assumption is still valid. The moments of X_n are $E[X_n] = \mu_n = \frac{1}{L}\sum_{l=1}^{L} p_{n,l}$ and $E[(X_n - E[X_n])^2] = \sigma_n^2 = \frac{1}{L-1}\sum_{l=1}^{L}(p_{n,l} - E[X_n])^2$. As we introduced in total N Gaussian random variables, the moments of the multidimensional Gaussian distribution are $\boldsymbol{\mu} = E[\boldsymbol{X}] = [E[X_1], E[X_2], ..., E[X_N]]^T$ and $\Sigma = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]$.

Since the constraint of the optimization problem (cf. Eq. 1) has to be fulfilled, an accurate sensitivity estimation has to be iteratively performed.

6.1 Sensitivity Estimation

The formula for calculating the classifier's false-negative selection probability is given in Eq. 5.

$$p_{FN} = \frac{N_{FN}}{N_{FN} + N_{TP}} \leq p_{FN,MAX} \tag{5}$$

However, $\hat{p}_{FN} = \frac{\hat{N}_{FN}}{\hat{N}_{FN}+N_{TP}}$ has to be estimated, since the number of mistakenly deselected failing test cases N_{FN} is unknown, and thus it is estimated by $\hat{N}_{FN}$. The number of already detected failing test cases is given by N_{TP}. Before a decision can be taken on whether a test case t_i can be deselected, the currently allowed risk of taking a wrong decision has to be estimated in advance. The estimation of $\hat{N}_{FN}$ has to be adjusted by the term $xP(\hat{H} = H_0|H_1)$, where x is the failure probability of t_i and $P(\hat{H} = H_0|H_1)$ is the estimated false-negative probability if t_i is deselected. Accordingly, the recursive formulation $\hat{N}_{FN,new} = \hat{N}_{FN,old} + xP(\hat{H} = H_0|H_1)$ is continuously updated whenever an arbitrary test case is deselected. As $\hat{p}_{FN} \leq p_{FN,MAX} \Leftrightarrow \frac{\hat{N}_{FN,new}}{\hat{N}_{FN,new}+N_{TP}} \leq p_{FN,MAX}$ is required, the maximum allowed false-negative probability for deselecting the next test case t_i is given in Eq. 6.

$$\begin{aligned} P(\hat{H} = H_0|H_1) &\leq \frac{\frac{N_{TP} \cdot p_{FN,MAX}}{1-p_{FN,MAX}} - \hat{N}_{FN,old}}{x} \\ &=: p_{FN,Limit} =: \frac{p_{FN,Bound}}{x} \end{aligned} \tag{6}$$

6.2 Univariate Stochastic Model

We model the Bayesian Network that consists of the random variables X, H and $\hat{H}$, in Fig. 3. In the univariate stochastic model, the focus is on modeling of only one failure probability distribution. Thus the random variable X stands for the previously defined X_1 and its realization x is given by $p_{1,l}$ where l is the index of that sub-cluster $C_{1,l}$ that fulfills $t_i \in C_{1,l}$.

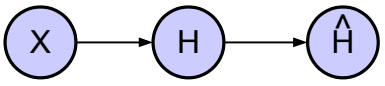


Figure 3: Bayesian Network consisting of the random variables X, H and $\hat{H}$.

H and $\hat{H}$ are binary random variables for modeling

test case states and classifier decisions. As the state of a test case t_i is a-priori unknown, it needs to be modeled by a corresponding random variable. According to the realization of X, a pass or a fail of the corresponding test case t_i, whose failure probability distribution is modeled by X, is expected. Finally, $\hat{H}$ takes a decision for t_i based on its failure probability x:

$$\hat{H}(x) = \begin{cases} H_0, & \text{if } x \in \mathcal{X}_0 = [0; p_{TH}[\quad (7) \\ H_1, & \text{if } x \in \mathcal{X}_1 = [p_{TH}; 1] \quad (8) \end{cases}$$

According to $\hat{H}$'s selection rule a very simple *hyperplane* $y(x) = x - p_{TH}$ is derived where in the case of $y(x) \geq 0$ a selection decision is taken. A particularly important factor is the definition of the threshold probability p_{TH}, as its setting determines the classifier's sensitivity and specificity. The common ways of estimating false-negative and false-positive probabilities are given in equations 9 and 10, respectively.

$$\hat{p}_{FN} = \int_{\mathcal{X}_0} p(X|H_1)dx \quad (9)$$

However, we could only estimate the probability density function (pdf) $p(X)$ in contrast to the conditional probability density functions $p(X|H_0)$ and $p(X|H_1)$. The reason for this is that pdf estimations are based on mean calculations of test case evaluations. Hence passing and failing test cases are both considered in calculating average failure probabilities. Thus, $p(X)$ is a distribution over failure probabilities of passing as well as failing test cases. Accordingly, $p(X|H_0)$ and $p(X|H_1)$ cannot be estimated and in conclusion, $\hat{p}_{FN}$ and $\hat{p}_{FP}$ cannot be estimated as in Eq. 9 and 10.

$$\hat{p}_{FP} = \int_{\mathcal{X}_1} p(X|H_0)dx \quad (10)$$

Fig. 4 shows the important probability distribution functions that are used for estimating $\hat{p}_{FN}$ and $\hat{p}_{FP}$.

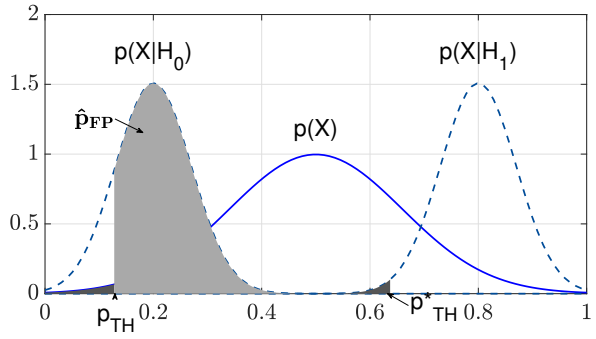


Figure 4: Considered probability distribution functions in the univariate stochastic model.

The threshold probability p_{TH} is calculated based on the estimation of $\hat{p}_{FN}$ as the following relation in Eq. 11 holds.

$$\hat{p}_{FN} = P(X < p^*_{TH}|H_1) \quad (11)$$

As Eq. 11 cannot be directly estimated, the following relation in Eq. 12 is used for estimating $\hat{p}_{FN}$ and finally for p_{TH}.

$$\hat{p}_{FN} = P(X < p_{TH}|H_1) \leq P(X < p_{TH}) \leq p_{FN,Limit} \quad (12)$$

$\hat{p}_{FN}$ is estimated in Eq. 12 according to the assumption that the quantiles of $p(X|H_1)$ are larger than the quantiles of $p(X)$. By solving Eq. 12 the threshold probability is computed as given in Eq. 13

$$p_{TH} = \text{erfinv}(2p_{FN,Limit} - 1)\sigma\sqrt{2} + \mu \quad (13)$$

with $\mu = E[X]$ and $\sigma = \sqrt{VAR(X)}$. As a result, the classifier's sensitivity is larger than $1 - p_{FN,Limit}$, since its false-negative selection probability is smaller than $p_{FN,Limit}$. Finally, the decision regions of the linear classifier are defined (cf. Eq. 7 and 8) by determining p_{TH}.

Furthermore, the minimization of the classifier's specificity is required by the definition of the constraint optimization problem (cf. Eq. 1). Accordingly, the classifier's false-positive selection probability is estimated as given in Eq. 14 and shown in Fig. 4.

$$\hat{p}_{FP} = P(X \geq p_{TH}|H_0) \quad (14)$$

However, $p(X|H_0)$ is not given and thus $\hat{p}_{FP}$ cannot reasonably be estimated. Furthermore, $\hat{p}_{FP}$ cannot be reasonably minimized as an optimization parameter is not defined; consequently, we need a so-called multivariate stochastic model for performing this. In the first instance, the idea of minimizing $\hat{p}_{FP}$ and hence gaining testing efficiency by regarding several distribution functions is explained.

6.3 Preliminaries

Let us assume that two dependent Gaussian random variables X and X' are given. The focus is again on estimating $\hat{p}_{FN}$ and $\hat{p}_{FP}$. Fig. 5 shows the probability distribution functions $p(X)$, $p(X|H_0)$ and $p(X|H_1)$ as in Fig. 4. Additionally, the a-posteriori failure probability distribution function $p(X|X')$ is shown.

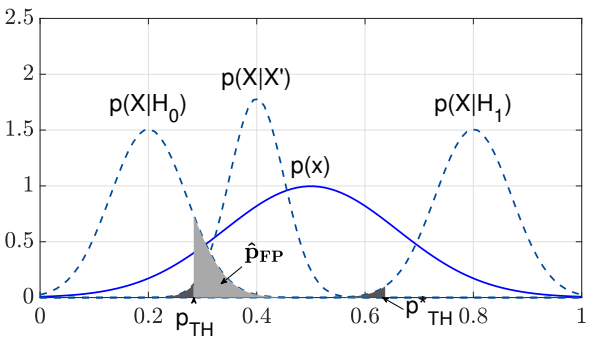


Figure 5: Qualitatively minimizing $\hat{p}_{FP}$ by introducing $p(X|X')$ and hence by using conditional informations.

The main idea is to use several observations of distinct dependent random variables to achieve a consid-

erably more representative a-posteriori failure probability distribution function that is relatively narrow within a certain range. So $p(X|X')$ is considered as the more representative distribution for the failure probabilities and hence $\hat{p}_{FN}$ and p_{TH} are estimated by using this distribution function. Comparing Figs. 4 and 5, it can easily be seen that $\hat{p}_{FP}$ is basically minimized, since the risk of a false-negative selection probability is computed based on $p(X|X')$, which allows a more representative risk estimation.

All in all, by regarding a set of dependent Gaussian random variables and by using the information about their observations, a more representative a-posteriori failure probability distribution function is achieved, which allows a more precise risk estimation. Accordingly, the probability of false-positive selection can be minimized. As a result, a multivariate stochastic model is created to exploit the dependency information between random variables for finally achieving testing efficiency.

6.4 Multivariate Stochastic Model

By using the dependency between the random variables $X_n, 1 \leq n \leq N$, a considerably more accurate estimation of $\hat{p}_{FN}$ is achieved and hence $\hat{p}_{FP}$ is minimized. Fig. 6 shows the modeled Bayesian network consisting of the random variables $X_n, 1 \leq n \leq N$, H and $\hat{H}$.

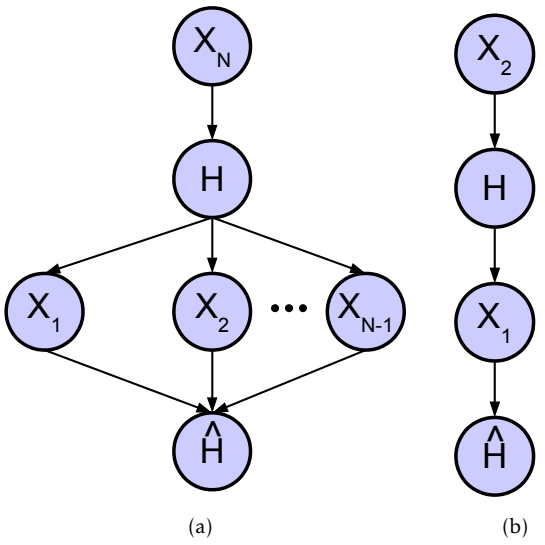


Figure 6: Bayesian Network consisting of an (a) indefinite and (b) definite number of random variables $X_n, 1 \leq n \leq N$.

The focus is on taking a selection decision for an arbitrary test case t_i. X_N now models the failure probability distribution of t_i. In our previous example, as shown in Fig. 1 c), the failure probability distribution of test case t_4 was calculated based on the empirically evaluated failure probabilities of test cases inside Fig. 1 c). Thus t_4 was an element of C_2 ($N = 2$), and its failure probability was modeled by X_2. Analogously, X_1 was defined by the empirically evaluated failure probabilities of test cases inside Fig. 1 b). In the interest of simplification, we always assume that the currently focused test case t_i is an element of cluster C_N and thus X_N models its failure probability distribution.

Furthermore, we can calculate the dependency among $X_n, 1 \leq n \leq N$. However, the Bayesian network in Fig. 6 models the statistical dependency between H and further random variables $X_n, 1 \leq n \leq N-1$. These dependencies cannot be calculated, but have to be modeled for estimating $\hat{p}_{FN}$ and $\hat{p}_{FP}$.

First of all, we model the classifier $\hat{H}(\cdot)$ as follows

$$\hat{H}(x_{ML}) = \begin{cases} H_0, & \text{if } x_{ML} \in \mathcal{X}_0 = [0; p_{TH}[\quad (15) \\ H_1, & \text{if } x_{ML} \in \mathcal{X}_1 = [p_{TH}; 1] \quad (16) \end{cases}$$

where $x_{ML} = \underset{x_N}{\operatorname{argmax}}\Big[\ln[\mathcal{L}(x_N|x_1,...,x_{N-1})]\Big]$ (consult [1]) is the maximum likelihood estimation. Accordingly, the likelihood estimation is a weighted sum as given in Eq. 17

$$x_{ML} =: \sum_{n=1}^{N-1} w_n(x_n - \mu_n) + \mu_N \tag{17}$$

with weights $w_n, 1 \leq n \leq N-1$ as given in Eq. 18.

$$w_n = -\frac{(\Sigma^{-1})_{n,N}}{(\Sigma^{-1})_{N,N}} \tag{18}$$

Further, p_{TH} has to be calculated based on a precise estimation of $\hat{p}_{FN}$. Thus we derive a calculation formula for $\hat{p}_{FN}$ for the case $N = 2$, but we will also provide a general calculation formula of $\hat{p}_{FN}$ for an arbitrary number N of random variables.

6.4.1 Derivation of probability distribution functions

In the following, some probability distributions are driven that are used for estimating $\hat{p}_{FN}$. First of all, the joint pdf $p(\hat{H}X_1HX_2)$

$$p(\hat{H}X_1HX_2) = p(\hat{H}|X_1)p(X_1|H)p(H|X_2)p(X_2) \tag{19}$$

and the conditional pdf $p(\hat{H}X_1H|X_2)$ are given in Eq. 19 and 20, respectively.

$$\begin{aligned} p(\hat{H}X_1H|X_2) &= \frac{p(\hat{H}X_1HX_2)}{p(X_2)} \\ &= \underbrace{p(\hat{H}|X_1)p(X_1|H)}_{p(\hat{H}X_1|H)}p(H|X_2) \end{aligned} \tag{20}$$

In the next step Eq. 21 is obtained by setting the equation $p(\hat{H}|X_1)p(X_1|H) = p(\hat{H}X_1|H)$ into Eq. 20.

$$p(\hat{H}X_1|H) = \frac{p(\hat{H}X_1H|X_2)}{p(H|X_2)} \tag{21}$$

Furthermore, the most important relation $p(\hat{H}|H) \leq \frac{p(\hat{H}|X_2)}{p(H|X_2)}$ is driven in Eq. 22.

$$p(\hat{H}X_1|H) \leq p(\hat{H}|H) \leq \frac{p(\hat{H}H|X_2)}{p(H|X_2)} \leq \frac{p(\hat{H}|X_2)}{p(H|X_2)} \tag{22}$$

Thus, the probability calculation $P(\hat{H} = H_0|H_1)$ can be estimated by using the relation in Eq. 22 as given in Eq. 23.

$$\hat{p}_{FN} = P(\hat{H} = H_0|H_1) \leq \frac{P(\hat{H} = H_0|X_2 = x_2)}{P(H_1|X_2 = x_2)} \tag{23}$$

Since $\hat{p}_{FN}$ cannot be directly estimated, as the conditional pdf $p(\hat{H}|H)$ is not given for performing the probability calculation $P(\hat{H} = H_0|H_1)$, the relation in Eq. 23 is used for estimating an upper bound for $\hat{p}_{FN}$. However, the linear classifier's actual false-negative deselection probability would be smaller than the calculated upper bound.

Since the constraint in Eq. 24 has to be fulfilled,

$$P(\hat{H} = H_0|H_1) \leq p_{FN,Limit} \tag{24}$$

we solve the inequality in Eq. 25.

$$\frac{P(\hat{H} = H_0|X_2 = x_2)}{P(H_1|X_2 = x_2)} \leq p_{FN,Limit} \tag{25}$$

As $x_2 = P(H_1|X_2 = x_2)$ holds, the following inequality is finally solved.

$$P(\hat{H} = H_0|X_2 = x_2) \leq p_{FN,Bound} \tag{26}$$

Eq. 26 is driven for the case $N = 2$ but in the general case, where the number of random variables $X_n, 1 \leq n \leq N$ is given by an arbitrary N, the following inequality has to be solved.

$$P(\hat{H} = H_0|X_N = x_N) \leq p_{FN,Bound} \tag{27}$$

By solving Eq. 27, the threshold probability

$$p_{TH} = -w_N(x_N - \mu_N) - w_0 + \mu_N \tag{28}$$

is obtained with weights

$$w_N = -\frac{e^{2I} - 1}{e^{2I}} \tag{29}$$

and

$$w_0 = -\frac{\sqrt{2}\sigma_N\sqrt{e^{2I} - 1}\,\mathrm{erfinv}(2p_{FN,Bound} - 1)}{e^{2I}} \tag{30}$$

Thus, the differential mutual information is defined in Eq. 31.

$$\mathcal{I} := \mathcal{I}(X_1, ..., X_{N-1}; X_N) \tag{31}$$

6.4.2 Conditional Independence

We have already motivated and introduced the following dependent random variables $X_n, 1 \leq n \leq N$. We have explained the fact that test case failure probabilities are correlated, since test cases are executed on the same system, and thus they show a dependent behavior.

However, the random variables $X_n, 1 \leq n \leq N$ are conditionally independent. This means that the information about a test case evaluation dominates such that a test case's originally calculated failure probability becomes irrelevant after observation of its state. Accordingly, the dependency among failure probabilities vanishes after observation of test case evaluations. This means that a fail of a test case t_m is actually expected based on the information about the evaluation of another test case t_n and no longer on t_n's originally calculated failure probability. Thus the remaining random variables $X_n, 1 \leq n \leq N-1$ become independent of the random variable X_N after observation of H's realization (cf. Fig. 6).

6.4.3 Specificity Estimation

The specificity is given by the term $1 - \hat{p}_{FP}$. As extensive mathematical derivations are needed for obtaining a calculation formula of $\hat{p}_{FP}$, these derivation steps are given in the appendix and in what follows here only the result is given.

Theorem 1 (False-Positive Probability Estimation)**.** *$\hat{p}_{FP}$ is estimated as given in Eq. 32*

$$\hat{p}_{FP} = P(Z \geq -w_0 + \boldsymbol{w}^T \Delta\boldsymbol{\mu}) = \frac{1}{2}\left[1 - erf\left(\frac{-w_0 + \boldsymbol{w}^T \Delta\boldsymbol{\mu}}{\sqrt{2}\sigma_Z}\right)\right] \tag{32}$$

with $\sigma_Z = [w_1 \cdots w_{N-1}]\Sigma_{1,1}[w_1 \cdots w_{N-1}]^T + w_N^2\sigma_N^2$ and $\Delta\boldsymbol{\mu} = \boldsymbol{\mu} - \boldsymbol{\mu}_{H_0}$. The conditional moment has the following definition $E[\boldsymbol{X}|H_0] = \boldsymbol{\mu}_{H_0}$

For the case $N = 2$, Eq. 32 can be simplified; after several calculation steps the following Eq. 54 results

$$\hat{p}_{FP} = \frac{1}{2}\left[1 - erf\left(\psi\right)\right] \tag{33}$$

with

$$\psi = \frac{\sqrt{e^{2\mathcal{I}} - 1}\,erfinv(2p_{FN,Bound} - 1) + \frac{\sqrt{e^{4\mathcal{I}} - e^{2\mathcal{I}}}}{\sigma_1\sqrt{2}}\Delta\mu_1 - \frac{\sqrt{e^{2\mathcal{I}} - 1}}{\sigma_2\sqrt{2}}\Delta\mu_2}{\sqrt{2e^{4\mathcal{I}} - 3e^{2\mathcal{I}} + 1}} \tag{34}$$

6.4.4 Small Dimension Validation

Fig. 7 shows five plots of $\hat{p}_{FP}$ for different values of displacements $\Delta = \mu_n - \mu_{n,H_0}$. Indeed, the actual value of Δ is unknown. However, the focus is on the minimization of $\hat{p}_{FP}$. Accordingly, $\hat{p}_{FP}$ decreases in each sub-figure of Fig. 7. The actual value of Δ only determines how fast $\hat{p}_{FP}$ decreases. So we can solve the optimization problem (cf. Eq. 1) by minimizing $\hat{p}_{FP}$. We considered two random variables X_1 and X_2, as in our example in Sec. 4 where we created two clusters. A very important factor here is the underlying strategy for clustering test cases. As the distribution of the random variables X_1 and X_2 is directly related to the clustering strategy the main focus is on the maximization of the differential mutual information $\mathcal{I}(X_1; X_2)$. Accordingly, $\mathcal{I}$ is an optimization parameter for effectively reducing $\hat{p}_{FP}$. Lastly, we chose the

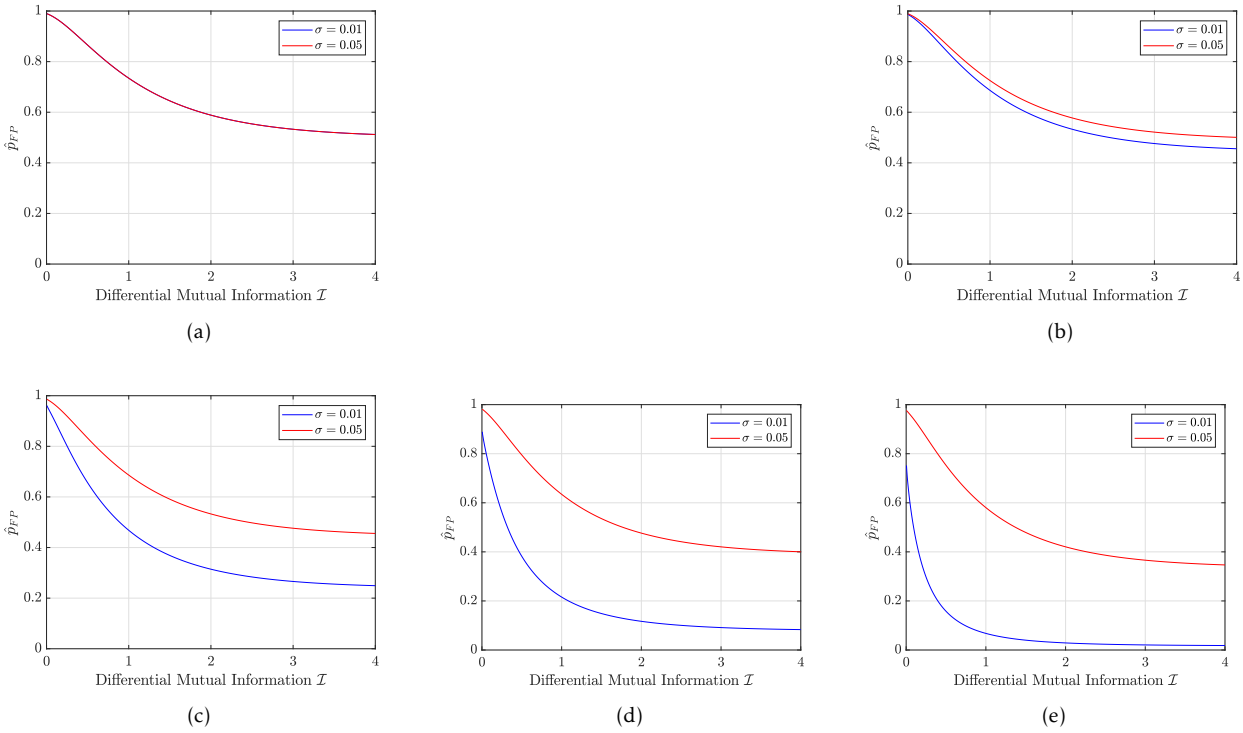


Figure 7: Estimation of $\hat{p}_{FP}$ for different values of $\boldsymbol{\Delta}$ and σ with $\boldsymbol{\Delta} = \alpha \cdot \boldsymbol{v} \cdot 10^{-3}$. The value of α is 0, 1, 5, 10 and 15 in Fig. 7 a), Fig. 7 b), Fig. 7 c), Fig. 7 d) and Fig. 7 e) respectively.

following values $\{0.01, 0.05\}$ for $\sigma = \sigma_1 = \sigma_2$ and selected the following displacements $\boldsymbol{\Delta} = \alpha \cdot \boldsymbol{v} \cdot 10^{-3}$ with $\alpha \in \{0; 1; 5; 10; 15\}$ and $\boldsymbol{v} = [1, -1]^T$.

7 Optimization

The first strategy is to optimize the feature selection. Optimal features are learned in an unsupervised learning session where an evolutionary optimization framework is applied to search for optimal features. The next strategy is to improve the labeling of test cases through an active learning strategy.

7.1 Evolutionary Optimization

Clustering (and sub-clustering) of test cases is performed based on features. Therefore, different clusterings for different selections of feature subsets (Φ_m, Φ_s) are possible. Accordingly, a different statistical model is obtained, as it reflects the failure frequencies in clusters. Furthermore, the differential mutual information (cf. Eq. 31) depends on the statistical dependencies and thus changes for different clusterings.

Sec. 6 proposed a calculation formula for the weights $w_n, 0 \leq n \leq N$, of a linear classifier. However, those formulas still depend on the differential mutual information I. A desired sensitivity has to be guaranteed, and thus the hyperplane is adjusted according to the value of I. It can be shown that for small values of I, the position of the hyperplane still guarantees a desired sensitivity but the false-positive selection probability increases. To minimize the false-positive selection probability, the differential mutual information has to be maximized, which is the final strategy for solving the constrained optimization problem (cf. Eq. 1).

First, clustering depends on the history-tuples of test cases as, for example, the length of the history-tuples determines the maximum number $|S|^K$ of clusters. Second, feature selection is optimized. All in all, we have summarized that K (number of considered previous regressions) is an optimization parameter and $\boldsymbol{b}_s$ (for coding selected and main features) is an optimization matrix. However, this is a large-scale high dimensional optimization problem, as there exist many possible settings for K and $\boldsymbol{b}_s$. Thus, [3] and [4] suggest that the high dimensional optimization problem can be solved in a reasonable time by using evolutionary algorithms. Accordingly, an evolutionary optimization framework is applied for solving the mentioned high dimensional optimization problem. As each setting for K and $\boldsymbol{b}_s$ is one possible solution for clustering test cases, which is the basis for derivation of a stochastic model, the fitness of this solution can be evaluated by calculating the extracted information $\mathcal{I}$ in Eq. 31. Thus, the optimal parameter and matrix setting with the best fitness will survive and will be returned by the evolutionary optimization algorithm.

Fig. 8 shows the overall flow chart of the evolutionary optimization framework. First of all, a new population consisting of several genotypes is initialized. Each genotype stands for a possible setting of K and $\boldsymbol{b}_s$. In the next step, the corresponding phenotypes of the genotypes are derived. Hence each phe-

notype encodes a stochastic model. Accordingly, the population is evaluated, wherein the fitness of each phenotype is calculated. However, a *bad* fitness is also possible due to *bad* statistical properties of the underlying stochastic model. This means that statistical calculations based on the stochastic model that a phenotype encodes cannot guarantee desired statistical confidence bounds. This will be explained in more detail in Sec. 8. Those phenotypes with *bad* fitness cannot survive and hence are eliminated.

Accordingly, remaining genotypes (phenotypes) are stochastically selected, and successively new genotypes are generated due to *crossover* and *mutation* operations. After a certain number of iterations, the phenotype with the best fitness will be selected, and this will be used in the selection algorithm. However, if the population is empty since all phenotypes were of *bad* fitness, then the training mode is activated, in which test cases are still executed without running the selection algorithm.

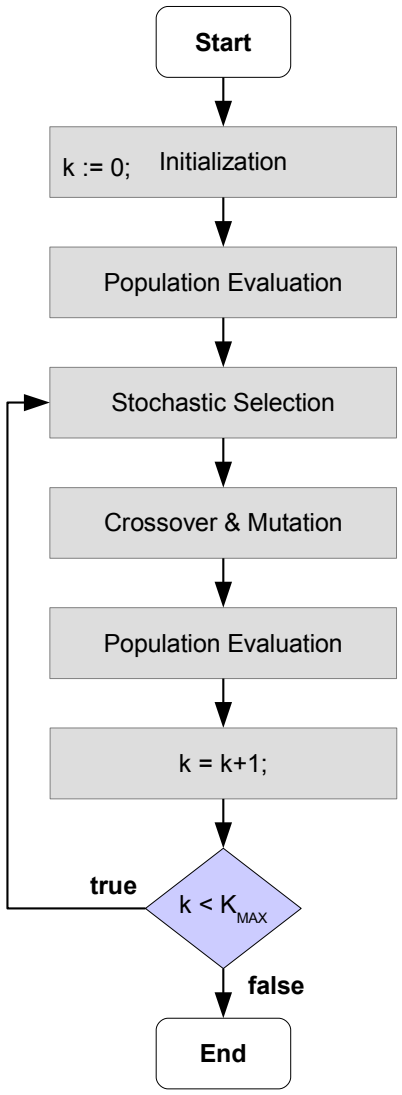


Figure 8: Evolutionary Optimization Framework.

7.2 Active Learning

Our classifier's conducted decisions can be regarded as *hard* or even as *soft* decisions. Once taken, *hard* decisions are never changed later on, in contrast to *soft* decisions. The test efficiency can be significantly increased by conducting *soft* decisions as opposed to *hard* decisions.

7.2.1 Hard Decision

[1] performs *hard decisions* since a selected test case is automatically executed and a once deselected test case is never selected again in the current regression test. In the following passages, the disadvantage of conducting *hard decisions* will be explained in detail in relation to classifier decisions.

The linear classifier's decision depends on the current estimation of $\hat{N}_{FN}$ as it calculates the allowed residual risk $p_{FN,Limit}$ (cf. Eq 6) of potentially taking a wrong decision. $\hat{N}_{FN}$ returns the number of supposedly unrevealed system failures that would be detected by those already deselected test cases that are elements of $T_{\overline{Exec}}$. Accordingly, the linear classifier's decision depends on the decisions it has already taken ($T_{\overline{Exec}}$) and hence it is memory driven.

Each deselected test case t_j has an individual additional contribution $\hat{N}_{FN,j}$ (cf. Eq. 36) to the overall estimation $\hat{N}_{FN}$ such that the relation in Eq. 35 holds.

$$\hat{N}_{FN} = \sum_{\forall t_j \in T_{\overline{Exec}}} \hat{N}_{FN,j} \tag{35}$$

$\hat{N}_{FN,j}$ is the product of t_j's failure probability x and the false-negative probability $P(\hat{H} = H_0|H_1)$ by deselecting t_j as given in Eq. 36.

$$\hat{N}_{FN,j} = P(H_1)P(\hat{H} = H_0|H_1) \tag{36}$$

Because of this fact, a deselection of an arbitrary test case can cause that the residual risk $p_{FN,Limit}$ reaches zero as $\hat{N}_{FN}$ increases (cf. Eq. 6). This means that no more risk ($p_{FN,Limit} = 0$) is allowed, and all remaining test cases have to be consequently selected.

Indeed, selecting test cases even if their deselection is allowed according to risk calculations is sometimes the better choice. In fact, this is the case if $p_{FN,Limit}$ is zero and thus it can be significantly increased by selecting and executing an already deselected test case in order to eliminate its risk. When this is done, $\hat{N}_{FN}$ decreases and hence $p_{FN,Limit}$ increases and thus a residual risk for further deselections is obtained.

However, the amount $\Delta\hat{N}_{FN}$ of how much $\hat{N}_{FN}$ can be decreased by selecting an arbitrary test case is significant. If later more than one test case can be deselected, and these deselected test cases add the same amount of expected unrevealed system failures $\Delta\hat{N}_{FN}$ to $\hat{N}_{FN}$ is in fact a gain in terms of reducing the regression test effort. So the strategy is to deselect primarily those test cases with fewer failure probabilities in order to increase testing efficiency.

As a result, the regression test efficiency can be increased. Therefore, the proposed selection method [1] is extended by a *soft decision* methodology. So each decision for deselecting a test case is now regarded as a *soft decision* that might be changed later. (We note here that the other way round is impossible since an already selected test case is automatically executed on the system under test and hence deselecting it later does not make sense).

7.2.2 Soft Decision

Fig. 9 shows the logic for managing soft selection decisions: Let us assume that t_i is the next test case that is analyzed by the linear classifier. If t_i is deselected, then it is queued into a priority queue whereby its priority is calculated as given in Eq. 37.

$$prio(t_i) = \hat{N}_{FN,i} = P(H_1)P(\hat{H} = H_0|H_1) \qquad (37)$$

In the other case, if t_i is selected then test cases deselected up to this point are analyzed to the end of improving the trade-off between the assumed risk and the total number of deselected test cases. As a consequence, the most probable failing test case t_j is obtained by taking the peek-operation on the priority queue. The priority of t_i and t_j is compared, and the test case with the higher priority is selected and executed on the system under test.

If t_j is executed, then it is removed from the set $T_{\overline{Exec}} \leftarrow T_{\overline{Exec}} \backslash t_j$ and added into the set $T_{Exec} \leftarrow T_{Exec} \cup t_j$. Furthermore, t_j's state is evaluated $eval(t_j)$ and accordingly the empirical failure probabilities of test cases are updated in algorithm 1. Since the calculated failure probabilities are averages of test case evaluations, the failure probabilities of those sub-clusters (see Eq. 4) have to be updated where t_j is an element of them. Accordingly, the failure probabilities of $\forall t_k \in T_{\overline{Exec}}$ are updated in algorithm 1.

Algorithm 1 Test case selection algorithm

```
procedure UPDATE STATISTICS(T_Exec_bar, t_j)        ▷ T_Exec_bar contains already deselected test cases; t_j is executed
    for each test case t_k ∈ T_Exec_bar do
        ∃!C_{n,l} ⟹ t_k ∈ C_{n,l}        ▷ Find sub-cluster of t_k and thus determine n and l
        if t_j ∈ C_{n,l} then
            p_{n,l} ← 1/|C_{n,l}| Σ_{∀t_i∈C_{n,l}} eval(t_i)        ▷ see Eq. 4
            P(H_1) ← p_{n,l}
            update t_k's priority: prio(t_k)        ▷ see Eq. 37
        end if
    end for
    if eval(t_j) == 1 then
        N_TP ← N_TP + 1
    end if
end procedure
```

The important point is that even the failure probability of t_i is computed again. In most cases, t_i would be deselected. Nevertheless, it could be possible that the execution of t_j has failed, such that a further system failure has been found. In such a case, even t_i's failure probability may have increased such that its deselection has to be checked again by the linear classifier.

All in all, testing efficiency can be significantly increased by performing *soft selection* decisions. The performance of both selection strategies (*hard decision* and *soft decision*) will be compared in Sec. 9.

8 Learning Phase

The learning phase is of essential importance due to the fact that during this phase, the system reliability is actually learned. Test case selection is a safety-critical binary classification task as probably system failures would remain undetected and hence, a corresponding quality measure of wrong decisions is required. Accordingly, risk estimations on probably undetected system failures due to deselection of test cases have to be as accurate as possible. The more the system is learned during a regression test, the more precise the risk estimations are. However, learning a system in terms of understanding its reliability is a costly process, as it requires test cases to be executed. The fundamentally important research question is how much training data is enough for safely selecting test cases with a desired sensitivity.

8.1 Statistical Sensitivity Estimation

We have already required a specific sensitivity in the constraint optimization problem (cf. Eq. 1). Accordingly, we define the following confidence level in Eq. 38, which is basically driven from the constraint of Eq. 1.

$$P\Big(\Psi \leq \gamma\Big) \geq 1 - \alpha \qquad (38)$$

Ψ is an estimator for the number of false-negatives $\hat{N}_{FN} = \sum_{t_i \in T_{\overline{EXEC}}} \hat{N}_{FN,i}$ and the bound is given as $\gamma = \frac{N_{TP} \cdot p_{FN,MAX}}{1 - p_{FN,MAX}}$. $\Psi = \sum_i \psi_i$ is composed of several random variables ψ_i standing for the distribution of each $\hat{N}_{FN,i}$. ψ_i's distribution is complex, since the individual contribution of a deselected test case t_i is given by $\hat{N}_{FN,i} = x_N \hat{p}_{FN}$ where x_N is t_i's failure probability and $\hat{p}_{FN}$ is the corresponding estimated false-negative probability: The following theorem is already proved in [1] and gives the formula for the false-negative probability estimation.

Theorem 2 (False-Negative Probability)**.** *For a given p_{th} the calculation formula of the false-negative probability $P(\hat{H} = H_0|H_1)$ has the form*

$$\hat{p}_{FN} = \frac{1}{2}\left(1 + erf\left(\frac{1}{\sqrt{2}} \frac{x_N - (x_N - p_{th})e^{2\mathcal{I}} - \mu_N}{\sigma_N \sqrt{e^{2\mathcal{I}} - 1}}\right)\right) \qquad (39)$$

where x_N is the failure probability of a test case, whereas μ_N, σ_N are parameters of the probability distribution function $\mathcal{N}(\mu_N, \sigma_N)$.

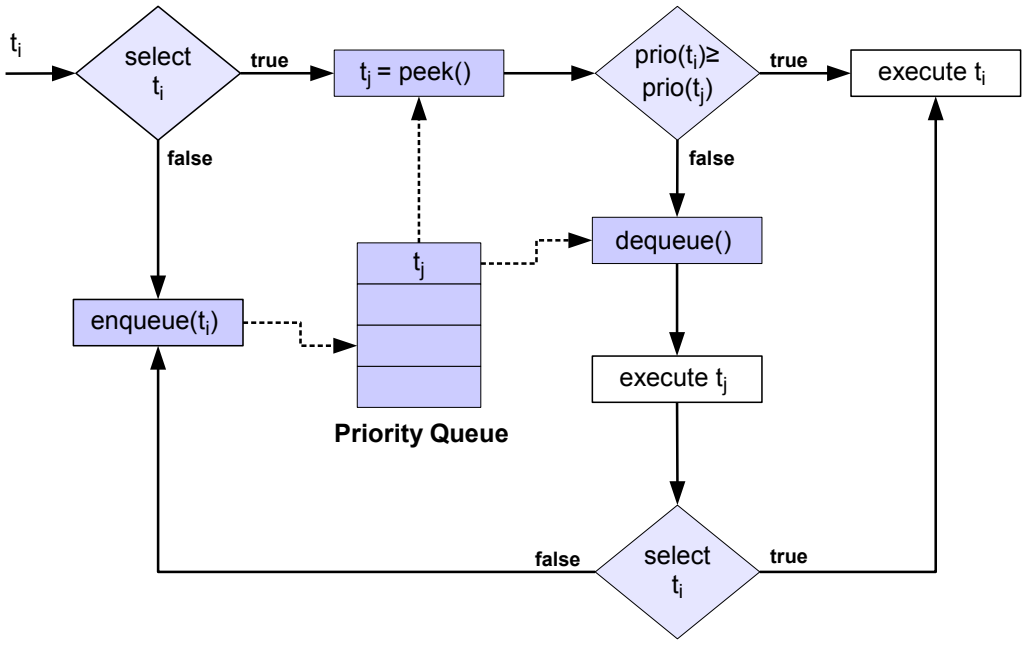


Figure 9: The extended concept for conducting soft decisions

These statistics are created on the basis of sample averages, such that the consideration of sample variances becomes inevitable during the estimation of confidence intervals. For instance, the false-negative probability estimation $\hat{p}_{FN}$ (see Eq. 39) is composed of the statistics x_N, μ_N, σ_N and I. Therefore, its variance depends on the individual variances of each statistic, including the sample variance of x_N. Especially the probability distribution of I is complex as it is nonlinearly composed of a set of multivariate distributed Gaussian random variables.

Therefore, we choose the following approach to solving Eq. 38. We simplify the definition of ψ_i as follows $\psi_i = \hat{p}_{FN} \cdot X_N$ where $\hat{p}_{FN}$ is assumed to be a constant value without any distribution. This step simplifies the calculation complexity of Eq. 38 significantly, as Ψ becomes simply a weighted sum of Gaussian random variables. However, the variance of $\hat{p}_{FN}$ is of course relevant and should not be easily neglected. Accordingly, we require a maximum confidence interval width for $\hat{p}_{FN}$ such that the estimated false-negative probability is quite accurate and hence can be assumed to be just like a constant value without any statistical deviation. We calculate the confidence interval $[\hat{p}_{FN}^{(l)}; \hat{p}_{FN}^{(u)}]$ and its width $\delta = \hat{p}_{FN}^{(u)} - \hat{p}_{FN}^{(l)}$ and require a maximal confidence interval width of δ_{max}.

The Wilson score interval [22] delivers confidence bounds for binomial proportions. Therefore, we calculate the following confidence intervals $[x_n^{(l)}; x_n^{(u)}]$ (confidence level: $1 - \alpha = 99\%$) for each failure probability estimation $x_n, 1 \leq n \leq N$. Each bound is consequently used for building the bounds of the composed statistics σ, Σ and I. By doing this, we obtain the following bounds: $\sigma^{(b)}$, $\Sigma^{(b)}$ and $I^{(b)}$ with $b = \{u, l\}$. Accordingly, we calculate $\hat{p}_{FN}^{(b)}$ by consequently inserting the bounds $\sigma^{(b)}$, $\Sigma^{(b)}$ and $I^{(b)}$ for the statistics σ, Σ and I respectively.

8.2 Criteria for Training

In order to guarantee a statistical bound on the sensitivity with a 99% confidence level, the following conditions have to be checked.

1. $\delta \leq \delta_{max}$
2. $P\left(\Psi \leq \gamma\right) \geq 1 - \alpha = 99\%$

If both conditions are fulfilled, then these risk calculations in the selection algorithm are reasonably accurate and hence selection decisions can be performed. However, if one condition is not fulfilled then the training mode is just active, such that test cases still have to be executed.

9 Industrial Case Study

A German premium car manufacturer constitutes each regression test as being a system release test, and thus the system test takes up to several weeks according to [5]. However, a first detected system failure makes a system release impossible so optimizing the current regression for achieving high efficiency in reducing the regression effort becomes justified.

It is often the case that close to the so-called start of production (SOP) of a vehicle, many electronic control units (ECU) have only some critical spots and thus each regression test is expected to be a system-release test. Since many test cases pass, a lot of time is spent in observing passing test cases. Therefore, reducing the number of executed passed test cases (since a system failure is detected and a system release

is no longer possible) and keeping the limited testing time back for fault-revealing test cases decreases the regression test effort significantly. In any case, a final regression test will succeed after further system updates have been conducted; this will be constituted as a final release-test that meets the high-quality standards of [5].

In our industrial case study, we applied our selection method to a production-ready controller that implements complex networked functionalities for the protection of passengers and other road users. Therefore its test effort is immense, and hence we apply our regression test selection method for accelerating its testing phase. In Fig. 10 the right-hand side of the well-known *V-Model* (see [23]) is shown, whereas the focus is on system testing in our case study.

A hardware-in-the-loop simulator (HiL) [24] is used for validating an ECU's networked complex functionalities as well as its I/O-interaction and its robustness during voltage drops, as it provides an effective platform for testing complex real-time embedded systems.

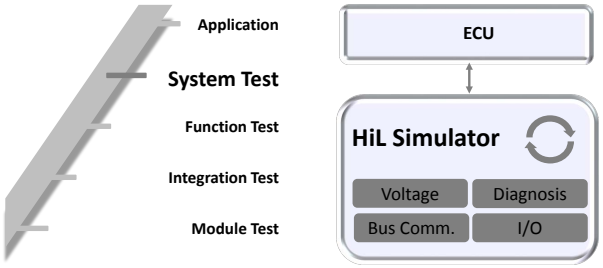


Figure 10: A HiL simulator is used for performing the system test.

Further, we selected for the following test case features for training the machine learning algorithm:

- Name of verified system parts
- Name of a function for which reliability is assured
- Number of totally involved electronic control units during the testing of a networked functionality
- Error type (broken wire etc.) in hardware robustness tests
- Set of checked diagnostic trouble codes, *DTC*s
- Number of checked diagnostic trouble codes, *DTC*s

Since the quality of our selection decisions is hedged on a stochastic level, it can appear that during different runs of our selection method, a statistical deviation of the false-positive probabilities could occur. Therefore, we constitute several independent runs of a regression test, where we set $p_{FN,MAX} = 1\%$. The boxplots and the quantiles of the false-positive probabilities are given in Fig. 11 and in Table 4, respectively.

Fig. 11 shows the overall boxplots of the false-positive probabilities achieved during the regression test replications. To compare the *hard* with the *soft decision* strategy we performed distinct regression test replications where we disabled and enabled the parameter for *'soft decision'*, respectively.

It can be seen from Fig. 11a) and Fig. 11b) that the average false-positive probability is about 74% and 23% for *hard* and enabled *soft decisions* respectively. As already mentioned, conducting *hard decisions* does not allow for global optimization of the trade-off between an already assumed risk and the corresponding number of totally deselected test cases. Global optimization hence requires the analysis of all test cases deselected thus far over and over again, and, if necessary, the selection of an already deselected test case. Therefore, test cases with a higher failure probability should be considered again for eventual selection in an ongoing regression test in order to potentially deselect further less risky test cases. As a result, the regression test effort can be reduced much more by applying *soft decisions*.

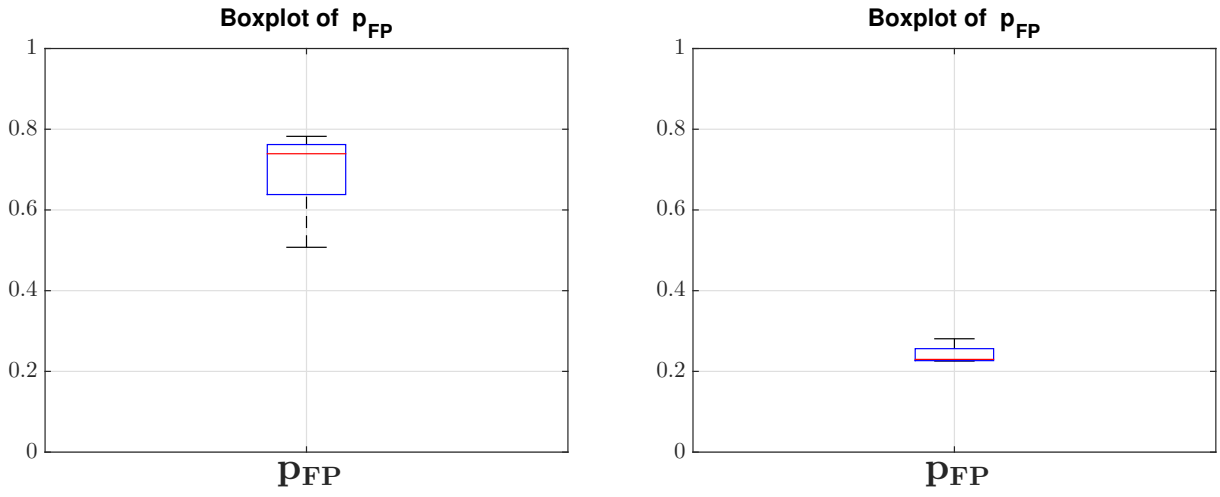
Furthermore, the condition in Eq. 40 on the false-negative probability p_{FN} or on the number of actually occurring false-negatives N_{FN} was fulfilled in all conducted regression test replications.

$$\begin{aligned} &p_{FN} \leq p_{FN,MAX} = 1\% \text{ or} \\ &N_{FN} \leq 1 \end{aligned} \tag{40}$$

Our implemented algorithm for selecting test cases runs on a desktop CPU that is specified in Table 5. We decided to conduct a multithreaded execution of the evolutionary algorithm such that the fitness of all phenotypes in a population is computed in a multithreaded manner (in total 32 threads). Thus the average CPU load is approximately 95% and the maximum memory allocation is about 4GB. We need a mean analysis time of 0.9s for deciding whether a test case should be selected or not.

10 Conclusion and Future Work

We proposed a holistic optimization framework for the safety assessment of systems during regression testing. To this end, we designed a linear classifier for (de-) selecting test cases according to a classification due to a risk-associated recognition. Therefore we defined an optimization problem, since the classifier's specificity has to be maximized whereas its sensitivity still has to exceed a certain threshold $1 - p_{FN,MAX}$. Accordingly, we developed a novel method for determining the weights of a linear classifier that solves the above optimization problem. We have theoretically shown that the classifier performance is directly interrelated with the success of selected relevant features of test cases. Lastly, we applied our method to a production-ready controller and analyzed the overall regression test effort subject to an active learning strategy. We have demonstrated that, in the regression testing of safety-critical systems, significant sav-



(a) hard decision

(b) soft decision

Figure 11: Boxplots of false-positive probabilities in case of a) hard and b) soft decisions

Table 4: Quantiles of the false-positive probabilities in each replicated regression test

	Quantiles of False-Positive Probability				
	0.025	0.25	0.5	0.75	0.975
Hard Decision	50.8%	63.8%	73.9%	76.2%	78.2%
Soft Decision	22.6%	22.7%	23%	25.6%	28.1%

ings can be achieved. As feature selection is a complex task, and thus an evolutionary optimization supposedly finds local optima, more thorough research in this field may indeed allow higher-order reductions of the classifier's false-positive selection probability.

11 Appendix

In the following, a detailed proof of Theorem 1 is given, relating to the proofs given in [1].

Proof of Theorem 1

Proof. According to [1], the maximum likelihood estimation $x_{N,ML}$ (abbreviated x_{ML} in the following) is given in Eq. 41.

$$x_{ML} = \mu_N - \frac{(\boldsymbol{x}_D - \boldsymbol{\mu}_D)^T \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T}{\boldsymbol{\phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T} \tag{41}$$

As $\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T = P_1\Sigma^{-1}P_2^T$ and $\boldsymbol{\phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T = P_2\Sigma^{-1}P_2^T = \left(\Sigma^{-1}\right)_{N,N}$ holds (consult Proof of Theorem 3 in [1]) x_{ML} can be written as given in Eq. 42.

$$x_{ML} =: \sum_{n=1}^{N-1} w_n(x_n - \mu_n) + \mu_N \tag{42}$$

with $w_n = -\frac{\left(\Sigma^{-1}\right)_{n,N}}{\left(\Sigma^{-1}\right)_{N,N}}$. Furthermore, the threshold probability p_{th} was calculated in [1] as given in Eq. 43.

$$p_{th} = x_N + \frac{\sqrt{2}\sigma_N\sqrt{e^{2\mathcal{I}}-1}\,\mathrm{erfinv}(2p_{FN,Bound}-1) + \mu_N - x_N}{e^{2\mathcal{I}}} \tag{43}$$

By introducing the definitions of the weights $w_0 := -\frac{\sqrt{2}\sigma_N\sqrt{e^{2\mathcal{I}}-1}\,\mathrm{erfinv}(2p_{FN,Limit}-1)}{e^{2\mathcal{I}}}$ and $w_N := -\frac{e^{2\mathcal{I}}-1}{e^{2\mathcal{I}}}$ the threshold probability can be written as $p_{TH} = -w_0 + \mu_N - w_N(x_N - \mu_N)$.

Eq. 44 derives the final definition of the hyperplane $y(\boldsymbol{x})$ and the acceptance region of the rival hypothesis H_1 by putting the definitions of x_{ML} and p_{TH} together.

$$\begin{aligned} x_{ML} &\geq p_{TH} \\ \mu_N + \sum_{n=1}^{N-1} w_n(x_n - \mu_n) &\geq -w_0 + \mu_N - w_N(x_N - \mu_N) \\ w_0 + \sum_{n=1}^{N} w_n(x_n - \mu_n) &\geq 0 \\ \underbrace{w_0 + \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu})}_{y(\boldsymbol{x})} &\geq 0 \end{aligned} \tag{44}$$

The estimation of the false-positive probability is

Table 5: Computational/memory requirements and algorithm performance

CPU	Intel Core i7-4800MQ @ 2.7 GHz, 8 Threads
CPU Load	95%
RAM	4 GB
# Allocated Threads	32 Threads
Response Time	0.9 seconds

performed by calculating $\hat{p}_{FP} = P(\hat{H} = H_1|H_0)$. However, as already explained in Sec. 6, the conditional pdf $p(\hat{H}|H)$ is not given and therefore $\hat{p}_{FP}$ cannot be directly estimated. Thus, we assume that the conditional pdf $p(\hat{H}|H)$ is of a Gaussian distribution with mean $E[\boldsymbol{X}|H_0] = \boldsymbol{\mu}_{H_0}$ and covariance matrix $E[(\boldsymbol{X} - \boldsymbol{\mu}_{H_0})(\boldsymbol{X} - \boldsymbol{\mu}_{H_0})^T|H_0] = \Sigma$. However, $\boldsymbol{\mu}_{H_0}$ is an unknown parameter vector; additionally, the second-order moments are assumed to be invariant of an event H_0. We will calculate $\hat{p}_{FP}$ in dependency on the unknown vector $\boldsymbol{\mu}_{H_0}$ and will qualitatively show that $\hat{p}_{FP}$ can be minimized independently of $\boldsymbol{\mu}_{H_0}$ due to an optimization strategy. In showing this, we demonstrate that the concept of our work is validated and mathematically proved.

First of all, Eq. 44 is given in Eq. 45 with an additional term.

$$\begin{aligned} w_0 + \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu}) &\geq 0 \\ w_0 + \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu}_{H_0} - \boldsymbol{\mu}_{H_0}) &\geq 0 \end{aligned} \tag{45}$$

Furthermore, by introducing the definition $\Delta\boldsymbol{\mu} = \boldsymbol{\mu} - \boldsymbol{\mu}_{H_0}$ Eq. 45 can be written as given in Eq. 46.

$$\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu}_{H_0}) \geq -w_0 + \boldsymbol{w}^T\Delta\boldsymbol{\mu} \tag{46}$$

Accordingly, $\hat{p}_{FP}$ is estimated as given in Eq. 47.

$$\hat{p}_{FP} = P\Big(\boldsymbol{w}^T(\boldsymbol{X} - \boldsymbol{\mu}_{H_0}) \geq -w_0 + \boldsymbol{w}^T\Delta\boldsymbol{\mu}\Big|H_0\Big) \tag{47}$$

By substituting $U := \sum_{n=1}^{N-1} w_n(X_n - \mu_{n,H_0})$ and $V := w_N(X_N - \mu_{N,H_0})$ with $\mu_{n,H_0} = \left(\boldsymbol{\mu}_{H_0}\right)_n$ the false-positive estimation is given as follows.

$$\hat{p}_{FP} = P\Big(U + V \geq -w_0 + \boldsymbol{w}^T\Delta\boldsymbol{\mu}\Big|H_0\Big) \tag{48}$$

Based on that fact, that H_0 is given U and V are conditionally independent (see explanation in subsection 6.4.2) of each other and hence the random variable $Z := U + V$ has the mean $E[Z] = E[U] + E[V]$ and the variance $\sigma_Z^2 = \sigma_U^2 + \sigma_V^2$. Since $E[X_n|H_0] = \mu_{n,H_0}$ holds the mean of U and V is zero ($E[U] = 0$; $E[V] = 0$).

The variances of U and V are given in Eq. 49

$$\sigma_U^2 = [w_1 \cdots w_{N-1}]\Sigma_{1,1}[w_1 \cdots w_{N-1}]^T \tag{49}$$

and Eq. 50 respectively.

$$\sigma_V^2 = w_N^2\sigma_N^2 \tag{50}$$

As $[w_1 \cdots w_{N-1}]^T = -\frac{\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T}{\boldsymbol{\phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T}$ holds and by substituting $\beta := \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T$ the term $\beta^T\Sigma_{1,1}\beta$ can be simplified in Eq. 51 as already explained in the proof of Theorem 3 in [1].

$$\boldsymbol{\beta}^T\Sigma_{1,1}\boldsymbol{\beta} = -\frac{\det(\Sigma_{1,1})}{\det(\Sigma)} + \sigma_N^2\left[\frac{\det(\Sigma_{1,1})}{\det(\Sigma)}\right]^2 \tag{51}$$

It can easily be seen that the variance of U is equal to $\sigma_U^2 = \frac{\beta^T\Sigma_{1,1}\beta}{\left[\left(\Sigma^{-1}\right)_{N,N}\right]^2}$. Furthermore, $\left(\Sigma^{-1}\right)_{N,N} = \frac{\det(\Sigma_{1,1})}{\det(\Sigma)}$ holds and thus the variance is further simplified and is given in Eq. 52.

$$\sigma_U^2 = -\frac{\det(\Sigma)}{\det(\Sigma_{1,1})} + \sigma_N^2 \tag{52}$$

Finally, the false-positive probability is estimated as follows:

$$\hat{p}_{FP} = P(Z \geq -w_0 + \boldsymbol{w}^T\Delta\boldsymbol{\mu}) = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{-w_0 + \boldsymbol{w}^T\Delta\boldsymbol{\mu}}{\sqrt{2}\sigma_Z}\right)\right] \tag{53}$$

For the case $N = 2$, Eq. 53 can be simplified and after several calculation steps the following Eq. 54 results

$$\hat{p}_{FP} = \frac{1}{2}\left[1 - \mathrm{erf}\Big(\psi\Big)\right] \tag{54}$$

with

$$\psi = \frac{\sqrt{e^{2\mathcal{I}} - 1}\,\mathrm{erfinv}(2p_{FN,Bound} - 1) + \frac{\sqrt{e^{4\mathcal{I}} - e^{2\mathcal{I}}}}{\sqrt{2}\sigma_1}\Delta\mu_1 - \frac{\sqrt{e^{2\mathcal{I}} - 1}}{\sqrt{2}\sigma_2}\Delta\mu_2}{\sqrt{2e^{4\mathcal{I}} - 3e^{2\mathcal{I}} + 1}} \tag{55}$$

□

References

[1] I. Alagöz, T. Herpel, and R. German, "A selection method for black box regression testing with a statistically defined quality level," in *2017 IEEE International Conference on Software Testing, Verification and Validation (ICST)*, March 2017, pp. 114–125, DOI: 10.1109/ICST.2017.18.

[2] F. Bürger and J. Pauli, "Representation optimization with feature selection and manifold learning in a holistic classification framework," in *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, 2015, pp. 35–44, DOI: 10.5220/0005183600350044.

[3] T. Back, *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford university press, 1996, DOI: 10.1108/k.1998.27.8.979.4.

[4] F. Bürger and J. Pauli, "Understanding the interplay of simultaneous model selection and representation optimization for classification tasks," in *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods*, 2016, pp. 283–290, DOI: 10.5220/0005705302830290.

[5] "ISO/DIS 26262-10 - Road vehicles — Functional safety," http://www.iso.org, Tech. Rep., 2012, DOI: doi:10.3403/30205385.

[6] G. Xie and Z. Dang, "Model-checking driven black-box testing algorithms for systems with unspecified components," *CoRR*, vol. cs.SE/0404037, 2004. [Online]. Available: http://arxiv.org/abs/cs.SE/0404037

[7] R. Grosu and S. A. Smolka, *Monte Carlo Model Checking*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 271–286, DOI: 10.1007/978-3-540-31980-1_18. [Online]. Available: https://doi.org/10.1007/978-3-540-31980-1_18

[8] A. Legay, B. Delahaye, and S. Bensalem, "Statistical model checking: An overview," in *International Conference on Runtime Verification*. Springer, 2010, pp. 122–135, DOI: 10.1007/978-3-642-16612-9_11.

[9] E. Elkind, B. Genest, D. Peled, and H. Qu, *Grey-Box Checking*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 420–435, DOI: 10.1007/11888116_30. [Online]. Available: https://doi.org/10.1007/11888116_30

[10] K. Sen, M. Viswanathan, and G. Agha, "Statistical model checking of black-box probabilistic systems," in *International Conference on Computer Aided Verification*. Springer, 2004, pp. 202–215, DOI: 10.1007/978-3-540-27813-9_16.

[11] D. Peled, M. Y. Vardi, and M. Yannakakis, *Black Box Checking*. Boston, MA: Springer US, 1999, pp. 225–240, DOI: 10.1007/978-0-387-35578-8_13. [Online]. Available: https://doi.org/10.1007/978-0-387-35578-8_13

[12] S. Yoo and M. Harman, "Regression testing minimization, selection and prioritization: a survey," *Software Testing, Verification and Reliability*, vol. 22, no. 2, pp. 67–120, 2012, DOI: 10.1002/stvr.430. [Online]. Available: http://dx.doi.org/10.1002/stvr.430

[13] G. Rothermel and M. J. Harrold, "A safe, efficient regression test selection technique," *ACM Trans. Softw. Eng. Methodol.*, vol. 6, no. 2, pp. 173–210, Apr. 1997, DOI: 10.1145/248233.248262. [Online]. Available: http://doi.acm.org/10.1145/248233.248262

[14] A. Orso, M. J. Harrold, D. Rosenblum, G. Rothermel, M. L. Soffa, and H. Do, "Using component metacontent to support the regression testing of component-based software," in *Proceedings IEEE International Conference on Software Maintenance. ICSM 2001*, 2001, pp. 716–725, DOI: 10.1109/ICSM.2001.972790.

[15] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, R. Anil, Z. Haque, L. Hong, V. Jain, X. Liu, and H. Shah, "Wide & deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, ser. DLRS 2016, vol. abs/1606.07792. New York, NY, USA: ACM, 2016, pp. 7–10, DOI: 10.1145/2988450.2988454. [Online]. Available: http://doi.acm.org/10.1145/2988450.2988454

[16] R. Langone, O. M. Agudelo, B. De Moor, and J. A. Suykens, "Incremental kernel spectral clustering for online learning of non-stationary data," *Neurocomputing*, vol. 139, pp. 246–260, 2014, DOI: 10.1016/j.neucom.2014.02.036 .

[17] S. Mehrkanoon, O. M. Agudelo, and J. A. Suykens, "Incremental multi-class semi-supervised clustering regularized by kalman filtering," *Neural Networks*, vol. 71, pp. 88–104, 2015, DOI: 10.1016/j.neunet.2015.08.001.

[18] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009. [Online]. Available: http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf

[19] C. Lin, M. Mausam, and D. Weld, "Re-active learning: Active learning with relabeling," 2016. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12500

[20] F. Bürger and J. Pauli, "Automatic representation and classifier optimization for image-based object recognition," in *VISAPP 2015 - Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Volume 2, Berlin, Germany, 11-14 March, 2015.*, 2015, pp. 542–550, DOI: 10.5220/0005359005420550. [Online]. Available: https://doi.org/10.5220/0005359005420550

[21] *Handbook of Mathematics*. Springer Berlin Heidelberg, 2007, DOI: 10.1007/978-3-540-72122-2. [Online]. Available: https://doi.org/10.1007%2F978-3-540-72122-2

[22] M. Thulin, "The cost of using exact confidence intervals for a binomial proportion," *Electron. J. Statist.*, vol. 8, no. 1, pp. 817–840, 2014, DOI: 10.1214/14-EJS909. [Online]. Available: https://doi.org/10.1214/14-EJS909

[23] I. Sommerville, *Software Engineering: (Update) (9th Edition) (International Computer Science)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2011.

[24] M. Schlager, *Hardware-in-the-Loop Simulation: A Scalable, Component-based, Time-triggered Hardware-in-the-loop Simulation Framework*. VDM Verlag Dr. Müller E.K., 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 49-57 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Interference Avoidance using Spatial Modulation based Location Aware Beamforming in Cognitive Radio IOT Systems

Jayanta Datta*,1, Hsin-Piao Lin[2]

[1]*Department of Electrical Engineering and Computer Science, National Taipei University of Technology, Taipei City 10608, Taiwan*

[2] *Department of Electronic Engineering, National Taipei University of Technology, Taipei City 10608, Taiwan*

ARTICLE INFO

Article history:
Received: 25 November, 2017
Accepted: 05 February, 2018
Online: 08 March, 2018

Keywords:
Spatial Modulation
Femto-Cell
Virtual Antenna Array

ABSTRACT

The Internet of Things (IOT) is a revolutionary communication technology which enables numerous heterogeneous objects to be inter-connected. In such a wireless system, interference management between the operating devices is an important challenge. Cognitive Radio (CR) seems to be a promising enabler transmission technology for the 5G-IOT system. The "sense-and-adapt" smart transmission strategy in CR systems can help to overcome the problem of multiple access interference (MAI) in IOT systems. In this paper, a 5G-IOT smart infrastructure system is arranged in the form of CR based virtual antenna array (VAA) system. In VAA based wireless system, knowledge of users' locations can help the transmitter to achieve interference avoidance by steering the main beam towards the intended recipient. This idea has been applied to the VAA-IOT system, where smart antenna array based location aware beamforming are applied at both transmitter and receiver cluster of smart sensors with the help of spatial modulation principle. The waveform of choice for the CR-IOT clusters is Generalized Frequency Division Multiplexing (GFDM) while corresponding waveform for the primary user (PU) cluster is conventional Orthogonal Frequency Division Multiplexing (OFDM). Computer simulation shows that under multipath fading conditions, the implemented system can reduce the interference to the primary user (PU) system, leading to better coexistence.

1. Introduction

This paper is an extension of work originally presented in 2017 International Conference of Applied System Innovation, held at Sapporo, Japan [1]. In this work, a beamforming scheme is presented which is based on transmitter activation by spatial modulation (SM) technique. Spatial Smoothing (SS) based Multiple Signal Classification (MUSIC) algorithm is applied for direction of arrival (DOA) estimation of the PU. Based on the DOA data, the combined spatial modulation-beamforming scheme is designed for avoiding interference to the PU. In [1], an overview of the methodology was presented. However, in this work, details about application of the SS-MUSIC as well as receiver side data detection are presented. Computer simulations are performed to demonstrate the effectiveness of the designed scheme for better interference avoidance in CR scenario.

Increasing demand for higher data rate and quality of mobile communications has led to the development of advanced multicarrier signaling schemes such as OFDM and GFDM among others. GFDM is a recently proposed non-orthogonal multicarrier waveform which is a potential candidate for 5G wireless technology. Its benefits are flexible carrier aggregation and low out-of-band (OOB) radiation [2-4]. Due to its flexibility in subcarrier allocation, the GFDM based user can sense the spectral bands in a CR environment and intelligently allocate the subcarriers to the bands where the interference temperature threshold can be satisfied. Apart from frequency bands, space and angle dimensions can be exploited too with the recent advances in multiple antenna technologies. In such a smart heterogeneous infra-structure, multi-antenna based GFDM system can lead to enhanced system performance and low cross-tier interference. This can be achieved with the application of beamforming technology at the Femto-Cell transceiver system. Moreover, cognitive radio technology can be combined with beamforming at

*Jayanta Datta, 1, Section 3, Zhongxiao Xinsheng, Taipei City 10608, Taiwan, +886-0988354648, Email: jdatta1@iit.edu

the Femto-Cell network enabling it to become a self-aware entity [5, 6]. The cognitive Femto-Cell system can use spatial modulation principle to encode the location estimates of the Macro-Cell PUs and select the Femto-Cell users closest to the Femto-Cell SU receiver system. The selected SU transmitters can perform null-steering based beamforming to avoid cross-tier interference by directing nulls towards Macro-Cell PUs. In this work, the CR-IOT [7-9] system is considered as a cognitive Femto-Cell under-laid with an existing main Macro-Cell infrastructure. This kind of cognitive interference management strategy can maximize Femto-Cell CR-IOT system performance while minimizing interference on the Macro-Cell PU system, leading to better spectral coexistence between the primary Macro-Cell user and the secondary Femto-Cell CR-IOT user.

2. System Model

It is assumed that the primary Macro-Cell and secondary Femto-Cell use the same uplink and downlink frequencies and the same bandwidth. The Macro-Cell uses OFDM for transmission, while the Femto-Cell uses GFDM as its signaling scheme. Figure 1 shows the system model.

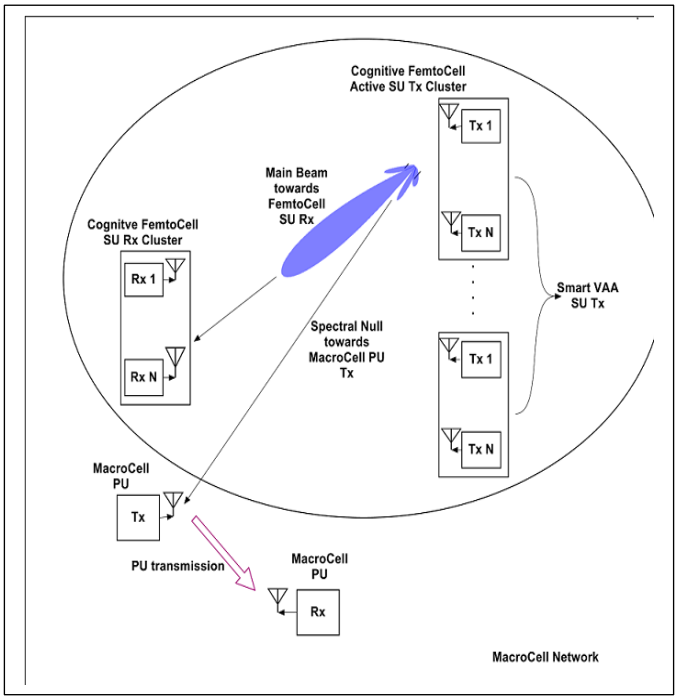


Figure 1. CR-IOT Femto-Cell VAA system under-laid with Macro-Cell PU system

The Femto-Cell based CR-IOT transmitter system is divided into multiple clusters. Each cluster contains a number of low power single antenna terminals arranged in the form of virtual antenna array (VAA) [10, 11]. The VAA based Femto-Cell CR-IOT system collects DOA information about the Macro-Cell PU and performs null-steering based beamforming to transmit its data. It uses spatial modulation principle to encode the DOA information of the PU in the form of bit blocks. This bit-block is used to identify a suitable transmit cluster which is then activated. The terminals in the active SU cluster transmit data to the SU receiver using null-steering based beamforming.

3. Transmitter Side Processing

Let there be 'N_c' SU transmitter clusters in total. Let there be 'N' terminals within each cluster. A central controller acts as the cluster head. The central controller has the role of the cognitive engine (CE). It is responsible for analyzing the received signal parameters and taking decisions on spectrum band occupancy status and estimated locations of the PUs and SUs. The transmit side processing operation can be divided into 2 phases, namely sensing phase and transmission phase.

3.1. Sensing Phase

The sensing phase operations performed at the SU transmitter side can be pictorially depicted as in Figure 2 below.

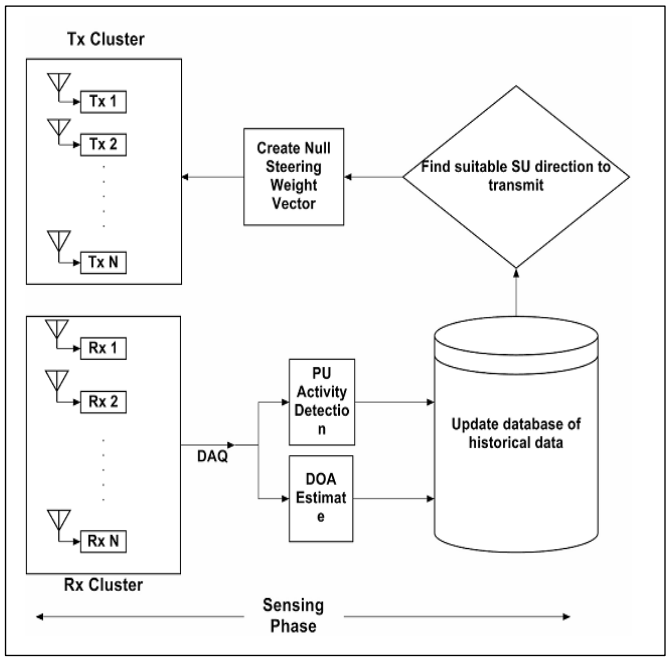


Figure 2. Sensing Phase at the Femto-Cell based CR-IOT transmitter cluster

1. PU parameters like Interference Temperature (IT), PU transmission power, available spectrum hole, PU location estimates (for e.g. DOA of received PU signal) etc. are collected and forwarded to the central controller, by the SU clusters.
2. Based on received parameters, the central controller calculates the distance of the clusters from the receiver SU. This can be achieved by localization algorithms like angle of arrival (AOA), received signal strength (RSS) estimation etc.
3. The central controller maintains a database of possible location estimates of the SU receiver clusters and PU terminals, as well as channel occupancy status within the spectrum band. The DOA estimates are matched with the database entries to verify the presence of PU activity at that location.

Based upon DOA estimation result of the received signal, the central controller creates a null-steering based beamforming weight vector. This is applied to the transmit signal during the transmission phase to steer spectral nulls towards the PU transmitter in order to avoid interference, whereas the main beam is steered towards the SU receiver cluster.

3.2. DOA Estimation by Spatial Smoothing MUSIC

SS-MUSIC [12, 13] is chosen for DOA estimation of the incoming signal under multipath propagation. Let the PU-OFDM signal be denoted by $x_{PU_OFDM}(t)$.

$$\boldsymbol{x_{PU_OFDM} = W^H . s} \tag{1}$$

In the above equation, 'W' indicates the discrete Fourier transform (DFT) matrix. 'W^H' indicates the inverse discrete Fourier transform matrix. The term's' indicates the data subcarriers which can be *M*-ary QAM or *M*-ary PSK modulated symbols. In general for an antenna array with N terminals, the received signal at all elements of the array can be expressed by the following mathematical equation:

$$\boldsymbol{r(t) = [r_1(t)\, r_2(t) \ldots\ldots r_N(t)]^T}$$

$$\boldsymbol{= \left[1\, e^{-j\omega\frac{dsinf}{c}} \ldots\ldots e^{-j\omega\frac{(N-1)dsinf}{c}}\right]^T . \left(H \otimes x_{PU_{OFDM}}(t)\right)}$$

$$\boldsymbol{= a(f).y(t), where\ y(t) = (H \otimes x_{PU_{OFDM}}(t))} \tag{2}$$

Suppose the antenna array comprising 'N' terminals be divided into 'L' overlapping sub-arrays. Each sub-array contains 'M' terminals, with M>K, K being the number of source directions. Then, N=M+L-1.

Let $x_i(t)$ be the received signal vector at the i^{th} sub-array. Let $a_M(\phi)$ be the steering vector associated with each sub-array.

$$\boldsymbol{x_i(t) = e^{-j\frac{\omega(i-1)dsinf}{c}} . a_M(f).y(t)\ , i = 1, 2, \ldots, M} \tag{3}$$

With 'K' DOAs, the received signal '$x_i(t)$' for the *i*th sub-array can be expressed by the following mathematical equation:

$$\boldsymbol{x_i(t) = [a_M(f_1) \ldots\ldots a_M(f_K)].\begin{bmatrix} e^{-j\frac{\omega(i-1)dsinf_1}{c}} & 0 & \cdots & 0 \\ \vdots & . & \ddots & \vdots \\ 0 & . & \cdots & 0 \\ 0 & . & . & e^{-j\frac{\omega(i-1)dsinf_K}{c}} \end{bmatrix} . y(t)} \tag{4}$$

$$\boldsymbol{A_M = [a_M(f_1) \ldots\ldots\ldots a_M(f_K)]} \tag{5}$$

$$\boldsymbol{D(i) = \begin{bmatrix} e^{-j\frac{\omega(i-1)dsinf_1}{c}} & 0 & \cdots & 0 \\ \vdots & . & \ddots & \vdots \\ 0 & . & \cdots & 0 \\ 0 & . & . & e^{-j\frac{\omega(i-1)dsinf_K}{c}} \end{bmatrix}} \tag{6}$$

$$\boldsymbol{x_i(t) = A_M . D_i . y(t)} \tag{7}$$

The correlation matrix at each sub-array can be computed as follows:

$$\boldsymbol{R_{x_i} = E\{x_i(t).x_i^H(t)\}}$$

$$\boldsymbol{= A_M . D_i . R_y . D_i^H . A_M^H + \sigma_0^2 . I}$$

$$\boldsymbol{= A_M . D_i . E\{y(t).y^H(t)\}. D_i^H . A_M^H + \sigma_0^2 . I} \tag{8}$$

The average of all the correlation matrices at all the sub-arrays can be expressed as follows:

$$\boldsymbol{R_{x_{avg}} = \frac{1}{L} . \sum_{i=1}^{L} E\{x_i(t).x_i^H(t)\}}$$

$$\boldsymbol{= A_M . [\frac{1}{L} . \sum_{i=1}^{L} D_i . R_y . D_i^H] . A_M^H + \sigma_0^2 . I} \tag{9}$$

The conventional MUSIC [12] algorithm can be applied on the matrix 'R_{x_avg}' to obtain the DOA estimate of the incoming PU-OFDM signal. The dimension of the matrix 'A_M' is M-by-M. The dimension of the matrix within square brackets is K-by-K. Hence, it is evident that the dimension of 'R_{x_avg}' is M-by-M.

By calculating the eigen-values of R_{x_avg}, 2 disjoint eigen-spaces can be identified. One is the signal subspace, and it consists of signal eigen-vectors on which noise is overlapped. The other one is the noise subspace which comprises eigen-vectors only due to noise. The signal eigen-vectors correspond to the 'K' (with K≤M-1) greatest eigen-values. The noise eigen-vectors correspond to the other (M-K) eigen-values. The MUSIC algorithm then searches for those steering vectors which are orthonormal to the noise subspace.

Let u_i be the *i*th eigen-vector of R_{x_avg} corresponding to the eigen-value σ_i^2.

Step 1: We perform eigen-decomposition of the M-by-M matrix 'R_{x_avg}'.

$$\boldsymbol{R_{x_{avg}} . u_i = \left[A_M . R_{s_{avg}} A_M^H + \sigma_0^2 . I\right] . u_i = \sigma_i^2 . u_i, i = 1, 2, \ldots, M}$$

$$\boldsymbol{where\ \sigma_i^2 > \sigma_0^2, i = 1, 2, \ldots, K; \sigma_i^2 = \sigma_0^2, i = K+1, \ldots, M} \tag{10}$$

Step 2: The above expression can be re-expressed as follows:

$$\boldsymbol{A_M . R_{s_{avg}} . A_M^H . u_i = (\sigma_i^2 - \sigma_0^2) . u_i; i = 1, 2 \ldots K}$$

$$\boldsymbol{= 0 \quad ; i = K+1, \ldots, M} \tag{11}$$

Step 3: We partition the M-dimensional vector space into signal subspace U_s and noise subspace U_n:

$$[\boldsymbol{U_s}\ \boldsymbol{U_n}] = [\boldsymbol{u_1} \dots \dots \dots \boldsymbol{u_K}\ \boldsymbol{u_{K+1}} \dots \dots \boldsymbol{u_M}],$$

$$\boldsymbol{where}\ \boldsymbol{U_s} = [\boldsymbol{u_1} \dots \dots \dots \boldsymbol{u_K}] : \boldsymbol{\sigma_i^2, i = 1, 2, \dots, K}$$

$$\boldsymbol{and}\ \boldsymbol{U_n} = [\boldsymbol{u_{K+1}} \dots \dots \dots \boldsymbol{u_M}] : \boldsymbol{\sigma_i^2, i = K+1, 2, \dots, M} \quad (12)$$

Step 4: The steering vector associated with the DOA of the PU-OFDM signal is in the signal subspace, the latter being orthogonal to the noise subspace. Hence, the MUSIC algorithm searches through all the angles 'ϕ' and plots the spatial spectrum given by the following equation:

$$\boldsymbol{P(f)} = \frac{\boldsymbol{1}}{\boldsymbol{a^H(f).U_n}} \quad (13)$$

Whenever the scanned angle $\phi=\phi_i$, the DOA of the PU-OFDM signal, the MUSIC spatial spectrum will exhibit a peak, providing an estimate of the DOA of the PU-OFDM signal.

3.3. Data Transmission Phase

This phase consists of 2 sub-phases, namely VAA cluster index modulated CR-IOT transmit signal generation and null-steering based transmit side beamforming.

i. VAA cluster index modulated CR-IOT transmit signal generation

Multiple-antenna techniques constitute a key technology for modern wireless communications, which trade-off superior error performance and higher data rates for increased system complexity and cost. The basic idea of SM-Beamforming is to map a block of information bits into information carrying units. Each information carrying unit comprises a symbol chosen from a complex constellation diagram and a unique cluster index that is chosen from a set of transmitter clusters. Each transmitter cluster contains a VAA of transmitters arranged in the form of a uniform linear array (ULA). As a consequence, we have a hybrid modulation and MIMO technique [14, 15], in which the modulated signal belongs to a tridimensional constellation diagram [15-17], which jointly combines signal and spatial information (cluster index). At the transmitter side, the bit-stream emitted from the central controller is divided into blocks containing ($\log_2 N_c + N_M.\log_2 M$) bits each, with $\log_2 N_c$ and $\log_2 M$ being the number of bits needed to identify a transmitter cluster and a transmitter symbol from the symbol constellation diagram respectively. Here, 'N_M' denotes the total number of information symbols to be transmitted for a particular time slot. Each block is then processed by a SM mapper, which splits each of them into two sub-blocks of $\log_2 (N_c)$ and $N_M.\log_2 (M)$ bits each. The spatial mapper comprises 2 separate tables, namely active cluster index (ACI) mapping table and information symbol mapping table. The ACI mapping table contains a list of the angular locations of the SU transmitter clusters, and bit combinations of size $\log_2 (N_c)$ corresponding to each transmitter cluster. The incoming bit sub-block of size $\log_2 (N_c)$ selects that SU transmitter cluster index which is nearest in distance to the SU receiver. It is assumed that the CE already has prior knowledge of the distances based on prior location estimation. The other bit sub-block of size $N_M.\log_2 (M)$ selects N_M symbols according to the information symbol mapping table. The ACI mapping table is depicted in Table 1 below:

Table 1. Active Cluster Index (ACI) Mapping Table.

SU Tx Cluster Index	SU Tx Cluster Locations	Bit Pattern
1	30 degrees	00
2	45 degrees	01
3	60 degrees	10
4	75 degrees	11

The information symbol mapping table is shown in Table 2 below:

Table 2. Information Symbol Mapping Table.

Symbol	Bit Pattern
1+j	00
1-j	01
-1+j	10
-1-j	11

According to the above scheme, only one cluster of SU transmitters is activated at a particular frame instant while the other clusters remain silent. This helps to avoid the inter-cluster interference. In other SM schemes [15-18], only one transmitter antenna is active at a particular time instant, thereby transmitting only one constellation symbol per instant. However in our proposed transmission scheme, multiple information symbols are transmitted by the antenna array leading to improved spectral efficiency.

The cluster index modulated symbol stream is passed through the GFDM modulator, after which null-steering based beamforming is performed by the array of SU terminals within the activated cluster. The process can be pictorially depicted in Figure 3 below:

Details of baseband GFDM transceiver implementation can be obtained from [2].

i. *Null-Steering Transmit Side Beamforming*

The digital baseband signal snapshots $x=[x_1, x_2 \dots x_N]^T$ in a transmit beamforming cluster with N single antenna sensors (virtual antenna array) can be represented by the following equation:

$$\boldsymbol{x} = \boldsymbol{w}(\emptyset_{SU}).\boldsymbol{x}_{SU_GFDM} \quad (14)$$

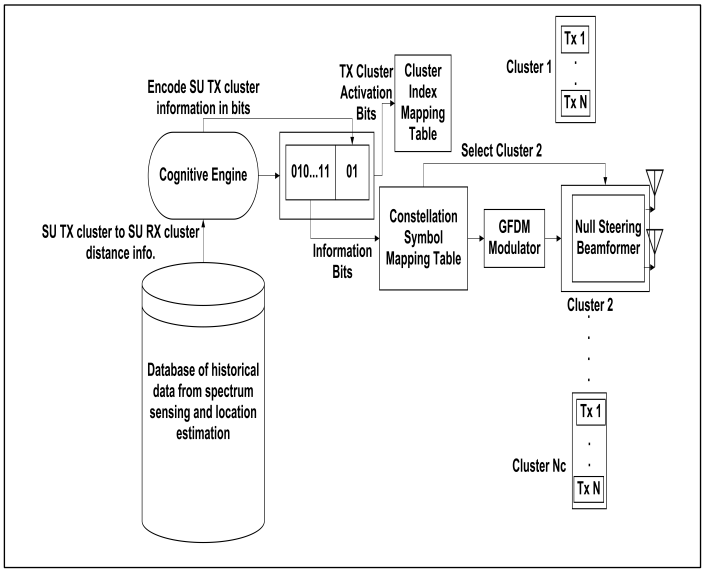


Figure 3. SM-Beamforming operation in CR-IOT GFDM Transmitter

In the above equation, $w(\phi_{SU}) = [w_1(\phi_{SU}), w_2(\phi_{SU}), \ldots.., w_N(\phi_{SU})]^T$ represents the beamforming weight vector under a specific optimization criteria towards the desired direction ϕ_{SU} [19],[20]. In other words, ϕ_{SU} is the direction in which the intended SU receiver is located. 'x_{SU_GFDM}' refers to the spatially modulated symbol transmitted by the activated SU cluster. It is assumed that the sensors within each transmit cluster are closely spaced. Under this assumption, the effective channels between the different transmit antennas inside the cluster and the receiver antenna differ by phase shift included in the steering vector of the transmit cluster. The steering vector for ULA in direction 'ϕ' can be represented by the following equation [21]:

$$\boldsymbol{a}(\emptyset) = [1, e^{j.d.k.cos\emptyset}, e^{j.2d.k.cos\emptyset}, \ldots.., e^{j.(N-1)d.k.cos\emptyset}]^T \quad (15)$$

In the above equation, the constant 'k'=2π/λ, where λ is the wavelength of the RF signal frequency. The distance between the elements of the cluster is expressed as'd'. The received snapshot of the signal received at a direction 'ϕ' is represented by the following equation:

$$\boldsymbol{y}(\emptyset) = \boldsymbol{a}^H(\emptyset).\boldsymbol{w}(\emptyset_{SU}).\boldsymbol{x}_{SU_GFDM} + \boldsymbol{n} \quad (16)$$

Suppose there is 'M' number of PUs operating in the same frequency band as the SU transmitter. Let the angles of the PUs from the SU transmitter be denoted by $\phi_{PU,1}$, $\phi_{PU,2}$,…., $\phi_{PU,M}$. Let the null-steering matrix be denoted by 'A'. Then the matrix 'A' consists of the steering vectors to all the PU directions.

$$\boldsymbol{A} = [\boldsymbol{a}(\emptyset_{PU,1}), \boldsymbol{a}(\emptyset_{PU,2}), \ldots\ldots\ldots, \boldsymbol{a}(\emptyset_{PU,M})] \quad (17)$$

The N-by-N matrix, P_A, refers to the orthogonal projection matrix onto the subspace spanned by the columns of A. It is expressed mathematically as follows:

$$\boldsymbol{P}_A = \boldsymbol{A}.[\boldsymbol{A}^H.\boldsymbol{A}]^{-1}.\boldsymbol{A}^H \quad (18)$$

Conventional null-steering beamforming problem can be formulated as follows [19]:

$$max_H |\boldsymbol{w}^H.\boldsymbol{a}(\emptyset_{SU})|^2 \; subject\; to\; \boldsymbol{w}^H.\boldsymbol{A} = 0 \; and \; \boldsymbol{w}^H.\boldsymbol{w} = c\,,$$

$$c: constant\; value \quad (19)$$

The solution to the above optimization problem is given as follows [19]:

$$\boldsymbol{w}(\emptyset_{SU}) = \frac{c}{||(\boldsymbol{I}-\boldsymbol{P}_A).\boldsymbol{a}(\emptyset_{SU})||}.(\boldsymbol{I}-\boldsymbol{P}_A).\boldsymbol{a}(\emptyset_{SU}) \quad (20)$$

In the above equation, 'I' refers to N-by-N identity matrix. The radiation pattern of the activated SU cluster represents the spatial response of the VAA performing the null-steering beamforming with the derived beamforming weights 'w (ϕ_{SU})'. Mathematically it can be expressed as follows:

$$\boldsymbol{R}(\emptyset) = |\boldsymbol{a}^H(\emptyset).\boldsymbol{w}(\emptyset_{SU})|^2 \quad (21)$$

4. Receiver Side Processing

4.1. Spatial Signature based Matched Filtering

In general, the wireless channel impulse response between the i^{th} terminal in the s^{th} SU transmitter cluster and the j^{th} terminal in the receiver cluster can be represented by the following mathematical equation [22]:

$$h_{s,i,j}(t,\tau) = \sum_{l=0}^{L_i(t)} \alpha_{i,l}(t).\delta(t-\tau_{i,l}(t)) \quad (22)$$

In the above equation, '$\alpha_{i,l}(t)$' denotes the complex channel amplitude. The number of multipath components is denoted by 'L_i (t)'.The path delay of the l^{th} multipath component is denoted by '$\tau_{i,l}(t)$'.Let '$H_{s,i,j}$' denote the frequency domain channel coefficients between the i^{th} transmitter antenna of the s^{th} SU transmitter cluster and j^{th} receiver antenna.

Let '$H^{diag}{}_{s,i,j}$' denote a diagonal matrix containing the elements of '$H_{s,i,j}$' along the main diagonal.

$$H_{i,j}^{diag} = \begin{bmatrix} H_{i,j}(1) & 0 & 0 & \cdots & 0 \\ \vdots & H_{i,j}(2) & 0 & \ddots & \vdots \\ . & . & . & . & . \\ . & . & . & . & 0 \\ 0 & 0 & 0 & \cdots & H_{i,j}(N_{sub}) \end{bmatrix} \quad (23)$$

Let '$\mathbf{x}^i{}_{SU_GFDM}$', i=1, 2…N_T , denote the transmitted GFDM signal from the i^{th} transmitter antenna, before application of the null-steering beamforming weight vector. Its mathematical expression is same as that for baseband GFDM signal shown in [2]. Let

'$\mathbf{S}^{i}_{SU_GFDM}$', i=1, 2…N_T , denote the pre –IFFT GFDM signal as expressed below. As mentioned earlier, details of the mathematical expression for '$\mathbf{x}^{i}_{SU_GFDM}$' can be found in [2].

$$\boldsymbol{S}^{i}_{SU_{GFDM}} = \boldsymbol{W}_{N_{sub}}.\boldsymbol{x}^{i}_{SU_{GFDM}}$$
$$= \boldsymbol{W}_{N_{sub}}.\left\{\boldsymbol{W}^{H}_{N_{sub}}.\sum_{k=1}^{K}\boldsymbol{P}_k.\boldsymbol{\Gamma}_{diag}.\boldsymbol{R}^{L}.\boldsymbol{W}_M.d_k\right\}$$
$$= \sum_{k=1}^{K}\boldsymbol{P}_k.\boldsymbol{\Gamma}_{diag}.\boldsymbol{R}^{L}.\boldsymbol{W}_M.d_k\ ; i = 1,2,\dots,N_T \quad (24)$$

Further, '$\mathbf{S}^{i}_{SU_GFDM}$' can be represented by the column vector as shown below:

$$\boldsymbol{S}^{i}_{SU_GFDM} = [S^{i}_{SU_{GFDM}}(1)\ S^{i}_{SU_{GFDM}}(2)\ \dots\dots\ S^{i}_{SU_{GFDM}}(N_{sub})]^T \quad (25)$$

It is assumed that the cyclic prefix (CP) is longer than the maximum delay spread of the channel. So after removal of the CP, the channel appears circular to the symbols. As such, application of FFT converts the circulant channel matrix to diagonal matrix, with frequency domain channel coefficients along the main diagonal.

The null-steering weight vector can be expressed as a column vector as follows:

$$w(\emptyset_{SU}) = [w^{(1)}(\emptyset_{SU})\ w^{(2)}(\emptyset_{SU})\ \dots\dots\dots w^{(N_T)}(\emptyset_{SU})]^T \quad (26)$$

The received signal at the *j*th receiver antenna can be expressed as follows:

$$\boldsymbol{y}_j = \left[\boldsymbol{H}^{diag}_{1,j}\ \boldsymbol{H}^{diag}_{2,j}\ \dots\dots\dots\boldsymbol{H}^{diag}_{N_T,j}\right].\begin{bmatrix} w^{(1)}(\emptyset_{SU}).\boldsymbol{S}^{1}_{SU_GFDM} \\ w^{(2)}(\emptyset_{SU}).\boldsymbol{S}^{2}_{SU_GFDM} \\ \vdots \\ w^{(N_T)}(\emptyset_{SU}).\boldsymbol{S}^{N_T}_{SU_GFDM} \end{bmatrix} + \boldsymbol{N} \quad (27)$$

The above equation can be re-expressed as:

$$y_j = \left(\boldsymbol{w}(\emptyset_{SU}) \odot \left[\boldsymbol{H}^{diag}_{1,j}\ \boldsymbol{H}^{diag}_{2,j}\ \dots\dots\boldsymbol{H}^{diag}_{N_T,j}\right]\right).\begin{bmatrix} \boldsymbol{S}^{1}_{SU_GFDM} \\ \boldsymbol{S}^{2}_{SU_GFDM} \\ \vdots \\ \boldsymbol{S}^{N_T}_{SU_GFDM} \end{bmatrix} + \boldsymbol{N}$$
$$= \left(\boldsymbol{w}(\emptyset_{SU}) \odot \left[\boldsymbol{H}^{diag}_{1,j}\ \boldsymbol{H}^{diag}_{2,j}\ \dots\dots\boldsymbol{H}^{diag}_{N_T,j}\right]\right).\boldsymbol{S}^{all}_{SU_GFDM} + \boldsymbol{N}$$
$$\boldsymbol{where}\ \boldsymbol{S}^{all}_{SU_GFDM} = [\boldsymbol{S}^{1}_{SU_GFDM}\ \boldsymbol{S}^{2}_{SU_GFDM}\ \dots\dots\boldsymbol{S}^{N_T}_{SU_GFDM}]^T \quad (28)$$

In (28), the sign 'Θ' indicates element-wise multiplication.

Let 'b_j' denote the spatial signature of the *j*th receiver antenna.

$$b_j = \boldsymbol{w}(\emptyset_{SU}) \odot [\boldsymbol{H}^{diag}_{1,j}\ \boldsymbol{H}^{diag}_{2,j}\ \dots\dots\boldsymbol{H}^{diag}_{N_T,j}] \quad (29)$$

Let '**b**' denote the matrix of spatial signatures at all the receiver antennas.

$$\boldsymbol{b} = [b_1\ b_2\ \dots\dots b_{N_R}]^T \quad (30)$$

Then, the received signal at 'N_R' receiver antennas can be expressed in terms of the transmitted signal and spatial signature associated with each receiver antenna as follows:

$$\boldsymbol{y} = [y_1\ y_2\ \dots\dots y_{N_R}]^T = \boldsymbol{b}.\boldsymbol{S}^{all}_{SU_GFDM} + \boldsymbol{N} \quad (31)$$

It is assumed that the channel is temporally slowly fading and that each post-FFT sample of the received GFDM signal encounters independent and identical Rayleigh distribution. Moreover, the channel stays constant over a number of symbols and the receiver has perfect knowledge of the slowly varying channel. The receiver computes the pseudo-inverse of the spatial signature matrix '**b**' to recover the transmitted spatially modulated GFDM signal.

4.2. Active Transmitter Cluster Index Estimation

While it is assumed that the receiver has perfect knowledge of the spatial signatures, the receiver does not have knowledge which transmitter cluster is active for that particular frame instant. The receiver performs spatial signature based matched filtering on the received signal using spatial signature belonging to each SU transmitter cluster and stores the result in vector form in a buffer. Next, the receiver computes the squared absolute value of each stored vector in the buffer and compares the result to find out the maximum value. The index of the maximum value gives an estimate of the active transmitter SU cluster.

The procedure is tabulated in Table 3 below:

Table 3. Spatial Signature based Matched filtering

Inputs:
Received Signal: y, Spatial Signature of s^{th} transmitter cluster: b_s ; s=1,2,……..,N_c
Output:
Index of Active Transmit SU cluster: s
(i) $\boldsymbol{y}^{matched}_{s} = (\boldsymbol{b}^{H}_{s}.\boldsymbol{b}_s)^{-1}.\boldsymbol{b}^{H}_{s}.\boldsymbol{y}, s = 1,2\dots., N_c$

(ii) $a_s = ||\boldsymbol{y}_s^{matched}||^2, s = 1,2,\dots,N_c$

(iii) $s = \arg max_s\, a_s, s = 1,2,\dots.,N_c$

5. Simulation Results and Analysis

In this section, simulation results are provided to evaluate the performance of the proposed system under multipath Rayleigh fading channel environment. It is assumed that the clusters are closely spaced. Inter-element distance within each cluster is d=0.5λ, where 'λ' is the wavelength. Cluster elements are arranged in the form of ULA. The center frequency of operation of both the CR-IOT Femto-Cell and primary Macro-Cell is f_c=2.3 GHz. The number of transmitter and receiver terminals within each cluster in the CR-IOT Femto-Cell network is 16. The multicarrier signaling scheme chosen for the Femto-Cell CR-IOT system is GFDM, which uses *M*-ary QAM constellation. The Macro-Cell PU system uses OFDM waveform, which also uses *M*-ary QAM constellation. The PU-OFDM transmitter is assumed to be at 45 degrees.

The following table shows the specifications of the Femto-Cell CR-IOT GFDM waveform.

Table 4. GFDM system parameters

Simulation Parameters for SM-beamformer GFDM system	
Total number of time slots	5
Active Subcarrier Indices 100:200	
Pulse Shaping Filter	RRC
RRC Overlap Factor	0.1
Number of SU Tx clusters	4
Number of SU Tx terminals per cluster	16
Number of SU Rx clusters	1
Number of SU Rx terminals per cluster	16

In Figure 4 showing antenna array pattern, null is steered at 45 degrees which is the Macro-Cell PU-OFDM transmitter's location. The SS-MUSIC algorithm provides the DOA estimate of the PU-OFDM transmitter, which helps to steer null towards the PU location. As evident from Figure 4, the red dotted line at 45 degrees indicates the direction of the Macro-Cell PU-OFDM transmitter.

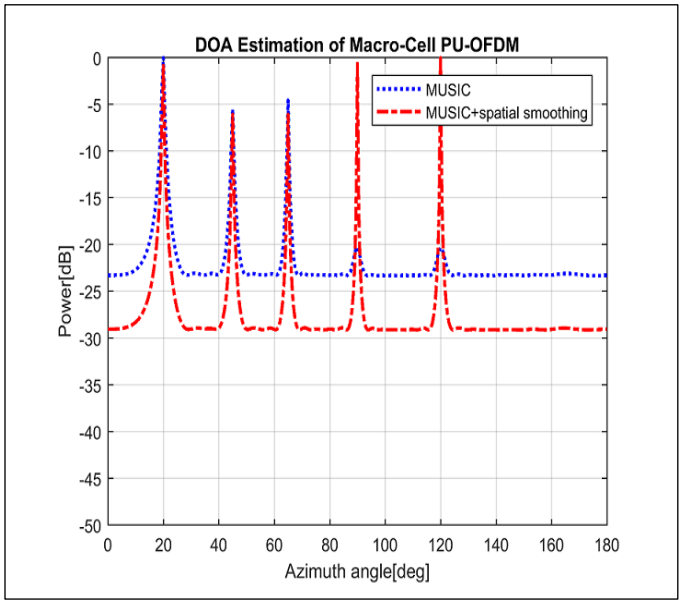


Figure 4. Cognitive Null Steering towards Macro-Cell PU-OFDM user

SS-MUSIC is preferred over conventional MUSIC primarily because of the fact that under multipath propagation, the latter does not perform well in identifying all the sources. This is because under multipath propagation environment, 2 or more DOAs might belong to the same source .i.e. for 'K' number of DOAs, actually the number of sources might be less than 'K' under multipath environment.

Figure 5 shows DOA estimation performance comparison between SS-MUSIC and conventional MUSIC for 2 sources arriving at the receiver from different directions, under multipath propagation conditions.

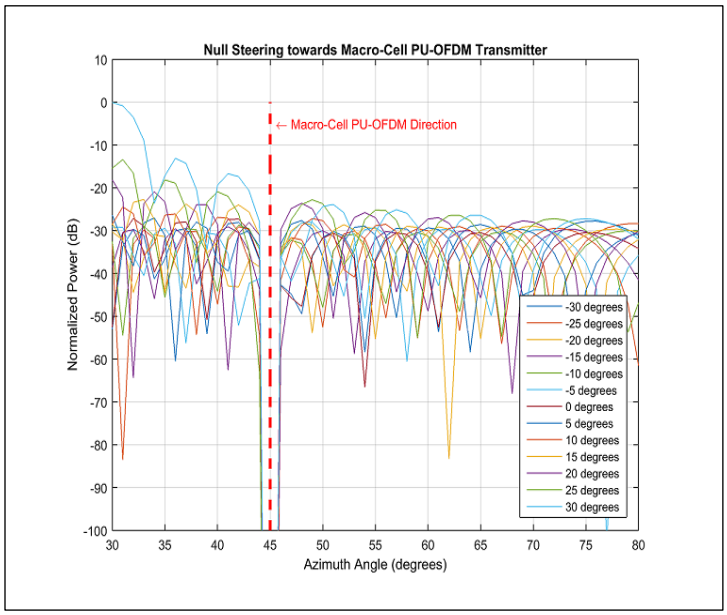


Figure 5. SS-MUSIC vs conventional MUSIC

As evident from Figure 5, for signals with multipath components too close to the main direction, the estimation capability of SS-MUSIC is better due to the application of the smoothing procedure. In the above figure, resolution of the main DOA (45º) and the multipath DOAs (20º, 65º, 90º and 120º) by the SS-MUSIC is better than that with MUSIC.

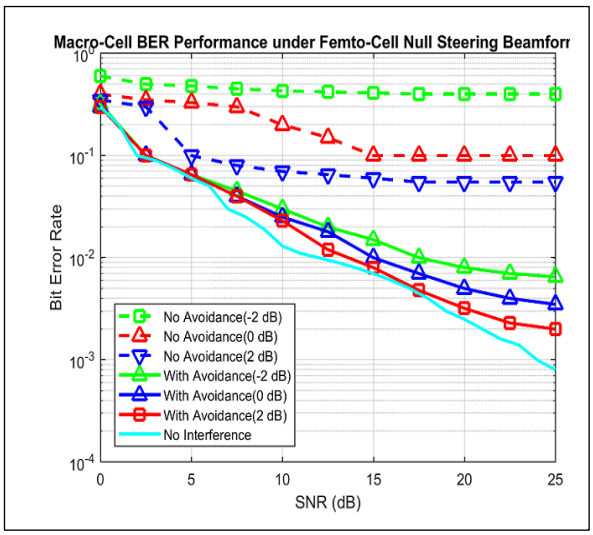


Figure 6. BER performance of Macro-Cell based PU-OFDM

Figure 6 compares the BER curves as a function of SNR for different signal-to-interference ratio (SIR) values for the Macro-Cell PU-OFDM user under conditions of no interference avoidance and with interference avoidance. As evident from the above figure, the Macro-Cell PU suffers significant performance loss under no interference avoidance scenario; however considerable performance improvement is achieved with the application of the cognitive null-steering based beamforming at the Femto-Cell CR-IOT GFDM transmitter.

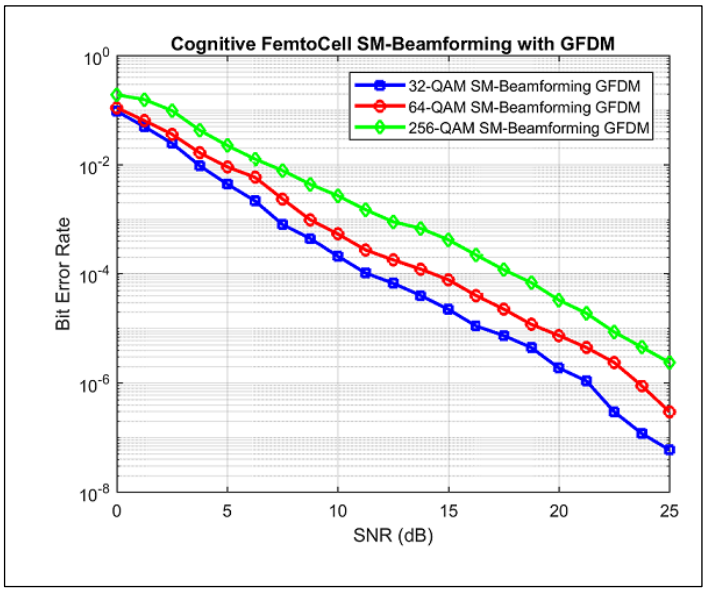


Figure 7. BER performance of Femto-Cell based CR-IOT GFDM

The proposed SM-beamforming system is very well capable of avoiding interference to the PU system for signal-to-noise ratio (SNR) less than or equal to 15 dB. At higher SNR values, it is observed that there is presence of residual interference. This happens due to imperfect steering of nulls at higher SNRs, for different SIR values.

Figure 7 shows the BER performance of the Femto-Cell based CR-IOT GFDM employing SM-beamforming scheme. As expected, BER for 256-QAM system is higher than that of 32-QAM and 64-QAM system under multipath Rayleigh fading channel conditions.

6. Conclusion

This paper introduces the application of SM based null-steering beamforming in VAA based CR-IOT system, under multipath fading channel conditions. Computer simulation results show that the proposed scheme can lead to good BER performance of the PU system by appropriate design of the null-steering beam-former. Moreover, the application of beamforming at the spatial modulation based SU transmitter also provides good BER performance of the SU system under multipath channel conditions. Spatial Smoothing based MUSIC is selected for DOA estimation due to its superiority over conventional MUSIC under multipath propagation. However, performance analysis of other DOA estimation algorithms needs to be evaluated for multicarrier signaling case. As a future work, transmit-receive beamforming schemes with DOA estimation will be investigated under full duplex spatial modulation based transmission conditions.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by National Taipei University of Technology, Taiwan.

References

[1] J. Datta, H.P. Lin, D.B.Lin, “Spatial modulation based location aware beam-forming in GFDM modulated cognitive radio systems”, in International Conference on Applied System Innovation, Sapporo Japan, 2017. https://doi.org/ 10.1109/ICASI.2017.7988357

[2] I. Gaspar, N. Michailow, A. Navarro, E. Ohlmer, S. Krone, G. Fettweis, “Low Complexity GFDM Receiver Based on Sparse Frequency Domain Processing” in Proceedings of the IEEE 77th Vehicular Technology Conference, Dresden Germany, 2013. https://doi.org/ 10.1109/VTCSpring.2013.6692619

[3] N. Michailow, M. Matthé, I.S. Gaspar, A. Navarro, L.L. Mendes, A. Festag, “Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks” IEEE Trans. Comm.., **62(9)**, 3045-3061, 2014. https://doi.org/10.1109/TCOMM.2014.2345566

[4] N. Michailow, I. Gaspar, S. Krone, M. Lentmaier, G. Fettweis, “Generalized frequency division multiplexing: Analysis of an alternative multi-carrier technique for next generation cellular systems” in Proceedings of the IEEE International Symposium on Wireless Communication Systems, Paris France, 2012. https://doi.org/ 10.1109/ISWCS.2012.6328352

[5] G. Gur, S. Bayhan, F. Alagoz, “Cognitive femtocell networks: an overlay architecture for localized dynamic spectrum access.” IEEE Mag. Wireless Comm, **17**(4), 62-70, 2010. https://doi.org/ 10.1109/MWC.2010.5547923

[6] M.Z. Shakir, R. Atat, M.S. Alouini, “On the interference suppression capabilities of cognitive enabled femtocellular networks”, Proceedings of the IEEE International Conference on Communications and Information

Technology, Hammamet Tunisia, 2012. https://doi.org/ 10.1109/ICCITechnol.2012.6285835
[7] P. Rawat, K.D. Singh, J.M. Bonnin, "Cognitive radio for M2M and Internet of Things: A survey", Comp Comm., **94**, 1-29, 2016. https://doi.org/10.1016/j.comcom.2016.07.012
[8] K. Katzis, H. Ahmadi, "Challenges Implementing Internet of Things (IoT) Using Cognitive Radio Capabilities in 5G Mobile Networks", Internet of Things (IoT) in 5G Mobile Technologies, **8**, 55-76, 2016.
[9] R. Fantacci, D. Marabissi, "Cognitive Spectrum Sharing: An Enabling Wireless Communication Technology for a Wide Use of Smart Systems", Future Internet, **8**, 23, 2016. https:// doi:10.3390/fi8020023
[10] M. Dohler, E. Lefranc, H. Aghvami, "Space-time block codes for virtual antenna arrays", Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Lisboa, Portugal, 2002. https://doi.org/ 10.1109/PIMRC.2002.1046733
[11] M. Dohler, E. Lefranc, H. Aghvami, "Link capacity analysis for virtual antenna arrays", Proceedings of the IEEE 56th Vehicular Technology Conference, Vancouver, Canada, 2002. https://doi.org/ 10.1109/VETECF.2002.1040381
[12] B.D. Rao, K.V.S. Hari, "MUSIC and spatial smoothing: A statistical performance analysis", Proceedings of the Twenty-Third Asilomar Conference on Signals, Systems and Computers, Pacific Grove, USA, 1989. https://doi.org/ 10.1109/ACSSC.1989.1201035
[13] A. Paulraj, V.U. Reddy, T.J. Shan, T. Kailath, "Performance Analysis of the Music Algorithm with Spatial Smoothing in the Presence of Coherent Sources", Proceedings of the IEEE Military Communications Conference, Monterey, USA, 1986. https://doi.org/ 10.1109/MILCOM.1986.4805849
[14] R, Schmidt, "Multiple emitter location and signal parameter estimation", IEEE Trans. Antennas and Propagation, **34** (3), 276-280, 1986. 10.1109/TAP.1986.1143830
[15] M. D. Renzo, H. Haas, P.M. Grant, "Spatial modulation for multiple-antenna wireless systems: a survey", IEEE Communications Mag., **49** (12), 182-191, **2011**. https://doi.org/10.1109/MCOM.2011.6094024
[16] M.D. Renzo, H. Haas, A. Ghrayeb, S. Sugiura, L. Hanzo, "Spatial Modulation for Generalized MIMO: Challenges, Opportunities, and Implementation", Proceedings of the IEEE, **102** (1), 56-103, 2014. https://doi.org/ 10.1109/JPROC.2013.2287851
[17] R. Mesleh, H. Haas, C.W. Ahn, S. Yun, "Spatial Modulation-A low Complexity Spectral Efficiency Enhancing Technique", in Proceedings of First International Conference on Communications and Networking in China, Beijing, China, 2006. https://doi.org/ 10.1109/CHINACOM.2006.344658
[18] J. Jeganathan, A. Ghrayeb, L. Szczecinski, "Spatial Modulation: Optimal Detection and Performance Analysis", IEEE Communications Letters, **12**(8), 545-547, 2008. https://doi.org/ 10.1109/LCOMM.2008.080739
[19] K. Zarifi, S. Affes, A. Ghrayeb, "Collaborative Null-Steering Beamforming for Uniformly Distributed Wireless Sensor Networks", IEEE Trans. Signal Process. **58** (3), 1889-1903, 2010. https://doi.org/ 10.1109/TSP.2009.2036476
[20] B. Friedlander, B. Porat, "Performance analysis of a null-steering algorithm based on direction-of-arrival estimation", IEEE Trans. Acoust., Speech, Signal Process. **37** (4), 461 –466, 1989. https://doi.org/ 10.1109/29.17526
[21] L. Godara, "Application of antenna arrays to mobile communications. II. beam-forming and direction-of-arrival considerations", Proc. IEEE **85** (8), 1195 –1245, 1997. https://doi.org/ 10.1109/5.622504
[22] R.B. Ertel, P. Cardieri, K.W. Sowerby, T.S. Rappaport, J.H. Reed, " Overview of spatial channel models for antenna array communication systems", IEEE Pers. Comm., **5** (1), 10-22, 1998. https://doi.org/ 10.1109/98.656151

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 6, 56-62 (2017)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

A New Identification Approach of MIMO Hammerstein Model with Separate Nonlinearities

Chekib Ghorbel *, Zeineb Rayouf, Naceur Benhadj Baraiek

Advanced System Laboratory, Polytechnic School of Tunisia BP 743, 2078, La Marsa, Tunisie.

ARTICLEINFO

Article history:
Received: 20 September, 2017
Accepted: 09 October, 2017
Online: 23 October, 2017

Keywords:
Parametric identification
Hammerstein model
Separate nonlinearity
Quadruple-tank process

ABSTRACT

A new coupled structure identification of Multi-Input Multi-Output (MIMO) Hammerstein models with separate nonlinearities is proposed. It is based on the use of the Recursive Least Squares (RLS) algorithm. A comparative study between a decoupled and coupled structures identification of MIMO Hammerstein models is discussed. A quadruple-tank process is used to illustrate the e ectiveness of the new structure.

1 Introduction

This paper is an extension of work originally presented in Conference on Science of Electronics, Technologies of Information and Telecommunications (SETIT 2016) [1].
The latter work presented two methods of parametric identification of decoupled multivariable Hammerstein model.
It consists of one nonlinear static block and one linear dynamic block. Many process (chemical and biological process, signal processing, *etc.*) have this structure, for example: pH neutralization processes [2], polymerization reactor [3], distillation columns [4] and dryer process [5].
Many system identification methods have been used to identify the single-input single-output (SISO) model. They can be divided into stochastic methods [6,7], iterative methods [8], over-parameterization methods [9], separable least squares methods [10,11], blind identification methods [12] and frequency domain methods [13].
It is possible to transform the SISO Hammerstein model to MISO and MIMO model which is linear in the parameters [14]. Serval approches have been proposed to identify MIMO Hammerstein model. In [15,16,17], neuronal networks and fuzzy logic have been used to deal with more general nonlinearities. An approach based on multivariable cardinal cubic spline functions to model the static nonlinearities have been proposed in [18]. The Least Squares Support Vector Machines (LS-SVMs) have been presented in [19,20]. A generalized Hammerstein model consisting of a static polynomial function in series with time-varying linear model is developed in order to model the Hammerstein-like multivariable processes whose linear dynamics vary over the operating space in [21]. In this work, we propose a new coupled structure identification of MIMO model with separate nonlinearities. It is organized as follows: SISO Hammerstein system is presented in part 1 of section 2. A new coupled structure for MIMO Hammerstein system is developed in part 2 of section 2. Simulation results of a quadruple-tank process is given in section 3. Finally, a conclusion is made.

2 Parametric identification of Hammerstein model

2.1 Parametric identification of SISO Hammerstein model

Assume that the Hammerstein model of Figure 1 is composed of a nonlinear block $F(.)$ associated with a linear sub-system $\frac{B(q^{-1})}{A(q^{-1})}$. It is described by:

$$\begin{cases} y_k = \frac{B(q^{-1})}{A(q^{-1})} v_k + w_k \\ v_k = F(u_k) \end{cases} \quad (1)$$

*Corresponding Author: Chekib Ghorbel, Advanced System Laboratory, Polytechnic School of Tunisia, BP 743, 2078, La Marsa, Tunisie.
Contact No: +21698722239.
Email: chekib.ghorbel@enicarthage.rnu.tn

https://dx.doi.org/10.25046/aj020607

with:

$$A\left(q^{-1}\right) = 1 + a_1 q^{-1} + a_2 q^{-2} + \ldots + a_{n_A} q^{-n_A}$$

$$B\left(q^{-1}\right) = b_1 q^{-1} + b_2 q^{-2} + \ldots + b_{n_B} q^{-n_B}$$

$$v_k = \lambda_1\, u_k + \lambda_2\, u_k^2 + \ldots + \lambda_N\, u_k^N$$

q^{-1} delay operator, u_k input of the system, y_k output, v_k the unmeasurable internal signal and w_k represents the modeling error, external disturbances, *etc*.

In order to have a unique parameterizations of the Hammerstein model structure, the first coefficient of the nonlinear function $F(.)$ equals to 1, $\lambda_1 = 1$ [11,22].

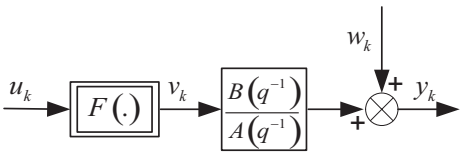


Figure 1: SISO Hammerstein model

The output y_k is given by:

$$y_k = -\sum_{i=1}^{n_A} a_i\, y_{k-i} + \sum_{i=1}^{n_B} b_i \left(u_{k-i} + \sum_{\rho=2}^{N} \lambda_\rho\, \mathrm{u}^\rho{}_{k-i} \right) + w_k \quad (2)$$

Equation (2) can be written in the following form:

$$y_k = \Phi_k^T\, \theta_k + w_k \quad (3)$$

with:

$$\Phi_k = \begin{pmatrix} Y_k \\ U_k \end{pmatrix},\ \theta = \begin{pmatrix} a \\ b \\ s \end{pmatrix},$$

Φ_k and $\theta_k \in R^{n_R}$ where $n_R = n_A + N\, n_B$,

$Y_k = \begin{pmatrix} -y_{k-1}, -y_{k-2}, \ldots, -y_{k-n_A} \end{pmatrix} \in R^{n_A}$,

$U_k = \begin{pmatrix} U_{1k}, U_{2k}, \ldots, U_{Nk}, \end{pmatrix} \in R^{N n_B}$,

$U_{j_k} = \begin{pmatrix} u^j{}_{k-1}, u^j{}_{k-2}, \ldots, u^j{}_{k-n_B} \end{pmatrix} \in R^{n_B}$,

for $j = 1, 2, \ldots, N$

$a = \begin{pmatrix} a_1, a_2, \ldots, a_{n_A} \end{pmatrix} \in R^{n_A}$,

$b = \begin{pmatrix} b_1, b_2, \ldots, b_{n_B} \end{pmatrix} \in R^{n_B}$,

$s = \begin{pmatrix} \lambda_2 b, \lambda_3 b, \ldots, \lambda_N b \end{pmatrix} \in R^{N n_B}$.

The parameter vector θ can be estimated using the RLS algorithm. It is described by the following equations:

$$\begin{cases} \hat{\theta}_k = \hat{\theta}_{k-1} + P_k\, \Phi_k\, \varepsilon_k \\ P_k = P_{k-1} - \frac{P_{k-1}\, \Phi_k\, \Phi_k^T\, P_{k-1}}{1 + \Phi_k^T\, P_{k-1}\, \Phi_k} \\ \varepsilon_k = y_k - \hat{\theta}_{k-1}^T\, \Phi_k \end{cases} \quad (4)$$

where P_k is the adaptation gain matrix, Φ_k is the observation vector and θ_k is the parameters vector.

2.2 Parametric identification of MIMO Hammerstein model

Two structures are used to identify MIMO Hammerstein models in the literature: with separate nonlinearities [21] or with combined nonlinearities [23]. The second case is the most general, but it can cause a very challenging parameter estimation problem because of the large number of parameters to be estimated.

In this paper, we developed a new coupled structure for MIMO Hammerstein model with separate nonlinearities. They are presented in Figure 2 and 4 where u_j, $v_{i,j}$, y_i for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, m$ are the system input, internal signal and system output. $F_{i,j}(.)$ are nonlinear function.

2.2.1 A decoupled structure identification of MIMO Hammerstein model

The decoupled structure of MIMO Hammerstein model is given in Figure 2. Each output $y_{i,k}$, $i = 1, 2, \ldots p$, of the multivariable system corresponds to a linear model, which at its input are introduced a nonlinear functions, Figure 3.

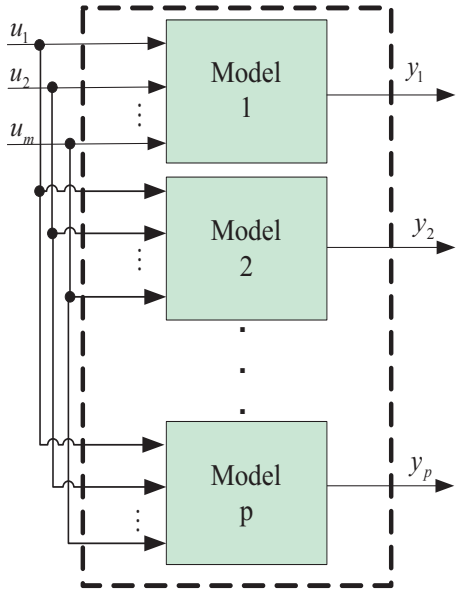


Figure 2: Decoupled structure of MIMO Hammerstein model

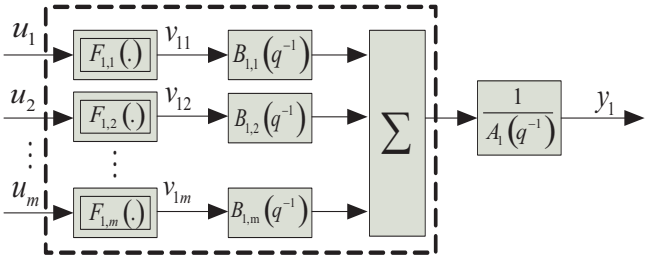


Figure 3: Structure of first model

Each output $y_{i,k}$, $i = 1, 2, \ldots p$, of the MIMO Hammerstein model is proposed as:

$$\begin{cases} A_i\left(q^{-1}\right)y_{i,k} = \sum_{j=1}^{m} B_{i,j}\left(q^{-1}\right)v_{j,k} \\ v_{i,j,k} = F_{i,j}\left(u_{j,k}\right) = u_{j,k} + \sum_{\rho=2}^{N} \lambda_{i,j,\rho}\, u_{j,k}^{\rho} \end{cases} \quad (5)$$

with:

$A_i\left(q^{-1}\right) = 1 + a_{i,1}q^{-1} + a_{i,2}q^{-2} + \ldots + a_{i,n_{A_i}}q^{-n_{A_i}}$

$B_{i,j}\left(q^{-1}\right) = b_{i,j,1}q^{-1} + b_{i,j,2}q^{-2} + \ldots + b_{i,j,n_{B_{i,j}}}q^{-n_{B_{i,j}}}$

System (5) can be rewritten as:

$$\begin{aligned} y_{i,k} = &-\sum_{\tau=1}^{n_{A_i}} a_{i,\tau}\, y_{i,k-\tau} + \sum_{j=1}^{m}\sum_{\tau=1}^{n_{B_{i,j}}} b_{i,j,\tau}\, u_{j,k-\tau} \\ &+ \sum_{j=1}^{m}\sum_{\tau=1}^{n_{B_{i,j}}}\sum_{\rho=2}^{N} b_{i,j,\tau}\,\lambda_{i,j,\rho}\, u_{j,k-\tau}^{\rho} \end{aligned} \quad (6)$$

then:

$$\begin{aligned} y_{i,k} = &-\sum_{\tau=1}^{n_{A_i}} a_{i,\tau}\, y_{i,k-\tau} + \sum_{j=1}^{m}\sum_{\tau=1}^{n_{B_{i,j}}} b_{i,j,\tau}\, u_{j,k-\tau} \\ &+ \sum_{j=1}^{m}\sum_{\tau=1}^{n_{B_{i,j}}}\sum_{\rho=2}^{N} s_{i,j,\tau,\rho}\, u_{j,k-\tau}^{\rho} \end{aligned} \quad (7)$$

Equation (7) can be written in the following form:

$$y_{i,k} = {\Phi_{i,k}}^{T}\ \theta_i \quad (8)$$

with:

$$\Phi_{i,k} = \begin{pmatrix} Y_{i,k} \\ U_k \\ \varphi_k \end{pmatrix},\ \theta_i = \begin{pmatrix} A_i \\ B_i \\ S_i \end{pmatrix}$$

$\Phi_{i,k}$ and $\theta_i \in R^{n_R}$; $n_R = n_{A_i} + \sum_{j=1}^{m} N n_{B_{i,j}}$

$$Y_{i,k} = \left(-y_{i,k-1},\ -y_{i,k-2},\ \ldots,\ -y_{i,k-n_{A_i}}\right) \in R^{n_{A_i}}$$

$$U_k = (U_{1,k},\ U_{2,k},\ \ldots,\ U_{m,k}) \in R^{m\times n_{B_{i,j}}}$$

$$U_{j,k} = \left(u_{j,k-1},\ u_{j,k-2},\ \ldots,\ u_{j,k-n_{B_{i,j}}}\right) \in R^{n_{B_{i,j}}}$$

$$\varphi_k = (\varphi_{1,k},\ \varphi_{2,k},\ \ldots,\ \varphi_{m,k}) \in R^{(N-1)\times m\times n_{B_{i,j}}}$$

$$\varphi_{j,k} = \left(\varphi_{j,1,k},\ \varphi_{j,2,k},\ \ldots,\ \varphi_{j,n_{B_{i,j}},k}\right) \in R^{(N-1)\times n_{B_{i,j}}}$$

$$\varphi_{j,\tau,k} = \left(u_{j,k-\tau}^{2},\ u_{j,k-\tau}^{3},\ \ldots,\ u_{j,k-\tau}^{N}\right) \in R^{(N-1)}$$

$$A_i = \left(a_{i,1},\ a_{i,2},\ \ldots,\ a_{i,n_{A_i}}\right) \in R^{n_{A_i}}$$

$$B_i = (B_{i,1},\ B_{i,2},\ \ldots,\ B_{i,m}) \in R^{m\times n_{B_{i,j}}}$$

$$B_{i,j} = \left(b_{i,j,1},\ b_{i,j,2},\ \ldots,\ b_{i,j,n_{B_{i,j}}}\right) \in R^{n_{B_{i,j}}}$$

$$S_i = (S_{i,1},\ S_{i,2},\ \ldots,\ S_{i,m}) \in R^{(N-1)\times m\times n_{B_{i,j}}}$$

$$S_{i,j} = \left(S_{i,j,1},\ S_{i,j,2},\ \ldots,\ S_{i,j,n_{B_{i,j}}}\right) \in R^{(N-1)\times n_{B_{i,j}}}$$

$$S_{i,j,\tau} = \left(s_{i,j,\tau,2},\ s_{i,j,\tau,3},\ \ldots,\ s_{i,j,\tau,N}\right) \in R^{(N-1)}$$

$$s_{i,j,\tau,\rho} = b_{i,j,\tau}\,\lambda_{j,\rho}$$

for $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, m$, and $\tau = 1, 2, \ldots, n_{B_{i,j}}$

The steps of the identification scheme are summarized as follows:

1. choosing an initial values for the adaptation matrix,
2. acquiring the input and output of the system and form the vector data as shown in (11) using the present and past values of the input u, output y, and u power,
3. solving the estimate parameter A_i, B_i and S_i using the algorithm RLS,
4. solving $\lambda_{i,j,\rho}$ using the estimated values $b_{i,j,\tau}$ and $s_{i,j,\tau,\rho}$ as:

$$\lambda_{i,j,\rho} = \left(\sum_{\tau=1}^{n_{B_{i,j}}} b_{i,j,\tau}^{2}\right)^{-1}\left(\sum_{\tau=1}^{n_{B_{i,j}}} b_{i,j,\tau}\, s_{i,j,\tau,\rho}\right) \quad (9)$$

2.2.2 A new coupled structure identification of MIMO Hammerstein system

The structure of this method is given in Figure 5. Each output of the system is depended on inputs and all other system outputs, Figure 4.

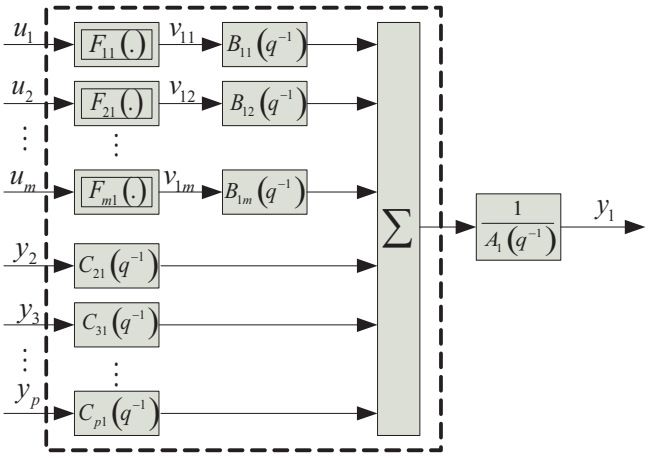


Figure 4: Structure of first model

Each output $y_{i,k}$, $i = 1, 2, \ldots p$, of the MIMO Hammerstein model is described by:

$$\begin{cases} A_i\left(q^{-1}\right)y_{i,k} = \sum_{j=1}^{m} B_{i,j}\left(q^{-1}\right)v_{j,k} + \sum_{\substack{l=1 \\ l\neq i}}^{p} C_{i,l}\left(q^{-1}\right)y_{l,k} \\ v_{i,j,k} = F_{i,j}\left(u_{j,k}\right) = u_{j,k} + \sum_{\rho=2}^{N} \lambda_{i,j,\rho}\, u_{j,k}^{\rho} \end{cases} \quad (10)$$

with:

$$A_i\left(q^{-1}\right) = 1 + a_{i,1}q^{-1} + a_{i,2}q^{-2} + \ldots + a_{i,n_{A_i}}q^{-n_{A_i}}$$
$$B_{i,j}\left(q^{-1}\right) = b_{i,j,1}q^{-1} + b_{i,j,2}q^{-2} + \ldots + b_{i,j,n_{B_{i,j}}}q^{-n_{B_{i,j}}}$$
$$C_{i,l}\left(q^{-1}\right) = c_{i,l,1}q^{-1} + c_{i,l,2}q^{-2} + \ldots + c_{i,l,n_{C_{i,l}}}q^{-n_{C_{i,l}}}$$

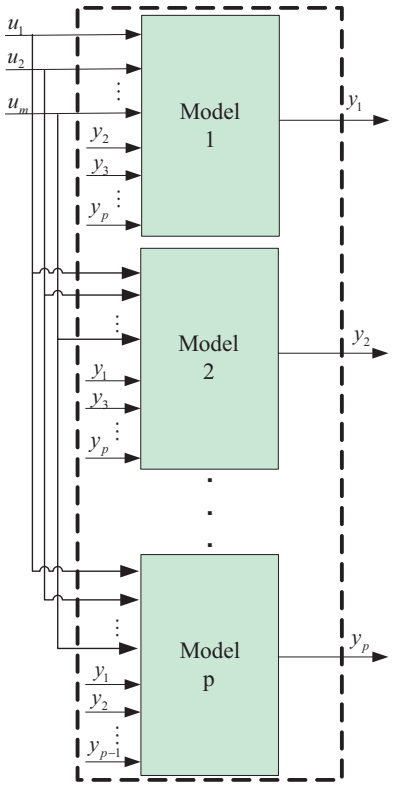


Figure 5: A proposed structure of MIMO Hammerstein model

System (10) can be rewritten as:

$$y_{i,k} = -\sum_{\tau=1}^{n_{A_i}} a_{i,\tau}\, y_{i,k-\tau} + \sum_{j=1}^{m}\sum_{\tau=1}^{n_{B_{i,j}}} b_{i,j,\tau}\, u_{j,k-\tau} + \sum_{j=1}^{m}\sum_{\tau=1}^{n_{B_{i,j}}}\sum_{\rho=2}^{N} b_{i,j,\tau}\,\lambda_{i,j,\rho}\, u_{j,k-\tau}^{\rho} + \sum_{\substack{l=1\\ l\neq i}}^{p}\sum_{\tau=1}^{n_{C_{i,l}}} c_{i,l,\tau}\, y_{l,k-\tau} \tag{11}$$

then:

$$y_{i,k} = -\sum_{\tau=1}^{n_{A_i}} a_{i,\tau}\, y_{i,k-\tau} + \sum_{j=1}^{m}\sum_{\tau=1}^{n_{B_{i,j}}} b_{i,j,\tau}\, u_{j,k-\tau} + \sum_{j=1}^{m}\sum_{\tau=1}^{n_{B_{i,j}}}\sum_{\rho=2}^{N} s_{i,j,\tau,\rho}\, u_{j,k-\tau}^{n} + \sum_{\substack{l=1\\ l\neq i}}^{p}\sum_{\tau=1}^{n_{C_{i,l}}} c_{i,l,\tau}\, y_{l,k-\tau} \tag{12}$$

Equation (12) can be written in the following form:

$$y_{i,k} = \Psi_{i,k}^{T}\theta_i^{new} \tag{13}$$

with:

$$\Psi_{i,k} = \begin{pmatrix} Y_{i,k} \\ U_k \\ \varphi_k \\ Y_{L,k} \end{pmatrix},\quad \theta_i^{new} = \begin{pmatrix} A_i \\ B_i \\ S_i \\ C_i \end{pmatrix}$$

$$\Phi_{i,k} \text{ and } \theta_i \in R^{n_R};\ n_R = n_{A_i} + \sum_{j=1}^{m} N n_{B_{i,j}} + n_{C_{i,l}} \times (p-1)$$

$$Y_{\substack{L,k\\ L\neq i}} = \left(Y_{1,k},\ Y_{2,k},\ \ldots, Y_{p,k}\right) \in R^{(p-1)\times n_{C_{i,l}}}$$

$$Y_{l,k} = \left(-y_{l,k-1},\ -y_{l,k-2},\ \ldots,\ -y_{l,k-n_{C_{i,l}}}\right) \in R^{n_{C_{i,l}}}$$

$$C_i = \left(C_{i,1},\ C_{i,2},\ \ldots,\ C_{i,p}\right) \in R^{(p-1)\times n_{C_{i,l}}}$$

$$C_{i,j} = \left(c_{i,j,1},\ c_{i,j,2},\ \ldots,\ c_{i,j,n_{C_{i,l}}}\right) \in R^{n_{C_{i,l}}}$$

for $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, m$, and $\tau = 1, 2, \ldots, n_{B_{i,j}}$, and $l = 1, 2, \ldots, p$, such as $l \neq i$.

Parametric identification consists to determine the parameters of the system based on vectors of inputs and outputs, using the recursive least squares algorithm.

3 Application: quadruple-tank process

A quadruple-tank process is used to illustrate the performance of the proposed structures of the MIMO Hammerstein model.

3.1 System description

The system setup is a model of a chemical plant fragment. Very often tanks are coupled through pipes and the reactant level and flow has to be controlled. The type of the experiments was performed on the 33-041 Coupled Tanks System of Feedback Instruments [24]. This plant, a variant of the quadruple tank process originally proposed in [25], is a model of a fragment of a chemical plant, Figure 6.



Figure 6: Plant of four coupled tanks [26]

The line diagram of the reservoir system is shown in Figure 7. The coupled tanks unit consists of four tanks placed on a rig. Another reservoir tank is placed at the bottom. In the reservoir two submersible pumps are placed, which pump the water on command to the tanks. The water flows freely to the bottom tanks through the configurable orifice. The way

the water flows through the setup can be configured in many ways with manual valves label led (MVA, MVB, MVC, MVD, MVE, MVF, MVG, MV1, MV2, MV3 and MV4).

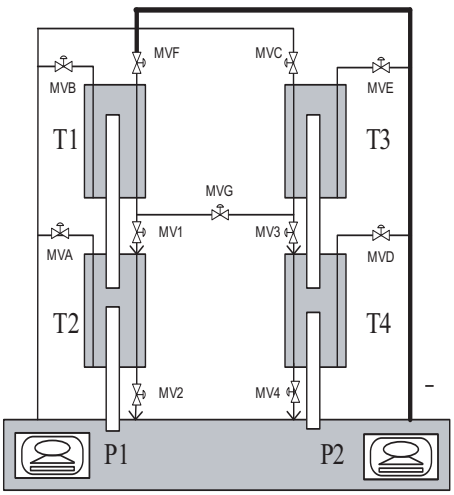


Figure 7: Line diagram for quadruple-tank process

3.2 Mathematical modelling

The quadruple-tank process admits the following nonlinear model [25] which has been assembled in simulink:

$$\begin{cases} \dot{h}_1(t) = \eta v_1(t) - \frac{a_1}{A}\sqrt{2gh_1(t)} - \frac{a_{13}}{A}\sqrt{2g(h_1(t)-h_3(t))} \\ \dot{h}_2(t) = \frac{a_1}{A}\sqrt{2gh_1(t)} - \frac{a_2}{A}\sqrt{2gh_2(t)} \\ \dot{h}_3(t) = \eta v_2(t) - \frac{a_3}{A}\sqrt{2gh_3(t)} - \frac{a_{13}}{A}\sqrt{2g(h_1(t)-h_3(t))} \\ \dot{h}_4(t) = \frac{a_3}{A}\sqrt{2gh_3(t)} - \frac{a_4}{A}\sqrt{2gh_4(t)} \end{cases} \quad (14)$$

where h_i, for $i = 1,2,3,4$, denote the water level in the corresponding tank and v_i, for $i = 1,2$, are voltages applied to the pumps. a_i, for $i = 1,2,3,4$, are the outlet area of the tanks, a_{13} is the outlet area betwixt tanks 1 and 3; η constant relating the control voltage with the water flow from the pump, A is the cross-sectional area of the tanks and g is the gravitational constant.

3.3 Simulation and results

The proposed structures has been tested with the model parameters presented in table 1.

	Value	Unit	Description
h_i	$0-25$	cm	Water level of tank i
v_i	$0-5$	V	Voltage level of pump i
S	0.014	m^2	Cross-sectional area
a_i	$5e-5$	m^2	Outlet area of tank i
a_{13}	$5e-5$	m^2	Outlet area betwixt $T1$ and $T3$
η	$2e-3$	$\frac{m^3}{V.s}$	Water level of tank i
g	9.81	m^{-2}	Gravitational constant

Table 1: Parameters of the plant

The inputs u_1 and u_2 are show in Figure 8. They are set to $[0\ldots+5V]$. The sample time used for all simulations is $T_e = 1s$.

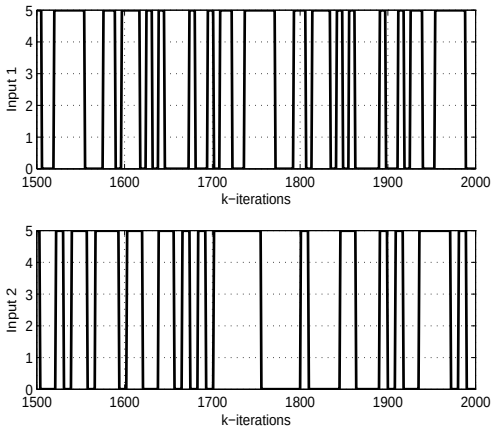


Figure 8: Inputs u_1 and u_2

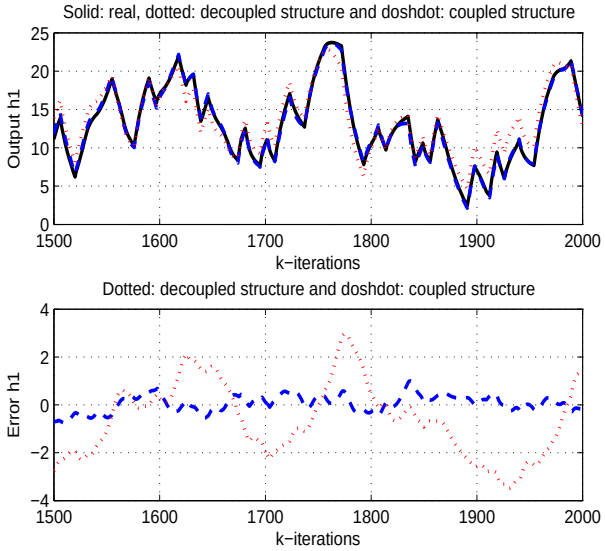


Figure 9: Response of the real and estimated h_1

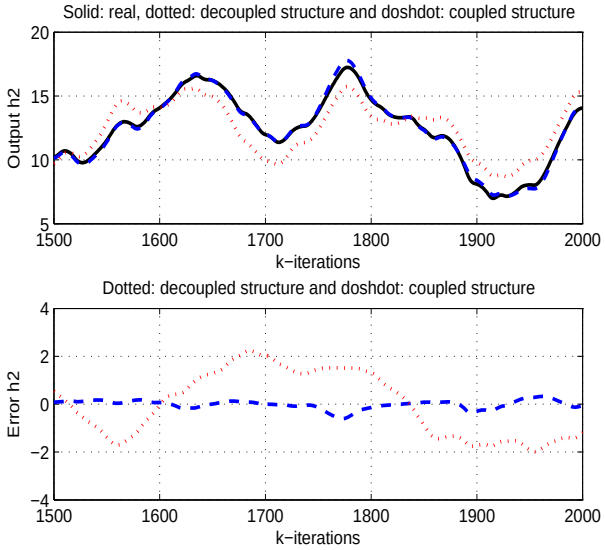


Figure 10: Response of the real and estimated h_2

Figures 9 to 12 show the results of the experiments with two method from a comparative point of view. Each of them shows a superposition of the actual output and the two outputs estimated as their two error curves. Solid line represents the real output, dotted line represents the estimated systems output signals with the decoupled structure and the dash dot lines represent the estimated systems output signals with the coupled structure in all four graphs. The responses of the original system and the results of the proposed method are very similar. It is clear that the error corresponds to the proposed structure is smaller than that corresponding to the decoupled structure. On inspection, the proposed structure is seen to work better than the decoupled structure. Thus, the feasibility and superiority of this proposed identification method are validated.

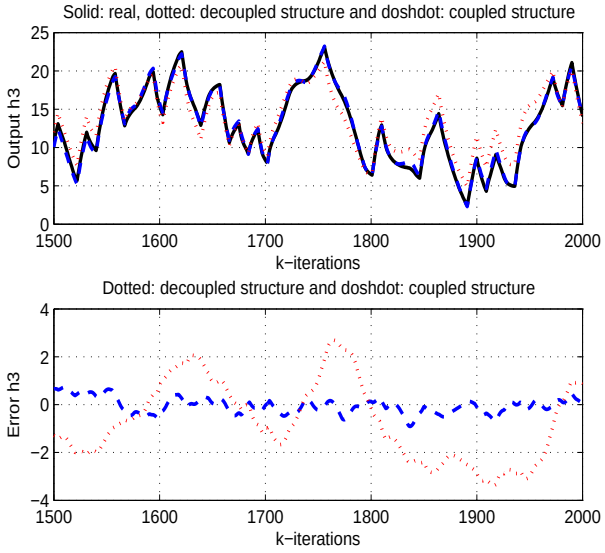


Figure 11: Response of the real and estimated h_3

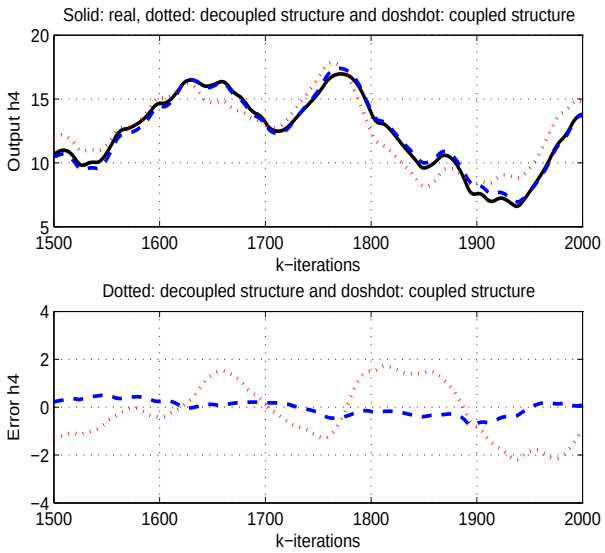


Figure 12: Response of the real and estimated h_4

4 Conclusion

In this study, a new coupled structure identification of MIMO Hammerstein model with separate nonlinearities has been developed. The method is based on RLS algorithm. A comparative study using simulation results for quadruple-tank process, between the proposed structure and the decoupled structure is discussed. Simulation results reveal the performance and effectiveness of the proposed method.
As a perspective, we will judge the performance of the method used in the presence of perturbations

References

1. Z. Rayouf, C. Ghorbel, and N. B. Braiek, "An extended identification methods for multivariable Hammerstein system" in 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, Hammamet Tunisia, 2016. https://doi.org/10.1109/SETIT.2016.7939921
2. Z. O. U. Zhiyun, Y. U. Meng, W. A. N. G. Zhizhen, L. I. U. Xinghong, G. U. O. Yuqing, F. ZHANG, and G. U. O. Ning, "Nonlinear model algorithmic control of a pH neutralization process" Chinese Journal of Chemical Engineering, **21**(4), 395–400, 2013. https://doi.org/10.1016/S1004-9541(13)60479-6
3. D. Z. Li, Y. X. Jia, Q. S. Li, and Q. B. Jin, "Identification and nonlinear model predictive control of MIMO Hammerstein system with constraints" Journal of Central South University, **24**(2), 448-458, 2017. https://doi.org/10.1007/s11771-017-3447-3
4. R. Kanthasamy, H. Anwaruddin, and S. K. Sinnadurai, "A new approach to the identification of distillation column based on hammerstein model" Modelling and Simulation in Engineering, **2014**(11), 1-7, 2014. http://dx.doi.org/10.1155/2014/813757
5. D. K. Rollins, N. Bhandari, A. M. Bassily, G. M. Colver, and S. T. Chin, "A continuous-time nonlinear dynamic predictive modeling method for hammerstein processes" Industrial and engineering chemistry research, **42**(4), 860-872, 2003. https://doi.org/10.1021/ie020169
6. D. Wang, and F. Ding, " Extended stochastic gradient identification algorithms for Hammerstein-Wiener ARMAX systems" Computers and Mathematics with Applications, **56**(12), 3157–3164, 2008. https://doi.org/10.1016/j.camwa.2008.07.015
7. F. Ding, "Hierarchical multi-innovation stochastic gradient algorithm for Hammerstein nonlinear system modeling" Applied Mathematical Modelling, **37**(4), 1694–1704, 2013. https://doi.org/10.1016/j.apm.2012.04.039
8. R. Su, B. Wang, and S. Liu, "The iterative identification method for hammerstein nonlinear channel" in 10th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM), Beijing, China, 2014. https://doi.org/10.1049/ic.2014.0075
9. J. Wang, Q. Zhang, and L. Ljung. "Revisiting Hammerstein system identification through the two-stage algorithm for bilinear parameter estimation" Automatica, **45**(11), 2627-2633, 2009. https://doi.org/10.1016/j.automatica.2009.07.033
10. D. Westwick, and R. Kearney, "Separate least squares identification of nonlinear Hammerstein models: application to stretch reflex dynamics" Annals of Biomedical Engineering, **29**(8), 707-718, 2001. https://doi.org/10.1114/1.1385806
11. E. Bai, and M. Fu, "Identification of linear systems with hard input nonlinearities of known structure" IEEE Transactions on Signal Processing, **38**(5), 853-860, 2002. https://doi.org/10.1016/S0005-1098(01)00281-3

12. E. Bai, and M. Fu, "A blind approach to Hammerstein model identification" IEEE Transactions on Signal Processing, **50**(7), 1610-1619, 2002. https://doi.org/10.1109/TSP.2002.1011202

13. E. Bai, "Frequency domain identification of Hammerstein models" IEEE Transactions on Automatic Control, **48**(4), 530-542, 2003. https://doi.org/10.1109/TAC.2003.809803

14. Y. Liu, J. Sheng, and R. Ding, "Convergence of stochastic gradient estimation algorithm for multivariable ARX-like systems" Computers and Mathematics with Applications, **59**(8), 2615-2627, 2010. https://doi.org/10.1016/j.camwa.2010.01.030

15. J. Abonyi, R. Babuška, M. A. Botto, F. Szeifert, and L. Nagy, "Identification and control of nonlinear systems using fuzzy Hammerstein models" Industrial and engineering chemistry research, **39**(11), 4302–4314, 2000. https://doi.org/10.1021/ie990629e

16. L. Jia, M. S. Chiu, and S. S. Ge, "A noniterative neuro-fuzzy based identification method for Hammerstein processes" Journal of Process Control, **15**(7), 749-761, 2005. https://doi.org/10.1016/j.jprocont.2005.03.006

17. A. Jawad, A. M. Butt, M. T. Riaz1, S. Bhutta1, M. Z. khan1, and I. U. Haq "Multiclass Myoelectric Identification of Five Fingers Motion using Artificial Neural Network and Support Vector Machine" Journal of Advances in Science, Technology and Engineering Systems, **2**(3), 1026–1033, 2017. https://doi.org/10.25046/aj0203130

18. K. H. Chan, J. Bao, and W. J. Whiten, "Identification of MIMO Hammerstein systems using cardinal spline functions" Journal of Process Control, **16**(7), 659-670, 2006. https://doi.org/10.1016/j.jprocont.2006.01.004

19. I. Goethals, K. Pelckmans, J. A. Suykens, and B. De Moor, "Identification of MIMO Hammerstein models using least squares support vector machines" Automatica, **41**(7), 1263–1272, 2005. https://doi.org/10.1016/j.automatica.2005.02.002

20. H. Liang, and B. Wang, "Identification of MIMO Hammerstein Models Using Support Vector Machine" In International Symposium on Neural Networks, Springer, Berlin, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-72395-0-51

21. Y. M. Hlaing, M. S. Chiu, and S. Lakshminarayanan"Modeling and control of multivariable process using generalized Hammerstein model" Chemical Engineering Research and Design Transactions, **85**(4), 445-454, 2007. https://doi.org/10.1205/cherd06210

22. F. Ding, and T. Chen, "Identification of Hammerstein nonlinear ARMAX system" IEEE Transactions Automatic Control, **41**(9), 1479-1489, 2005. https://doi.org/10.1016/j.automatica.2005.03.026

23. K. H.Chan, J. Bao, and W. J. Whiten, "Identification of MIMO Hammerstein systems using cardinal spline functions" Journal of Process Control, **16**(7), 659–670, 2006. https://doi.org/10.1016/j.jprocont.2006.01.004
http://dx.doi.org/10.1080/00207721.2014.951419

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 58-68 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Enhanced Outdoor to Indoor Propagation Models and Impact of Different Ray Tracing Approaches at Higher Frequencies

Muhammad Usman Sheikh[*,1], Kimmo Hiltunen[2], Jukka Lempiainen[3]

[1]*Tampere University of Technology, Department of Electronics and Communications Engineering, Finland.*

[2]*Ericsson Research, Helsinki, Finland.*

[3]*Tampere University of Technology, Department of Electronics and Communications Engineering, Finland.*

ARTICLE INFO

Article history:
Received: 26 November, 2017
Accepted: 05 February, 2018
Online: 08 March, 2018

Keywords:
Outdoor to Indoor Propagation
Angular loss
Wall loss model
Building penetration loss
Ray tracing

ABSTRACT

The main target of this article is to study the provision of indoor service (coverage) using outdoor base station at higher frequencies i.e. 10 GHz, 30 GHz and 60 GHz. In an outdoor to indoor propagation, an angular wall loss model is used in the General Building Penetration (GBP) model for estimating the additional loss at the intercept point of the building exterior wall. A novel angular wall loss model based on a separate incidence angle in azimuth and elevation plane is proposed in this paper. In the second part of this study, an Extended Building Penetration (EBP) model is proposed, and the performance of EBP model is compared with the GBP model. In EBP model, the additional fifth path known as the "Direct path" is proposed to be included in the GBP model. Based on the evaluation results, the impact of the direct path is found significant for the indoor users having the same or closed by height as that of the height of the transmitter. For the indoor users located far away from the exterior wall of building, a modified and enhanced approach of ray tracing type is proposed in this article. In the light of acquired simulation results, the impact of a modified ray tracing approach is emphasized.

1. Introduction

This article is an extension of research work originally presented at International Wireless Communication and Mobile Computing (IWCMC'17) conference [1]. In reference [1], studies were made at 10 GHz only; whereas in this article the impact of different propagation models is also analyzed at 30 GHz and 60 GHz. Additionally, two different ray tracing approaches are also analyzed in this article.

Outdoor to indoor propagation in a small cell environment involves Line of Sight (LOS) and Non-LOS (NLOS) propagation. It includes path loss computation, determination of reflection and diffraction loss, penetration loss, and other indoor losses. The penetration loss can be divided into four major categories i.e. building exterior wall penetration loss also known as building penetration loss, floor (ceiling) penetration loss, indoor propagation (room penetration) loss, and angular wall loss [2]. Several propagation models are given in literature e.g. General Building Penetration (GBP) model provides and evaluates the candidate paths, Berg's recursive method for micro cell path loss [3], and the linear attenuation model for indoor propagation. The frequency dependent penetration loss models are presented at [4].

The angular wall loss model presented in [2] and used in [4] depends on a single three-dimensional incidence angle. This paper presents a new angular wall loss model based on a separate incidence angle in azimuth and elevation plane. An Extended Building Penetration (EBP) model is proposed in this article. Path gain is used as a metric to compare the performance of different angular wall loss models and building penetration loss models. Section II provides the details about the GBP model, frequency

[*] Muhammad Usman Sheikh, Tampere University of Technology
Email: muhammad.sheikh@tut.fi

dependent penetration models, angular wall loss model, and ray tracing. Whereas Section III explains the proposed angular wall loss model and extended building penetration model. Section IV gives the description of the simulation environment and provides the details about the assumptions and simulation parameters. Section IV discusses the results in detail. Finally, Section V concludes the paper.

2. Background Theory

2.1. General Building Penetration (GBP) Model

The General Building Penetration (GBP) model is based on the COST231 building penetration model presented at [2]. The model considers a path from each exterior wall of the building. A top view of a single building scenario with a Transmitter (TX) located in an outdoor environment is shown in Figure 1. For each Receiver (RX) point located inside the building, there exist four candidate paths. A path coming from the front face of the building without any diffraction is called the "LOS path", and the paths which are reaching the receiver point after diffracting from the corners of the building are known as "NLOS paths". For the NLOS paths, the Berg's recursive model is used to capture the additional path loss due to the diffraction around the building, before entering the building, whereas the outdoor to indoor and other indoor losses are determined from the COST231 building penetration model as done in [4]. Each multipath has two propagation parts; the first part comprises free space propagation between the transmitter and the building's exterior wall intercept point. The second part is the propagation of a path in an indoor environment after penetrating through the exterior wall to the receiver point.

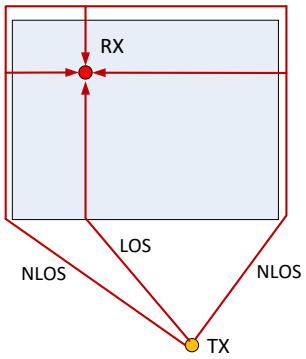


Figure 1. Illustration of propagation paths from outdoor base station to indoor location (top view).

A well-known Friis free space path loss model is used to find the path loss in the first free space propagation part. For the second part of the propagation path, along with the free space path loss model, the building penetration loss model, an indoor wall loss model, and an angular wall loss model is used to compute the additional loss due to indoor propagation. The received signal strength at the receiver location inside the building is the sum of four paths as shown in Figure 1.

Figure 2 shows the side view of the paths entering the building for the two receiver points RX1 and RX2 located on the Floor 1 and Floor 4, respectively. The signal path intercepts the building wall at the same height as that of a receiver height; therefore the ceiling penetration loss is not taken into account. It is important here to mention that in GBP model both the LOS path and NLOS paths do not penetrate through the ceiling.

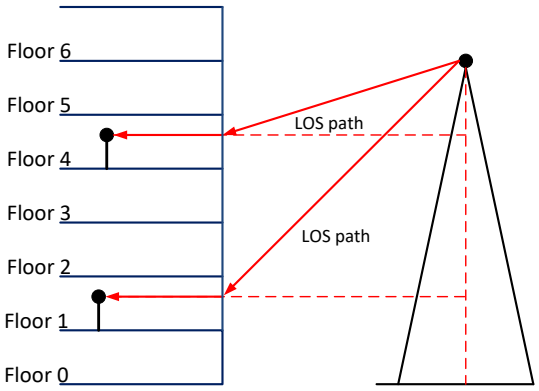


Figure 2. Illustration of LOS paths entering the building (side view).

2.2. Building Penetration Loss (BPL)

The signal experiences a penetration loss while penetrating from the outdoor environment to the building. Outdoor to indoor penetration loss is generally termed as Building Penetration Loss (BPL). The building penetration loss relies heavily on the frequency and on the material characteristics of the building; therefore the BPL can be significantly different for different material types at different frequencies. Generally, the old houses are composed of plane standard glass windows and concrete wall, while the Infrared Reflective (IRR) glass windows are commonly used in the new modern energy saving houses. In reference [4], the old buildings are assumed to have 30 % of the standard glass windows and 70 % of the concrete wall. Similarly, the assumption for new modern building type corresponds to the 70 % of the IRR glass windows and 30 % of the concrete wall.

A simple model structure has been proposed in [4] to model a single material frequency dependent penetration loss. The penetration loss for different material types is provided at references [5-9].

$$L_{\text{Single glass,dB}} = 0.1 * \text{Frequency}_{\text{GHz}} + 1, \quad (1)$$

$$L_{\text{Double glass,dB}} = 0.2 * \text{Frequency}_{\text{GHz}} + 2, \quad (2)$$

$$L_{\text{IRR glass,dB}} = 0.3 * \text{Frequency}_{\text{GHz}} + 23, \quad (3)$$

The penetration loss for the concrete wall is modeled as

$$L_{\text{Concrete,dB}} = 4 * \text{Frequency}_{\text{GHz}} + 5, \quad (4)$$

As the buildings are composite of windows and concrete wall, the building penetration loss for old buildings and new buildings is modeled as shown in (5) and (6), respectively [10].

$$L_{\text{Old building,dB}} = -10\text{Log}_{10}\left[0.3 * 10^{\frac{-L_{\text{Double glass,dB}}}{10}} + 0.7 * 10^{\frac{-L_{\text{Concrete,dB}}}{10}}\right], \quad (5)$$

$$L_{\text{New building,dB}} = -10\text{Log}_{10}\left[0.7 * 10^{\frac{-L_{\text{IRR glass,dB}}}{10}} + 0.3 * 10^{\frac{-L_{\text{Concrete,dB}}}{10}}\right], \quad (6)$$

The building penetration loss as a function of frequency for different types of building is shown in Figure 3.

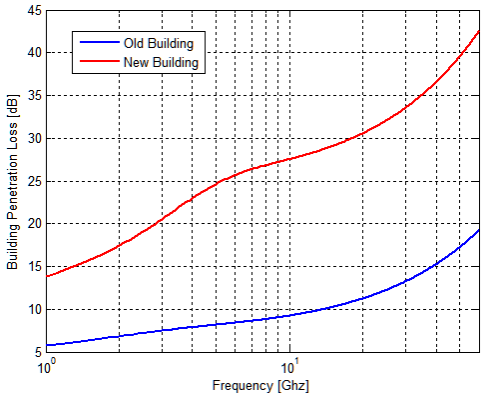


Figure 3. Building penetration loss as a function of frequency.

2.3. Indoor Propagation Loss

In an indoor environment, generally the indoor walls are made up of standard glass alternatively plaster. In [4], two different indoor wall loss models are presented as a function of the frequency assuming an average wall distance of 4 m. The Indoor Loss Model 1 assumes an indoor wall of standard glass, whereas Indoor Loss Model 2 is based on the measurements performed in [5]. Two indoor wall loss models are modeled as shown in (7) and (8).

$$L^{(1)}_{\text{Wall loss,dB/m}} = L_{\text{Single glass,dB}}, \quad (7)$$

$$L^{(2)}_{\text{Wall loss,dB/m}} = 0.2 * \text{Frequency}_{\text{GHz}} + 1.7, \quad (8)$$

Indoor loss as a function of frequency for two different indoor wall loss models, expressed as db/m is shown in Figure 4.

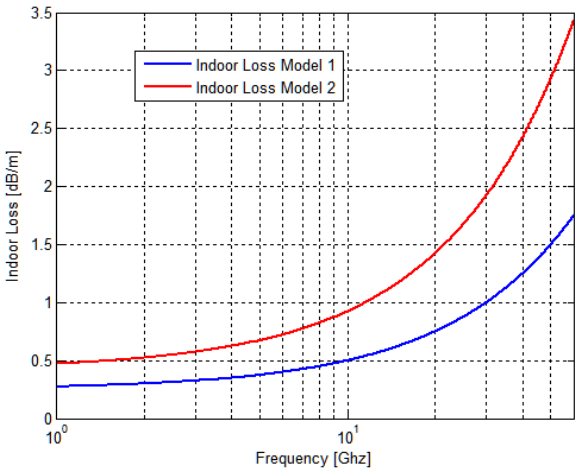


Figure 4. Indoor loss models as a function of frequency.

2.4. Body Loss as Function of Frequency

In reference [10], the frequency dependency of the body loss is modeled as given in (9).

$$L_{\text{Body,dB}} = \frac{\text{Frequency}_{\text{GHz}}}{60} + 3, \quad (9)$$

Equation (9) shows that frequency has a negligible impact on a considered body loss model at 10 GHz or lower frequencies, as traditionally the body loss is assumed to be 3 dB. However, equation (9) gives additional 0.5 dB and 1 dB body loss at 30 GHz and 60 GHz, respectively.

2.5. Angular Wall Loss Model Based on a Single Three-dimensional Incidence Angle

In addition to the building penetration loss and indoor wall loss, there exists an angular wall loss. An angular wall loss model presented at [2, 4] is used to include the angular loss that can be experienced at the building's exterior wall intercept point. The angular wall loss model is given by (10),

$$L_{\text{angular,dB}} = 20 * [1 - \text{Cos}(\theta_i)]^2, \quad (10)$$

where θ_i is the single three-dimensional incidence angle.

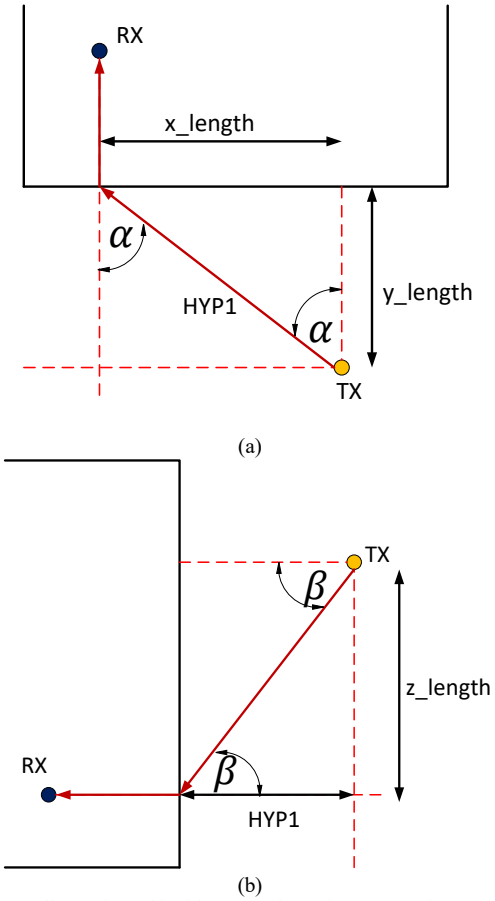


Figure 5. Illustration of incidence angles at intercept point on wall in, a) Azimuth (horizontal) plane, and b) Elevation (vertical plane).

To understand the geometry and the computation of θ_i, consider an example scenario in an azimuth plane (top view) as shown in Figure 5(a). The receiver point is located inside the

building and the transmitter is located in an outdoor environment. The distance between the transmitter and the receiver point along the x-axis and the y-axis is given by x_length and y_length, respectively. HYP1 is the distance between the transmitter and the incidence point on the wall, and is given in (11).

$$\mathrm{HYP1} = \sqrt{\mathrm{x_length}^2 + \mathrm{y_length}^2}, \tag{11}$$

In Figure 5(b), z_length is the difference of height between the transmitter and the receiver point, and HYP2 is the three dimensional distance between the transmitter and the wall intercept point. HYP2 is calculated as follows:

$$\mathrm{HYP2} = \sqrt{\mathrm{HYP1}^2 + \mathrm{z_length}^2}, \tag{12}$$

The single three-dimensional incidence angle θ_i can be computed as

$$\theta_i = \cos^{-1}\left(\frac{\mathrm{y_length}}{\mathrm{HYP2}}\right), \tag{13}$$

Considering the geometry shown in Figure 5, the equation given in (13) can also be re-written in a simplified form as

$$\mathrm{L_{angular,dB}} = 20 * \left[1 - \frac{\mathrm{y_length}}{\mathrm{HYP2}}\right]^2, \tag{14}$$

2.6. Ray Tracing (RT)

Ray tracing techniques can be used for estimating the received signal level and for the characterization of the radio propagation environment. Finding the multipath components between the transmitter and receiver is the first step towards the computation of the received electric field or power at the receiver point. By using Image Theory (IT) algorithm, all multipaths with the given finite number of reflections and diffractions can theoretically be found between the transmitter and receiver. An Image theory algorithm shows a high level of accuracy and precision. Determination of multipath components by image based ray tracing technique may require large computation time. The complexity and the computational time of the ray tracing algorithm increases with the increase in number of supported reflections and diffractions [11-13]. Reflection losses are determined by reflection coefficients. The reflection coefficient depends upon the polarization and on the material permittivity. For perpendicular and parallel polarization, the reflection coefficients are given in (15) and (16), respectively. β is the angle between the incident ray and the reflected surface, and ε_r is the material permittivity of the reflecting surface.

$$|\Gamma_\perp| = \frac{\mathrm{Sin}(\beta) - \sqrt{\varepsilon_r - \mathrm{Cos}^2(\beta)}}{\mathrm{Sin}(\beta) + \sqrt{\varepsilon_r - \mathrm{Cos}^2(\beta)}} \tag{15}$$

$$|\Gamma_\parallel| = \frac{-\varepsilon_r\,\mathrm{Sin}(\beta) + \sqrt{\varepsilon_r - \mathrm{Cos}^2(\beta)}}{\varepsilon_r\,\mathrm{Sin}(\beta) + \sqrt{\varepsilon_r - \mathrm{Cos}^2(\beta)}} \tag{16}$$

There are several ways for computing the diffraction loss; however a recursive method proposed in [14] is used in this article to compute the diffraction loss.

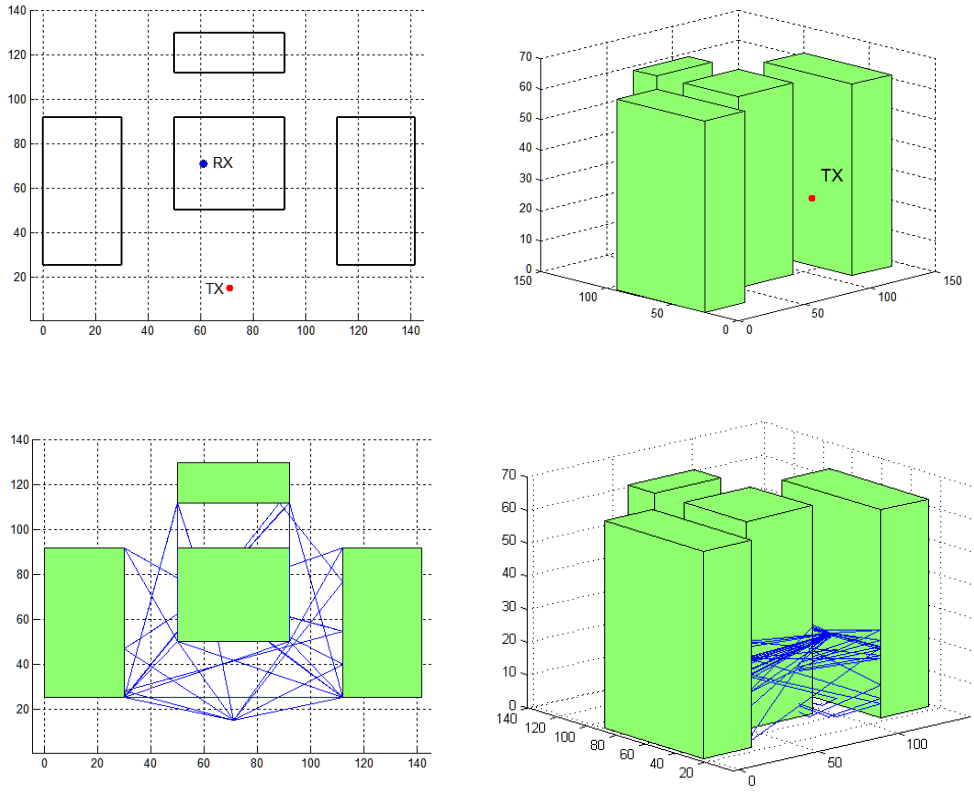


Figure 6. Illustration of ray tracing, (a) Two dimensional view of sample building scenario, (b) Three dimensional view of sample building scenario, (c) Two dimensional view of ray tracing, and (d) Three dimensional view of ray tracing.

Figure 6(a) shows the sample building scenario in two-dimensional environment with transmitter and receiver locations marked with TX and RX, respectively. Figure 6(b) shows the three-dimensional view of a sample scenario with TX height at 31.5 meter. The receiver point is located inside the central building at the height of 10.5 m. Figure 6(c) illustrates the ray tracing in two-dimensional view, whereas Figure 6(d) shows the three-dimensional view of ray tracing. Blue lines show different propagation paths with finite number of reflections and diffractions, and LOS path.

3. Proposed Models

3.1. Angular Wall Loss Model based on Separate Incidence Angle in Azimuth and Elevation Plane

In this article, a novel angular wall loss model based on a separate incidence angle in azimuth and elevation plane is proposed. It is proposed to compute the total angular loss as a sum of the angular loss in an azimuth and elevation plane. Unlike the single three-dimensional angle, in this proposed approach the incidence angle in the horizontal plane and vertical plane should be computed separately. The azimuth angular loss and elevation angular loss is given in (17) and (18), respectively.

$$L_{\text{Azimuth ang,dB}} = 10 * [1 - \text{Cos}(\alpha)]^2, \quad (17)$$

$$L_{\text{Elevation ang,dB}} = 10 * [1 - \text{Cos}(\beta)]^2 \quad (18)$$

The angles α and β used in (17) and (18), respectively, are shown in Figure 5. The angles α and β are the incidence angles computed separately in an azimuth and elevation plane. Now, the total angular loss is given by (19).

$$L_{\text{angular,dB}} = L_{\text{Azimuth ang,dB}} + L_{\text{Elevation ang,dB}} \quad (19)$$

While considering a fixed x_length of 50 m for a geometry shown in Figure 5, the Figure 7 shows the angular loss based on a single three-dimensional incidence angle as a function of y_length and z_length.

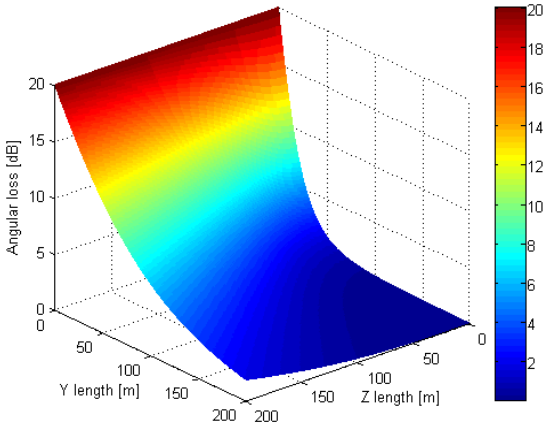


Figure 7. The angular loss based on single 3D incidence angle.

Figure 8 shows the angular loss based on a separate incidence angle in azimuth and elevation plane. The first approach of the angular wall loss model based on a single three-dimensional incidence angle gives higher angular loss at the lower values of y_length and z_length. The angular wall loss model based on a separate incidence angle in azimuth and elevation plane considers the change of angle of an incident ray in both domains separately, and therefore it can be seen in Figure 8 that the angular loss increases with the increase in z_length which is not the case in Figure 7. The impact of change of z_length especially at the lower values of y_length, is minimal on the angular loss based on a single three-dimensional approach, whereas Figure 8 shows that dual-angle model has greater dependency on the change of angle in the elevation plane compared with the single three-dimensional angle approach.

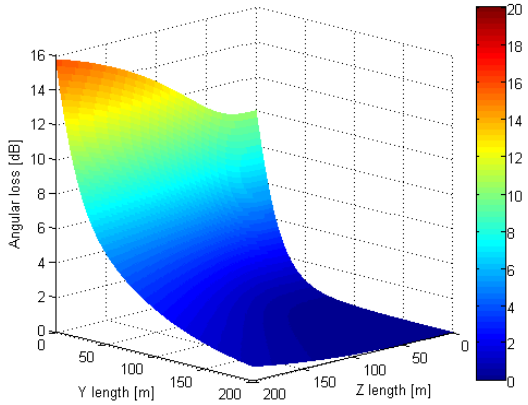


Figure 8. The angular loss based on separate incidence angle in an azimuth (horizontal) and elevation (vertical) plane.

3.2. Extended Building Penetration (EBP) Model

The extended building penetration model is the extension of the general building penetration model. In this model, the additional fifth path known as the "Direct path" is proposed to include as shown in Figure 9. In 3D environment, the direct path is the shortest path between the transmitter and the receiver and it can penetrate through the building wall and through the ceilings of the floors. Figure 10 shows two receiver points RX1 and RX2 located inside the building at Floor 1 and Floor 4, respectively.

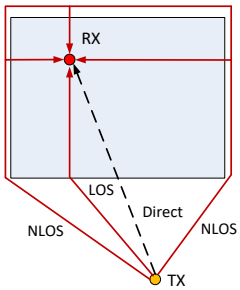


Figure 9. Illustration of direct path along with other paths (top view).

In Figure 10, the LOS paths are shown with the red arrows and the direct paths are shown with the black arrows. It can be seen that the direct ray follows the shortest path between the transmitter and the receiver, and intercepts the building wall at different height compared with that of a receiver height. In Figure 10, it can also be seen that the direct path to RX1 penetrates through the single ceiling, whereas the direct path to RX2 penetrates through the ceilings of two floors. Therefore, the ceiling penetration loss needs to be taken in to account for the direct path.

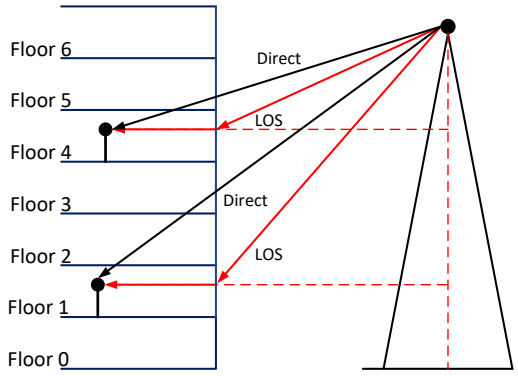


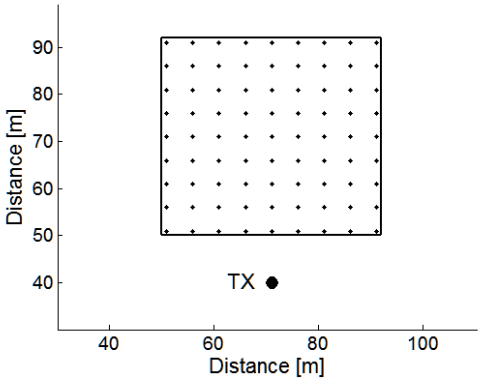
Figure 10. Ray path entering the building while penetrating through the ceiling (side view).

Generally, the ceilings are made up of concrete. Therefore, the ceiling penetration loss model is assumed to follow the concrete penetration loss model as given by (20). The given ceiling penetration loss model provides the ceiling penetration loss of 45 dB, 85 dB, and 125 dB at 10 GHz, 30 GHz and 60 GHz, respectively.

$$L_{Ceiling,dB} = 4 * Frequency_{GHz} + 5, \quad (20)$$

4. Simulation Environment and Simulation Results

For the first and second part of research work, a single twenty-one stories building with an average floor height of 3 m is considered for simulation as shown in Figure 11. An outdoor base station is located at a distance of 10 m away from the building as shown in Figure 11.

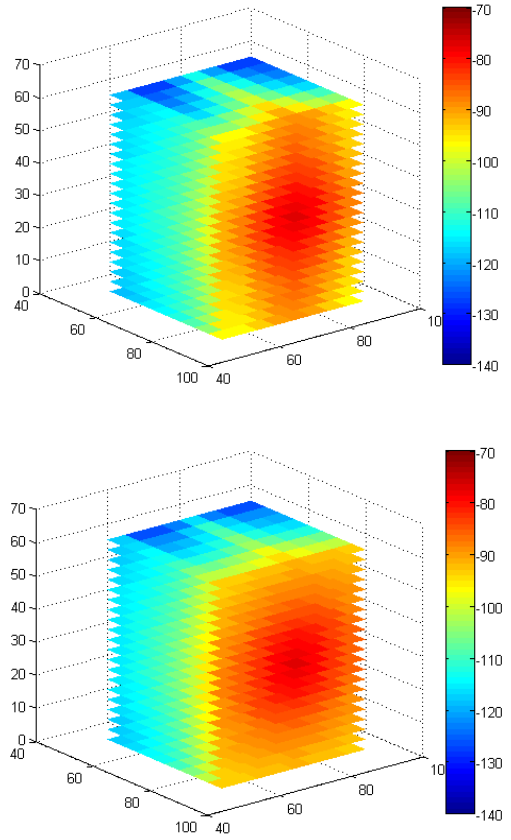


(a)

Figure 11. Two-dimensional map of a single building scenario with transmitter 10 m away from the building.

The transmission power and the height of the base station antenna are 33 dBm and 31.5 m, respectively. The simulations are performed at the frequency of 10 GHz. The indoor location points are placed on each floor with the separation of 5 m among them. An old building type was assumed with indoor wall loss model 1.

In the first part of this research work, the impact of angular wall loss models is analyzed. In case of an angular wall loss model based on a single three-dimensional incidence angle, the incidence angle θ_i was set to 60° for the NLOS paths. Similarly, in case of an angular wall loss model based on a separate incidence angle in azimuth and elevation plane, the incidence angle in an azimuth plane 'α' and the incidence angle in an elevation plane 'β' were also set to 60° for NLOS paths. Therefore, both models give similar results for NLOS paths.



(b)

Figure 12. Path gain, a) Angular wall loss model based on single three-dimensional incidence angle, and (b) Angular wall loss model based on separate incidence angle in azimuth and elevation.

Figure 12 shows the heat map of a path gain achieved with two angular wall loss models. The difference between the two angular wall loss models is evident and visible. In Figure 12(b) as we move away from the centre point of the building wall, the path gain starts to fade (deteriorate) in a perfect circular pattern. However, in Figure 12(a) the change of the path gain with the change in an azimuth and elevation angle is not in a perfectly circular way.

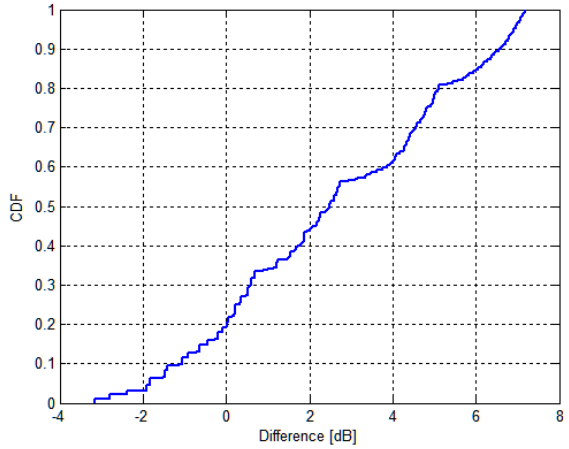


Figure 13. Difference of angular wall loss models.

The difference between the two angular wall loss models is computed by subtracting the received signal power at each point using angular loss model based on separate incidence angle in azimuth and elevation plane from the received signal power calculated using angular wall loss model based on a single three-dimensional incidence angle. The difference was computed assuming only LOS path. The NLOS paths were neglected for calculating the difference, as both models show similar results for NLOS paths. Figure 13 shows the CDF plot of difference between the two angular wall loss models. It can be seen in Fig. 10 that the difference between the two angular wall loss models has values up to 7 dB. The CDF curve with the large number of positive values shows that the angular wall loss model based on a single three-dimensional angle provides aggressive angular loss compared with the dual-angle based angular loss model.

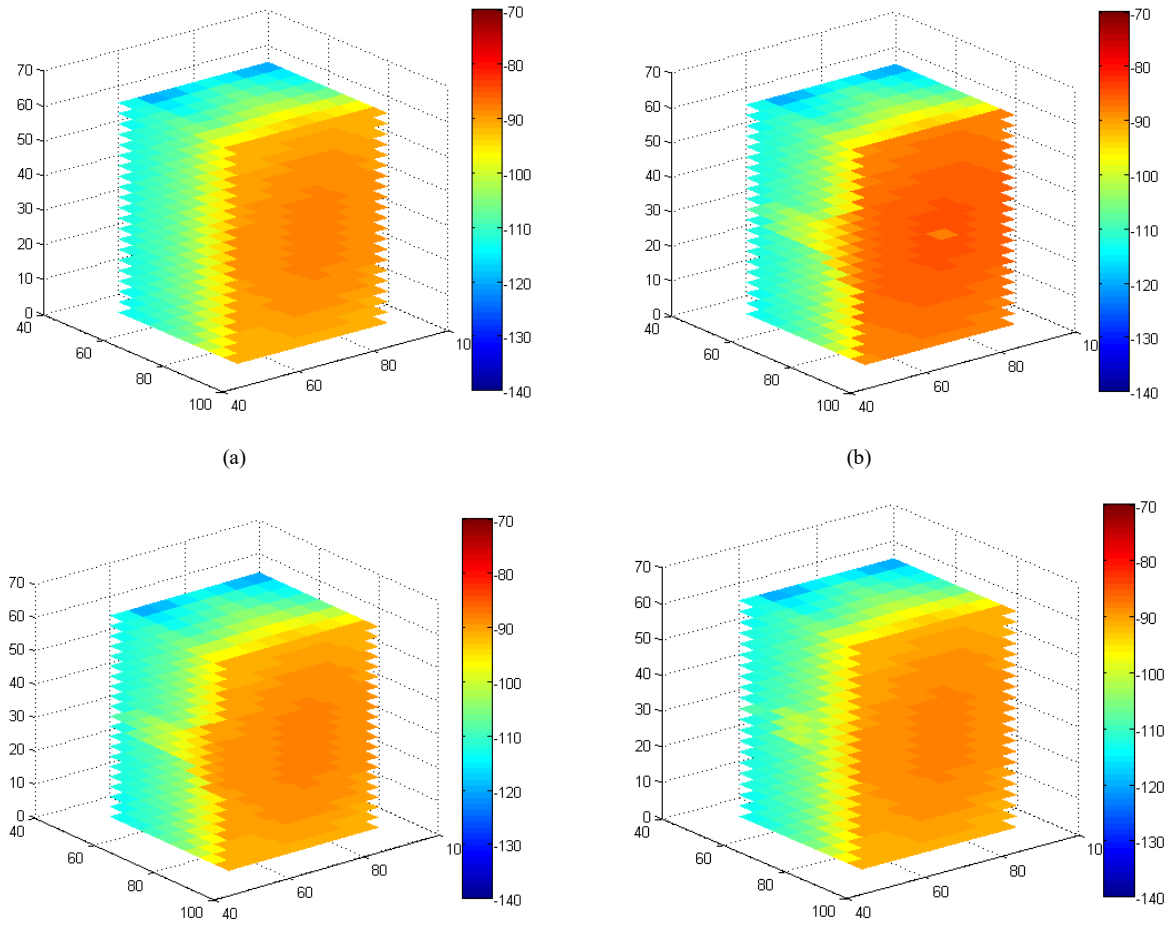


Figure 14. Path gain at 10 GHz, a) GBP model, (b) EBP model without any filter, (c) EBP model with 3 dB filter, and (d) EBP model with 7 dB filter.

In the second part of this research work, the impact of the direct path in the extended building penetration model is analyzed, and the performance of the general building penetration model is compared with the extended building penetration model. The angular wall loss model based on a separate incidence angle in horizontal and vertical plane is assumed, and the separation between the building wall and transmitter is set to 35 m.

Figure 14 shows a heat map of path gain for different cases. Figure 14(a) shows a path gain for GBP model, and while going from front to the back of the building an almost similar path gain is achieved on all the floors of the building. In Figure 14(b), a direct path is included. Now, due to an additional path a higher path gain is obtained. It is interesting to see that an additional direct path improves the received signal level of the users located on the floors which are close to the height of the outdoor transmitter

It can be seen in Figure 14(b) that due to higher ceiling penetration loss at higher frequency, the impact of a direct path diminishes as the difference between the outdoor antenna height and the user height increases. However, the impact of a direct path is clearly evident to the users located deep inside the building at the same floor height as an outdoor antenna, or on one floor above and below. On the other hand, it can also be observed that in Figure 14(b) the path gain is overestimated at the front face of a building, as the “LOS path” and the “Direct path” are almost identical, and they are both included in the modeling. Therefore, a filtering threshold of 3 dB is used in Figure 14(c) to filter out the direct path if the difference between the LOS path and the direct path is less than or equal to 3 dB. An impact of simple 3 dB filter can be seen in Figure 14(c) as the front face of a building is showing now almost similar results as in Figure 14(a). Similarly, a filtering threshold of 7 dB was used in Figure 14(d). The direct path has quite significant impact on the floors close to the outdoor antenna

height even with 3 dB filtering threshold. Therefore, it will improve the model of outdoor to indoor propagation by including a direct path along with the certain filtering threshold, as we have used in this study.

For ray tracing simulations in order to acquire reflected and diffracted multipaths a multiple building scenario is considered as shown in Figure 15.

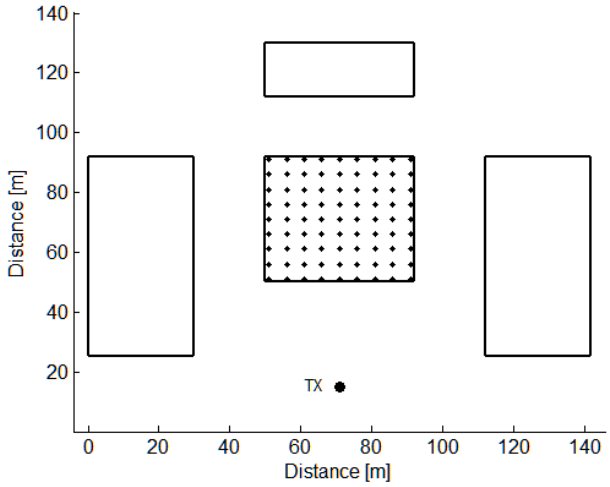


Figure 15. Two-dimensional map of a multiple building scenario with transmitter 35 m away from the building.

In this scenario, a single twenty stories building surrounded by four other buildings of the same height are considered. However, the focus area is the central building, and therefore indoor location test points are distributed in the central building only. An outdoor base station is located at a distance of 35 m away from the central building as in the case of a single building case.

In the third part of this research work, two different ray tracing approaches are considered for the indoor users:

Ray tracing type 1 (RT type 1): In this case, an Image Theory (IT) based ray tracing technique is used to find the propagation paths between the outdoor transmitter and the indoor receiver point. A smooth building surface is assumed, which acts as a perfect reflecting surface. For the transmitter and the receiver points at different heights, the incidence ray path does not intercept the building wall at receiver height. Therefore, for the paths which pass through the ceiling of a floor as shown in Figure 10, the ceiling penetration loss model is used to include the ceiling penetration loss.

Ray tracing type 2 (RT type 2): It is the modified approach of ray tracing type 1. In this case, the propagation paths between the transmitter and receiver in an azimuth plane are found by using the image theory. However, in an elevation plane for the transmitter and receiver at different heights the incidence ray path "always" intercepts the building wall at receiver height. It means that the ray path would never penetrate through the ceiling of a floor. Therefore, the ceiling penetration loss model is not used in this case.

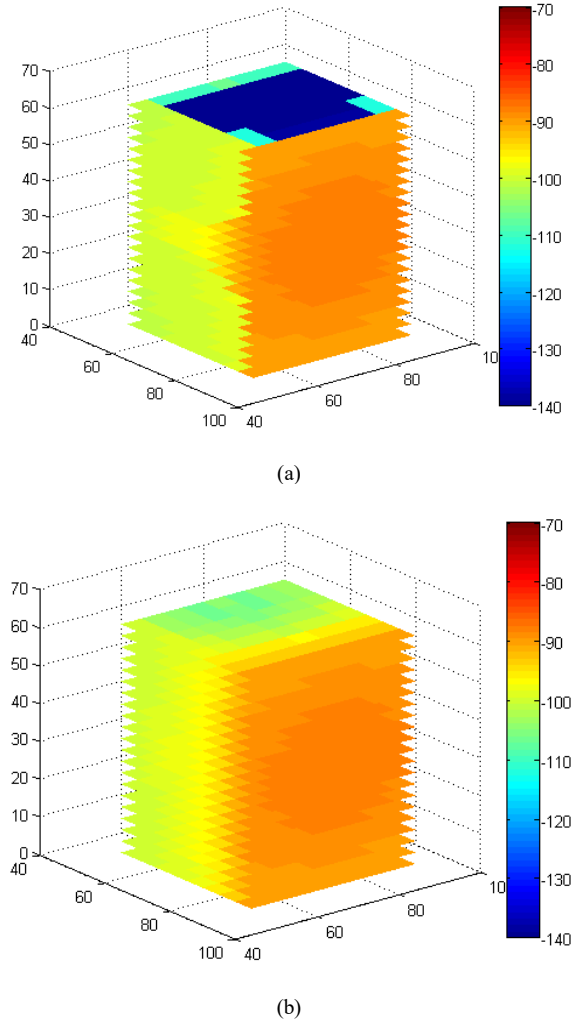


Figure 16. Path gain at 10 GHz, a) Ray tracing type 1, (b) Ray tracing type 2.

Figure 16 shows a path gain map for different cases. It is interesting to see the results obtained by ray tracing type 1. In Figure 16(a), the acquired results from ray tracing type 1 for the indoor users located 6 m or more deep inside the building are quite pessimistic in comparison with the results obtained with ray tracing type 1. As explained earlier, that in case of ray tracing type 1 approach, the ray path entering the building intercepts the building wall at different height compared with that of a receiver height. For the users located deep inside the building the most of the ray paths reach at the receiver point after penetrating through a single or through the multiple ceilings of the floor. Due to high ceiling penetration loss, the signal power attenuates significantly after passing through the ceilings.

Therefore, the ray tracing type 1 is found not suitable for the indoor users located far away from the exterior wall of the building. However, the results obtained with ray tracing type 2 seem more optimistic and realistic in comparison with the ray tracing type 1. It is critical here to mention that in ray tracing type 2, the incident ray always intercepts the building wall at the same height as that of a receiver height, and therefore the ceiling penetration loss is ignored in ray tracing type 2.

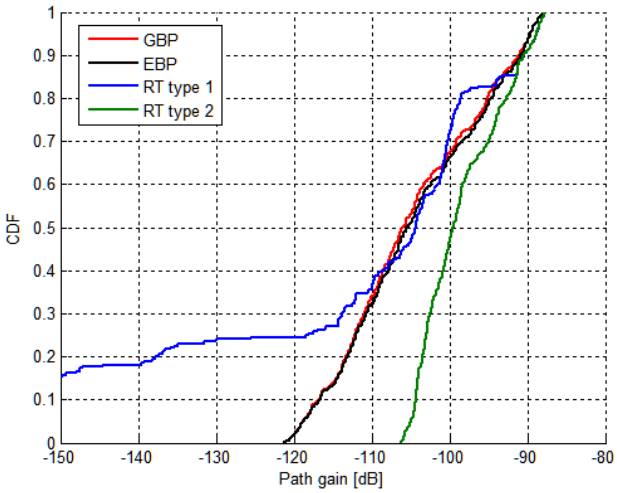


Figure 17. CDF plot of path gain for different propagation models at 10 GHz.

Figure 17 shows the CDF plot of path gain for different considered cases. For an extended building penetration model, a 3 dB filtering threshold was used. For the top nearly 14.5 % of the samples, the ray tracing type 1 shows an almost identical result as ray tracing type 2. Those 14.5 % of the samples mainly represent the receiver locations from the front row and few from the second row i.e. receiver points at the front face of the building. Again, for the same top 14.5 % of the samples the general building penetration model provides an almost similar result like that of an extended building penetration model and other ray tracing approaches. Then, the path gain starts to drop significantly for the receiver points located deep inside the building. A large portion of samples have the extremely pessimistic values of the path gain with a ray tracing type 1. For the considered cases, the ray tracing type 2 was found the most optimistic approach. For ray tracing simulations, all the possible ray paths with two reflections and two diffractions were considered.

Therefore, there is a large number of considered ray paths in case of ray tracing compared with the GBP and EBP. The results provided in Figure 14 and Figure 16 reveal that ray tracing type 2 is a better and realistic approach compared with ray tracing type 1 for estimating the received signal level, especially for the indoor users located deep inside the building.

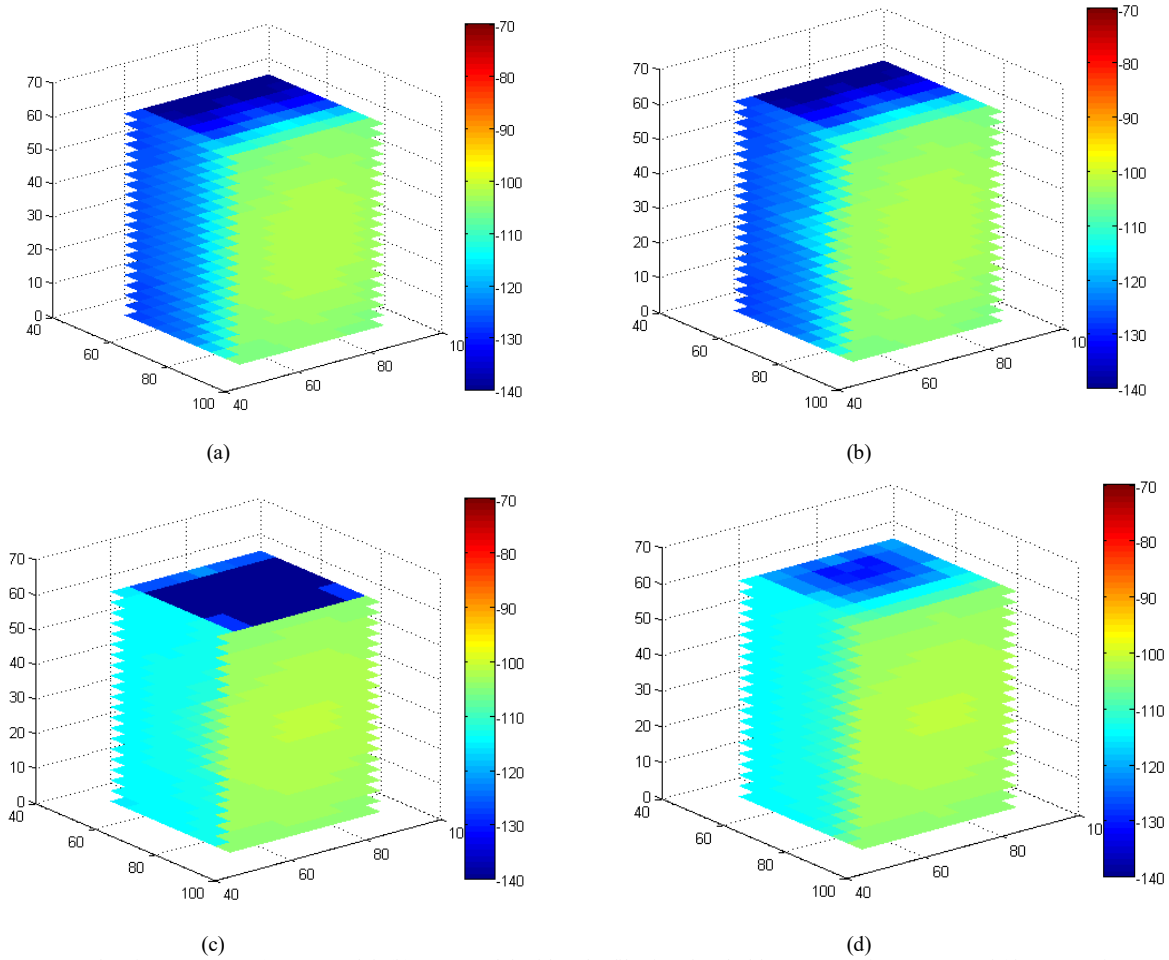


Figure 18. Path gain at 30 GHz, a) GBP model, (b) EBP model with 3 dB filtering threshold, c) Ray tracing type 1, and (d) Ray tracing type 2.

Figure 18 shows a heat map of path gain for different models assuming old building type at 30 GHz and indoor wall loss model 1. Path gain is the function of frequency and a significant impact of frequency of operation can be seen between the results presented in Figure 16 (at 10 GHz) and Figure 18 (at 30 GHz). In Figure 18(b), the path gain for EBP model with 3 dB filtering threshold is shown, and the impact of the direct path is even

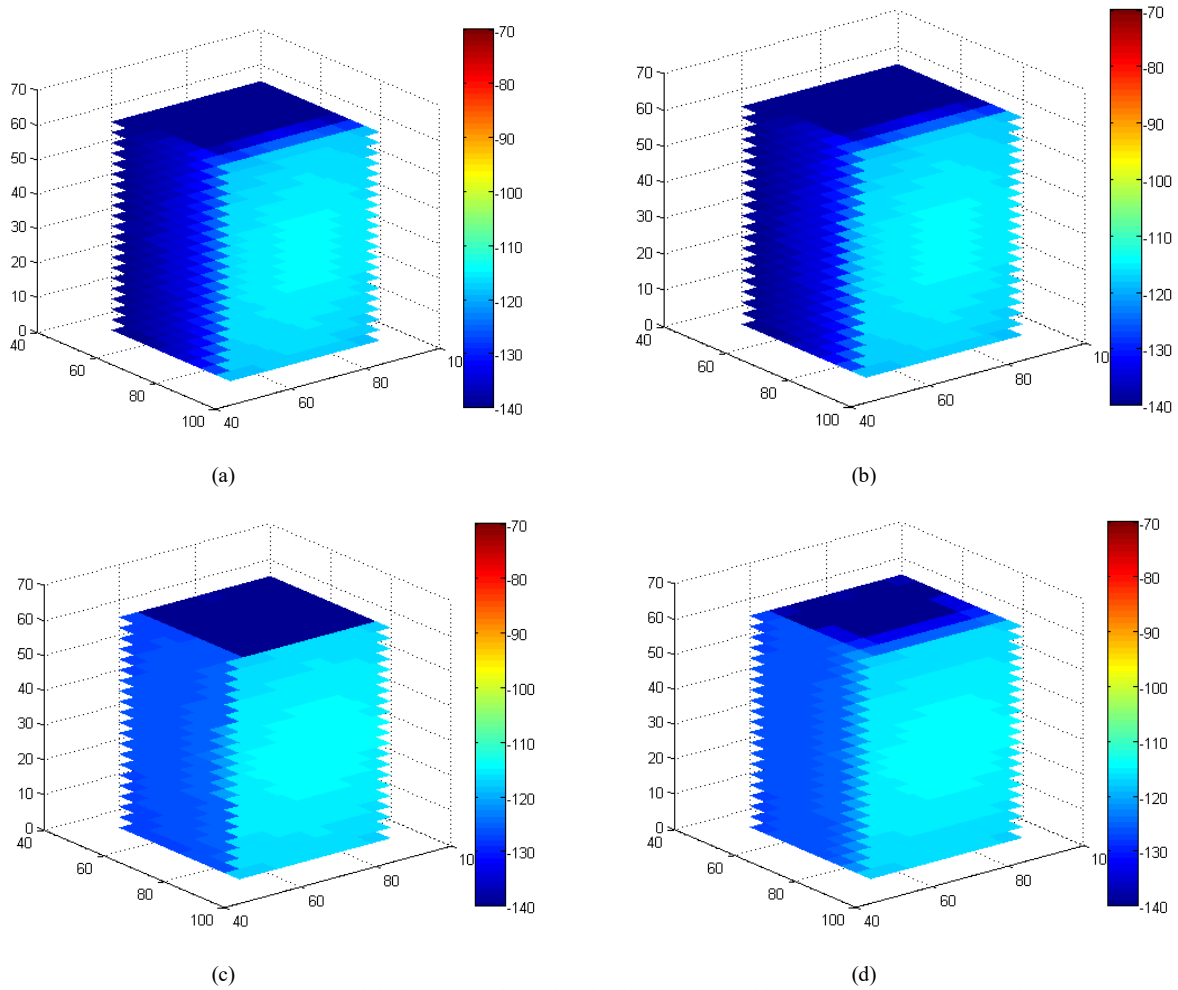


Figure 19. Path gain at 60 GHz, a) GBP model, (b) EBP model with 3 dB filtering threshold, c) Ray tracing type 1, and (d) Ray tracing type 2.

visible at 30 GHz frequency for the center floors. The impact of a direct path is not prominent in overall results. Again, the ray tracing type 1 is found as an inadequate approach for estimating the path gain, and it provides the extremely low values of path gain at 30 GHz. However, it is interesting to see the results acquired with ray tracing type 2. It is fascinating to compare the results of two ray tracing approaches, as ray tracing type 1 clearly shows lack of indoor coverage with the outdoor base station except the front wall of the building. On the other hand, the simulation results of ray tracing type 2 show that the low indoor coverage can be provided with the outdoor base station. Similarly, GBP and EBP models show lack of indoor coverage at 30 GHz. These results leave an open discussion that the results acquired from the ray tracing type 2 are more realistic or the results acquired with ray tracing type 1 are more realistic. As we don't have any measured reference data therefore it is hard to make a clear statement.

Figure 19 shows a heat map of path gain for old building type at 60 GHz assuming indoor wall loss model 1. At 60 GHz, all of the considered propagation models show that due to high propagation and penetration losses the indoor coverage cannot be provided with the outdoor base station. Higher antenna gain at higher frequencies can compensate the high penetration and propagation losses. Otherwise, an indoor coverage with the outdoor base station at 60 GHz can be provided if new buildings are constructed with the material having better signal penetration properties.

5. Conclusion

In this paper, the outdoor to indoor propagation at higher frequency is studied. Furthermore, a novel angular wall loss model based on a separate incidence angle in azimuth and elevation plane is proposed. In the proposed angular wall loss model, the total angular loss is the sum of loss due to the change of angle of an incident ray in both azimuth and elevation plane. The simulation results show that the angular wall loss model based on a single three-dimensional angle provides aggressive angular loss compared with the proposed dual-angle angular wall loss model. In the second part of this study, an Extended Building Penetration (EBP) model is proposed, and the performance of EBP model is compared with the COST231 building penetration model. In an extended building penetration model a fifth path known as a "Direct path" is added. Through the simulation results, it is shown that the direct path has a significant impact on the indoor receiver points located close to the height of the transmitter. Therefore, the extended building penetration model can be considered as a better

approach for modelling the outdoor to indoor propagation at higher frequencies.

The ray tracing is a promising and precise technique for estimating the received signal level and for channel modeling. In this study, a performance comparison between the general building penetration model and the ray tracing model is done. It was found that the traditional ray tracing technique provides good approximation for the users located close to the wall of the building. However, the traditional ray tracing approach was found in-efficient for modeling the outdoor to indoor propagation especially for the indoor users located far away from the exterior wall of the building. A new approach of ray tracing is also proposed in this study. The simulation results show that ray tracing performance is significantly improved by the proposed recommendation. With the proposed ray tracing approach the results of a path gain were fairly better compared with the other models. Later, the performance of different propagation models was compared at 10 GHz, 30 GHz and 60 GHz frequency of operation. It was found that for the considered simulation scenario the adequate coverage can be provided for the indoor users with the outdoor base station at 10 GHz. However at 30 GHz, the general penetration model clearly showed the lack of indoor coverage. Whereas the proposed ray tracing type 2 approach showed a fair indoor coverage at 30 GHz. The propagation and building penetration losses were quite high at 60 GHz; therefore the simulation results show no indoor service at 60 GHz.

Acknowledgment

Authors would like to thank European Communications Engineering (ECE) Ltd, TAKE-5 project, and Oy L M Ericsson Ab for supporting this research work.

References

[1] M. U. Sheikh, K. Hiltunen and J. Lempiäinen, "Angular wall loss model and Extended Building Penetration model for outdoor to indoor propagation," 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, 2017, pp. 1291-1296.

[2] E. Damosso and L. M. Correia, Eds., Digital mobile radio towards future generation systems, final report ed. European Commission, 1999.

[3] J. E. Berg, "A recursive method for street microcell path loss calculations," Proceedings of 6th International Symposium on Personal, Indoor and Mobile Radio Communications, Toronto, Canada, September 1995, pp. 140−143 vol. 1.

[4] E. Semaan, F. Harrysson, A. Furuskär and H. Asplund, "Outdoor-to-indoor coverage in high frequency bands", 2014 IEEE Globecom Workshops (GC Wkshps), Austin, TX, 2014, pp. 393−398.

[5] C. Larsson, F. Harrysson, B.-E. Olsson, and J. E. Berg, "An outdoor-to-indoor propagation scenario at 28 GHz," in 8th European Conference on Antennas and Propagation (EuCAP 2014), The Hague,The Netherlands, April 2014, pp. 3301−3304.

[6] W. C. Stone, "Electromagnetic signal attenuation in construction materials, "NIST Building and Fire Research Laboratory, Gaithersburg, Maryland, NISTIR 6055 Report No. 3 6055, Oct. 1997.

[7] L. M. Frazier, "Radar surveillance through solid materials," in Proceedings of the SPIE - The International Society for Optical Engineering, vol. 2938, Hughes Missile Syst. Co., Rancho Cucamonga, CA, USA, 1997, pp. 139−146.

[8] R. Wilson, "Propagation losses through common building materials 2.4 GHz vs 5 GHz," University of Southern California, CA, Tech. Rep. E10589, Aug. 2002.

[9] C. A. Remley, G. H. Koepke, C. L. Holloway, C. A. Grosvenor, D. G. Camell, J. M. Ladbury, R. Johnk, D. R. Novotny, W. F. Young, G. Hough, M. McKinley, Y. Becquet, and J. Korsnes, "Measurements to support modulated-signal radio transmissions for the public-safety sector," NIST, Boulder, CO, Tech. Rep. Tech. Note 1546, Apr. 2008.

[10] White paper on "5G Channel Model for bands upto 100 GHz".

[11] H.W. Son, and N.H. Myung, "A deterministic ray tube method for microcellular wave propagation prediction model," Antennas and Propagation, IEEE Transactions on , vol.47, no.8, pp.1344-1350, Aug 1999.

[12] D.N. Schettino, F.J.S. Moreira, C.G. Rego, "Efficient Ray Tracing for Radio Channel Characterization of Urban Scenarios," Magnetics, IEEE Transactions on , vol.43, no.4, pp.1305-1308, April 2007

[13] S. Soni, and A. Bhattacharya, "An efficient two-dimensional ray-tracing algorithm for modeling or urban microcellular environment", International Journal of Electronics and Communications (AEU), volume 66, issue 6, pp. 439-447, June 2012.

[14] J. E. Berg, "A recursive method for street microcell path loss calculation", 6th PIMRC' 95, Toronto, Canada, September 1995, pp. 140-143.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 63-69 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

ARMA Feeding Techniques for Isoflux Coverage from a Micro Satellite

Ali Siblini [1,2,*], Hussein Abou Taam[2], Bernard Jecko[1], Mohammad Rammal[2], Anthony Bellion[3], Eric Arnaud[1]

[1]*Systems-RF University of Limoges, 12, rue Gemini 87068 Limoges Cedex 03, France*

[2]*Lebanese University, Doctoral School of Science and Technology, Beirut*

[3]*CNES, 18 Avenue Edouard Belin, 31401 Toulouse Cedex 9, France*

ARTICLE INFO

Article history:
Received : 26 September 2017
Accepted : 18 October 2017
Online : 04 November 2017

Keywords :
Agile Radiating Matrix antenna
Multimode radiating beams
Spatial Telemetry

ABSTRACT

This paper deals with the design of a new reconfigurable beam antenna used to improve the efficiency of spatial telemetry links on Nano-Satellite. This agile beam antenna is not built on the well-known array concept AESA (Agile Electronically Scanned Array) but using a new one called ARMA (Agile Matrix Radiating Antenna); MARPEM in French. In this paper there is the design of the circularly polarized matrix antenna, the generation of two beam forming modes, the design of the distribution circuit and the polarization circuits.

1. Introduction

The needs in high data rate of future Nano-satellite missions are incompatible with telemetry applications in VHF or S bands unlike the X-band. Moreover, most of applications require earth wide angle coverage in addition to a long time visibility, implying a theoretical isoflux radiation pattern particularly for LEO satellites. But this theoretical limited gain radiation pattern in nadir direction reduces the global efficiency of the PDTM (Payload Data Telemetry). A solution to maximize the PDTM is to use an agile beam antenna to increase the data rate efficiency by introducing several kinds of radiation patterns depending on the satellite position.

These links are produced by beams of various shapes delivered by an agile antenna working on several radiation modes. To limit the complexity of the architecture and to be accommodated on a Nano-satellite, these requirements lead to a dual mode circularly polarized antenna: one mode characterized by isoflux earth coverage and another one introducing high gain directive beam around the axial direction (nadir). To realize this PDTM link in X-band, a very compact antenna using dual mode radiation pattern has been developed with the following requirements:

- [8 – 8.4]GHz.
- To be accommodated on a 100mm x 100mm surface.
- Gain > 2dB for θ [-60°: 60°] and φ [-180°:180°].
- Axial Ratio < 4dB for θ [-60°:60°] and φ [-180°:180°].

The beam forming procedure and the switching process between two or more working modes need to use agile beam antennas. There are two kinds of electronically agile beam antennas; the first one is the well-known AESA (Agile Electronically Scanned Array) technique and the second, a new one, called ARMA (Agile Radiating Matrix Antenna) which introduces some advantages for such applications [1], [2] as larger angles beam scanning, higher power balance, side lobes minimization [3]-[6], lower couplings between adjacent elements and high gain solutions [7], [2] with smaller number of elements and less feeding network complexity.

In the context of this paper, the new ARMA concept is defined and the circularly polarized pixel with the polarization circuits design is presented. Then the ARMA matrix with circular polarization is used to generate two beams, the wide Isoflux and the high gain axial beams. Beam forming law is optimized to solve

*Corresponding Author: Ali Siblini, Lebanese University Beirut Lebanon.
Email: ali.siblini.84@hotmailcom

the problem of Isoflux dissymmetry in the different azimuthal planes,

2. The ARMA Concept

ARMA principle introduces new sampling procedure for the radiating surface of the antenna that is sampled or pixelated using the pixel. Brief comparison of AESA and ARMA principles is found in [3]. ARMA approach uses jointed pixels as shown in Figure 1 to sample the radiating surfaces instead a lot of punctual sources for the array technique [4]. Each pixel generates a uniform radiating surface that never overlaps outside its dimensions [3].

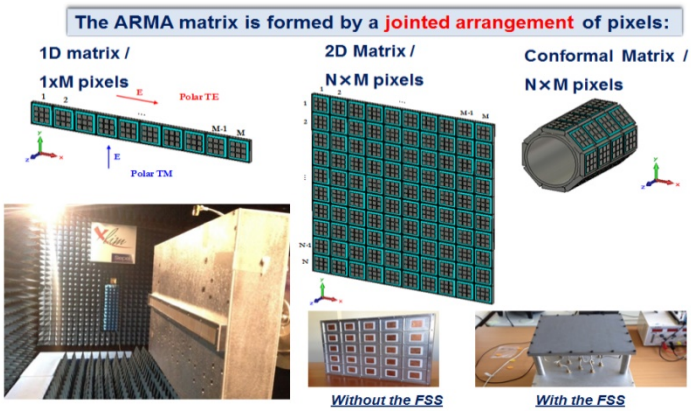


Figure 1. ARMA antennas architectures.

The most important point in ARMA approach is to be able to design a pixel which can generate a uniform *EM* field on its surface. This technique, deduced from the knowledge of EBG antennas properties, was extensively developed and validated in many papers [8]-[11] for linear polarization; it is the key point of the ARMA solution. Example of a pixel with one feeding probe is shown in Figure 2 with the corresponding radiating surface on the top.

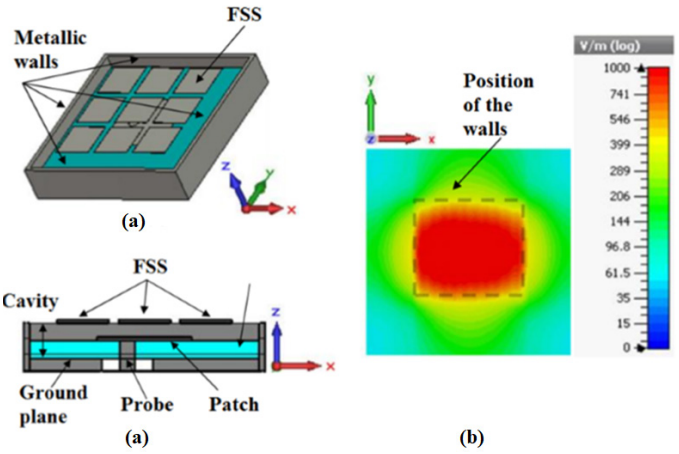


Figure 2. Pixel design: (a) FSS on top, patch inside the cavity and metallic walls, (b) the corresponding radiating surface.

For low profile antennas, a planar radiating surface *S* is built with pixels with a square radiating surfaces as shown in Figure 3.

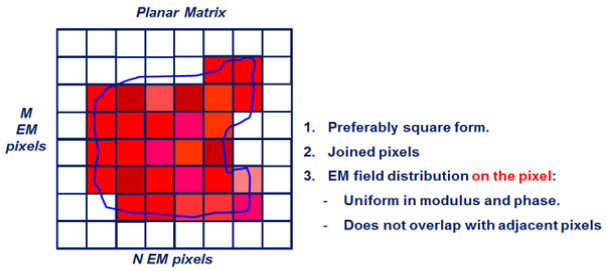


Figure 3. Radiating surface sampled by jointed square pixels.

3. Circularly Polarized Pixel Design

3.1. Pixel Design and Axial Ratio

For circular polarization (two orthogonal linear polarizations), the EBG cavity is usually fed by a patch probe generating a local circularly polarized field in the cavity. A good circular polarization can be obtained by two or four ports feeding circuit, shown in Figure 4. Of course, for a pixel considered alone, the axial ratio is better [12] [13] with the four ports technique (Figure 5 and Figure 6), but with higher manufacturing complexity and price.

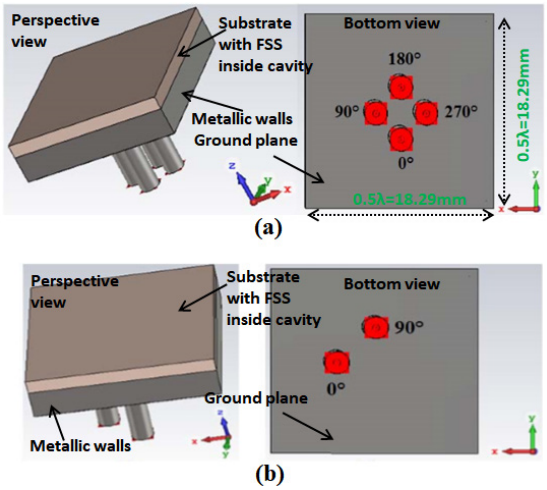


Figure 4. Pixel with two and four feeding ports for circular polarization generation.

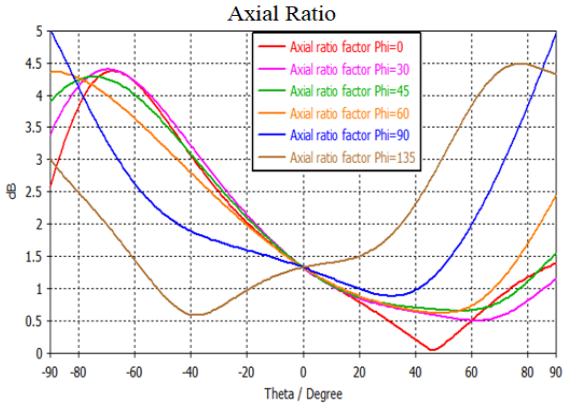


Figure 5. Axial ratio comparison obtained with a 2 ports ideal feeding procedure.

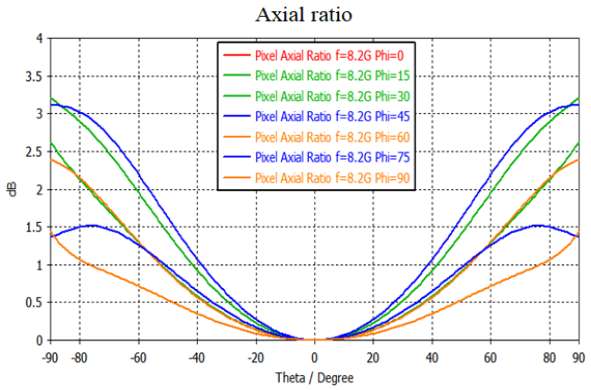


Figure 6. Axial ratio comparison obtained with a 4 ports ideal feeding procedure.

3.2. Polarization Circuits and Area Limitations

The main limitation in the design of the polarization circuit is the small area behind the pixel which is 18.29mmx18.29mm (0.5λx0.5) (Figure 4). This small area must include also the pad to solder the connector for the input power. For that the substrate TMM10i with permittivity 9.8 and thickness 0.381mm was used to print the circuit on it. The results of the axial ratio are not very modified by introducing the circuit to the pixel design: one hybrid coupler for the two ports case and 3 hybrid couplers for the 4 ports case as shown in Figure 7 and Figure 8.

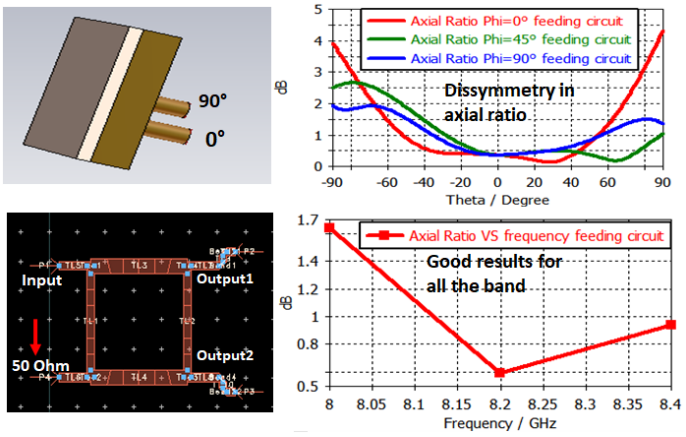


Figure 7. Axial ratios obtained with a one hybrid coupler for the two ports.

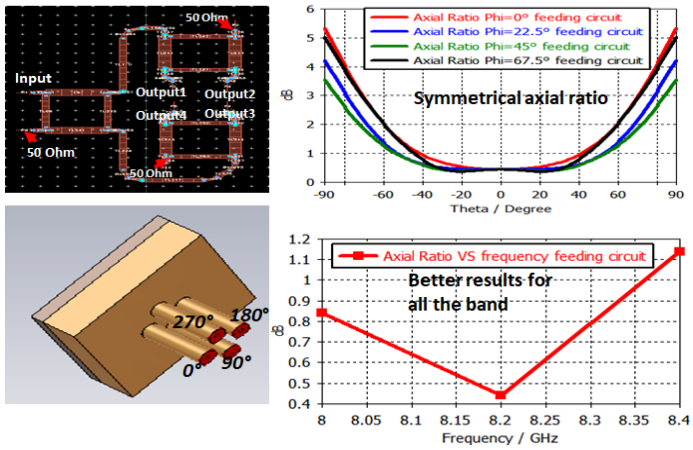


Figure 8. Axial ratios obtained with a d a 3 hybrid couplers for the 4 ports.

4. PDTM Application

4.1. Antenna Design

On this way, a square low profile ARMA antenna was built with 25 pixels where only nine are fed (Figure 9) by the BFN (Beam Forming Network); the other ones, located at the edges of the antenna, are only used to keep a good axial ratio for the whole matrix antenna.

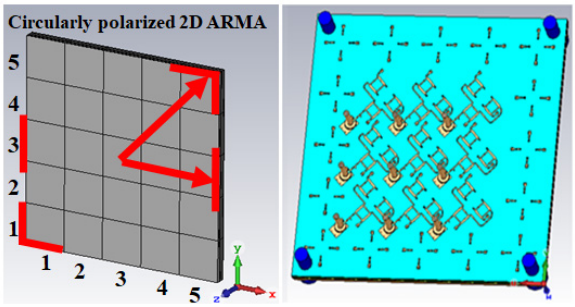


Figure 9. 25 pixels matrix antenna (left) with nine feeding ports and nine circular polarization circuits (Right).

4.2. Isoflux Pattern and Dissymmetry Problem

The isoflux coverage (first mode) is obtained by feeding the pixels with a weighting law deduced from a Sinc (sinus cardinal) function [14] [15]. This approach leads to weighting tables and to the corresponding matrix radiation patterns and axial ratios shown in Figure 10 and Figure 11.

Table of amplitudes $A_{i,j}$

	1	2	3	4	5
1	0	0	0	0	0
2	0	0.03	0.18	0.03	0
3	0	0.18	1	0.18	0
4	0	0.03	0.18	0.03	0
5	0	0	0	0.002	0

Table of $\varphi_{i,j}$

	1	2	3	4	5
1	0	0	0	0	0
2	0	180°	180°	180°	0
3	0	180°	0°	180°	0
4	0	180°	180°	180°	0
5	0	0	0	0	0

Figure 10. Sinc circular feeding law.

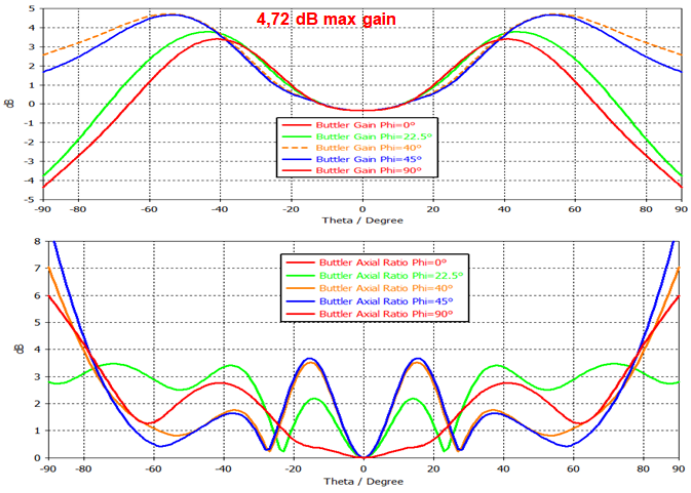


Figure 11. Gain an axial ratio patterns.

The pattern in Figure 11 shows a dissymmetry in the isoflux maximum gain positions (e.g. planes φ=0° and φ=45°); this is due to the non-symmetrical physical structure of the matrix antenna, where the distance from the center to the edges is not the same for all directions (Figure 9).

4.3. Feeding Law Optimization for Symmetrical isoflux

To compensate the dissymmetry resulted from the square matrix structure, the feeding law is optimized to modify the isoflux radiation pattern [15]. The value "A" is the weight applied on each pixel surrounding the central one, "A" is changed as shown in Figure 12, Figure 13 and Figure 14. This procedure, called Sinc-circular symmetrical, will be very useful to match easily the two modes together. It is also important to notice that the part of the following curves located between θ=±20° is not useful in the isoflux mode because for theses directions the matrix is working on the axial mode.

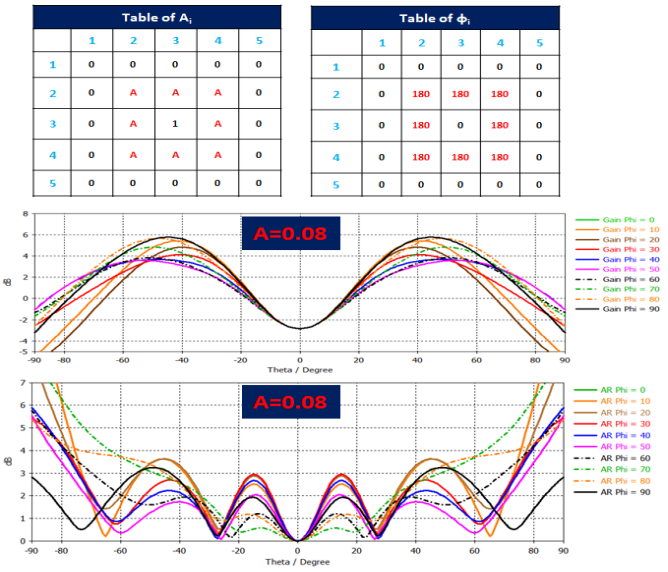


Figure 12. Gain and axial ratio evolution as a function of the elevation angle "θ" for different azimuth planes with A = 0.08.

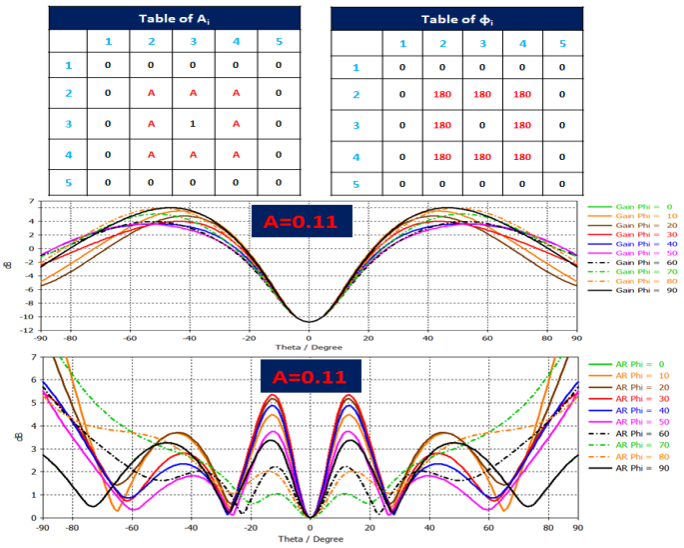


Figure 13. Gain and axial ratio evolution as a function of the elevation angle "θ" for different azimuth planes with A = 0.11.

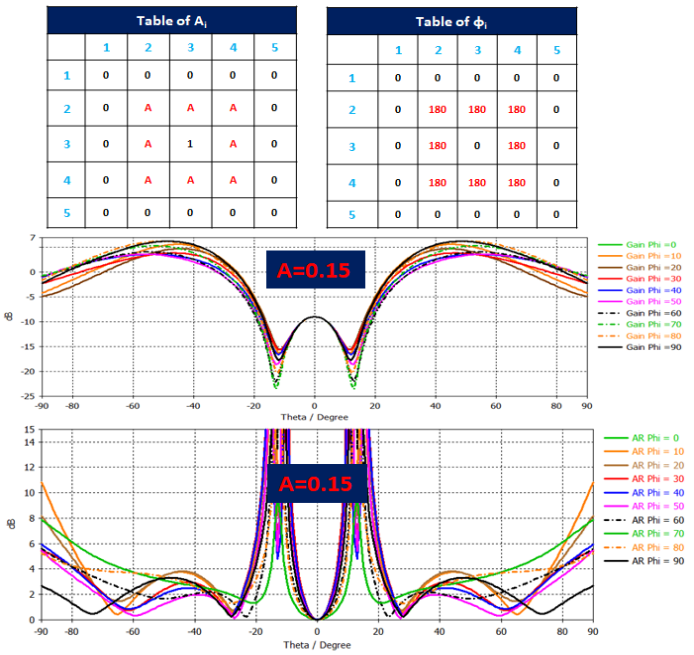


Figure 14. Gain and axial ratio evolution as a function of the elevation angle "θ" for different azimuth planes with A = 0.15.

Then with this symmetrical feeding solution the Isoflux gain pattern is more symmetrical with the position of the maximum gain located in the worst case at θ=±50° for the solution with A=0.15 (Figure 14). The axial ratio in this case is going very high for θ around ±20°, but these directions are not useful in this mode as mentioned above.

5. Directive Beam

The directive beam around the axial direction (second mode) is easy to obtain with the highest gain with all the weights equal to 1 (Figure 15). But, like for an array, it is also possible to optimize the radiation pattern and the axial ratio by modifying theses weights.

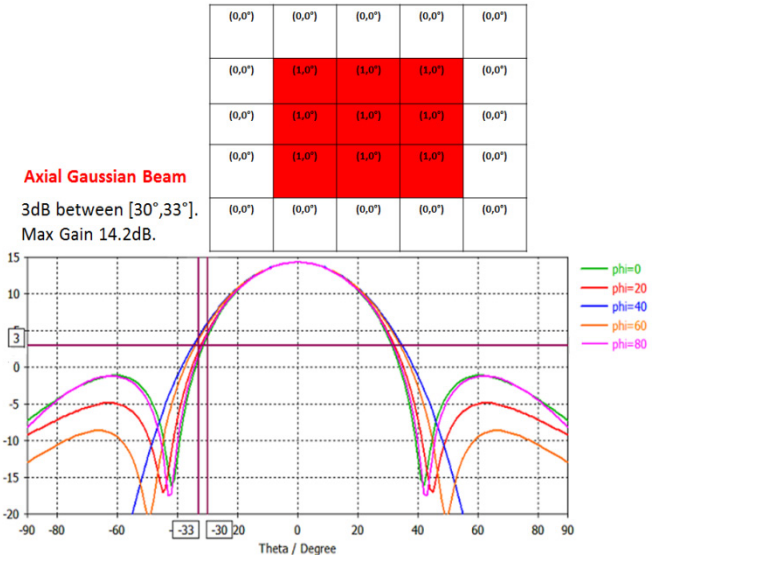


Figure 15. Radiation pattern of the axial mode as a function of the elevation angle θ for different azimuth planes φ.

6. Junction between Two Modes

The BFN (Beam Forming Network) architecture used to switch between the two modes is presented in Figure 16 for a 1D ARMA antenna.

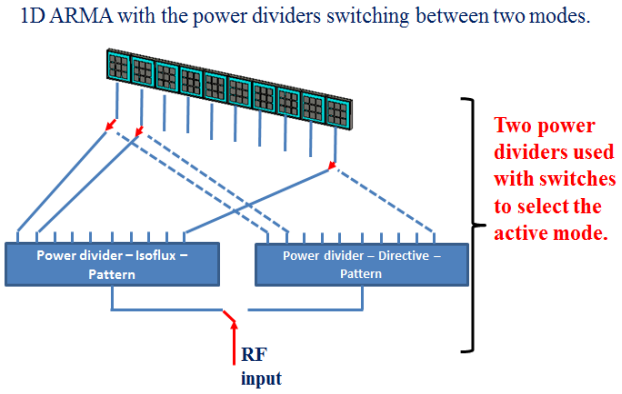


Figure 16. Switching procedure between the two modes.

The switching procedure [16] between the 2 modes can be easily performed for the elevation angle and the gain values expected by the users. For example it can be chosen to keep the gain above 3dB between -60° and +60° as shown in Figure 17.

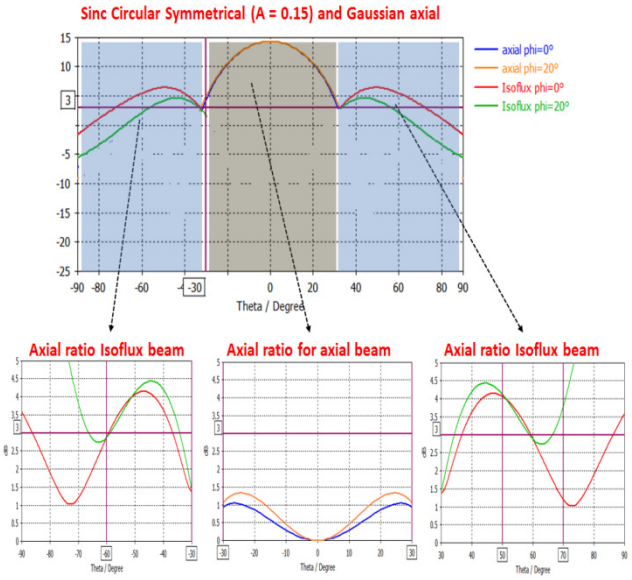


Figure 17. Two modes radiation patterns and axial ratios for the whole agile antenna for φ = 0° and φ = 20° planes.

7. Power Divider Design for Isoflux

The symmetrical feeding law with A=0.15 is considered for the Isoflux beam forming. To achieve this law a power divider was designed with one input and nine outputs as shown in Figure 18. The port labeled 5 is to be connected to the central pixel in the matrix antenna; the other output ports (1, 2, 3, 4, 6, 7, 8 and 9) will be connected to the other pixels around the central one. The return loss at the input of the power divider is shown in Figure 19. Using of the hybrid couplers at the outputs insure good matching (50Ω) when connecting the power divider to the polarization circuits that feed the patches inside the pixels cavities.

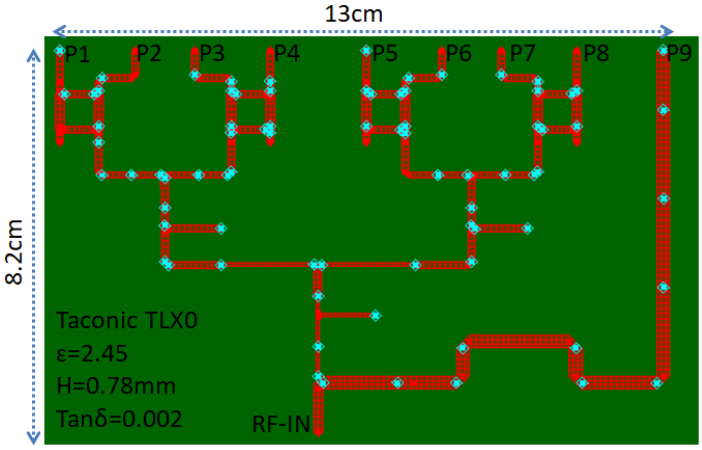


Figure 18. Power divider Momentum design for Isoflux beam forming.

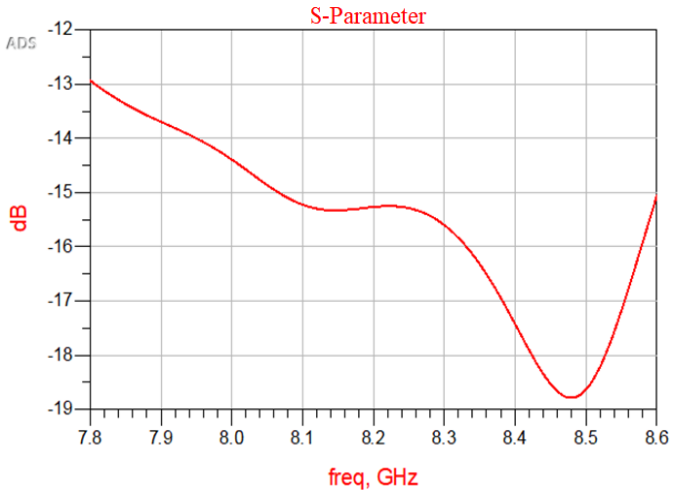


Figure 19. Return loss at the input of the power divider.

The output weights in magnitude and phase at 8.2GHz is presented in the table of Figure 20. These magnitudes are taken normalized with respect to the central port5 and the phases are the phase shift with the phase of port 5.

Port	P1	P2	P3	P4	P5	P6	P7	P8	P9
Magnitude	0.154	0.152	0.157	0.158	1	0.158	0.157	0.152	0.155
Phase	181°	182°	184°	181°	0°	181°	184°	182°	181°

Figure 20. Outputs of the power divider at 8.2GHz.

8. Pattern of the Whole Matrix

Polarization circuits and the power divider results files are taken from Momentum software and connected to the matrix antenna in CST (Figure 21). The realized gain and the axial ratio patterns are shown in Figure 22 at 8.2GHz. The position of the isoflux maximum is attained at approximately at fixed elevation angles and the difference in the gain between the different azimuth planes is 2.9dB.

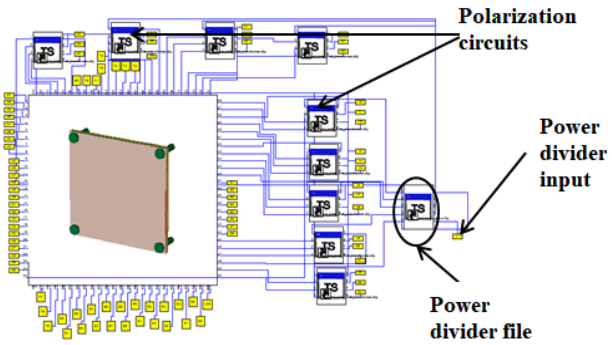


Figure 21. Matrix antenna with the polarization circuits and power divider.

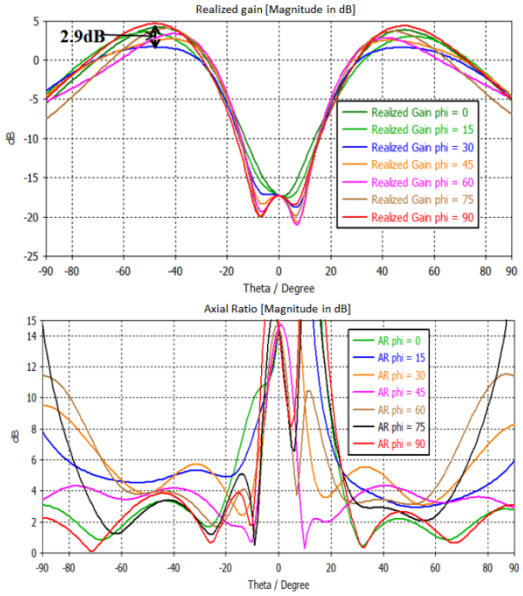


Figure 22. Realized gain and axial ratio patterns at 8.2GHz (whole antenna).

9. Testing the Antenna with the Nano-Satellite Support

The antenna is placed on the top of the Nano-satellite, so to check the effect of the satellite support the antenna was placed on the top of a metallic cube in CST (Figure 23) (cube with 130mm height and 100mm surface). In this case the antenna is tested as the real case for the satellite to be placed on it. Results of the antenna with and without the cube support show no significant modifications in the gain and axial ratio patterns (Figure 24).

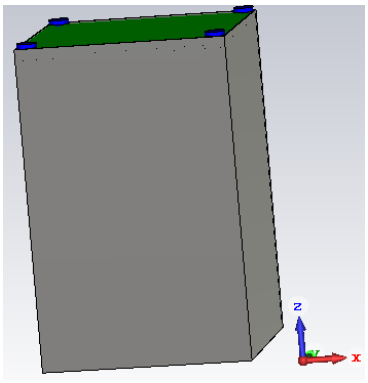


Figure 23. Antenna placed on the cube-sat support.

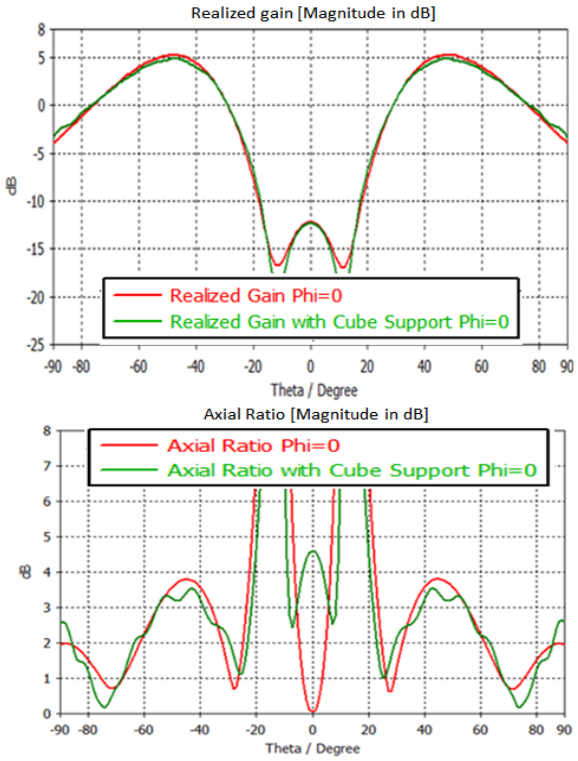


Figure 24. Realized gain and axial ratio patterns with and without cube-sat support.

10. Conclusion

A dual mode agile antenna based on the ARMA principle [1] [2] was presented in this paper. A prototype is under manufacturing and will be tested next month. The pixel with four ports with insures high stability of the axial ratio over wide elevation angles using the polarization circuit build by hybrid couplers. This pixel also insures the uniformity characteristics that are the basis of the new ARMA sampling procedure. The matrix ARMA was used to generate two different beams with and axial ratio for both below 3dB in their interesting regions. Also the antenna is tested on the Nano-satellite support with no problems.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has been funded by the French Spatial Agency (CNES).

References

[1] Patent: JECKO, B.; HAJJ, M.; CHANTALAT, R.; SALAH TOUBET, «Antenne élémentaire et antenne réseau mono ou bidimensionnelle correspondante» PCT Patent: PCT/EP2012/076509; French Patent: FR 11 62141.

[2] ABOU TAAM, H., ZAKKA EL NASHEF, G., SALAH TOUBET, M., et al. Scan performance and re-configurability of agile radiating matrix antenna prototype. International Journal of Antennas and Propagation, 2015, vol. 2015.

[3] ABOU TAAM, Hussein, SALAH TOUBET, Moustapha, MONEDIERE, Thierry, et al. A new agile radiating system called electromagnetic band gap

matrix antenna. International Journal of Antennas and Propagation, 2014, vol. 2014.

[4] TAAM, Hussein Abou, EL NASHEF, Georges Zakka, ARNAUD, Eric, et al. Design development and experimental validation of an EBG matrix antenna for tracking application. International Journal of Microwave and Wireless Technologies, 2015, p. 1-9.

[5] TAAM, Hussein Abou, NASHEF, El, ZAKKA, Georges, et al. Feeding technique for the experimental validation of EBG matrix antenna. Microwave and Optical Technology Letters, 2015, vol. 57, no 3, p. 533-537.

[6] TAAM, Hussein Abou, TOUBET, Moutapha Salah, MONEDIERE, Thierry, et al. Interests of a 1D EBG Matrix compared to a patch array in terms of mutual coupling and grating lobe. In: Antennas and Propagation (EuCAP), 2013 7th European Conference on. IEEE, 2013. p. 1045-1048.

[7] TAAM, Hussein Abou, TOUBET, Moutapha Salah, MONEDIERE, Thierry, et al. Radiation control of an agile Matrix antenna by using specific algorithm. In: Microwave Symposium (MMS), 2013 13th Mediterranean. IEEE, 2013. p. 1-4.

[8] TAAM, Hussein Abou, EL NASHEF, Georges Zakka, JECKO, Bernard, et al. Agile radiating matrix antenna system for radar applications. In : Radar Conference (Radar), 2014 International. IEEE, 2014. p. 1-4.

[9] TAAM, Hussein Abou, SIBLINI, Ali, EL NASHEF, Georges Zakka, et al. An agile electronically scanned EBG matrix antenna for monitoring target activity. In: Microwave Conference (EuMC), 2015 European. IEEE, 2015. p. 1435-1438.

[10] TAAM, H. Abou, EL NASHEF, G. Zakka, ARNAUD, E., et al. Experimental validation of an agile electromagnetic band gap matrix antenna. In: Antennas and Propagation (EuCAP), 2015 9th European Conference on. IEEE, 2015. p. 1-4.

[11] B. Jecko, E. Aranud, H. Abou Taam, A. Siblini The ARMA concept: Comparison of AESA and ARMA technologies for agile antenna design. Fermat Journal vol.20, 2016

[12] SIBLINI, Ali, TAAM, Hussein Abou, JECKO, Bernard, Rammal, Mohammad. Pixel and Patch Comparison. SENSET 2017 Sensors, Networks, Smart and Emerging Technologies, Beirut Lebanon.

[13] SIBLINI, Ali, TAAM, Hussein Abou, JECKO, Bernard, et al. Wideband circularly polarized pixel antenna using special EBG approach: Building of agile antenna ARMA for space applications. In: Antennas and Propagation (MECAP), 2016 IEEE Middle East Conference on. IEEE, 2016. p. 1-4.

[14] SIBLINI, Ali, TAAM, Hussein Abou, JECKO, Bernard, Rammal, Mohammad. Pixel and Patch Comparison. New Circularly Polarized Matrix Antenna for Space Applications. Wireless testing symposiums 2016 London.

[15] SIBLINI, Ali, TAAM, Hussein Abou, JECKO, Bernard, et al. New agile EBG matrix antenna for space applications. In: Microwave Conference (EuMC), 2016 46th European. IEEE, 2016. p. 874-877.

[16] SIBLINI, Ali, JECKO, Bernard, et ARNAUD, Eric. Multimode reconfigurable Nano-satellite antenna for PDTM application. In: Antennas and Propagation (EUCAP), 2017 11th European Conference on. IEEE, 2017. p. 542-545.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 66-75 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Numerical Solution of Fuzzy Differential Equations with Z-numbers using Fuzzy Sumudu Transforms

Sina Razvarz[1], Raheleh Jafari[*,2], Wen Yu[1]

[1]*Departamento de Control Automatico CINVESTAV-IPN (National Polytechnic Institute) Mexico City 07360, Mexico*

[2]*Department of Information and Communication Technology Agder University College, 4876 Grimstad, Norway*

ARTICLE INFO

Article history:
Received: 14 November, 2017
Accepted: 17 January, 2017
Online: 30 January, 2018

Keywords:
K fuzzy Sumudu transform
fuzzy differential equation
Z-number

ABSTRACT

The uncertain nonlinear systems can be modeled with fuzzy differential equations (FDEs) and the solutions of these equations are applied to analyze many engineering problems. However, it is very difficult to obtain solutions of FDEs.
In this paper, the solutions of FDEs are approximated by utilizing the fuzzy Sumudu transform (FST) method. Here, the uncertainties are in the sense of Z-numbers. Important theorems are laid down to illustrate the properties of FST. The theoretical analysis and simulation results show that this new technique is effective to estimate the solutions of FDEs.

1. Introduction

This paper is an extension of work originally presented in [1]. In many physical and dynamical processes, mathematical modeling leads to the deterministic initial and boundary value problems. In practical the boundary values may be different from crisp and displays in the form of unknown parameters [2]. When the parameters or the states of the differential equations are uncertain, they can be modeled with FDE. In recent days, many methods have used FDE for modeling and control of uncertain nonlinear systems [3-5]. The basic idea of the fuzzy derivative was first introduced in [Chang]. Then it is extended in [6]. The first-order fuzzy initial value problem, as well as fuzzy partial differential equation, have been studied in [7]. By generalizing the differentiability, [6] gave an analytical solution. The Lipschitz condition, as well as the theorem for existence and uniqueness of the solution related to FDEs, are discussed in [10-12]. In [13], the analytical solutions of second order FDE are obtained. The analytical solutions of third order linear FDE are found in [14]. By the interval-valued method, [15] examined the basic solutions of nonlinear FDEs with generalized differentiability.

A novel technique in order to solve FDEs is laid down based on the Sumudu transform. Sumudu transform along with broad applications has been utilized in the area of system engineering and applied physics [16-18]. In [19], some simple and deeper fundamental theorems, as well as properties of the Sumudu Transform, were generalized. In [20], Sumudu transform is applied to the system of differential equations. In [21], Sumudu transform is used in order to find the solution of the fuzzy partial differential equation. In [22], Sumudu transform has been used to solve fractional differential equations.

In this paper, we use FST to approximate the Z-number solutions of the FDEs. The FST reduces the FDE to an algebraic equation. A very important property of the FST is that it can solve the equation without resorting to a new frequency domain. The procedure of switching FDEs to an algebraic equation is cited in [10] and is stated as an operational calculus. We extend our previous work [1] by generating more theorems for describing the properties of FST and displaying the uncertainties with Z-numbers. The Z-number is a new concept that is subjected to a higher potential to demonstrate the information of the human being as well as to utilize in information processing [23]. Z-numbers can be regarded as to answer questions and carry out the decisions [24]. There exist few structure based on the theoretical concept of Z-numbers [25]. [26] gave an inception, which results in the extension of the Z-numbers. [27] generated a theorem to convert the Z-numbers to the usual fuzzy sets.

In this paper, initially, some preliminary definitions along with properties related to FST are demonstrated. After that, solving FDEs by using the methodology of FST has been discussed. At the end, two examples along with comparisons are utilized in order to demonstrate the effectiveness of our proposed method.

*Corresponding Author: Raheleh Jafari, jafari3339@yahoo.com

2. Preliminaries

Prior to the introduction of the FST, some concepts related to the fuzzy variables and Z-numbers are laid down in this section [28, 29].

Definition 1: A fuzzy number B is a function of $B \in E : R \to [0,1]$, in such a manner, 1) B is normal, (there exists $a_0 \in R$ in such a manner $B(a_0) = 1$; 2) B is convex, $B(\gamma a + (1-\gamma)c) \geq \min\{B(a), B(c)\}$, $\forall a, c \in R, \forall \gamma \in [0,1]$; 3) B is upper semi-continuous on R, i.e., $B(a) \leq B(a_0) + \varepsilon$, $\forall a \in N(a_0)$, $\forall a_0 \in R$, $\forall \varepsilon > 0$, $N(a_0)$ is a neighborhood; 4) The set $B^+ = \{a \in R, B(a) > 0\}$ is compact.

Definition 2: The r-level of the fuzzy number B is defined as follows

$$[B]^r = \{a \in R : B(a) \geq r\} \tag{1}$$

where $0 < r \leq 1$, $B \in E$.

Definition 3: Let $B_1, B_2 \in E$ and $\xi \in R$, the operations addition, subtraction, multiplication and scalar multiplication are defined as

$$[B_1 \oplus B_2]^r = [B_1]^r + [B_2]^r = [\underline{B}_1^r + \underline{B}_2^r, \overline{B}_1^r + \overline{B}_2^r] \tag{2}$$

$$[B_1 ? B_2]^r = [B_1]^r - [B_2]^r = [\underline{B}_1^r - \underline{B}_2^r, \overline{B}_1^r - \overline{B}_2^r] \tag{3}$$

$$[B_1 ? B_2]^r = \begin{pmatrix} \min\{\underline{B}_1^r \underline{B}_2^r, \underline{B}_1^r \overline{B}_2^r, \overline{B}_1^r \underline{B}_2^r, \overline{B}_1^r \overline{B}_2^r\} \\ \max\{\underline{B}_1^r \underline{B}_2^r, \underline{B}_1^r \overline{B}_2^r, \overline{B}_1^r \underline{B}_2^r, \overline{B}_1^r \overline{B}_2^r\} \end{pmatrix} \tag{4}$$

$$[\xi B_1]^r = \xi [B_1]^r = \begin{cases} (\xi \underline{B}_1^r, \xi \overline{B}_1^r), & \xi \geq 0 \\ (\xi \overline{B}_1^r, \xi \underline{B}_1^r), & \xi \leq 0 \end{cases} \tag{5}$$

Definition 4: The Hausdorff distance between two fuzzy numbers B_1 and B_2 is defined as [30,31]

$$D(B_1, B_2) = \sup_{0 \leq r \leq 1} \{\max(|\underline{B}_1^r - \underline{B}_2^r|, |\overline{B}_1^r - \overline{B}_2^r|)\} \tag{6}$$

$D(B_1, B_2)$ has the following properties

(i) $D(B_1 \oplus u, B_2 \oplus u) = D(B_1, B_2)$, $\forall B_1, B_2, u \in E$

(ii) $D(\xi B_1, \xi B_2) = |\xi| D(B_1, B_2)$, $\forall \xi \in R, B_1, B_2 \in E$

(iii) $D(B_1 \oplus B_2, u \oplus v) \leq D(B_1, u) + D(B_2, v)$, $\forall B_1, B_2, u, v \in E$

(iv) (D, E) is stated as complete metric space.

Definition 5: The function $\psi : [a_1, a_2] \to E$ is integrable on $[a_1, a_2]$, if it satisfies in the below mentioned relation

$$\int_{a_1}^{\infty} \psi(x)dx = \left(\int_{a_1}^{\infty} \underline{\psi}(x,r)dx, \int_{a_1}^{\infty} \overline{\psi}(x,r)dx\right) \tag{7}$$

If $\psi(x)$ be a fuzzy value function, as well as $q(x)$ be a fuzzy Riemann integrable on $[a_1, \infty]$ so $\psi(x) \oplus q(x)$ can be a fuzzy Riemann integrable on $[a_1, \infty]$. Therefore,

$$\int_{a_1}^{\infty} (\psi(x) \oplus q(x))dx = \int_{a}^{\infty} \psi(x)dx \oplus \int_{a}^{\infty} q(x)dx \tag{8}$$

According to the fuzzy concept or in the case of interval arithmetic, equation $B_1 = B_2 \oplus s$ is not equivalent with $s = B_1 ? B_2 = B_1 \oplus (-1)B_2$ or to $B_2 = B_1 ? s = B_1 \oplus (-1)s$ and this is the main reason in introducing the following Hukuhara difference (H-difference).

Definition 6: The definition of H-difference [32,9], is proposed by $B_1 ?_H B_2 = s \Leftrightarrow B_1 = B_2 \oplus s$. If $B_1 ?_H B_2$ prevails, its r-level is $[B_1 ?_H B_2]^r = [\underline{B}_1^r - \underline{B}_2^r, \overline{B}_1^r - \overline{B}_2^r]$.

Precisely, $B_1 ?_H B_1 = 0$ but $B_1 ? B_1 \neq 0$.

Definition 7: Suppose $\psi : [a_1, a_2] \to E$ and $x_0 = [a_1, a_2]$. ψ is strongly generalized differentiable at x_0, if for all $k > 0$ adequately minute, $\psi'(x_0) \in E$ exists in such a manner that (i) $\exists$ $\psi(x_0 + k) ?_H \psi(x_0)$, $\psi(x_0) ?_H \psi(x_0 - k)$ and

$$\lim_{k \to 0^+} \frac{\psi(x_0+k) ?_H \psi(x_0)}{k} = \lim_{k \to 0^+} \frac{\psi(x_0) ?_H \psi(x_0-k)}{k} = \psi'(x_0)$$

or (ii) $\exists\ \psi(x_0) ?_H \psi(x_0 + k)$, $\psi(x_0 - k) ?_H \psi(x_0)$ and

$$\lim_{k \to 0^+} \frac{\psi(x_0) ?_H \psi(x_0+k)}{(-k)} = \lim_{k \to 0^+} \frac{\psi(x_0-k) ?_H \psi(x_0)}{(-k)} = \psi'(x_0),$$

or (iii) $\exists\ \psi(x_0 + k) ?_H \psi(x_0)$, $\psi(x_0 - k) ?_H \psi(x_0)$ and

$$\lim_{k \to 0^+} \frac{\psi(x_0+k) ?_H \psi(x_0)}{k} = \lim_{k \to 0^+} \frac{\psi(x_0-k) ?_H \psi(x_0)}{(-k)} = \psi'(x_0)$$

or (iv) $\exists\ \psi(x_0) ?_H \psi(x_0 + k)$, $\psi(x_0) ?_H \psi(x_0 - k)$ and

$$\lim_{k \to 0^+} \frac{\psi(x_0) ?_H \psi(x_0+k)}{(-k)} = \lim_{k \to 0^+} \frac{\psi(x_0) ?_H \psi(x_0-k)}{k} = \psi'(x_0)$$

Remark 1:It is clear that case (i) is H-derivative. Furthermore, a function is (i)-differentiable only when it is H-derivative. *Remark 2*: It can be concluded from [32] that, the definition of differentiability is noncontradictory [33].

Let us consider $\psi : R \to E$ where $\psi(t)$ has a parametric form as $[\psi(t,r)] = [\underline{\psi}(t,r), \overline{\psi}(t,r)]$, for all $0 \le r \le 1$, thus (i) If ψ be (i)-differentiable, so $\underline{\psi}(t,r)$ and $\overline{\psi}(t,r)$ are differentiable functions, moreover $\psi'(t) = (\underline{\psi}'(t,r), \overline{\psi}'(t,r))$. (ii) If ψ be (ii)-differentiable, so $\underline{\psi}(t,r)$ and $\overline{\psi}(t,r)$ are differentiable functions, moreover $\psi'(t) = (\overline{\psi}'(t,r), \underline{\psi}'(t,r))$. Suppose $f : (a_1, a_2) \to R$ is differentiable on (a_1, a_2), furthermore ψ' has finite root in (a_1, a_2), and $m \in E$, therefore, $\psi(x) = mf(x)$ is strongly generalized differentiable on (a_1, a_2) along with $\psi'(x) = mf'(x)$, $\forall x \in (a_1, a_2)$.

Theorem 1: [9] Assume $\psi : R \times E \to E$ is taken to be a continuous fuzzy function. If $x_0 \in R$, the fuzzy initial value constraint

$$\begin{cases} \phi'(t) = \psi(x, \phi) \\ \phi(x_0) = \phi_0 \end{cases} \tag{9}$$

is incorporated with two solutions: (i)-differentiable, also (ii)-differentiable. Hence the successive iterations

$$\phi_{n+1}(x) = \phi_0 + \int_{x_0}^{x} \psi(t, \phi_n(t))dt, \qquad \forall x \in [x_0, x_1] \tag{10}$$

and

$$\varphi_{n+1}(x) = \varphi_0 \,?_H (-1) \int_{x_0}^{x} \psi(t, \varphi_n(t))dt, \qquad \forall x \in [x_0, x_1] \tag{11}$$

approaches towards the two solutions sequentially.

Definition 8: A Z-number has two components $Z = [B(a), \tilde{p}]$. The primary component $B(a)$ is restriction on a real-valued uncertain variable a. The secondary component $\tilde{p}$ is a measure of the reliability of B. $\tilde{p}$ can be reliability, the strength of belief, probability or possibility. When $B(a)$ is a fuzzy number and $\tilde{p}$ is the probability distribution of a, the Z-number is stated as Z^+-number. When $B(a)$ as well as $\tilde{p}$ are fuzzy numbers, the Z-number is stated as Z^- -number.

The Z^+ -number carries more information when compared with Z^- -number. In this paper, we utilize the definition of Z^+ -number, i.e., $Z = [B, \tilde{p}]$, B is a fuzzy number and $\tilde{p}$ is a probability distribution.

To express the fuzzy number the most common membership functions are utilized in this paper. The popular membership functions are the triangular function

$$\mu_B = F(\lambda_1, \lambda_2, \lambda_3) = \begin{cases} \frac{a-\lambda_1}{\lambda_2-\lambda_1} & \lambda_1 \le a \le \lambda_2 \\ \frac{\lambda_3-\zeta}{\lambda_3-\lambda_2} & \lambda_2 \le a \le \lambda_3 \end{cases} otherwise \mu_B = 0 \tag{12}$$

and trapezoidal function

$$\mu_B = F(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \begin{cases} \frac{a-\lambda_1}{\lambda_2-\lambda_1} & \lambda_1 \le a \le \lambda_2 \\ \frac{\lambda_4-a}{\lambda_4-\lambda_3} & \lambda_3 \le a \le \lambda_4 \\ 1 & \lambda_2 \le a \le \lambda_3 \end{cases} otherwise \mu_B = 0 \tag{13}$$

The probability measure is defined as

$$\tilde{P} = \int_R \mu_B(a)\tilde{p}(a)da \tag{14}$$

where $\tilde{p}$ is the probability density of a, also R is the restriction on $\tilde{p}$. For discrete Z-numbers

$$\tilde{P}(B) = \sum_{i=1}^{n} \mu_B(a_i)\tilde{p}(a_i) \tag{15}$$

Definition 9: The r -level of the Z-number $Z = (B, \tilde{P})$ is illustrated as

$$[Z]^r = \left([B]^r, [\tilde{p}]^r\right) \tag{16}$$

where $0 < r \le 1$. $[\tilde{p}]^r$ is computed by the Nguyen's theorem

$$[\tilde{p}]^r = \tilde{p}([B]^r) = \tilde{p}([\underline{B}^r, \overline{B}^r]) = \left[\underline{\tilde{P}}^r, \overline{\tilde{P}}^r\right] \tag{17}$$

where $\tilde{p}([B]^r) = \{\tilde{p}(a) \mid a \in [B]^r\}$. So $[Z]^r$ can be demonstrated as the form r -level of a fuzzy number

$$[Z]^r = \left(\underline{Z}^r, \overline{Z}^r\right) = \left(\left(\underline{B}^r, \underline{\tilde{P}}^r\right), \left(\overline{B}^r, \overline{\tilde{P}}^r\right)\right) \tag{18}$$

where $\underline{\tilde{P}}^r = \underline{B}^r \tilde{p}(\underline{a_i}^r)$, $\overline{\tilde{P}}^r = \overline{B}^r \tilde{p}(\overline{a_i}^r)$, $[a_i]^r = (\underline{a_i}^r, \overline{a_i}^r)$.

Similar with the fuzzy numbers, the Z-numbers are also incorporated with three primary operations, addition, subtraction, and multiplication. The operations in this paper are different definitions with [34]. The r -level of Z-numbers is applied to simplify the operations.

Suppose $Z_1 = (B_1, \tilde{p}_1)$ and $Z_2 = (B_2, \tilde{p}_2)$ be two discrete Z-numbers expressing the uncertain variables a_1 and a_2, $\sum_{\iota=1}^{n} \tilde{p}_1(a_{1\iota}) = 1$, $\sum_{\iota=1}^{n} \tilde{p}_2(a_{2\iota}) = 1$. The operations are displayed as

$$Z_{12} = Z_1 * Z_2 = (B_1 * B_2, \tilde{p}_1 * \tilde{p}_2) \tag{19}$$

where $* \in \{\oplus, ?, ?\}$.

For all $\tilde{p}_1 * \tilde{p}_2$ operations, we use convolutions for the discrete probability distributions

$$\tilde{p}_1 * \tilde{p}_2 = \sum_i \tilde{p}_1(a_{1,i})\tilde{p}_2(a_{2,(n-i)}) = \tilde{p}_{12}(a) \tag{20}$$

The above definitions satisfy the Hukuhara difference [35],

$$\begin{aligned} Z_1 \,?_H Z_2 &= Z_{12} \\ Z_1 &= Z_2 \oplus Z_{12} \end{aligned} \tag{21}$$

If $Z_1 ?_H Z_2$ prevails, the r -level is

$$[Z_1 ?_H Z_2]^r = [\underline{Z}_1^r - \underline{Z}_2^r, \overline{Z}_1^r - \overline{Z}_2^r] \quad (22)$$

Obviously, $Z_1 ?_H Z_1 = 0$, $Z_1 ? Z_1 \neq 0$.

Also the above definitions satisfy the generalized Hukuhara difference [7]

$$Z_1 ?_{gH} Z_2 = Z_{12} \Leftrightarrow \begin{cases} 1)\ Z_1 = Z_2 \oplus Z_{12} \\ 2)\ Z_2 = Z_1 \oplus (-1)Z_{12} \end{cases} \quad (23)$$

It is easy to display that 1) and 2) in combination are genuine if and only if Z_{12} is a crisp number. With respect to r -level we have

$[Z_1 ?_{gH} Z_2]^r = [\min\{\underline{Z}_1^r - \underline{Z}_2^r, \overline{Z}_1^r - \overline{Z}_2^r\}, \max\{\underline{Z}_1^r - \underline{Z}_2^r, \overline{Z}_1^r - \overline{Z}_2^r\}]$

and If and subsist, $Z_1 ?_H Z_2 = Z_1 ?_{gH} Z_2$. The circumstances for the inerrancy of $Z_{12} = Z_1 ?_{gH} Z_2 \in E$ are

$$1)\begin{cases} \underline{Z}_{12}^r = \underline{Z}_1^r - \underline{Z}_2^r and \overline{Z}_{12}^r = \overline{Z}_1^r - \overline{Z}_2^r \\ with\ \underline{Z}_{12}^r increa\sin g,\ \overline{Z}_{12}^r decrea\sin g,\ \underline{Z}_{12}^r \leq \overline{Z}_{12}^r \end{cases}$$
$$2)\begin{cases} \underline{Z}_{12}^r = \overline{Z}_1^r - \overline{Z}_2^r and\ \overline{Z}_{12}^r = \underline{Z}_1^r - \underline{Z}_2^r \\ with\ \underline{Z}_{12}^r increa\sin g,\ \overline{Z}_{12}^r decrea\sin g,\ \underline{Z}_{12}^r \leq \overline{Z}_{12}^r \end{cases} \quad (24)$$

where $\forall r \in [0,1]$

If B is a triangular function, the absolute value of the Z-number $Z = (B, \tilde{p})$ is defined as

$$|Z(a)| = (|\lambda_{11}| + |\lambda_{12}| + |\lambda_{13}|, \tilde{p}(|\lambda_{21}| + |\lambda_{22}| + |\lambda_{23}|)) \quad (25)$$

If B_1, as well as B_2 are triangular functions, the supremum metric for Z-numbers $Z_1 = (B_1, \tilde{p}_1)$ and $Z_2 = (B_2, \tilde{p}_2)$ is illustrated as

$$D(Z_1, Z_2) = d(B_1, B_2) + d(\tilde{p}_1, \tilde{p}_2) \quad (26)$$

in this case $d(\cdot,\cdot)$ is defined as the supremum metrics with fuzzy sets [3]. $D(Z_1, Z_2)$ is incorporated with the following possessions

$$D(Z_1 + Z, Z_2 + Z) = D(Z_1, Z_2)$$
$$D(Z_2, Z_1) = D(Z_1, Z_2)$$
$$D(\zeta Z_1, \zeta Z_2) = |\zeta| D(Z_1, Z_2)$$
$$D(Z_1, Z_2) \leq D(Z_1, Z) + D(Z, Z_2)$$

where $\zeta \in R$, $Z = (B, \tilde{p})$ is Z-number and B is triangle function, for proof refer to [36].

Definition 10: Suppose $\hat{Z}$ demonstrates the space of Z -numbers, then the $r-$ level of Z -number valued function $\Psi : [a_1, a_2] \to \hat{Z}$ is defined as

$$\Psi(\phi, r) = [\underline{\Psi}(\phi, r), \overline{\Psi}(\phi, r)]$$

where $\phi \in \hat{Z}$, and $r \in [0,1]$.

Based on the definition of Generalized Hukuhara difference, the gH-derivative of Ψ at ϕ_0 is defined as

$$\Psi'(\varphi_0) = \lim_{h \to 0} \frac{1}{h} [\Psi(\varphi_0 + h) ?_{gH} \Psi(\varphi_0)] \quad (27)$$

In (27), $\Psi(\phi_0 + h)$ and $\Psi(\phi_0)$ represents similar style with Z_1 and Z_2respectively defined in (laiop).

By implementing the $r-$ level (16) to initial value problem, $\phi'(t) = \psi(t, \phi(t))$, we generate two Z-number valued functions: $\underline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)]$ and $\overline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)]$.

The initial value problem can be equivalent to the following relation

$$i)\begin{cases} \underline{\phi}' = \underline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)] \\ \overline{\phi}' = \overline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)] \end{cases}$$
$$ii)\begin{cases} \underline{\phi}' = \overline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)] \\ \overline{\phi}' = \underline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)] \end{cases} \quad (28)$$

3. Fuzzy Sumudu transform

Fuzzy initial and boundary value problems can be resolved by utilizing fuzzy Laplace transform [10]. In this paper, the FST methodology for Z-number is illustrated, furthermore the properties of this methodology is stated. By applying the FST methodology, the FDE based on Z-number is reduced to an algebraic equation. The main advantage of the FST is that it can resolve the equation without resorting to a new frequency domain. The methodology of converting FDEs to an algebraic equation is expressed in [10].

Definition 11: Suppose $\psi(t)$ be a continuous Z-number valued function, also, $\psi(Bt) ? e^{-t}$ be an improper Z-number Riemann integrable on $[0, \infty)$. Accordingly, $\int_0^\infty \psi(Bt) ? e^{-t} dt$ is expressed as FST and it is defined by $\Omega(B) = \mathbf{S}[\psi(t)] = \int_0^\infty \psi(Bt) ? e^{-t} dt$, where $0 \leq B < K$, $K \geq 0$, also e^{-t} is a real-valued function. Based on the Theorem 3 we have the following relation

$$\int_0^\infty \psi(Bt) ? e^{-t} dt = (\int_0^\infty \underline{\psi}(Bt, r) e^{-t} dt, \int_0^\infty \overline{\psi}(Bt, r) e^{-t} dt) \quad (29)$$

Let

$$\mathbf{S}[\underline{\psi}(t,r)] = \int_0^\infty \underline{\psi}(Bt, r) e^{-t} dt$$
$$\mathbf{S}[\overline{\psi}(t,r)] = \int_0^\infty \overline{\psi}(Bt, r) e^{-t} dt \quad (30)$$

hence we obtain the following relation

$$\mathbf{S}[\psi(t)] = (\mathbf{S}[\underline{\psi}(t,r), \mathbf{S}\overline{\psi}(t,r)]) \quad (31)$$

Theorem 2: Suppose $\psi'(t)$ be a Z-number valued integrable function, as well as $\psi(t)$ be the primitive of $\psi'(t)$ on $[0,\infty)$. Therefore,

$$\mathbf{S}[\psi'(t)]=\frac{1}{B}?\mathbf{S}[\psi(t)]?(\frac{1}{B}?[\psi(0)]) \tag{32}$$

where is considered to be (i)-differentiable, or

$$\mathbf{S}[\psi'(t)]=\frac{-1}{B}?[\psi(0)]?(\frac{-1}{B}?\mathbf{S}[\psi(t)]) \tag{33}$$

where ψ is considered to be (ii)-differentiable.

Proof. For arbitrary fixed $r\in[0,1]$ we have

$$\begin{aligned}&\tfrac{1}{B}?\mathbf{S}[\psi(t)]?(\tfrac{1}{B}?\psi(0))\\&=(\tfrac{1}{B}\mathbf{S}[\underline{\psi}(t,r)]-\tfrac{1}{B}\mathbf{S}[\underline{\psi}(0,r)],\tfrac{1}{B}\mathbf{S}[\overline{\psi}(t,r)]-\tfrac{1}{B}\mathbf{S}[\overline{\psi}(0,r)])\end{aligned} \tag{34}$$

We have the following relations

$$\begin{aligned}\mathbf{S}[\overline{\psi}'(t,r)]&=\tfrac{1}{B}\mathbf{S}[\overline{\psi}(t,r)]-\tfrac{1}{B}[\overline{\psi}(0,r)]\\\mathbf{S}[\underline{\psi}'(t,r)]&=\tfrac{1}{B}\mathbf{S}[\underline{\psi}(t,r)]-\tfrac{1}{B}[\underline{\psi}(0,r)]\end{aligned} \tag{35}$$

Hence, we obtain

$$\frac{1}{B}?\mathbf{S}[\psi(t)]?(\frac{1}{B}?\psi(0))=(\mathbf{S}[\underline{\psi}'(t,r)],\mathbf{S}[\overline{\psi}'(t,r)]) \tag{36}$$

If is considered to be (i)-differentiable, so

$$\frac{1}{B}?\mathbf{S}[\psi(t)]?(\frac{1}{B}?\psi(0))=\mathbf{S}[\psi'(t)] \tag{37}$$

Let is (ii)-differentiable. For arbitrary fixed $\alpha\in[0,1]$ we obtain

$$\begin{aligned}&\tfrac{-1}{B}?[\psi(0)]?(\tfrac{-1}{B}?\mathbf{S}[\psi(t)])\\&=(\tfrac{-1}{B}\overline{\psi}(0,r)+\tfrac{1}{B}\mathbf{S}[\overline{\psi}(t,r)],\tfrac{-1}{B}\underline{\psi}(0,r)+\tfrac{1}{B}\mathbf{S}[\underline{\psi}(t,r)])\end{aligned} \tag{38}$$

The above equation can be written as the following relation

$$\begin{aligned}&\tfrac{-1}{B}?[\psi(0)]?(\tfrac{-1}{B}?\mathbf{S}[\psi(t)])\\&=(\tfrac{1}{B}\mathbf{S}[\overline{\psi}(t,r)]-\tfrac{1}{B}\overline{\psi}(0,r),\tfrac{1}{B}\mathbf{S}[\underline{\psi}(t,r)]-\tfrac{1}{B}\underline{\psi}(0,r))\end{aligned} \tag{39}$$

We obtain

$$\begin{aligned}\mathbf{S}[\overline{\psi}'(t,r)]&=\tfrac{1}{B}\mathbf{S}[\overline{\psi}(t,r)]-\tfrac{1}{B}\overline{\psi}(0,r)\\\mathbf{S}[\underline{\psi}'(t,r)]&=\tfrac{1}{B}\mathbf{S}[\underline{\psi}(t,r)]-\tfrac{1}{B}\underline{\psi}(0,r)\end{aligned} \tag{40}$$

So, we have

$$(\frac{-1}{B}\psi(0))?(\frac{-1}{B}?\mathbf{S}[\psi(t)])=(\mathbf{S}[\overline{\psi}'(t,r)],\mathbf{S}[\underline{\psi}'(t,r)]) \tag{41}$$

Hence

$$(\frac{-1}{B}\psi(0))?(\frac{-1}{B}?\mathbf{S}[\psi(t)])=\mathbf{S}([\overline{\psi}'(t,r)],[\underline{\psi}'(t,r)]) \tag{42}$$

Since ψ is (ii)-differentiable, therefore,

$$(\frac{-1}{B}\psi(0))?(\frac{-1}{B}?\mathbf{S}[\psi(t)])=\mathbf{S}[\psi'(t)] \tag{43}$$

Theorem 3: Taking into consideration that Sumudu transform is a linear transformation, so if $\psi(t)$ and $\vartheta(t)$ be continuous Z-number valued functions, moreover k_1 as well as k_2 be constant, therefore the following relation can be obtained

$$\mathbf{S}[(k_1?\psi(t))\oplus(k_2?\vartheta(t))]=(k_1?\mathbf{S}[\psi(t)])\oplus(k_2?\mathbf{S}[\vartheta(t)]) \tag{44}$$

Proof. We have

$$\begin{aligned}&\mathbf{S}[(k_1?\psi(t))\oplus(k_2?\vartheta(t))]\\&=\textstyle\int_0^\infty(k_1?\phi(Bt)\oplus k_2?\vartheta(Bt))?e^{-t}dt\\&=\textstyle\int_0^\infty k_1?\psi(Bt)?e^{-t}dt\oplus\int_0^\infty k_2?\vartheta(Bt)?e^{-t}dt\\&=k_1?(\textstyle\int_0^\infty\psi(Bt)?e^{-t}dt)\oplus k_2?(\int_0^\infty\vartheta(Bt)?e^{-t}dt)\\&=k_1?\mathbf{S}[\psi(t)]\oplus k_2?\mathbf{S}[\vartheta(t)]\end{aligned} \tag{45}$$

Therefore, we conclude

$$\mathbf{S}[(k_1?\psi(t))\oplus(k_2?\vartheta(t))]=(k_1?\mathbf{S}[\psi(t)])\oplus(k_2?\mathbf{S}[\vartheta(t)]) \tag{46}$$

Lemma 1: Assume that $\psi(t)$ is a continuous Z-number value function on $[0,\infty)$, also $\gamma\geq 0$, thus

$$\mathbf{S}[\gamma?\psi(t)]=\gamma?\mathbf{S}[\psi(t)] \tag{47}$$

Proof. Fuzzy Sumudu transform $\gamma?\psi(t)$ is defined as

$$\mathbf{S}[\gamma?\psi(t)]=\textstyle\int_0^\infty\gamma?\psi(Bt)?e^{-t}dt \tag{48}$$

furthermore, we have

$$\textstyle\int_0^\infty\gamma?\psi(Bt)?e^{-t}dt=\gamma?\int_0^\infty\psi(Bt)?e^{-t}dt \tag{49}$$

therefore,

$$\mathbf{S}[\gamma?\psi(t)]=\gamma?\mathbf{S}[\psi(t)] \tag{50}$$

Lemma 2: Assume that $\psi(t)$ is a continuous Z-number valued function, and $\vartheta(t)\geq 0$. Furthermore, if we suppose that $(\psi(t)?\vartheta(t))?e^{-t}$ is improper Z-number Reiman integrable on $[0,\infty)$, then

$$\begin{aligned}&\textstyle\int_0^\infty(\psi(Bt)?\vartheta(Bt))?e^{-t}dt\\&=(\textstyle\int_0^\infty\vartheta(Bt)\underline{\psi}(Bt,r)e^{-t}dt,\int_0^\infty\vartheta(Bt)\overline{\psi}(Bt,r)e^{-t}dt)\end{aligned} \tag{51}$$

Theorem 4: Suppose $\psi(t)$ is a continuous Z-number valued function, also $\mathbf{S}[\psi(t)]=D(B)$, therefore,

$$\mathbf{S}[e^{a_1t}?\psi(t)]=\frac{1}{1-a_1B}D(\frac{B}{1-a_1B}) \tag{52}$$

where e^{a_1t} is considered to be a real value function, also $1-a_1B>0$.

Proof. We have the following relation

$$\begin{aligned}&\mathbf{S}[e^{a_1 t}\,?\,\psi(t)]=\int_0^{\infty}e^{a_1 Bt}e^{-t}\psi(Bt)dt\\&=(\int_0^{\infty}e^{-(1-a_1B)t}\underline{\psi}(Bt,r)dt,\int_0^{\infty}e^{-(1-a_1B)t}\overline{\psi}(Bt,r)dt)\end{aligned}\tag{53}$$

Let us consider $z=1-a_1Bt$, then

$$\begin{aligned}&\mathbf{S}[e^{a_1 t}\,?\,\psi(t)]\\&=\tfrac{1}{1-a_1B}(\int_0^{\infty}\underline{\psi}(\tfrac{Bz}{1-a_1B},r)e^{-z}dz,\int_0^{\infty}\overline{\psi}(\tfrac{Bz}{1-a_1B},r)e^{-z}dz)\\&=\{\tfrac{1}{1-a_1B}\underline{D}(\tfrac{B}{1-a_1B}),\tfrac{1}{1-a_1B}\overline{D}(\tfrac{B}{1-a_1B})\}=\tfrac{1}{1-a_1B}D(\tfrac{B}{1-a_1B})\end{aligned}\tag{54}$$

4. Solving fuzzy initial value problem with fuzzy Sumudu transform method

Consider the following fuzzy initial value problem based on Z-numbers

$$\begin{cases}\phi'(t)=\psi(t,\phi(t)),\\\phi(0)=(\underline{\phi}(0,r),\overline{\phi}(0,r)),\qquad 0<r\le 1\end{cases}\tag{55}$$

where $\psi(t,\phi(t))$ is a Z-number function. By utilizing FST method for Z-numbers, we obtain

$$\mathbf{S}[\phi'(t)]=\mathbf{S}[\psi(t,\phi(t))]\tag{56}$$

The solving process of Eq. (56) is based on the following cases.

Case 1: Assume that $\phi'(t)$ is (i)-differentiable. Base on the Theorem 3 we extract

$$\phi'(t)=(\underline{\phi}'(t,r),\overline{\phi}'(t,r))\tag{57}$$

$$\mathbf{S}[\varphi'(t)]=(\frac{1}{B}?\mathbf{S}[\varphi(t)])?\frac{1}{B}\varphi(0)\tag{58}$$

Eq. (58) can be displayed as the following relation

$$\begin{cases}\mathbf{S}[\underline{\psi}(t,\phi(t),r)]=\frac{1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,\alpha)\\\mathbf{S}[\overline{\psi}(t,\phi(t),r)]=\frac{1}{B}\mathbf{S}[\overline{\phi}(t,r)]-\frac{1}{B}\overline{\phi}(0,r)\end{cases}\tag{59}$$

where

$$\begin{cases}\underline{\psi}(t,\phi(t),r)=\min\{\psi(t,B)\mid B\in(\underline{\phi}(t,r),\overline{\phi}(t,r))\}\\\overline{\psi}(t,\phi(t),r)=\max\{\psi(t,B)\mid B\in(\underline{\phi}(t,r),\overline{\phi}(t,r))\}\end{cases}\tag{60}$$

Accordingly, Eq. (60) can be resolved on the basis of the following assumptions

$$\mathbf{S}[\underline{\phi}(t,r)]=U_1(B,r)\tag{61}$$

$$\mathbf{S}[\overline{\phi}(t,r)]=U_2(B,r)\tag{62}$$

where $U_1(B,r)$, as well as $U_2(B,r)$ are the Z-number solutions of the Eq. (60). By applying inverse Sumudu transform, $\underline{\phi}(t,r)$ as well as $\overline{\phi}(t,r)$ are computed as

$$\underline{\phi}(t,r)=\mathbf{S}^{-1}[U_1(B,r)]\tag{63}$$

$$\overline{\phi}(t,r)=\mathbf{S}^{-1}[U_2(B,r)]\tag{64}$$

Case 2: Assume that $\phi'(t)$ is (ii)-differentiable. Based on the Theorem 3 we extract

$$\phi'(t)=(\overline{\phi}'(t,r),\underline{\phi}'(t,r))\tag{65}$$

$$\mathbf{S}[\varphi'(t)]=(\frac{-1}{B}?\varphi(0))?(\frac{-1}{B}?\mathbf{S}[\varphi(t)])\tag{66}$$

Eq. (66) can be displayed as the following relation

$$\begin{cases}\mathbf{S}[\underline{\psi}(t,\phi(t),r)]=\frac{1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,r)\\\mathbf{S}[\overline{\psi}(t,\phi(t),r)]=\frac{1}{B}\mathbf{S}[\overline{\phi}(t,r)]-\frac{1}{B}\overline{\phi}(0,r)\end{cases}\tag{67}$$

where

$$\begin{cases}\underline{\psi}(t,\phi(t),r)=\min\{\psi(t,B)\mid B\in(\underline{\phi}(t,r),\overline{\phi}(t,r))\}\\\overline{\psi}(t,\phi(t),r)=\max\{\psi(t,B)\mid B\in(\underline{\phi}(t,r),\overline{\phi}(t,r))\}\end{cases}\tag{68}$$

Accordingly, Eq. (68) can be resolved on the basis of the following assumptions

$$\begin{aligned}&\mathbf{S}(\underline{\phi}(t,r)=V_1(B,r)\\&\mathbf{S}(\overline{\phi}(t,r)=V_2(B,r)\end{aligned}\tag{69}$$

where $V_1(B,r)$, and $V_2(B,r)$ are the Z-number solutions of the Eq. (68). By applying inverse Sumudu transform, $\underline{\phi}(t,r)$ as well as $\overline{\phi}(t,r)$ are computed as

$$\begin{aligned}&\underline{\phi}(t,r)=\mathbf{S}^{-1}[V_1(B,r)]\\&\overline{\phi}(t,r)=\mathbf{S}^{-1}[V_2(B,r)]\end{aligned}\tag{70}$$

5. Examples

The following examples have been used to narrate the methodology proposed in this paper.

5.1. Example 1

A tank with a heating system is displayed in Figure 1, where $\widetilde{R}=0.5$, the thermal capacitance is $\widetilde{C}=2$ also the temperature is ψ . The model is formulated as follows[10, 37],

$$\begin{cases}\phi'(t)=-\frac{1}{\widetilde{R}\widetilde{C}}\phi(t),\qquad 0\le t\le T\\\phi(0)=[(\underline{\phi}(0,r),\overline{\phi}(0,r)),p(0.8,0.9,1)]\end{cases}\tag{71}$$

By utilizing the FST method based on Z-number we obtain

$$\mathbf{S}[\phi'(t)]=\mathbf{S}[-\phi(t)]\tag{72}$$

$$S[\phi'(Bt)]\odot e^{-t}dt\tag{73}$$

where $0\le B<K$. By considering case 1 for Z-numbers the following relation is obtained

$$\mathbf{S}[\varphi'(t)]=\frac{1}{B}?(\mathbf{S}[\varphi(t)]?\varphi(0))=\frac{1}{B}\mathbf{S}[\varphi(t)]?\frac{1}{B}\varphi(0) \quad (74)$$

Therefore

$$-\mathbf{S}[\varphi(t)]=\frac{1}{B}\mathbf{S}[\varphi(t)]?\frac{1}{B}\varphi(0) \quad (75)$$

Based on the Eq. (59), we have

$$\begin{cases}-\mathbf{S}[\bar{\phi}(t,r)]=\frac{1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,r)\\ -\mathbf{S}[\underline{\phi}(t,r)]=\frac{1}{B}\mathbf{S}[\bar{\phi}(t,r)]-\frac{1}{B}\bar{\phi}(0,r)\end{cases} \quad (76)$$

Therefore, the Z-number solution of Eq. (76) is extracted as $[(\mathbf{S}[\bar{\phi}(t,r)],\mathbf{S}[\underline{\phi}(t,r)]),p(0.8,0.94,1)]$ where

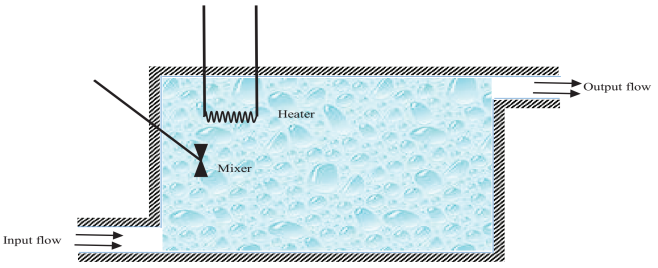


Fig. 1. A tank with a heating system

$$\begin{cases}\mathbf{S}[\bar{\phi}(t,r)]=(\frac{-1}{B^2-1})\bar{\phi}(0,r)+(\frac{B}{B^2-1})\underline{\phi}(0,r)\\ \mathbf{S}[\underline{\phi}(t,r)]=(\frac{-1}{B^2-1})\underline{\phi}(t,r)+(\frac{B}{B^2-1})\bar{\phi}(0,r)\end{cases} \quad (77)$$

By utilizing the inverse Sumudu transform for Z-numbers, we have

$$\begin{cases}\mathbf{S}[\bar{\phi}(t,r)]=\bar{\phi}(0,r)\mathbf{S}^{-1}(\frac{-1}{B^2-1})+\underline{\phi}(0,r)\mathbf{S}^{-1}(\frac{B}{B^2-1})\\ \mathbf{S}[\underline{\phi}(t,r)]=\underline{\phi}(0,r)\mathbf{S}^{-1}(\frac{-1}{B^2-1})+\bar{\phi}(0,r)\mathbf{S}^{-1}(\frac{B}{B^2-1})\end{cases} \quad (78)$$

where

$$\begin{cases}\bar{\phi}(t,r)=e^{t}(\frac{\bar{\phi}(0,r)-\underline{\phi}(0,r)}{2})+e^{-t}(\frac{\bar{\phi}(0,r)+\underline{\phi}(0,r)}{2})\\ \underline{\phi}(t,r)=e^{t}(\frac{\underline{\phi}(0,r)-\bar{\phi}(0,r)}{2})+e^{-t}(\frac{\underline{\phi}(0,r)+\bar{\phi}(0,r)}{2})\end{cases} \quad (79)$$

By considering case 2 for Z-numbers the following relation is obtained

$$\mathbf{S}[\varphi'(t)]=(\frac{-1}{B}\mathbf{S}[\varphi(t)])?(\frac{-1}{B}\varphi(0)) \quad (80)$$

Hence

$$-\mathbf{S}[\varphi(t)]=(\frac{-1}{B}\mathbf{S}[\varphi(t)])?(\frac{-1}{B}\varphi(0)) \quad (81)$$

Based on the above relations, Eq. (71) is illustrated as

$$\begin{cases}-\mathbf{S}[\underline{\phi}(t,r)]=\frac{1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,r)\\ -\mathbf{S}[\bar{\phi}(t,r)]=\frac{1}{B}\mathbf{S}[\bar{\phi}(t,r)]-\frac{1}{B}\bar{\phi}(0,r)\end{cases} \quad (82)$$

So, the Z-number solution of Eq. (82) is displayed as $[(\mathbf{S}[\underline{\phi}(t,r)],\mathbf{S}[\bar{\phi}(t,r)]),p(0.8,0.9,1)]$ where

$$\begin{cases}\mathbf{S}[\underline{\phi}(t,r)]=\underline{\phi}(0,r)(\frac{1}{B+1})\\ \mathbf{S}[\bar{\phi}(t,r)]=\bar{\phi}(t,r)(\frac{1}{B+1})\end{cases} \quad (83)$$

By utilizing the inverse Sumudu transform for Z-numbers, we have

$$\begin{cases}\underline{\phi}(t,r)=\underline{\phi}(0,r)\mathbf{S}^{-1}(\frac{1}{B+1})\\ \bar{\phi}(t,r)=\bar{\phi}(0,r)\mathbf{S}^{-1}(\frac{1}{B+1})\end{cases} \quad (84)$$

where

$$\begin{cases}\underline{\phi}(t,r)=e^{-t}\underline{\phi}(0,r)\\ \bar{\phi}(t,r)=e^{-t}\bar{\phi}(0,r)\end{cases} \quad (85)$$

If the initial condition is taken to be a symmetric triangular Z-number as $\phi(0)=[(-a(1-r),a(1-r)),p(0.8,0.9,1)]$, so

Case 1 :

$$\begin{cases}\underline{\phi}(t,r)=e^{t}(-a(1-r))\\ \bar{\phi}(t,r)=e^{t}(a(1-r))\end{cases} \quad (86)$$

Case 2:

Table 1. Approximation errors based on Z-numbers

α	FST	Average Euler
0	[(0.0078,0.0195),p(0.8,0.86,0.94)]	[(0.0138,0.0215),p(0.7,0.8,0.87)]
0.2	[(0.0085,0.0169),p(0.75,0.8,0.9)]	[(0.0188,0.0286),p(0.7,0.8,0.87)]
0.6	[(0.0058,0.0115),p(0.8,0.9,1)]	[(0.0182,0.0198),p(0.7,0.8,0.92)]
0.8	[(0.0091,0.0123),p(0.7,0.75,0.8)]	[(0.0148,0.0189),p(0.6,0.7,0.8)]
1	[(0.0132,0.0132),p(0.7,0.8,0.9)]	[(0.0710,0.0710),p(0.6,0.75,0.87)]

$$\begin{cases}\underline{\phi}(t,r)=e^{-t}(-a(1-r))\\ \bar{\phi}(t,r)=e^{-t}(a(1-r))\end{cases} \quad (87)$$

Approximation errors based on Z-numbers are shown in Table 1. These errors are the differences between the exact and the approximation solutions, for two different methods: FST and Average Euler method [38].

The following formula is utilized to transfer the Z-numbers to fuzzy numbers,

$$\sigma=\frac{\int\phi\pi_P(\phi)d\phi}{\int\pi_P(\phi)d\phi}$$

By taking in to consideration $Z=(B,\tilde{p})=[(0.0078,0.0195),p(0.8,0.86,0.94)]$, we obtain $Z^{\sigma}=[0.0078,0.0195;0.86]$ accordingly $Z'=[\sqrt{0.86}\,0.0078,\sqrt{0.86}\,0.0195]$. Approximation errors based on fuzzy numbers are shown in Table 2.

Figure 2 shows the corresponding error plots based on fuzzy numbers. FST is more accurate than the Average Euler method. Figure 3 and Figure 4 demonstrate the corresponding solution plots based on fuzzy numbers.

By implementing Z-numbers the degree of reliability of the information can be increased [39, 40]. The comparison between the Z-number $Z=[(0.0078,0.0195),p(0.8,0.86,0.94)]$ and fuzzy number $[0.0072,0.0180]$ is displayed in Figure 5. It can be seen that the Z-number incorporates with several information, also the solution related to the Z-number is more precise. The membership function regarding the restriction in the Z-number is considered to be $\mu_{B_Z}=[0.0078,0.0195]$. It can be in probability form.

Table 2. Approximation errors based on fuzzy numbers

α	FST	Average Euler
0	[0.0072,0.0180]	[0.0123,0.0192]
0.2	[0.0076,0.0151]	[0.0168,0.0255]
0.6	[0.0055,0.0109]	[0.0162,0.0177]
0.8	[0.0078,0.0106]	[0.0123,0.0158]
1	[0.0118,0.0118]	[0.0614,0.0614]

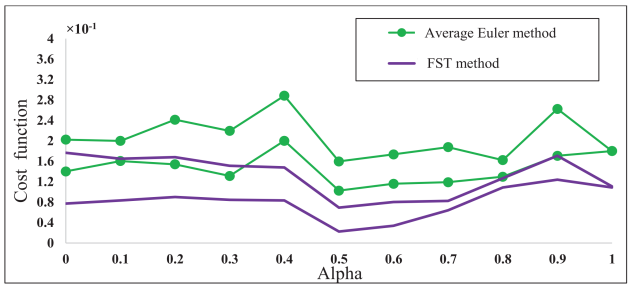


Fig. 2. The lower and upper bounds of absolute errors based on fuzzy numbers

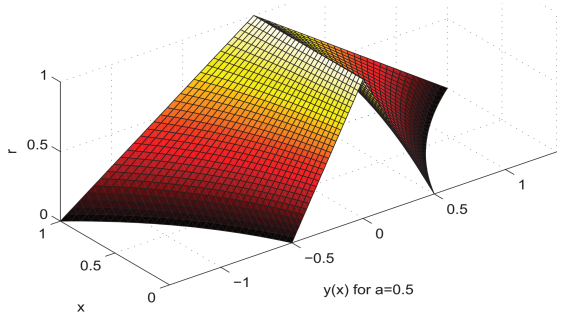


Fig. 3. The solution of FDE under case 1 consideration based on fuzzy numbers

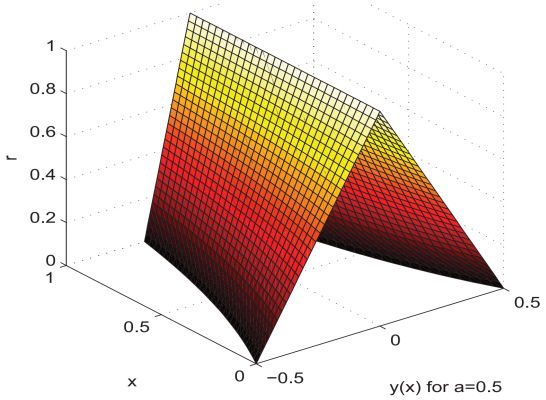


Fig. 4. The solution of FDE under case 2 consideration based on fuzzy numbers

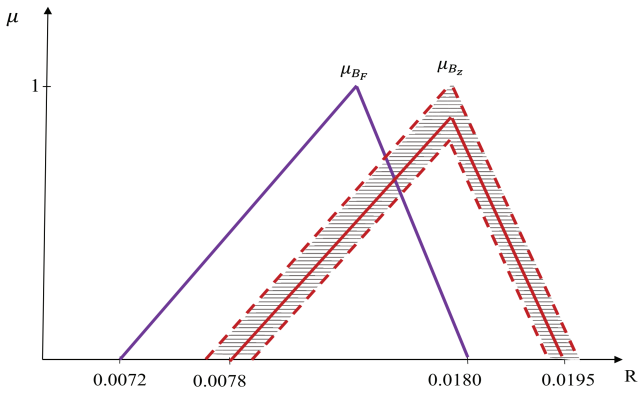


Fig. 5. The comparison between the Z-number and fuzzy number

5.2. Example 2

A tank system is displayed in Figure 6. Suppose $I=t+1$ is inflow disturbances of the tank that generates vibration in liquid level ϕ, also $H=1$ is the flow obstruction, which can be curbed utilizing the valve. $Q=1$ is the cross-section of the tank. The liquid level is illustrated as following relation [41, 42],

$$\begin{cases} \phi'(t)=-\frac{1}{QH}\phi(t)+\frac{I}{Q}, & 0\le t\le T \\ \phi(0)=[(\underline{\phi}(0,r),\overline{\phi}(0,r)),p(0.85,0.91,1)] \end{cases} \quad (88)$$

By utilizing the FST method based on Z-number we obtain

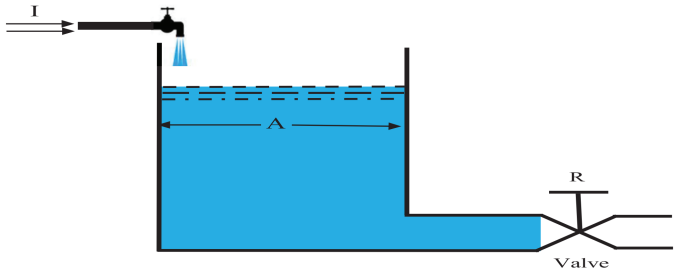


Fig. 6. Liquid tank system

$$-\mathbf{S}[\varphi(t)]=(\frac{1}{B}?\mathbf{S}[\varphi(t)])?(\frac{1}{B}\mathbf{S}[\varphi(0)]) \quad (89)$$

$$\mathbf{S}[\phi'(t)](\int_{a_1}^{\infty}\phi(Bt)e^{'-t}dt \quad (90)$$

By considering case 2 for Z-numbers the following relation is obtained

$$\mathbf{S}[\varphi'(t)]=(\frac{-1}{B}?\mathbf{S}[\varphi(t)])?(\frac{-1}{B}\mathbf{S}[\varphi(0)]) \quad (91)$$

So

$$-\mathbf{S}[\varphi(t)]+\mathbf{S}[t]+\mathbf{S}[1]=(\frac{-1}{B}?\mathbf{S}[\varphi(t)])?(\frac{-1}{B}\mathbf{S}[\varphi(0)]) \quad (92)$$

Based on the Eq. (59), we have

$$\begin{cases} -\mathbf{S}[\underline{\phi}(t,r)]+\mathbf{S}[t]+\mathbf{S}[1]=\frac{1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,r) \\ -\mathbf{S}[\overline{\phi}(t,r)]+\mathbf{S}[t]+\mathbf{S}[1]=\frac{1}{B}\mathbf{S}[\overline{\phi}(t,r)]-\frac{1}{B}\overline{\phi}(0,r) \end{cases} \quad (93)$$

Therefore, the Z-number solution of Eq. (93) is extracted as

$$\begin{cases} \mathbf{S}[\underline{\phi}(t,r)]=\mathbf{S}[t]+\mathbf{S}[1]+\frac{-1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,r) \\ \mathbf{S}[\overline{\phi}(t,r)]=\mathbf{S}[t]+\mathbf{S}[1]+\frac{1}{B}\mathbf{S}[\overline{\phi}(t,r)]-\frac{1}{B}\overline{\phi}(0,r) \end{cases} \quad (94)$$

hence,

$$\begin{cases} \mathbf{S}[\underline{\phi}(t,r)]=(\frac{1}{B+1})\underline{\phi}(0,r)+B \\ \mathbf{S}[\overline{\phi}(t,r)]=(\frac{1}{B+1})\overline{\phi}(t,r)+B \end{cases} \quad (95)$$

By utilizing the inverse Sumudu transform for Z-numbers, we obtain

$$\begin{cases} \underline{\phi}(t,r)=\underline{\phi}(0,r)\mathbf{S}^{-1}(\frac{1}{B+1})+\mathbf{S}^{-1}(B) \\ \overline{\phi}(t,r)=\overline{\phi}(0,r)\mathbf{S}^{-1}(\frac{1}{B+1})+\mathbf{S}^{-1}(B) \end{cases} \quad (96)$$

Where $$\begin{cases} \underline{\phi}(t,r)=e^{-t}\underline{\phi}(0,r)+t \\ \overline{\phi}(t,r)=e^{-t}\overline{\phi}(0,r)+t \end{cases} \quad (97)$$

If the initial condition is taken to be a symmetric triangular Z-number as $\phi(0)=[(-a(1-r),a(1-r)),p(0.85,0.9,1)]$, , so

$$\begin{cases} \underline{\phi}(t,r)=e^{-t}(-a(1-r))+t \\ \overline{\phi}(t,r)=e^{-t}(a(1-r))+t \end{cases} \quad (98)$$

Approximation errors based on Z-numbers are shown in Table 3. These errors are the differences between the exact and the approximation solutions, for two different methods: FST and Max-Min Euler method [38]. Figure 7 shows the Corresponding solution plot based on fuzzy numbers.

Table 3. Approximation errors based on Z-numbers

α	FST	Max-Min Euler
0	[(0.0053,0.0128),p(0.8,0.85,0.9)]	[(0.0108,0.0195),p(0.8,0.9,1)]
0.2	[(0.0019,0.0127),p(0.7,0.81,0.9)]	[(0.0072,0.0138),p(0.6,0.7,0.8)]
0.4	[(0.0031,0.0089),p(0.8,0.9,1)]	[(0.0113,0.0123),p(0.7,0.8,0.91)]
0.6	[(0.0008,0.0027),p(0.8,0.9,1)]	[(0.0043,0.0078),p(0.7,0.8,0.9)]
0.8	[(0.0049,0.0079),p(0.7,0.8,0.86)]	[(0.0069,0.0099),p(0.6,0.7,0.8)]
1	[(0.0109,0.0109),p(0.7,0.8,0.9)]	[(0.0587,0.0587),p(0.6,0.75,0.8)]

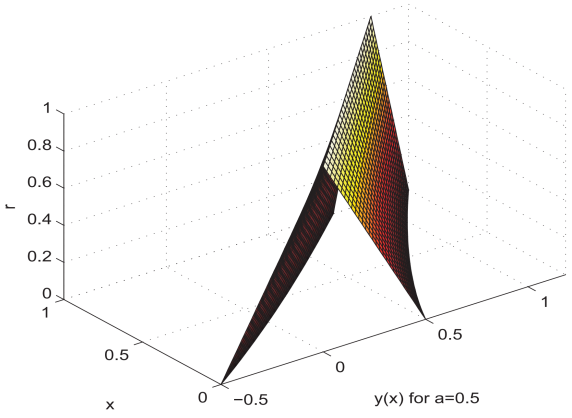


Fig. 7. The solution of FDE under case 2 consideration based on fuzzy numbers

6. Conclusion

In this paper, a novel method based on the FST is proposed in order to find the solution of the first order FDEs on the basis of the Z-numbers. The new method is clarified by utilizing the concept of strongly generalized differentiability. By using the FST method, the FDE converts to an algebraic problem. Some essential theorems are laid down in order to demonstrate the properties of the FST. Two real examples are applied to demonstrate the effectiveness of the proposed technique. This work has a significant contribution in initializing a superior starting point for such extensions.

References

[1] R. Jafari, S. Razvarz, "Solution of Fuzzy Differential Equations using Fuzzy Sumudu Transforms" IEEE International Conference on Innovations in Intelligent Systems and Applications, 84-89, 2017. DOI: 10.1109/INISTA.2017.8001137

[2] D. Zwillinger, "Handbook of differential equations" Gulf Professional Publishing. 1998.

[3] R. Jafari, W. Yu, "Fuzzy Control for Uncertainty Nonlinear Systems with Dual Fuzzy Equations" Journal of Intelligent and Fuzzy Systems. **29**(3), pp.1229-1240, 2015. DOI: 10.3233/IFS-151731

[4] R. Jafari, W. Yu, "Fuzzy Differential Equation for Nonlinear System Modeling with Bernstein Neural Networks" IEEE Access., 2017. doi:10.1109/ACCESS.2017.2647920

[5] R. Jafari, W. Yu, "Uncertainty Nonlinear Systems Modeling with Fuzzy Equations" Mathematical problems in Engineering. **2017**, 2017. doi:10.1155/2017/8594738

[6] S.L. Chang, L.A. Zadeh, "On fuzzy mapping and control" IEEE Trans. Syst. Man Cybern. **2**(1), 30-34, 1972. DOI: 10.1109/TSMC.1972.5408553

[7] D. Dubois, H. Prade, "Towards fuzzy differential calculus: part 3 differentiation" Fuzzy Sets Syst. **8**(3), 225-233, 1982. https://doi.org/10.1016/S0165-0114(82)80001-8

[8] P. Kloeden, "Remarks on Peano-like theorems for fuzzy differential equations" Fuzzy Set Syst. **44**(1), 161-164, 1991. https://doi.org/10.1016/0165-0114(91)90041-N

[9] B. Bede, I.J. Rudas, A.L. Bencsik "First order linear fuzzy differential equations under generalized differentiability" Inform. Sci. **177**(7), 1648-1662, 2007. https://doi.org/10.1016/j.ins.2006.08.021

[10] T. Allahviranloo, M.B. Ahmadi, "Fuzzy Laplace Transform" Soft Computing. **14**, 235-243, 2010.

[11] Z. Ding, M. Ma, A. Kandel, "Existence of the solutions of fuzzy differential equations with parameters" Inform Sci. **99**(3-4), 205-217, 1997. https://doi.org/10.1016/S0020-0255(96)00279-4

[12] V.A. Truong, V.H. Ngo, D.P. Nguyen, "Global existence of solutions for interval-valued integro-differential equations under generalized H-differentiability" Advances in Difference Equations. **1**, 217-233, 2013. https://doi.org/10.1186/1687-1847-2013-217

[13] T. Allahviranloo, N.A. Kiani, M. Barkhor- dari, "Toward the exiistence and uniqueness of solution of second-order fuzzy differential equations" Inform. Sci. **179**, 1207-1215, 2009. https://doi.org/10.1016/j.ins.2008.11.004

[14] F. Hawrra, K.H. Amal, "On fuzzy Laplace transform for fuzzy differential equations of the third order" Journal of Kerbala University, **11**(3), 251-256, 2013.

[15] L. Stefanini, B. Bede, "Generalized Hukuhara differentiability of interval-valued functions and interval differential equations" J. Nonlinear Anal. (TMA). **71**(3-4), 1311-1328, 2009. https://doi.org/10.1016/j.na.2008.12.005

[16] F.B.M. Belgacem, A.A. Karaballi, S.L. Kalla, "Analytical investigations of the Sumudu transform and applications to integral

production equations" Math. Prob. Eng. **103**(3), 103-118, 2003. http://dx.doi.org/10.1155/S1024123X03207018.

[17] Y. Liu, W. Chen, "A new iterational method for ordinary equations using Sumudu transform" Advances in Analysis, **1**(2), 89-94, 2016. DOI: 10.22606/aan.2016.12004

[18] H.M. Srivastava, A. Khalili Golmankhaneh, D. Baleanu, X.J Yang, "Local Fractional Sumudu Transform with Application to IVPs on Cantor Sets" Abstract and Applied Analysis. Article ID 620529, **2014**.1-7, 2014. http://dx.doi.org/10.1155/2014/620529, 2014.

[19] F.B.M. Belgacem, A.A. Karaballi, "Sumudu transform fundamental properties investigations and applications" J. Appl. Math. Stoch. Anal. **2006**, 1-23, 2006. http://dx.doi.org/10.1155/JAMSA/2006/91083

[20] A. Kiliçman, H. Eltayeb, R.P. Agarwal, "On Sumudu Transform and System of Differential Equations" abstr. Appl. Anal, **2010**, 1-11, 2010. doi:10.1155/2010/598702.

[21] N.A. Abdul Rahman, M.Z. Ahmad, "Fuzzy Sumudu transform for solving fuzzy partial differential equations" J. Nonlinear Sci. Appl. **9**, 3226-3239, 2016.

[22] Q.D. Katatbeh, F.B.M. Belgacem, "Applications of the Sumudu transform to fractional differential equations" Nonlinear Studies. **18**(1), 99-112, 2011.

[23] L.A. Zadeh, "Generalized theory of uncertainty (GTU)-principal concepts and ideas" Computational Statistics and Data Analysis, **51**(1), 15-46, 2006. https://doi.org/10.1016/j.csda.2006.04.029

[24] B. Kang, D. Wei, Y. Li, Y. Deng, "A method of converting Z-number to classical fuzzy number" Journal of Information and Computational Science, **9**(3), 703-709, 2012.

[25] L.A. Gardashova, "Application of operational approaches to solving decision making problem using Z-Numbers" Journal of Applied Mathematics, **5**(9), 1323-1334, 2014. DOI: 10.4236/am.2014.59125

[26] R.A. Aliev, A.V. Alizadeh, O.H. Huseynov, "The arithmetic of discrete Z-numbers" Inform. Sci. **290**, 134–155, 2015. https://doi.org/10.1016/j.ins.2014.08.024

[27] B. Kang, D. Wei, Y. Li, Y. Deng, "Decision making using Z-Numbers under uncertain environment" Journal of Computational Information Systems, **8**(7), pp.2807-2814, 2012.

[28] A. Jafarian, R. Jafari, M. Mohamed Al Qurashi, D. Baleanud, "A novel computational approach to approximate fuzzy interpolation polynomials" Springer Plus, **5**, pp.14-28, 2016. https://doi.org/10.1186/s40064-016-3077-5

[29] R. Jafari, W. Yu, X. Li, "Numerical Solution of Fuzzy Equations with Z-numbers using Neural Networks" Intelligent Automation and Soft Computing, 2017. doi: 10.1080/10798587.2017.1327154.

[30] M.L. Puri, D. Ralescu, "Fuzzy random variables" J. Math. Analysis Appl. **114**(2), 409-422, 1986. https://doi.org/10.1016/0022-247X(86)90093-4

[31] H.-C. Wu, "The improper fuzzy Riemann integral and its numerical integration" Inform. Sci. **111**(1-4), 109-137, 1999. https://doi.org/10.1016/S0020-0255(98)00016-4

[32] B. Bede, S.G. Gal, "Generalizations of the differentiability of fuzzy-number-valued functions with applications to fuzzy differential equations" Fuzzy Set Syst. **151**(3), 581-599, 2005. https://doi.org/10.1016/j.fss.2004.08.001

[33] Y. Chalco-Cano, H. Roman-Flores, "On new solutions of fuzzy differential equations" Chaos Solitons Fractals. **38**(1), 112-119, 2006. https://doi.org/10.1016/j.chaos.2006.10.043

[34] L.A. Zadeh, "Toward a generalized theory of uncertainty (GTU) an outline" Inform. Sci. **172** (1-2), 1-40, 2005. https://doi.org/10.1016/j.ins.2005.01.017

[35] R.A. Aliev, W. Pedryczb, V. Kreinovich, O.H. Huseynov, "The general theory of decisions" Inform. Sci. **327**(10), 125-148, 2016. https://doi.org/10.1016/j.ins.2015.07.055

[36] R.A. Aliev, O.H. Huseynov, R.R. Aliyev, A.V. Alizadeh, "The arithmetic of Z-numbers" Theory and applications World Scientific, Singapore, 2015.

[37] R.H. Pletcher, J.C. Tannehill, D. Anderson, "Computational Fluid Mechanics and Heat Transfer", Taylor and Francis, 1997.

[38] S. Tapaswini, S. Chakraverty, "Euler-based new solution method for fuzzy initial value problems" Int. J. Artificial. Intell. Soft. Comput. **4**(1), 58-79, 2014. https://doi.org/10.1504/IJAISC.2014.059288

[39] S. Razvarz, M. Tahmasbi, "Fuzzy equations and Z-numbers for nonlinear systems control", Procedia Computer Science, **120**, 923-930, 2017. https://doi.org/10.1016/j.procs.2017.11.327

[40] R. Jafari, W. Yu, "Uncertain nonlinear system control with fuzzy differential equations and Z-numbers", 18th IEEE International Conference on Industrial Technology, Canada, 890-895, doi:10.1109/ICIT.2017.7915477, 2017.

[41] V.L. Streeter, E.B. Wylie, K.W. Bedford, "Fluid mechanics" New York: McGraw Hill, 1998.

[42] R. Jafari, W. Yu, X. Li, S. Razvarz, "Numerical Solution of Fuzzy Di□erential Equations with Z-numbers Using Bernstein Neural Networks" International Journal of Computational Intelligence Systems, **10**, 1226-1237, 2017. doi:10.2991/ijcis.10.1.81

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 69-73 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

TPMTM: Topic Modeling over Papers' Abstract

Than Than Wai[*], *Sint Sint Aung*

Web Mining, University of Computer Studies, Mandalay, ZIP Code 05071, Myanmar

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 05 February, 2018
Online: 08 March, 2018

Keywords:
Frequent pattern mining
Latent Dirichlet Allocation
Entropy

ABSTRACT

Probabilities topic models are active research area in text mining, machine learning, information retrieval, etc. Most of the current statistical topic modeling methods, such as Probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). They are used to build models from unstructured text and produce a term-based representation to describe a topic by choosing single words from multinomial word distribution. There are two main weaknesses. First, popular or common words are different topics, often causing ambiguity for understanding the topics; Second, lack of consistent semantics for single words to be represented correctly. To address these problems, this paper proposes a model (A Two-Phase Method for Constructing Topic Model, TPMTM) that combines statistical modeling (LDA) with frequent pattern mining and produces better presentations of rich topics and semantics. Empirical evaluation shows that the results of the proposed model are better than LDA.

1. Introduction

Topic models are Bayesian statistical models that are structured in accordance with a hidden theme, usually called unstructured data in a set of textual documents, topics with multiple distributions of words. Due to a collection of unstructured text documents, the topic model assumes that the collection of documents (corpus) has a certain number of hidden topics and that each document contains more than one topic in different sizes. Researchers have developed several topic models such as Latent Semantic Indexing (LSA) [1], Probability Latent Semantic Analysis (PLSA) [2] and Latent Dirichlet Allocation (LDA) [3]. Topic modeling automatically selects topics from the text and identifies topics over time [4], explores the connection between topics[5], supervised the topics [6], recommendation [7], and so on.

LDA or unsupervised generation probabilistic methods for modeling the document collection (corpus), is the most commonly used topic modeling method. The LDA, each document can be described as a probabilistic distribution for latent topics, and that the topic distribution of all documents is distributed a common Dirichlet prior. Within each topic in the LDA model is described as a probabilistic distribution over-represented as a probabilistic distribution of words and words distributions of topics distribute the same Dirichlet prior. Each latent topic in the LDA model is also distributions of topics distribute a common Dirichlet prior as well. A corpus D consists of M documents, with document d having N_d words ($d \in \{1,..., M\}$), LDA models D as stated in the following process[8].

LDA concludes the following generative process for each document w in a D corpus:

1. Choose N ~ Poisson (ξ).

2. Choose θ ~Dir (α).

3. For each of the N words w_n:

(a) Choose a topic z_n~ Multinomial (θ).

(b) Choose a word w_n from $p(w_n|z_n.\beta)$, a multinomial probability conditioned on the topic Z_n.

The discovered variables are words in documents although others are hidden variables (θ and ϕ) and hyperparameters (β and α).To provide hyperparameters and hidden variables, the probability of discovered data D and maximized as follows:

$$p\,(D|\alpha,\beta)= \prod_{d=1}^{M} \int p(\theta_d\,|\alpha)\Big(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(\omega_{dn}|z_{dn},\beta)\Big)d\theta_d \quad (1)$$

The LDA model has three levels of the representation. In corpus level, there are two parameters (α and β) that are involved

[*]Corresponding Author: Than Than Wai, Web mining, University of Computer Studies, Mandalay, ZIP Code 05071, Myanmar | Email: thanwai85@gmail.com

in the process of building a corpus. Document-level variables are the variables θ_d, which are sampled once per document. Finally, word-level variables $z_{d,n}$ and $w_{d,n}$ are the variables that are collected for each word in each document. The current statistical topic modeling techniques make multinomial distributions in words to represent topics in a given collection of texts. For example, Table 1 displays a sample of multinomial distributions used to describe the three themes of a scientific collection of publications.

Table 1: A Sample of Topic Presentation on AAAI Dataset

TopicId	Words
7	problem, result,order ,solver, present
12	algorithm,show, state,find,result
18	behavior ,system ,agent, develop ,result

From the above results in Table1, a sample of word distributions used to present three themes of a scientific paper collection of AAAI dataset. The term "result" is a general term and very general term in showing research papers in all different fields. The general words cause ambiguous to the topic presentation. So, a new model is required to solve these problems. The new method should take higher special representations and explore latent associations under multinomial word distributions.

The LDA and other topic models are portions of the better field of probabilistic modeling. Generative probabilistic topic modeling is a method for unsupervised classification of documents, by modeling each document as a mix of themes and each theme as a mix of words. But there exist the problems of word uncertainty and semantic integrity [8].

Text mining is a technique that supports users' assets effective information from a variety of digital documents. Most text mining methods are keyword-based strategies that need single words to show the documents. Based on the theory that the phrase may have more linguistic meaning than the keyword, strategies for using phrases instead of keywords are also suggested. However, surveys have shown that phrase-based techniques are not always better than keyword-based techniques [9, 10]. Many strategies in the field of data mining are used in patterns to mine useful documents that yield encouraging results [11, 12].

Topic modeling provides a convenient way to analyze large classified text collections while extracting interesting features to express collections in text mining. Thus, it advances the proposed model to improve the accuracy and relevance of the topic's representations by using the techniques of text mining, especially frequent itemset mining techniques.

In this model, Latent Dirichlet Allocation (LDA) is integrated into data mining techniques and achieves successful accuracy for the collection documents (corpus). The proposed model is composed of two phases: 1) LDA is applied to accomplish first topic models and 2) the frequent itemset (pattern) mining method is applied to obtain further particular patterns to produce topics of the document collections. Furthermore, the frequent itemsets (patterns) often explain information about the structure of the relationship between words that provide topics that are understandable, relevant and broad.

2. Related Work

Probabilistic topic modeling is expanded to capture more interesting features [13], but they show topics through the distribution of multinomial words. The papers [14, 15] are a widespread way to express the linguistic meaning of the topics as mentioned in the introduction. The authors [16] show a way to calculate the similarities between given themes and a known hierarchy then select the most grant labels to show the topics. But, the weakness of existing methods is that they are strictly limited to resource candidates and are limited to linguistic coverage. The proposed model is a work extension originally presented at 16th IEEE/ACIS International Conference on Computer and Information Science, [17]. This paper [18] discusses the topic-related model phrase by Markov dependencies in word order based on LDA structure, related to this paper. The results provided on [19, 20, 21] show that the topics described by the phrases are easier to interpret than their LDA. But phrases may contribute low-level events in documents, which cannot be accomplished with efficient retrieval performance.

3. Phase1: Topic Presentation Propagation Using LDA

The Latent Dirichlet Allocation is an algorithm that automatically detects the themes that are present in the collection of documents. At the LDA, each document can be viewed as a mix of different topics. The LDA provides the topic using word distribution and representation of the document using the topic distribution. The description of the topic means which words are important to which the topic and representation of the document in which topics are important in the documents [22-26].

Let $D = \{d_1, d_2, ..., d_M\}$ be a corpus. Each document is considered as a mixture of themes and each theme can be defined as a distribution over frozen vocabulary words composed of documents using (1). In general, the proposed model has $\theta=\emptyset_1, \emptyset_2, \dots, \emptyset_V$ for all topics.

Example results of LDA, the topic presentation is shown in Table 2. At document level, table 3 shows LDA's example results, document representation and Table 4 also shows word-topic assignments results of LDA's example.

Table 2: Topic representations – probability distribution over words

Topic	θ
$\emptyset_7$	problem:$\frac{1}{3}$, result:$\frac{4}{12}$,order: $\frac{2}{12}$,solver$\frac{2}{12}$, present : $\frac{2}{12}$
$\emptyset_{12}$	algorithm$\frac{2}{15}$,show : $\frac{4}{15}$, state: $\frac{1}{3}$,find: $\frac{2}{15}$,result : $\frac{2}{15}$
$\emptyset_{18}$	behavior : $\frac{4}{15}$,system : $\frac{2}{15}$,agent : $\frac{1}{15}$, develop: $\frac{2}{15}$, result: $\frac{2}{15}$

LDA contributions are the representation of the topic that uses the word distribution that the words are important to what the topic matter is, and the representation of the document by distributing the topic which themes are important for a particular document. These representations are used for obtaining information, document classification, text mining, machine learning, etc.

Table 3: Document representation-probability distribution over topics

Document	z_7	z_{12}	z_{18}
d_7	0.385	0.123	0.108
d_{12}	0.464	0.118	0.073
d_{18}	0.305	0.11	0.098

Then, specified topics also indicate which words are important in which topics, similar to the representation on the subject. The proposed model performs word-topic emphasis on the LDA for more precise or more specific topic presentation for a given corpus.

4. Phase2: Topic Presentation Expansion

Words with high probability in topic distributions are selected to represent topics in most LDA based applications. For example, the top 5 words for the 3 topics, as shown in Table 2, are: problem, result, order ,solver, present for topic 7, algorithm, show, state, find, result for topic 12 and behavior ,system ,agent, develop ,result for topic 18. So, they are likely to represent the general concepts or common concepts of the three topics and can not describe the three topics that are noteworthy. Furthermore, the words in representations on the topics formed by the LDA are individual single words. Single words provide very little information about relationships between words and very limited language definitions to make the topic matter clear.

In this section, we propose a method based on frequent patterns (itemset) mining techniques, detailed in the following sub-sections, aimed at reducing the above-mentioned problems.

4.1. Frequent Itemsets based on LDA

The frequent itemset mining is the method of mining data in a set of items or words in large data sets. These patterns are usually described in various forms such as frequent itemsets, sequential patterns, or substructures.

In this paper, the proposed model uses frequent itemsets. Typical itemset generally indicates that a set of items often happens together in a transaction dataset. The methods of using frequent itemsets are categorized into three basic skills: horizontal data format, vertical data format, and expected database strategy.

We use the vertical data format in the mine frequent itemsets in this paper and believe that frequent pattern mining based representation can be more meaningful and more accurate to describe topics. In addition, frequent pattern based representations contain structural information that shows the relationship between the words.

Create Transactional Dataset: The aim of the proposed frequent pattern-based method is to discover related words (i.e., frequent itemsets) from the words assigned by LDA to the topics. From this purpose, we build a set of words from each word-topic assignment instead of using the order of words, because, for frequent pattern mining, the frequency of a word within a transaction is less important. A topical document transaction is a set of words without any duplicates. Let $D=\{d_1,d_2,\ldots,d_M\}$be the original document collection, the transactional dataset for topic Z_j is defined as T_j .

For the topics in D, we can develop V transactional datasets. An example of the transactional datasets is described in Table 5, made from the example in Table 4.

Generate Frequent Pattern-based Representation: The basic idea of the proposed method is to reduce frequent itemsets generated from each transactional dataset T_j to represent Z_j. Here the frequent items in each transactional dataset are taken. Then documents associated with each item are converted into the vertical format and a number of documents containing each item are counted. Items are sorted according to their number of hits (i.e. number of documents of each item). Removing noise items that the number of hits is greater than the user provided threshold. If document-set is a subset of another documents-sets in each transactional dataset then items are merged into enhanced frequent itemsets. For example, $w_1{:}d_1d_2{:}3$, $w_2{:}d_1d_3{:}2$, $w_3{:}d_1d_4{:}3$, $w_1{:}d_2d_3{:}2$, $w_8{:}d_1d_2d_3{:}2$, $w_7{:}d_1d_2d_4{:}3$ can be compressed as the enhanced frequent itemsets $w_1,w_2,w_8{:}d_1d_2d_3{:}2$, w_1,w_3, $w_7{:}d_1d_2d4{:}3$ (format - "items" : "document" : "topic (transactional dataset)"). Such meaningless itemsets may harm document filtering tasks using frequent itemsets because it has duplicates the same frequent itemsets. So, we can regard the proposed method as lossless compression because we can cover all the removed frequent itemsets with the exact topic and it can effectively filter out the redundant itemsets. Finally, frequent itemsets in each transactional dataset are reconstructed. ("itemset" and "pattern" are interchangeable in this thesis). Frequent itemsets are the most widely used patterns created from transactional datasets to illustrate useful or interesting patterns. The main idea of the proposed frequent itemset method is to use of frequent patterns generated from each transactional dataset to represent topic Zj. For a given minimal support threshold δ, and itemset p in T_j is frequent if supp (p) >= δ where supp(p) is the support of p where the number of transactions containing p. Take T_7 as an example, which is the transactional dataset for topic Z_7. For a minimal support threshold δ = 2, all the frequent itemsets generated from are given in Table 6. Patterns represent words related to specific and recognizable meanings.

Table 6: The Frequent Itemsets Explored from T_7

Frequent Patterns	minimal support threshold δ
<result>	3
<algorithm,show,state,find>	2

5. Evaluation

We made experiments to evaluate the performance of the proposed method. In this section, we show the results of the evaluation.

5.1. Datasets

Two datasets are used in experiments, containing abstracts of papers published in the AAAI from 2013 to 2014 and NSF from 1990 to 2003. The two datasets contain 548 and 129000 abstracts. The abstracts are obtained from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml) and by using Porter's stemmer in Java (http://www.tartarus.org/~martin/PorterStemmer).

Table 4: Word-topic Assignments

Document	z_7		z_{12}		z_{18}	
	Proportion of Topic	***Terms***	***Proportion of Topic***	***Terms***	***Proportion of Topic***	***Terms***
d_7	0.385	problem, result, order ,solver, present	0.123	algorithm ,state, find ,algorithm	0.108	behavior , system ,agent
d_{12}	0.464	algorithm,show, state,find,result	0.118	spars ,distribute ,find	0.073	class, complex
d_{18}	0.305	algorithm,show, state,find,result	0.11	algorithm, state	0.098	behavior, system ,system

Table 5: Transactional datasets

Document	$z_7(T_7)$		$z_{12}(T_{12})$		$z_{18}(T_{18})$	
	Proportion of Topic	***Terms***	***Proportion of Topic***	***Terms***	***Proportion of Topic***	***Terms***
d_7	0.385	problem, result, order ,solver, present	0.123	algorithm,state, find	0.108	behavior , system ,agent
d_{12}	0.464	algorithm,show, state,find,result	0.118	spars,distribute, find	0.073	class, complex
d_{18}	0.305	algorithm,show, state,find,result	0.11	algorithm,state	0.098	behavior, system

5.2. Settings

Firstly, we are preparing datasets from the UCI Machine Learning Repository and preprocessing of all documents by removing stop and stemming words.

Secondly, we apply the LDA model to construct a topic model with V = 20 topics for each data collection, using the MALLET topic modeling toolkit (http://mallet.cs.umass.edu//index.php). Our experiments show that an inadequate number of topics will mainly lead to abundant expanded patterns in the topic model. We run Gibbs sampling for 1000 iterations, the LDA hyperparameters are α = 50 / V and β = 0.01.

Thirdly, topical transaction datasets for optimizing topic representations are developed.

Finally, frequent itemsets based on topic representations using the proposed method presented in Section 4 are developed. We used 10% of the documents for testing purpose and trained the resting model at 90%.

5.3. Baseline Model

In order to compare the suggested method, the LDA chose the baseline model in the experiments. Examples of the results of the two models (the LDA model and the frequent itemset based on model) are shown in Table 7. The top 10 terms or frequently itemsets in each of the topic presentations produced by two models are shown in Table 7.

Table 7: Topic Presentations of All Models Using the AAAI Dataset

Topic8		**Topic4**	
LDA	***Frequent itemsets***	***LDA***	***Frequent itemsets***
change	algorithm,method	parallel	markov
system	gener	semant	agent
data	differ	complex	popular,demonstre
unsupervise	propos	algorithm	analyze,learn
input	mani	mathemat	design
compact	consid	condit	relationship,topic
benefit	effici	represent	tempor
superior	sever,train	tool	recommend,exploit
state	imag	regular	adapt
trust	target,problem	improve	time

5.4. Results

The objective of the proposed approach and other current topic modeling methodologies is to show the topics of a collection of documents as specific potential. For existing topic modeling methods and the proposed methodology, the topic representations are word distributions or patterns with probabilities. The more specific selected words or patterns are in the representation of the topic, the more precise representation of the topic matter becomes. The performance of the proposed method is evaluated by the use of information entropy. The higher the entropy, the more the proposed model is disordered.

Table 8: Comparison of All Models in Information Entropy Using a collection of documents of AAAI and NSF

Datasets	**Latent Dirichlet Allocation**	**Frequent Itemsets (patterns)**
AAAI	4.86159	2.68171
NSF	28.3747	20.7532

From the above results in Table 8, the suggested method based on frequent itemsets has lower entropy rates than the baseline model. Thus, it can make more exact presentations of the topics of a corpus.

6. Conclusion

This paper proposes a model to produce more discriminative and semantic rich representations for modeling topics in a given collection of documents. The main contribution of this paper is the novel approach of incorporating the pattern mining method and topic modeling method (Latent Dirichlet Allocation) to produce representations based on the pattern for modeling the topics. The test results show that representations based on patterns are more specific than representations developed by the Latent Dirichlet Allocation. In the future, we will examine the structure of the patterns and discover the relationship between words that represent topics at a more granular level.

References

[1] Blei, David M. "Probabilistic topic models." Communications of the ACM 55, no. 4 (2012): 77-84.

[2] Blei, David M., and D. Jon. "McAuliffe. supervised topic models." Advances in Neural Information Processing Systems 20 (2007): 121128.

[3] Blei, David M., and John D. Lafferty. "A correlated topic model of science." The Annals of Applied Statistics (2007): 17-35.

[4] Blei, David M., and John D. Lafferty. "Dynamic topic models." In Proceedings of the 23rd international conference on Machine learning, pp. 113-120. ACM, 2006.

[5] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[6] Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. "Indexing by latent semantic analysis." Journal of the American society for information science 41, no. 6 (1990): 391.

[7] Fürnkranz, Johannes. "A study using n-gram features for text categorization." Austrian Research Institute for Artifical Intelligence 3, no. 1998 (1998): 1-10.

[8] Gao, Yang, Yue Xu, Yuefeng Li, and Bin Liu. "A two-stage approach for generating topic models." In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 221-232. Springer, Berlin, Heidelberg, 2013.

[9] Gupta, Shivani D., and B. P. Vasgi. "Effective Pattern Discovery and Retrieving Relevant Document for Text Mining."

[10] Hofmann, Thomas. "Unsupervised learning by probabilistic latent semantic analysis." Machine learning 42, no. 1 (2001): 177-196.

[11] Lau, Jey Han, David Newman, Sarvnaz Karimi, and Timothy Baldwin. "Best topic word selection for topic labelling." In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 605-613. Association for Computational Linguistics, 2010.

[12] Lau, Jey Han, Karl Grieser, David Newman, and Timothy Baldwin. "Automatic labelling of topic models." In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 1536-1545. Association for Computational Linguistics, 2011.

[13] Magatti, Davide, Silvia Calegari, Davide Ciucci, and Fabio Stella. "Automatic labeling of topics." In Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on, pp. 1227-1232. IEEE, 2009.

[14] Moran, Kelly, Byron C. Wallace, and Carla E. Brodley. "Discovering Better AAAI Keywords via Clustering with Community-Sourced Constraints." In AAAI, pp. 1265-1271. 2014.

[15] Náther, Peter. "N-gram based Text Categorization." Lomonosov Moscow State Univ (2005).

[16] Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34, no. 1 (2002): 1-47.

[17] Wai, Than Than, and Sint Sint Aung. "Enhanced frequent itemsets based on topic modeling in information filtering." In Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on, pp. 155-160. IEEE, 2017.

[18] Wang, Chong, and David M. Blei. "Collaborative topic modeling for recommending scientific articles." In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 448-456. ACM, 2011.

[19] Wang, Xuerui, Andrew McCallum, and Xing Wei. "Topical n-grams: Phrase and topic discovery, with an application to information retrieval." In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, pp. 697-702. IEEE, 2007.

[20] Westergaard, D., Stærfeldt, H.H., Tønsberg, C., Jensen, L.J. and Brunak, S., 2017. "Text mining of 15 million full-text scientific articles." bioRxiv, p.162099.

[21] Wu, Sheng-Tang, Yuefeng Li, and Yue Xu. "Deploying approaches for pattern refinement in text mining." In Data Mining, 2006. ICDM'06. Sixth International Conference on, pp. 1157-1161. IEEE, 2006.

[22] Zeng, J., 2012. "A topic modeling toolbox using belief propagation." Journal of Machine Learning Research, 13(Jul), pp.2233-2236.

[23] Zhang, W., Ma, D. and Yao, W., 2014. "Medical Diagnosis Data Mining Based on Improved Apriori Algorithm." JNW, 9(5), pp.1339-1345.

[24] Zhao, Wayne Xin, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. "Topical keyphrase extraction from twitter." In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, pp. 379-388. Association for Computational Linguistics, 2011.

[25] Zhu, X., Ming, Z., Hao, Y. and Zhu, X., 2015, June. "Tackling Data Sparseness in Recommendation using Social Media based Topic Hierarchy Modeling." In IJCAI (pp. 2415-2423).

[26] Zhu, J., Wang, K., Wu, Y., Hu, Z. and Wang, H., 2016. "Mining User-Aware Rare Sequential Topic Patterns in Document Streams." IEEE Transactions on Knowledge and Data Engineering, 28(7), pp.1790-1804.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 70-75 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Selective Electron Beam Melting Manufacturing of Electrically Small Antennas

Saad Mufti[*, 1], Christopher Smith[2], Alan Tennant[1], Luke Seed[1]

[1]*Department of Electronic and Electrical Engineering, University of Sheffield, S10 2TN, UK*

[2]*Department of Materials Science and Engineering, University of Sheffield, S10 2TN, UK*

A R T I C L E I N F O

Article history:
Received : 01 September, 2017
Accepted: 03 November, 2017
Online: 08 November, 2017

Keywords :
3D Printing
Additive Manufacturing
Conformal Antennas
Electrically Small Antennas
Inverted-F Antennas

A B S T R A C T

Real estate pressures in modern electronics have resulted in the need for electrically small antennas, which have subsequently garnered interest amongst researchers and industry alike. These antennas are characterized by their largest dimensions translating to a fraction of the operating wavelength; such a diminutive size comes at the expense of reduced gain and efficiency, and a worse overall match to a corresponding power source. In order to compensate for this deterioration in performance, antenna designers must turn towards increasingly complex and voluminous geometries, well beyond the capabilities of traditional manufacturing techniques. We present voluminous metal antennas, based on a novel inverted-F design, and fabricated using the emergent selective electron beam melting manufacturing technique, a type of powder bed fusion process. As predicted by small antenna theory, simulation results presented show in increase in the antenna's efficiency as it is voluminously expanded into the third dimension. Measurement results illustrate that key trends observed from simulations are upheld; however, further understanding of the electromagnetic properties of raw materials, in particular how these change during the printing process, is needed. Nevertheless, this type of additive manufacturing technique is suitable for rapid prototyping of novel and complex antenna geometries, and is a promising avenue for further research and maturation.

1. Introduction

With the continued trend of miniaturization in the design and manufacture of modern, 'smart' electronics, antenna designers are challenged with delivering compact, low-cost, and easy to fabricate antennas, often operating at wavelengths many times their largest dimension. A further key requirement is that these antennas are made efficient, so as to prolong battery life. This paper is an extension of work originally presented in the 2017 International Workshop on Antenna Technology in Athens, Greece [1]. The discourse on additively manufactured (AM) voluminous antennas is expanded to provide an overview of electrically small antennas (ESA) in literature fabricated using novel techniques, the three-dimensional (3D) powder bed fusion (PBF) printing process, the antenna design process, as well as the implications of simulated and measured results on practical antenna design.

Formally, an antenna is classed as electrically small if the product of the wavenumber, k (= $2\pi/\lambda$), and the radius of an imaginary sphere circumscribing the maximum dimension of the antenna, a, is less than or equal to 0.5; note that the parameter ka is unit-less [2-4]. It is widely accepted that as the electrical size of an antenna is reduced, its performance (in terms of gain, efficiency, and bandwidth) deteriorates. Generally, the radiation resistance decreases, while the reactive component of its impedance increases, leading to a poor match with the feed line or network. As such, there is a compromise between miniaturization and performance; small antenna theory dictates that a favourable compromise is reached when the antenna fully occupies a volume defined by the radius a [2-8]. Consequently, there is tremendous potential in the successful manufacture of novel, 3D antenna geometries – hitherto too complex to fabricate with traditional processes – using emergent technologies such as holographic photolithography, 3D printing, direct-write printing, direct transfer patterning, thermal transfer printing, and aerosol jet printing [9-18].

[*]Saad Mufti, Department of Electronic and Electrical Engineering, University of Sheffield, S10 2TN, UK, Email: saadmuftism@sheffield.ac.uk

PBF-AM is a particularly attractive technology, as it has the capability to print with metallic raw materials, simplifying the production line compared to some of the other additive techniques cited above. Techniques such as stereolithography and aerosol jet printing entail first printing a plastic or ceramic component, and then subsequently coating it with a layer of metal. Direct-write techniques are serial in nature, and thus slower and unsuitable for mass manufacture. Patterning techniques are suspect to wear and tear of the mechanical stamping parts. All of these processes require a substrate for subsequent metal patterning, whereas no such limitation applies to PBF-AM.

Simulated results for radiation efficiency are presented for a type of planar inverted-F antenna (PIFA), introduced in [19], as it is voluminously extended into the 3rd dimension. The results are validated with three PBF-AM fabricated antennas: flat, part-spherical (~66% fill-factor), and full-spherical (~100% fill-factor), using the Wheeler cap method [20-22]. Radiation pattern measurements are provided for two full-spherical prototypes to compare the effects of fabrication tolerances. The main aims of this research are to present measurement results on pioneering 3D antennas fabricated using AM, and to assess the viability of this emergent technique, in its current state, as a possible fabrication method for inherently 3D antennas. In particular, two popular PBF-AM techniques – which make use of a laser and an electron beam as the heat source, respectively – are compared and the relative strengths and weaknesses assessed.

2. Background on Additive Manufacture

Additive manufacture, also known as 3D printing, is the umbrella term given to a range of processes where raw materials are formatively shaped, layer by layer, to form a desired object or part [23, 24]. A generic illustration of a PBF printer is provided in Figure 1; the technology can be used on a wide range of plastics and metals, and has the obvious advantages of speed, reduced waste, and greatly simplified production lines. In particular, PBF-AM is an attractive technology for antenna manufacture as it is one the few subcategories of 3D printing which can handle metallic raw materials. Compared to other AM techniques, PBF is also advantageous with respect to operating costs, albeit at the expense of higher power consumption and lower build volumes. Two common heat sources used in PBF are lasers and electron beams, leading to the respective processes termed as selective laser melting (SLM) and selective electron beam melting (SEBM).

For this research, the laser based Renishaw AM250 and the electron beam based Arcam A2 printers were used. The difference in heat source leads to some inherent differences in the printing process. In an SLM printer, the chamber is typically filled with inert gas in order to minimize oxidation and degradation of the powdered material. A pressurized vacuum environment is necessary in the SEBM printer, otherwise the bombarded electrons would interact with atoms in the gas and thus get deflected. The electron beam can also be moved a lot faster than the laser, leading to shorter build times. Energy delivered by the electron beam is, however, more diffuse than a laser, leading to a greater area affected by the heat. Consequently, the minimum feature size, resolution, and surface-finish of the SEBM process are typically worse off compared to SLM.

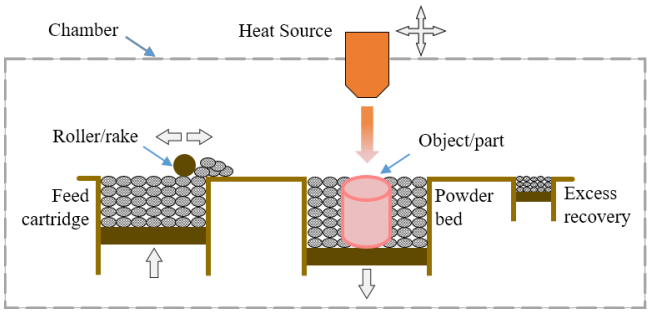


Figure 1. Illustration of generic PBF-AM printer.

In both processes, the powder in the build platform is maintained at an elevated temperature, typically just below its melting point and/or glass transition temperature; the build platform itself may be heated with the help of resistive heaters. For SLM, this elevated temperature allows for lowering the power rating of the laser, and prevents faults that might arise due to non-uniform thermal expansion and contraction. In the SEBM process, the powder is further heated to slightly sinter the particles prior to bombardment from the electron beam. This is done primarily to increase conductivity throughout the powder, to allow the negatively charged electrons to swiftly affect the desired object region. This eliminates the need for any sacrificial support structures under overhanging features, which are sometimes required in SLM printers.

3. Experimental Procedure

3.1. Antenna Design

Figure 2 shows the front profile view of the PIFA variant, based on the original design reported in [19]; the metallization is approximately 0.7 mm thick, resting on a hollow substrate with the same wall thickness, with the entire back side metallized to form the 'ground plane'. The centrally placed digitated structure contributes to a lowering of the resonance frequency of the antenna through additional capacitance and inductance. In fact, there are two current paths on this structure, contributing to orthogonally polarized modes, which may be tuned closer together with the relative positioning of the feed point and shorting pin, thereby increasing the effective bandwidth of the antenna. However, this was not a focus of the current research, as uncertainties in material properties and manufacturing tolerances meant that the focus was shifted purely to comparative efficiency measurements. Note than an extensive parametric study on antenna design and performance is presented in the reference [19], and is applicable to the voluminous antennas presented here.

This template is used for three variations of antennas with different voluminous expansions: flat, part-spherical, and full-spherical, as shown in Figure 3. For each antenna, the maximum radius, and parameter a, is 15.0 mm. Simulated results for the resonance frequency and efficiency as a function of the parameter h (as described in Figure 3) are provided in Figure 4 and Figure 5, respectively. It can be seen from Figure 5 that the 2nd resonance varies more as a function of the voluminous expansion, h, as compared to the 1st resonance, with the general trend that the respective resonance frequencies reduce as h increases, since the currents have a longer effective length on a more curved surface. The corresponding radiation efficiency at resonance also exhibits

a general trend, where the efficiency increases as h is increased and the antenna occupies more and more of the Chu sphere. For the 2nd resonance, the radiation efficiency of the flat antenna is 14%, increasing to 68% for the full-spherical antenna. Establishing an overall trend for this particular antenna geometry, at the dimensions specified, a unit (1 mm) change in height h results in an increase in radiation efficiency of 1.6 percentage points. Thus, the PIFA based antennas may be made more efficient, subsequently delivering power savings in appliances, at the cost of increased cross-sectional profile.

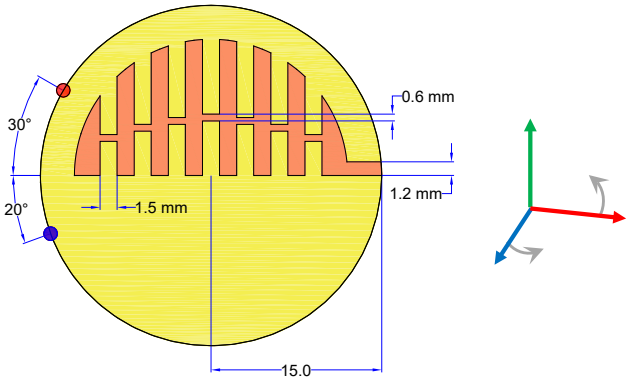


Figure 2. Schematic of top metallization for flat profile. Red dot is feed point; blue dot is shorting point. Back is fully metallicized.

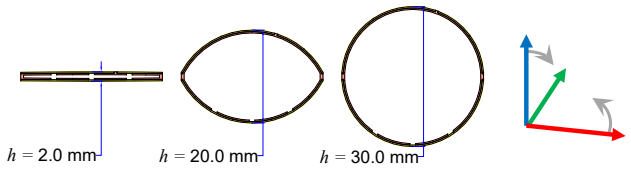


Figure 3. Cross-sectional view of CAD models simulated in CST Microwave Studio ®, for progressively greater voluminous expansion.

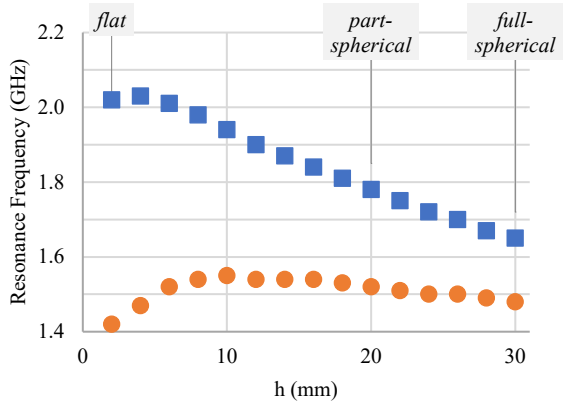


Figure 4. Simulated resonance frequency of (orange) 1st and (blue) 2nd resonances as a function of h. Fabricated antennas are annotated.

3.2. Fabricated Prototypes

The SLM process was trialed first, on a Renishaw AM250 printer, in order to take advantage of its superior surface finish. However, the process is much less capable of manufacturing overhanging features such as the centrally populated digits. Where this limitation is usually overcome with the use of sacrificial support structures analogous to scaffoldings used in building construction, the removal of such structures proved too damaging to the delicate digits.

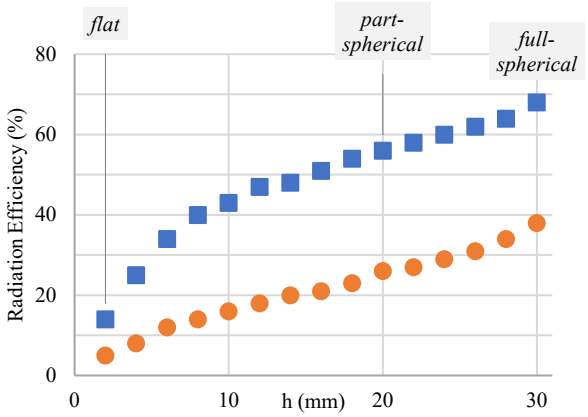


Figure 5. Simulated radiation efficiency of (orange) 1st and (blue) 2nd resonances as a function of h. Fabricated antennas are annotated.

A photograph of an antenna part fabricated with SLM is shown in Figure 6a, with the support structures partly removed. To alleviate this limitation, the antenna parts were subsequently fabricated using the SEBM based Arcam A2 printer. There is no need for support structures in this technique, as the metallic powder is pre-sintered before being melted by the electron beam. However, the poorer surface finish is evident from the photograph in Figure 6b. This is due to the use of coarser powder layers with the SEBM process (40-105 µm compared to 15-45 µm for the laser process), and the comparatively larger area affected at any one time by the electron beam compared to a heating laser.



(a) (b)

Figure 6. Photographs of metallic antenna parts fabricated using (a) SLM, and (b) SEBM.

In total, three variations of the same antenna design were fabricated: flat, part-spherical, and full-spherical; two of these are pictured in Figure 7. The antenna radiator and ground plane were fabricated separately using a Titanium alloy, Ti-6Al-4V ($\sigma \approx 10^5$ Sm^{-1}). Hollow plastic substrates fabricated from the plastic nylon12 were used to provide support. Copper tape was subsequently used to short the top metallization to the ground plane, and to provide contact pads for soldering Sub Miniature version A (SMA) connectors for interfacing to a power source. For this batch of antennas, a total of 18 parts were printed in just under two and a half hours; the build specifications for these antennas using the Arcam A2 printer are as follows:

- the printer was run under a vacuum of about 0.0001 mbar, with a slight amount of helium inserted into the chamber to help dissipate excess charge induced into the powder bed by the electron beam,
- a stainless-steel base plate was heated to 730 °C by the electron beam before any powder was deposited,
- for each layer of powder deposited (50 µm thick), the beam was first defocused and scanned across the layer at a high speed (30 ms^{-1}) with 2.3 kW power, for around 15 seconds, to sinter the powder,
- each part area was subsequently melted using a focused beam with a back and forth raster pattern at a speed of around 0.35 ms^{-1} with 300 W power,
- thereafter, successive powder layers were deposited and the previous sintering and melting steps repeated until the antenna parts were fully realized.



(a) (b)

Figure 7. Photographs of SEBM PBF-AM fabricated antenna prototypes, shown assembled; (a) flat, and (b) full-spherical.

4. Results and Discussion

Return loss measurements for the flat, part-spherical, and full-spherical antennas were conducted using the Agilent E5071B vector network analyser. Radiation efficiency (for 20 MHz bandwidth) was subsequently computed using the using the constant-loss-resistor Wheeler cap method described in [22]. Note that efficiency results were computed for the second resonance only as this is the dominant resonance from measurements, and that repeat measurements per prototype exhibit an error margin of ±10 %.

The corresponding simulations were carried out using CST Microwave Studio ®, using the frequency-domain solver. The titanium alloy was approximated using copper (Cu) metal properties in the CST material library, with reduced conductivity, σ. A notable limitation of using PBF-AM for manufacturing antenna parts is the relative lack of knowledge regarding the electromagnetic properties of the raw materials; this is in fact an active area of current research in the AM community [23, 25-27]. As such, these were iteratively refined in simulations. However, the general trends in performance observed can be interpreted as a validation of small antenna theory, as well as the usefulness of AM techniques in the swift prototyping of non-conventional antenna geometries.

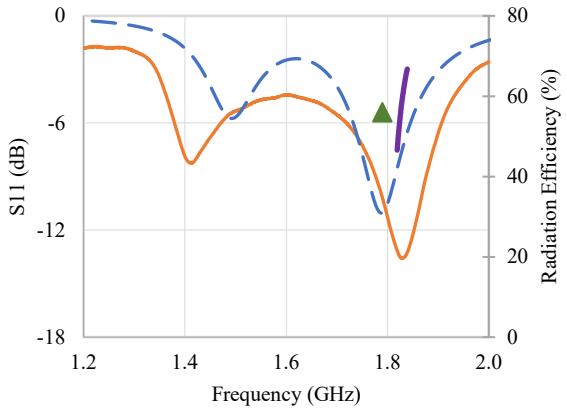


Figure 8. S_{11} results for PBF-AM part-spherical prototype; (orange) measured, (blue) simulated. Radiation efficiency at respective resonance frequencies, for the 2nd resonance; (purple) measured, (green) simulated.

Figure 9-11 compare the measured and simulated S_{11} parameter for the flat, part-spherical and two full-spherical prototypes, respectively. Note that S_{11} is the scattering parameter as obtained from a vector network analyser and is relatable to other reflectivity figures of merit such as return loss and reflection coefficient, and is typically presented in decibels (dB) [28]. Simulated efficiency at resonance and measured efficiency for 20 MHz bandwidth about the resonance frequency are also displayed. Despite differences in the absolute values for S_{11} and radiation efficiency, it can be seen that general trends are corroborated, as expected from small antenna theory. Computed radiation efficiency values for the 2nd resonances exhibit an increase as the voluminous expansion of the PIFA is increased.

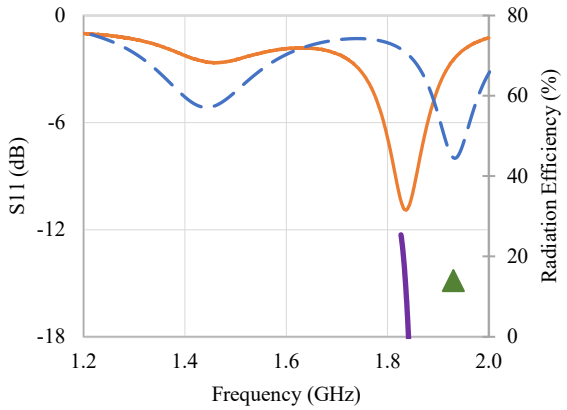


Figure 9. S_{11} results for PBF-AM flat prototype; (orange) measured, (blue) simulated. Radiation efficiency at respective resonance frequencies, for the 2nd resonance; (purple) measured, (green) simulated.

For the two full-spherical prototypes, despite the S_{11} curves following a similar curve, there is a difference of 50 MHz in the 2nd resonance. The mean radiation efficiency at resonance for the two prototypes is 68 %, with a discrepancy of 16 percentage points. Again, these differences are primarily due to inherent manufacturing differences in the antenna parts, as well as manual tolerances for the application of solder and copper tape contact pads. The differences in absolute values between simulations and measurements can be attributed to a number of factors, including

unknown changes in material properties during the printing process, the presence of solder and copper tape for electrical contact to the SMA connector, as well as the rough surface finish on the antenna parts.

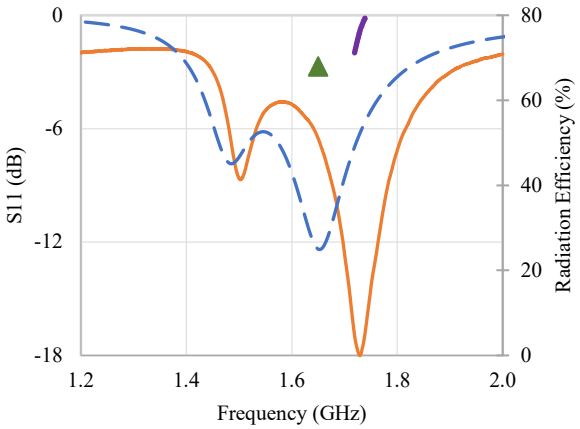


Figure 10. S_{11} results for PBF-AM full-spherical prototype-A; (orange) measured, (blue) simulated. Radiation efficiency at respective resonance frequencies, for the 2nd resonance; (purple) measured, (green) simulated.

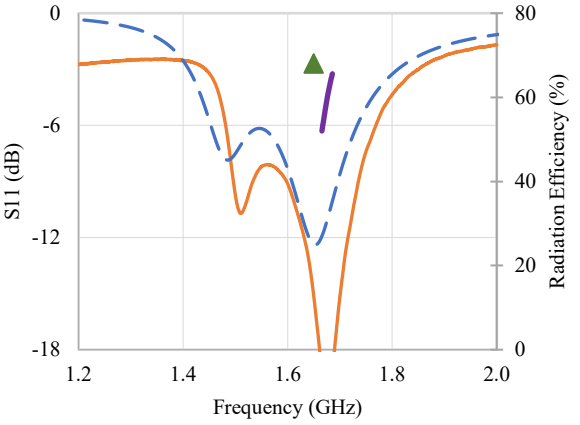


Figure 11. S_{11} results for PBF-AM full-spherical prototype-B; (orange) measured, (blue) simulated. Radiation efficiency at respective resonance frequencies, for the 2nd resonance; (purple) measured, (green) simulated.

Lastly, the realized gain for the full-spherical antenna prototypes was measured in an anechoic chamber, with an Agilent E5071C vector network analyser. Due to the particular feed location for this antenna, partial power gains were measured with respect to vertical and horizontal polarizations, and the gain transfer method [29] used subsequently to compute the total realized gain. Fig 6 and Fig. 7 plot the results for azimuth and elevation planes, respectively, measured at 1.7 GHz; the absolute gain is presented here in units of decibels-over-isotropic (dBi). It can be seen from these plots that simulation and measurement results largely agree, and are expectedly low directivity; there are some artifacts due to the setup (antenna holder) in the azimuth plane towards the rear of the antenna. Key results for the 2nd resonance of each antenna are summarized in Table 1.

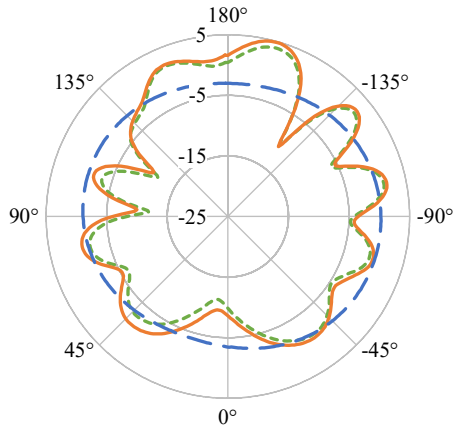


Figure 12. Realized gain (dBi) for the full-spherical ESA at 1.7 GHz; (orange) measured – sample A, (green) measured – sample B, (blue) simulated; azimuth (*z-x*) plane. Combined absolute gain (*φ and θ* components).

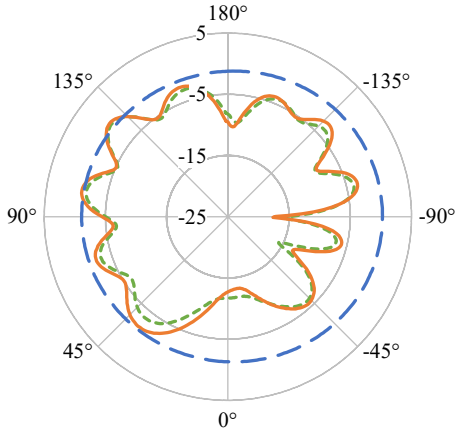


Figure 13. Realized gain (dBi) for the full-spherical ESA at 1.7 GHz; (orange) measured – sample A, (green) measured – sample B, (blue) simulated; elevation (*x-y*) plane. Combined absolute gain (*φ and θ* components).

Table 1. Simulated and measured resonant frequency and radiation efficiency at 2nd resonance for PBF-AM fabricated antennas. Measured and simulated gain at 1.7 GHz is provided for the two full-spherical antennas.

	Flat	*Part-spherical*	*Full-spherical (A)*	*Full-spherical (B)*
Simulated frequency (GHz)	1.93	1.79	1.65	-
Measured frequency (GHz)	1.84	1.83	1.73	1.68
Simulated efficiency (%)	14	56	68	-
Measured efficiency (%)	11	59	76	60
Simulated average gain; azimuth plane (dBi)	-	-	–1.7	-
Measured average gain; azimuth plane (dBi)	-	-	–2.2	–3.2
Simulated average gain; elevation plane (dBi)	-	-	–0.8	-
Measured average gain; elevation plane (dBi)	-	-	–5.9	–6.4

5. Conclusions and Future Work

This paper presents work done on voluminous, additive manufactured antennas based on a novel PIFA design. Measured results for return loss, radiation efficiency, and radiation pattern are presented for four prototypes manufactured using the SEBM based Arcam A2 printer. The metallization was realized from a Titanium alloy, Ti-6Al-4V, and hollow support substrates printed from nylon12. Copper tape was subsequently used to solder SMA connectors, and to short the top metallization to the respective antenna ground planes.

Radiation efficiency results corroborate the trends expected from simulations and small antenna theory. The mean radiation efficiency of 68% for the full-spherical antennas represents an improvement of 57 percentage points over their flat counterpart. The next stage of this project is to obtain an improved understanding of the material properties of metallic powders used in the printing process, as it has been demonstrated that the PBF-AM technique is viable for quick prototyping of complex and voluminous antenna geometries too difficult to realize using traditional fabrication techniques. Finally, the use of laser as the heat source would greatly improve the surface finish, but comes at the cost of limitations on the complexity of antenna geometry

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. Mufti, A. Tennant, L. Seed, C. J. Smith, "Efficiency measurements of additive manufactured electrically small antennas" in 2017 International Workshop on Antenna Technology: Small Antennas, Innovative Structures, and Applications, Athens, Greece, 2017. https://doi.org/10.1109/IWAT.2017.7915397

[2] S. R. Best, "Small and fractal antennas" in Modern Antenna Handbook, C. A. Balanis, Ed. John Wiley & Sons, 2008, pp. 475–528.

[3] J. L. Volakis, C. C. Chen, K. Fujimoto, Small Antennas: Miniaturization Techniques and Applications, McGraw-Hill, 2010.

[4] K. Fujimoto, H. Morishita, Modern Small Antennas, Cambridge University Press, 2014.

[5] L. J. Chu, "Physical limitations of omni-directional antennas" J. Appl. Phys., **19**(12), 1163–1175, 1948. https://doi.org/10.1063/1.1715038

[6] J. S. McLean, "A re-examination of the fundamental limits on the radiation Q of electrically small antennas" IEEE T. Antenn. Propag., **44**(5), 672–676, 1996. https://doi.org/10.1109/8.496253

[7] A. D. Yaghjian, S. R. Best, "Impedance, bandwidth, and Q of antennas" IEEE T. Antenn. Propag., **53**(4), 1298–1324, 2005. https://doi.org/10.1109/TAP.2005.844443

[8] C. Pfeiffer, "Fundamental Efficiency Limits for Small Metallic Antennas" IEEE T. Antenn. Propag., **65**(4), 1642–1650, 2017. https://doi.org/10.1109/TAP.2017.2670532

[9] A. Maiden, "Lithography in three dimensions using computer-generated holograms," Ph.D. Thesis, Durham University, 2005.

[10] D. E. Anagnostou, A. A. Gheethan, A. K. Amert, K. W. Whites, "A direct-write printed antenna on paper-based organic substrate for flexible displays and WLAN applications" J. Disp. Technol., **6**(11), 558–564, 2010. https://doi.org/10.1109/JDT.2010.2045474

[11] J. J. Adams *et al*., "Conformal printing of electrically small antennas on three-dimensional surfaces" Adv. Mater., **23**(11), 1335–1340, 2011. https://doi.org/10.1002/adma.201003734

[12] J. Hoerber, J. Glasschroeder, M. Pfeffer, J. Schilp, M. Zaeh, J. Franke, "Approaches for additive manufacturing of 3D electronic applications" Proc. CIRP, **17**, 806–811, 2014. https://doi.org/10.1016/j.procir.2014.01.090

[13] J. J. Toriz-Garcia *et al*., "Fabrication of a 3D electrically small antenna using holographic photolithography" J. Micromech. Microeng., **23**(5), no. 055010, 2013. https://doi.org/10.1088/0960-1317/23/5/055010

[14] C. Pfeiffer, X. Xu, S. R. Forrest, A. Grbic, "Direct transfer patterning of electrically small antennas onto three-dimensionally contoured substrates" Adv. Mater., **24**(9), 1166–1170, 2012. https://doi.org/10.1002/adma.201104290

[15] M. Kgwadi, C. J. Vourch, D. J. Harrison, T. D. Drysdale, "On-demand printing of antennas for TV white-space communications" in 2014 Loughborough Antennas and Propagation Conference, Loughborough, UK, 2014. https://doi.org/10.1109/LAPC.2014.6996448

[16] I. T. Nassar, T. M. Weller, "An electrically-small, 3-D cube antenna fabricated with additive manufacturing" in 2013 IEEE Radio and Wireless Symposium, Austin, USA, 2013. https://doi.org/10.1109/RWS.2013.6486708

[17] O. S. Kim, "3D printing electrically small spherical antennas" in 2013 IEEE Antennas and Propagation Society International Symposium, Orlando, USA, 2013. https://doi.org/10.1109/APS.2013.6711047

[18] H. Yuxiao, C. Oakley, P. Chahal, J. Albrecht, J. Papapolymerou, "Aerosol Jet printed 24 GHz end-fire quasi-Yagi-Uda antenna on a 3-D printed cavity substrate" in 2017 International Workshop on Antenna Technology: Small Antennas, Innovative Structures, and Applications, Athens, Greece, 2017. https://doi.org/10.1109/IWAT.2017.7915352

[19] S. Mufti, A. Tennant, L. Seed, "Electrically small modified planar inverted-F antenna" in 9th European Conference on Antennas and Propagation, Lisbon, Portugal, 2015.

[20] E. Newman, P. Bohley, C. Walter, "Two methods for the measurement of antenna efficiency" IEEE T. Antenn. Propag., **23**(4), 457–461, 1975. https://doi.org/10.1109/TAP.1975.1141114

[21] D. M. Pozar, B. Kaufman, "Comparison of three methods for the measurement of printed antenna efficiency" IEEE T. Antenn. Propag., vol. **36**(1), 136–139, 1988. https://doi.org/10.1109/8.1084

[22] D. Agahi, W. Domino, "Efficiency measurements of portable-handset antennas using the Wheeler cap" Applied Microwave and Wireless, **36**(1), 34–42, 2000.

[23] I. Gibson, D. W. Rosen, B. Stucker, Additive Manufacturing Technologies: Rapid Prototyping to Direct Digital Manufacturing, Springer, 2010.

[24] F. Calignano *et al*. "Overview on Additive Manufacturing Technologies" P. IEEE, **105**(4), 593–612, 2017. https://doi.org/10.1109/JPROC.2016.2625098

[25] Y. Huang, M. C. Leu, J. Mazumder, A. Donmez, "Additive manufacturing: current state, future potential, gaps and needs, and recommendations" J. Manuf. Sci. E-T ASME, **137**(1), no. 014001, 2015. https://doi.org/10.1115/1.4028725

[26] J. A. Slotwinski, E. J. Garboczi, "Metrology needs for metal additive manufacturing powders" JOM, **67**(3), 538–543, 2015. https://doi.org/10.1007/s11837-014-1290-7

[27] S. K. Everton, M. Hirsch, P. Stravroulakis, R. K. Leach, A. T. Clare, "Review of in-situ process monitoring and in-situ metrology for metal additive manufacturing" Mater. Design, **95**, 431–445, 2016. https://doi.org/10.1016/j.matdes.2016.01.099

[28] D. M. Pozar, Microwave Engineering, 4th ed, John Wiley & Sons, 2011.

[29] C. A. Balanis, Antenna Theory: Analysis and Design, 4th ed, John Wiley & Sons, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 74-84 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

MPC-based energy efficiency improvement in a pusher type billets reheating furnace

Silvia Maria Zanoli[*,1], Francesco Cocchioni[2], Crescenzo Pepe[2]

[1]*Università Politecnica delle Marche, Dipartimento di Ingegneria dell'Informazione, Ancona (AN), 60131, Italy*

[2]*i.Process S.r.l., Falconara Marittima (AN), 60015, Italy*

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 21 December, 2017
Online: 08 March, 2018

Keywords:
Model Predictive Control
Energy efficiency
Billets reheating furnace

ABSTRACT

The research reported in this paper proposes an Advanced Process Control system, denoted "i.Process | Steel – RHF", oriented to energy efficiency improvement in a pusher type billets reheating furnace located in an Italian steel plant. A tailored control method based on a two-layer Model Predictive Control strategy has been created that involves cooperating modules. Different types of linear models have been combined and an overall furnace global linear model has been developed and included in the controller formulation. The developed controller allows handling all furnace conditions, guaranteeing the fulfillment of the defined specifications. The reliability of the proposed approach has been tested through significant simulation scenarios. The controller has been installed on the considered real plant, replacing local standalone controllers manually conducted by plant operators. Very satisfactory field results have been achieved, both on process control and energy efficiency improvement. Optimized trade-offs between energy saving, environmental impact decreasing, product quality improvement and production maximization have been guaranteed. Consequently, Italian energy efficiency certificates have been obtained. The formulated steel industry reheating furnaces control method has been patented.

1. Introduction

Process industries were recently observing significant changes, due to the need to increase the automation features. This requirement is strictly related to the increasingly stringent environmental and pollution standards that are connected to the need to improve the energy efficiency level [1]. For this purposes, Advanced Process Control (APC) solutions are receiving the attention of process engineers [2]. Significant benefits can be obtained with respect to standalone controllers, both from process control and energy efficiency achievement and improvement point of view [3,4]. The magnitude of these results is directly tied to the amount of energy required for the considered processes.

Steel industry represents a high-energy intensive process industry: it is characterized by different complex phases that have to be carefully managed. Figure 1 briefly depicts the production chain of a steel industry [5]. Initially, raw materials (e.g. waste materials) are processed (Figure 1, *Raw Materials Processing Phase*) in order to obtain steel bars at an intermediate stage of manufacture, e.g. *billets*. Billets are then introduced in a reheating furnace in order to be suitably reheated (Figure 1, *Reheating Phase*). There are different typologies of reheating furnaces, distinguished by the furnace movement methodology: for example, in a *pusher type* reheating furnace, billets are moved along the furnace through the action of pushers. Billets can enter the furnace at different temperatures. In the case study that will be proposed in the present paper, i.e. a pusher type billets reheating furnace located in an Italian steel plant, the billets inlet temperature range is 20 [°*C*] – 700 [°*C*]. Through the combustion reactions triggered by some burners, i.e. air/fuel burners, billets are exposed to heating reactions during their path along the furnace. The billets must be reheated in order to fulfill the specifications required for the subsequent plastic deformation phase (Figure 1, *Rolling Phase*). In the proposed case study, a range example for billets discharge (furnace outlet) temperature is 1000 [°*C*] – 1100 [°*C*]. From the final phase, the finished products, e.g. iron rods or tube rounds, are obtained.

[*]Corresponding Author: Silvia Maria Zanoli, Università Politecnica delle Marche, Ancona (AN), 60131, Italy. Email: s.zanoli@univpm.it

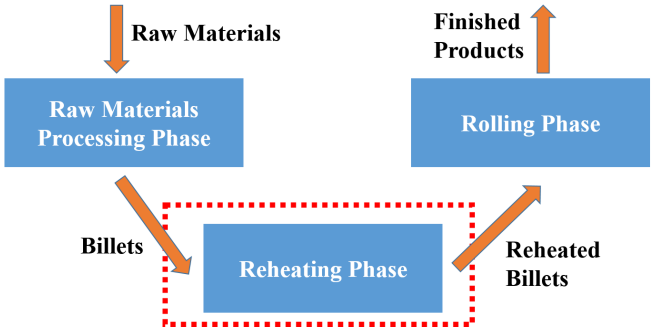


Figure 1. General production chain of a steel industry.

In a steel industry, the previously described *Reheating Phase* represents a crucial phase from an energy efficiency and products quality point of view [6]. In order to guarantee the desired trade-offs between the conflicting challenges of the steel industry *Reheating Phase*, i.e. energy saving and environmental impact decreasing versus product quality and production maximization, different approaches have been proposed in the control literature, ranging from classic to innovative techniques. In [7], a nonlinear optimization problem is formulated through a genetic algorithms approach; the minimization of fuel cost and the satisfaction of a desired discharge temperature represent the control objectives. In [8], two models and two related tracking control systems are proposed. In [9], a mixed neural network/heat transfer model approach is exploited, achieving an integrated intelligent control method. In [10], a nonlinear predictive approach is exploited for controlling a steel slabs continuous reheating furnace. A first principles mathematical model is developed that allows the controller to define the appropriate local furnace temperatures required for the achievement of the desired slabs discharge temperatures. The enormous potential of model-based control and optimization for steel reheating furnaces is described in [11], where one of the proposed control approaches is based on a transient nonlinear furnace model.

This paper is an extension of work originally presented in the 2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP) [12]. In [12], the authors have presented an APC framework for the optimization of a pusher type billets reheating furnace (*Reheating Phase*) located in an Italian steel plant. Some details about the process features, the controller formulation and field results have been provided, highlighting the achieved benefits with respect to the previous control system, based on local temperature controllers managed by plant operators. In this paper, in addition to the previously reported literature analysis, additional aspects about the process modellization are provided. Furthermore, more details on simulation results are given through specific scenarios. Finally, results related to the installation of the developed APC system, denoted "i.Process | Steel – RHF", on the Italian industrial plant are depicted and discussed, analyzing process control and energy efficiency aspects.

The paper organization is the following: Section 2 details the main process features, focusing on the control specifications and on the formulated process model. Details on the obtained relationships between the main process variables are provided. Section 3 depicts the control architecture, detailing the formulation of the two MPC modules. A significant plant scenario is reported in Section 4, where a typical condition is simulated and managed with the activation of the developed APC system. In Section 5, results related to the installation on the Italian steel industry are reported, focusing on process control performances and on energy efficiency evaluation with respect to the computed baseline. Section 6 reports the conclusions.

2. Process description and modellization

This section reports the description of the considered case study, i.e. a pusher type billets reheating furnace located in an Italian steel plant. Furthermore, the formulated process model is described.

2.1. Pusher Type Reheating Furnace

As previously described, a pusher type billets reheating furnace located in an Italian steel plant represents the case study that is proposed in this paper for the illustration of the main peculiarities of "i.Process | Steel – RHF" APC system. Figure 2 reports a synoptic of the developed Graphical User Interface (GUI) where a schematic representation of the considered reheating furnace is given. Note the three furnace areas: *Preheating Area* (green rectangle), *Heating Area* (yellow rectangle), *Soaking Area* (red rectangle). The main features of the furnace areas have been reported in Table 1. The furnace billets capacity is 136 (m_b=136); the billets are moved along the furnace based on the defined furnace production rate (up to 120 [*t/h*]). Billets enter the furnace through the *Preheating Area* (Figure 2, left side). In their path along this furnace area, they cross a unique zone (*tunnel*, 4.733 [*m*] length) that is not equipped with an own burners set. Subsequently, billets are moved towards the *Heating Area* that is constituted by zone 6 (3.477 [*m*] length), zone 5 (6.4 [*m*] length), and zone 4 (3.2 [*m*] length). All these furnace zones have air/fuel burners (Figure 2, yellow circles). Finally, billets enter the last furnace area, i.e. *Soaking Area* (Figure 2, right side). Three zones constitute this furnace area, characterized by an own burners set (Figure 2, red circles): zone 3 (4.546 [*m*] length), zone 2 (3.2 [*m*] length), and zone 1 (3.2 [*m*] length). Zone 2 and zone 1 are vertically disposed. Table 2 contains some billets features related disposed. Table 2 contains some billets features related to the considered case study.

Flowmeters detect air and fuel (natural gas) flow rates. The furnace zones temperatures are measured through thermocouples. Furnace and air pressures are measured by manometers placed near the furnace inlet. Billets temperature is measured by optical pyrometers only at the furnace inlet and after the billets furnace discharge (in the rolling mill area). Before the installation of "i.Process | Steel – RHF" APC system, no information was available about billets heating profile within the furnace. Plant operators managed the local Proportional Integral Derivative (PID) temperature controllers; they try to ensure the desired billets discharge temperature exploiting their experience and skills. However, due to the multivariable, nonlinear and time- varying characteristics of the considered process, the observed billets furnace discharge temperature was often higher than the minimum required temperature.

Through deepened process preliminary studies, significant energy efficiency margins (fuel specific consumption minimization) have been identified. For this reason, an APC

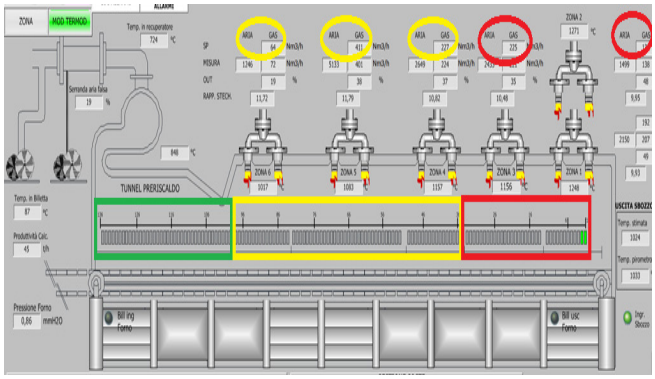


Figure 2. Pusher type reheating furnace synoptic.

Table 1: Details of furnace areas

Area	Billets Number / Area Length	Temperature	Acronym [*Units*]
Preheating	38 / 4.733 [m]	Tunnel Temp.	Tun [$°C$]
Heating	64 / 13.077 [m]	Zone 6 Temp.	$Temp_6$ [$°C$]
		Zone 5 Temp.	$Temp_5$ [$°C$]
		Zone 4 Temp.	$Temp_4$ [$°C$]
Soaking	34 / 7.746 [m]	Zone 3 Temp.	$Temp_3$ [$°C$]
		Zone 2 Temp.	$Temp_2$ [$°C$]
		Zone 1 Temp.	$Temp_1$ [$°C$]

Table 2: Billets details

Billets Feature	Case Study Billets Feature Value
Mass	2.2815 [t]
Length	9 [m]
Section	0.2 [m] × 0.16 [m]
Inlet Temperature	20 [$°C$] – 700 [$°C$]
Discharge Temperature	1000 [$°C$] – 1100 [$°C$]

system customized for the analyzed process has been developed. In particular, as it will be described in Section 3, a Model Predictive Control (MPC) strategy based on linear models has been formulated [13-15].

2.2. Process Modellization

In order to design an APC system, the availability of billets temperature estimations during their path within the furnace has been evaluated as a crucial milestone to satisfy. For this purpose, a virtual sensor has been developed [16]. The virtual sensor implements, for each billet that at each control instant lies in the furnace, a first principles adaptive nonlinear model. In this model, the involved heat phenomena and the billets movement information have been included. The inputs of the model are represented by the temperatures of the five furnace zones that are closer to the furnace inlet (tunnel, zone 6, zone 5, zone 4, and zone 3) and by the mean ($TempM_{21}$ [$°C$]) of the temperatures related to the two furnace zones that are closer to the furnace outlet (zone 2 and zone 1). The developed model has been based on the conduction model:

$$\dot{Q}_{cond} = -\lambda A \frac{dT}{dx} \quad [W] \tag{1}$$

In (1), A [m^2] represents the area related to the billet section that is normal to the heat transfer direction, λ [$W/(m \cdot K)$] is the billet thermal conductivity and dT/dx [K/m] is the temperature variation along the considered layer direction. The model reported in (1) can be customized with the needed number of billet layers. The convection and radiation phenomena have been considered through the following equations [17]:

$$\dot{Q}_{conv} = hA(T_{bill} - T_{env}) \quad [W] \tag{2}$$

$$\dot{Q}_{rad} = \varepsilon\sigma A(T_{bill}^4 - T_{env}^4) \quad [W] \tag{3}$$

In (2)-(3), A [m^2] represents the area related to the exposed surface, h [$W/(m^2 \cdot K)$] is the convection heat transfer coefficient, T_{bill} [K] is the billet temperature and T_{env} [K] is the environment temperature of the fluid around the billet. σ is the Stefan-Boltzmann constant and ε is the emissivity coefficient [17].

Through the equations reported in (1)-(3), a discretized model has been formulated for each billet that, at each considered sampling instant, lies in the reheating furnace. An important remark is the lack of an exact knowledge of the involved heat transfer coefficients. Online adaptation procedures for them have been included, based on customized constrained optimization problems that have been formulated exploiting feedback information. Figures 3-5 show an example of the virtual sensor field results related to October 2016. In Figure 3, blue stars indicate the measurements provided by the optical pyrometer in the rolling mill area; green stars represents the related virtual sensor temperature estimations. In Figures 4-5, trends related to the inputs of the virtual sensor model and related to the furnace production rate have been reported, respectively. Using the Root Mean Square Error of Prediction (RMSEP) as a performance indicator for the virtual sensor estimation, an RMSEP less than 10 [$°C$] has been detected (about 1 [$\%$] of an optical pyrometer measurement range). In Figure 4, examples of zone temperatures ranges can be observed. Note the increasing monotonicity of the temperatures according to the proximity to the furnace outlet.

The virtual sensor nonlinear model has then been linearized, in order to include billets temperatures estimations within an MPC strategy based on linear models. A Linear Parameter-Varying (LPV) model has been obtained [18]. The billets temperatures have been included in an ad hoc group of Controlled Variables (CVs), denoted as bCVs (b) group. In the considered case study, the bCVs group is composed by 136 elements (m_b=136). All terms involved in the inputs vector related to the bCVs model, i.e. the temperatures of all furnace zones, have been included in an own group, called as *zones* Controlled Variables (zCVs, y) group. This group also includes temperature differences between adjacent furnace zones, smoke-exchanger temperature (T_{SE}, [$°C$]), and fuel valves opening position [12]. As typical in industrial APC applications, other two categories of measured input process variables have been defined: Manipulated Variables (MVs, u) and Disturbance Variables (DVs, d). The MVs group includes the six fuel flow rates ($Fuel_i$ (i=1-6), [Nm^3/h]) and the six stoichiometric ratios (R_i (i=1-6), []) related to the furnace zones equipped with an own burners set. In DVs group,

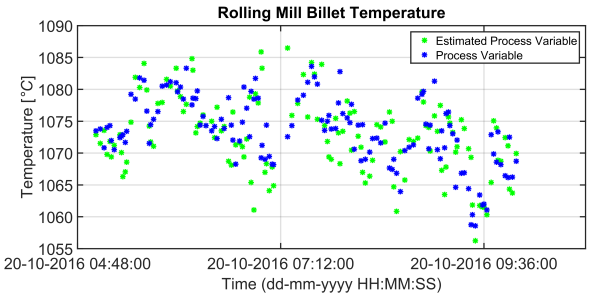


Figure 3. Virtual sensor results: rolling mill billets temperature.

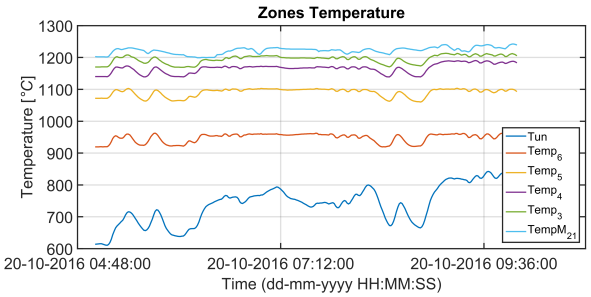


Figure 4. Virtual sensor results: inputs.

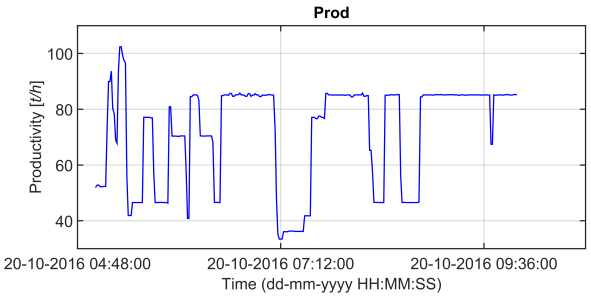


Figure 5. Virtual sensor results: furnace production rate.

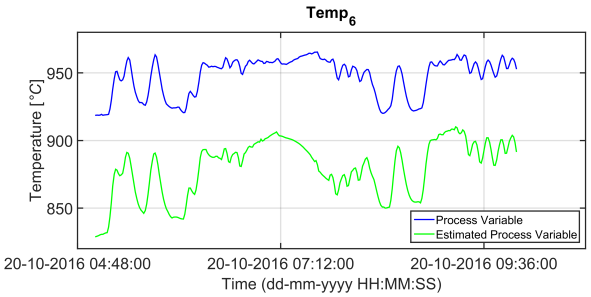


Figure 6. Zone 6 temperature model performances.

the furnace production rate (*Prod*, [*t/h*]), the furnace pressure (*FurnPress*, [*mmH2O*]) and the air pressure (*FurnPress*, [*mbar*]) have been included. The zCVs-MVs/DVs relationships have been modelled through asymptotically stable linear time invariant models without delays on the input-output channels. Steady-state gain signs related to some zCVs-MVs/DVs models have been reported in Tables 3-4. The empty boxes indicate the absence of a relationship.

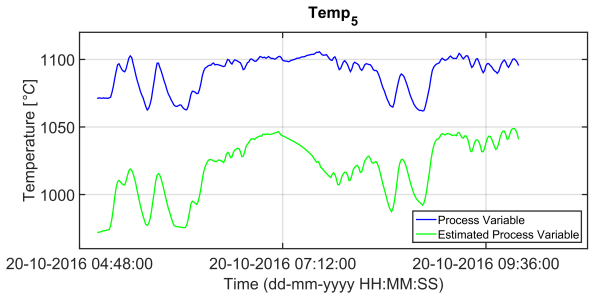


Figure 7. Zone 5 temperature model performances.

Table 3: zCVs-MVs models: steady-state gains signs

Acronym	*Fuel*$_6$	*Fuel*$_5$	*Fuel*$_4$	*Fuel*$_3$	*Fuel*$_2$	*Fuel*$_1$
Tun	+	+	+	+	+	+
Temp$_6$	+	+	+	+	+	+
Temp$_5$		+	+	+	+	+
Temp$_4$			+	+	+	+
Temp$_3$				+	+	+
Temp$_2$					+	+
Temp$_1$					+	+
T_{SE}	+	+	+	+	+	+

Table 4: zCVs-DVs models: steady-state gains signs

Acronym	*FurnPress*	*Prod*	*AirPress*
Tun	+	-	-
Temp$_6$	+	-	-
Temp$_5$	+	-	-
Temp$_4$	+	-	-
Temp$_3$	+	-	-
Temp$_2$	+	-	-
Temp$_1$	+	-	-
T_{SE}	+	-	-

Note, in Table 3, that the temperatures of the furnace zones located closer to the furnace inlet are influenced by the fuel flow rates of the downstream zones. Figures 6-7 report examples of the performances of the models related to zone 6 and zone 5 temperatures. Green lines represent the model prediction while blue lines represent the field process variables. The depicted trends refer to the same period taken into account in the scenario proposed in Figures 3-5.

3. "i.Process | Steel – RHF" APC system

In this section, "i.Process | Steel – RHF" APC system technology is described, focusing on the optimization problems solved by the two MPC modules and on tuning procedures.

3.1. APC Architecture Description

Figure 8 schematically describes "i.Process | Steel – RHF" APC system architecture. At each control instant *k*, a Supervisory Control and Data Acquisition (*SCADA*) system provides updated data; note the location of the developed billets temperature estimation virtual sensor. A *Data Conditioning & Decoupling Selector* (DC&DS) block receives updated process variables values (Figure 8, right side, *u(k-1), d(k), y(k), b(k)*). Furthermore, DC&DS block exploits other information, i.e. local control loops conditions (Figure 8, *Plant Signals & Parameters*), status information related to the selected process variables (Figure 8, right side, *u-d-y-b Status*), and control requirements related to MVs manipulation (Figure 8, right side, *Decoupling Matrix*) [19,20]. The status information related to the selected process variables defines which process variables have to be included in the MPC problem at each control instant. The *Decoupling Matrix* defines which MVs have to be moved by MPC strategy for the satisfaction of the specifications related to the zCVs. Table 5 shows the initial *Decoupling Matrix* provided to DC&DS block. For example, considering the zone 6 temperature, this zCVs is tied to all fuel flow rates (see Table 3); for this reason, without any expedients, all fuel flow rates could be moved by MPC system for the satisfaction of the zone 6 temperature specifications. According to some additional specifications that have been defined for the considered case study, only zone 6 fuel flow rate should be moved for the cited zCV. With regard to this aspect, observe the second row of Table 5 (related to zone 6 temperature): the "0" value indicates the MVs to be inhibited for its control [20].

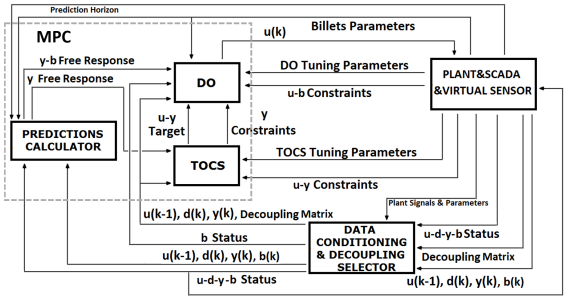


Figure 8. "i.Process | Steel – RHF" architecture.

Table 5: zCVs-MVs models: initial decoupling matrix

Acronym	$Fuel_6$	$Fuel_5$	$Fuel_4$	$Fuel_3$	$Fuel_2$	$Fuel_1$
Tun	1	1	0	0	0	0
$Temp_6$	1	0	0	0	0	0
$Temp_5$	1	1	0	0	0	0
$Temp_4$	1	1	1	0	0	0
$Temp_3$	1	1	1	1	0	0
$Temp_2$	1	1	1	1	1	0
$Temp_1$	1	1	1	1	0	1
T_{SE}	1	1	0	0	0	0

Based on the received information, DC&DS block performs different operations and checks, e.g. bad detection and data conditioning. From this phase, the possibly modified process variables values (Figure 8, left side, *u(k-1), d(k), y(k), b(k)*), the final status values for process variables (Figure 8, left side, *u-d-y-b Status*) and the final *Decoupling Matrix* (Figure 8, left side, *Decoupling Matrix*) resulted. The final *Decoupling Matrix* contains information about process variables final status values and about MVs action inhibition specifications [20].

MPC block, based on all the detailed parameters and information, computes the current optimal input to be applied to the plant (Figure 8, *u(k)*). *MPC* block has been based on a two-layer scheme formulated on linear models, constituted by an upper layer module called as Targets Optimizing and Constraints Softening (Figure 8, *TOCS*) and by a lower layer module called as Dynamic Optimizer (Figure 8, *DO*). A *Predictions Calculator* module cooperates with the two-layer scheme.

3.2. Two-Layer MPC Scheme

The MPC scheme exploits process variables predictions on a prediction horizon H_p. The zCVs and bCVs *free response* ([13]) is computed by *Predictions Calculator* module (Figure 8, *y Free Response* and *y-b Free Response*).

At the upper layer of the proposed MPC scheme, *TOCS* module solves a Linear Programming (LP) problem. A linear cost function is minimized, subject to linear constraints:

$$V_{TOCS}(k) = {c_u}^T \cdot \Delta\hat{u}_{TOCS}(k) + {\rho_{y_TOCS}}^T \cdot \varepsilon_{y_TOCS}(k) \qquad (4)$$

subject to

$$\begin{aligned} &\text{i. } lb_{du_TOCS} \le \Delta\hat{u}_{TOCS}(k) \le ub_{du_TOCS} \\ &\text{ii. } lb_{u_TOCS} \le \hat{u}_{TOCS}(k) \le ub_{u_TOCS} \\ &\text{iii. } lb_{y_TOCS} - \gamma_{lby_TOCS} \cdot \varepsilon_{y_TOCS}(k) \le \hat{y}_{TOCS}(k) \le \\ &\qquad \le ub_{y_TOCS} + \gamma_{uby_TOCS} \cdot \varepsilon_{y_TOCS}(k) \\ &\text{iv. } \varepsilon_{y_TOCS}(k) \ge 0 \end{aligned} \qquad (5)$$

In the LP problem represented by (4)-(5), minimization or maximization directions for MVs can be preferred through c_u term that multiplies the MVs steady-state move $\Delta\hat{u}_{TOCS}(k)$. $\Delta\hat{u}_{TOCS}(k)$ is constrained by lb_{du_TOCS} and ub_{du_TOCS}; with regard to MVs, the steady-state value $\hat{u}_{TOCS}(k)$ is constrained by lb_{u_TOCS} and ub_{u_TOCS}. zCVs steady state value $\hat{y}_{TOCS}(k)$ is constrained by lb_{y_TOCS} and ub_{y_TOCS}. *TOCS* MVs constraints have been considered as *hard* constraints: they can never be violated and their feasibility has been suitably imposed. On the other hand, *TOCS* zCVs constraints have been considered as *soft* constraints: they can be violated thanks to the slack variables contained in $\varepsilon_{y_TOCS}(k)$ term. This term contains two nonnegative slack variables for each zCV; it has been introduced in (4) through ρ_{y_TOCS} term and in (5) through γ_{lby_TOCS} and γ_{uby_TOCS} terms.

In Figure 8, lb_{du_TOCS}, ub_{du_TOCS}, lb_{u_TOCS}, ub_{u_TOCS} are among *u-y Constraints*. c_u, ρ_{y_TOCS}, γ_{lby_TOCS} and γ_{uby_TOCS} terms are among *TOCS Tuning Parameters*.

TOCS module formulation exploits zCVs-MVs/DVs models and predictions at the end of the prediction horizon H_p. Solving its LP problem, *TOCS* module computes MVs and CVs steady-state

targets (Figure 8, *u- y Target*) and zCVs constraints (Figure 8, *y Constraints*); these terms are provided to *DO* module [19,20].

At the lower layer of the proposed MPC scheme, *DO* module computes the H_u (H_u is denoted as control horizon [13]) optimal MVs moves, solving a Quadratic Programming (QP) problem. These moves are included in a $\Delta\hat{u}(k+M_i|k)$ vector ($i = 1,\dots,H_u$). M_i represent the MVs movement instant ($M_1 = 0$; $M_i < H_p$) [12]. A quadratic cost function is minimized, subject to linear constraints:

$$\begin{aligned} V_{DO}(k) = & \sum_{i=0}^{H_p-1} \|\hat{u}(k+i|k) - u_t(k+i|k)\|^2_{\mathcal{S}(i)} + \\ & + \sum_{i=1}^{H_p} \|\hat{y}(k+i|k) - y_t(k+i|k)\|^2_{Q(i)} + \\ & + \sum_{i=1}^{H_u} \|\Delta\hat{u}(k+M_i|k)\|^2_{\mathcal{R}(i)} + \left\|\varepsilon_y(k)\right\|^2_{\rho_y} + \\ & + \sum_{j=1}^{m_b} \left\|\hat{b}_j(k+e_j|k) - lb_{b_DO_j}\right\|^2_{T_j} + \|\varepsilon_b(k)\|^2_{\rho_b} \end{aligned} \quad (6)$$

subject to

$$\begin{aligned} & \text{i. } lb_{du_DO}(i) \le \Delta\hat{u}(k+M_i|k) \le ub_{du_DO}(i), i = 1,\dots,H_u \\ & \text{ii. } lb_{u_DO}(i) \le \hat{u}(k+M_i|k) \le ub_{u_DO}(i), i = 1,\dots,H_u \\ & \text{iii. } lb_{y_DO}(i) - \gamma_{lby_DO}(i)\cdot\varepsilon_y(k) \le \hat{y}(k+i|k) \le \\ & \qquad \le ub_{y_DO}(i) + \gamma_{uby_DO}(i)\cdot\varepsilon_y(k), i = 1,\dots,H_p \\ & \text{iv. } lb_{b_DO_j} - \gamma_{lbb_DO_j}\cdot\varepsilon_{b_j}(k) \le \hat{b}_j(k+e_j|k) \le \\ & \qquad \le ub_{b_DO_j} + \gamma_{ubb_DO_j}\cdot\varepsilon_{b_j}(k), j = 1,\dots,m_b \\ & \text{v. } \varepsilon_y(k) \ge 0;\ \varepsilon_b(k) \ge 0 \end{aligned} \quad (7)$$

In the QP problem represented by (6)-(7), tracking errors over H_p between MVs and zCVs reference trajectories ($u_t(k+i|k)$ and $y_t(k+i|k)$) and the related predicted values ($\hat{u}(k+i|k)$ and $\hat{y}(k+i|k)$) are weighted by $\mathcal{S}(i)$ and $Q(i)$ positive semidefinite matrices. $\hat{u}(k+i|k)$ and $\hat{y}(k+i|k)$ are constrained by lb_{u_DO}, ub_{u_DO}, lb_{y_DO} and ub_{y_DO}. *DO* MVs moves are weighted in (6) by positive definite matrices $\mathcal{R}(i)$ and they are constrained in (7) by lb_{du_DO} and ub_{du_DO}. *DO* MVs constraints have been considered as *hard* constraints: they can never be violated and their feasibility has been suitably imposed. On the other hand, *DO* zCVs constraints have been considered as *soft* constraints: they can be violated thanks to the slack variables contained in $\varepsilon_y(k)$ term. This term contains a nonnegative slack variable for each zCV; it has been introduced in (6) through ρ_y term and in (7) through γ_{lby_DO} and γ_{uby_DO} terms.

In Figure 8, lb_{du_DO}, ub_{du_DO}, lb_{u_DO}, ub_{u_DO} are among *u-b Constraints*. lb_{y_DO} and ub_{y_DO} terms are among *y Constraints*. $\mathcal{S}(i)$, $Q(i)$, $\mathcal{R}(i)$, ρ_y, γ_{lby_DO} and γ_{uby_DO} terms are among *DO Tuning Parameters*.

Taking into account the just described terms related to MVs and zCVs, and exploiting zCVs-MVs/DVs models, a first "i.Process | Steel – RHF" APC system control mode has been formulated, denoted *zones* APC mode.

Including terms related to bCVs in the *DO* QP problem represented by (6)-(7), a second control mode for "i.Process | Steel – RHF" APC system has been obtained, denoted *adaptive* APC mode. It constitutes the main control mode and it exploits, besides zCVs-MVs/DVs models, also first principles bCVs LPV model and billets virtual sensor information. In this way, an adaptive two-layer MPC strategy has been formulated. In (6)-(7), for the generic billet that lies on the j ($j = 1 \dots m_b$) furnace place at the current control instant k, its predicted furnace discharge instant is computed (e_j) taking into account the furnace production rate. The billets temperature predictions ($\hat{b}_j(k+e_j|k)$) at the related furnace discharge instants e_j ($j = 1 \dots m_b$) are constrained in (7) by $lb_{b_DO_j}$ and $ub_{b_DO_j}$ (Figure 8, *u-b Constraints*). These constraints have been considered as *soft* constraints: they can be violated thanks to the slack variables contained in $\varepsilon_b(k)$ term. This term contains a nonnegative slack variable for each bCV; it has been introduced in (6) through ρ_b term and in (7) through γ_{lbb_DO} and γ_{ubb_DO} terms (Figure 8, *DO Tuning Parameters*). Tracking option of desired values ($lb_{b_DO_j}$) for bCVs has been included in *DO* cost function (6), exploiting T_j nonnegative scalars.

3.3. Tuning Details

Tailored tuning methods have been developed for optimizing the controller performances in the two defined control modes. The control moves are computed by the APC system once a minute, according to the formulated process model.

With regard to the choice of the prediction horizon H_p, it varies based on the control mode that has to be exploited. Consequently, also the control horizon H_u and the MVs movement instants M_i are adapted. For example, in the simulation example that will be proposed in Section 4, the *adaptive* APC mode will be activated. In this case, a furnace movement time equal to about 95 [s] is assumed, which corresponds, for the present case study, to a furnace production rate equal to about 85 [t/h]. Accordingly, in order to guarantee the predicted reaching of the furnace outlet to the billet closer to the furnace inlet, a prediction horizon H_p of 216 [min] is set. In a parametric way, the control horizon H_u is set equal to 44 moves suitably spaced over the prediction horizon H_p.

In *TOCS* module formulation, the elements of c_u have been set as positive, in order to prefer, within the process variables defined constraints, minimization directions for fuel flow rates and stoichiometric ratios. Furthermore, ρ_{y_TOCS}, γ_{lby_TOCS} and γ_{uby_TOCS} terms have been set in order to guarantee the desired priority on constraints satisfaction. For example, constraints related to smoke-exchanger temperature are more important than those related to zones temperature: the related ρ_{y_TOCS},γ_{lby_TOCS} and γ_{uby_TOCS} terms have been set accordingly to this specification, taking into account also the magnitude of the involved process variables.

In *DO* module formulation, common tuning aspects between the two formulated control modes have been proposed. The priority of the soft constraints terms (if these constraints are included in the controller setup) ρ_y, γ_{lby_DO} and γ_{uby_DO} is guaranteed; furthermore, zCVs constraints are always present within *DO* module formulation. The option of tracking desired reference trajectories related to MVs is another similar tuning aspect, together with the need to take into account the magnitude of MVs moves. In *DO* module formulation related to the *adaptive* APC mode, billets (tracking and/or constraints satisfaction) specifications can be considered; optimal trade-offs between control and energy efficiency specifications must be ensured.

4. Simulation Results

This section proposes some simulation results related to "i.Process | Steel – RHF" APC system. In particular, a simulation scenario where the *adaptive* APC mode is activated is described.

4.1. A Simulated Scenario

The *adaptive* APC mode performances are shown through a simulated scenario: the zCVs-MVs/DVs plant model exploits the identified zCVs-MVs/DVs model and no measurement noise is assumed. The plant model for simulating the relationships between billets temperature and furnace zones temperature is based on the developed billets temperature nonlinear model that is exploited by the virtual sensor.

At the initial control instant of the proposed simulation, the *adaptive* APC mode is requested to be activated. The virtual sensor estimation gives reliable results and bCVs (billets temperature) can be included in the control problem. The zCVs reported in Table 1, the temperature differences between the bCVs model inputs and all fuel flow rates represent the other process variables that are considered in the simulation. The other MVs and all DVs are considered constant, so not influencing the proposed simulation scenario. Constraints related to zCVs and fuel flow rates have been reported in Tables 6-7. The temperature differences between the bCVs model inputs are constrained so as to ensure an increasing monotonicity of the temperatures along the furnace. As mentioned in Subsection 3.3, the furnace production rate is equal to about 85 [*t/h*]; the temperature of the 136 billets that initially lie within the furnace is in the range 20 [°*C*] - 1140 [°*C*], while the billets that will enter the furnace during the simulation are characterized by a temperature of about 550 [°*C*]. The *Rolling Phase* specifications require that the billets temperature in the rolling area must be in the range 1030 [°*C*] - 1045 [°*C*]. This specification can be converted in a temperature range of about 1051 [°*C*] - 1066 [°*C*] at the furnace outlet. The "i.Process | Steel – RHF" *adaptive* APC mode ensures that the billets temperature detected by the optical pyrometer in the rolling mill area converges towards the minimum required temperature (1030 [°*C*]), as can be observed in Figure 9.

Table 6: Simulation scenario: zCVs constraints

Acronym	Upper Constraint	Lower Constraint
Tun	950 [°*C*]	550 [°*C*]
$Temp_6$	1150 [°*C*]	800 [°*C*]
$Temp_5$	1150 [°*C*]	800 [°*C*]
$Temp_4$	1200 [°*C*]	800 [°*C*]
$Temp_3$	1250 [°*C*]	1000 [°*C*]
$Temp_2$	1250 [°*C*]	1000 [°*C*]
$Temp_1$	1250 [°*C*]	1000 [°*C*]

Table 7: Simulation scenario: MVs constraints

Acronym	Upper Constraint	Lower Constraint
$Fuel_6$	800 [Nm^3/h]	0 [Nm^3/h]
$Fuel_5$	1600 [Nm^3/h]	0 [Nm^3/h]
$Fuel_4$	650 [Nm^3/h]	0 [Nm^3/h]
$Fuel_3$	650 [Nm^3/h]	0 [Nm^3/h]
$Fuel_2$	250 [Nm^3/h]	0 [Nm^3/h]
$Fuel_1$	250 [Nm^3/h]	0 [Nm^3/h]

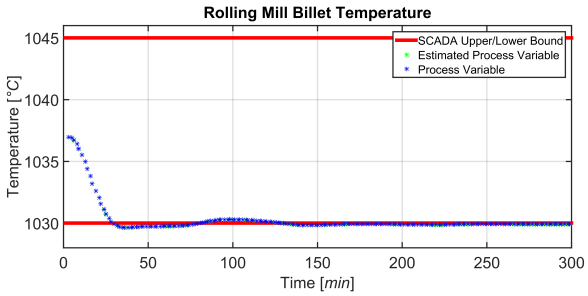


Figure 9. Simulation results: bCVs trends.

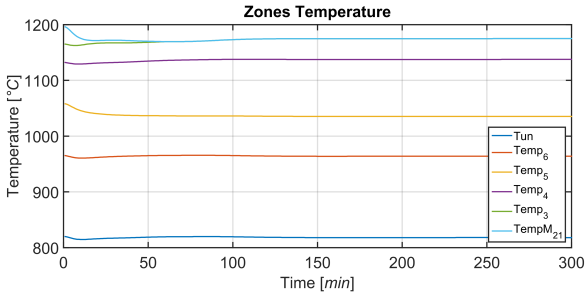


Figure 10. Simulation results: bCVs model inputs trends.

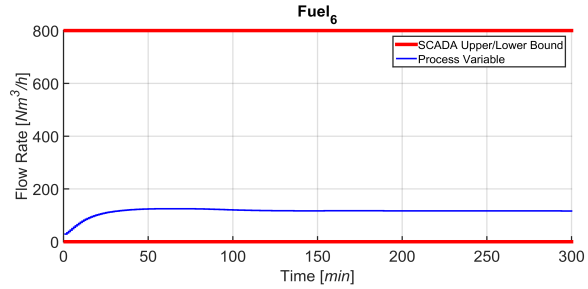


Figure 11. Simulation results: $Fuel_6$ trends.

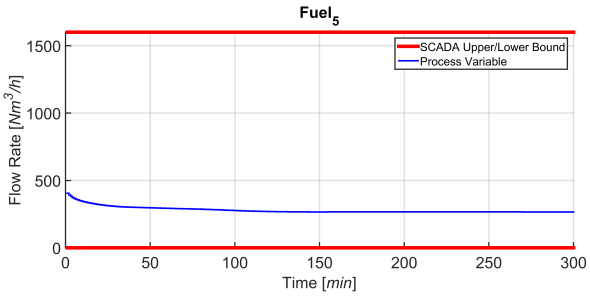


Figure 12. Simulation results: $Fuel_5$ trends.

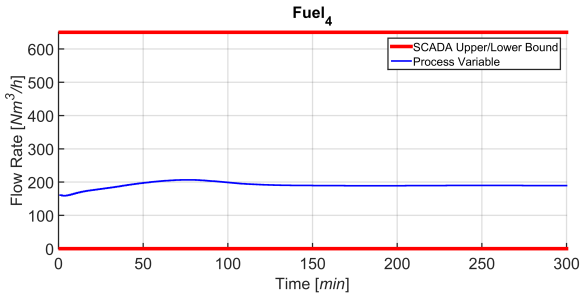


Figure 13. Simulation results: $Fuel_4$ trends.

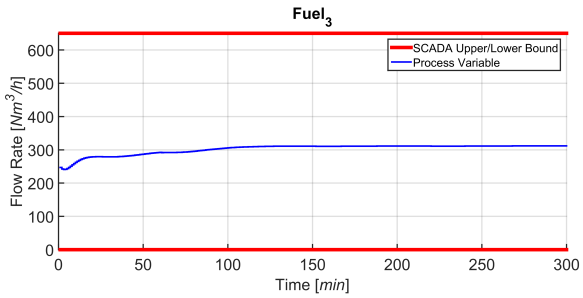


Figure 14. Simulation results: $Fuel_3$ trends.

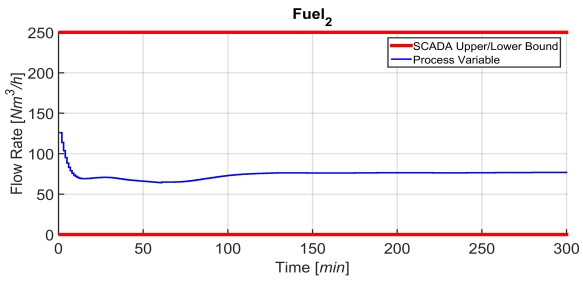


Figure 15. Simulation results: $Fuel_2$ trends.

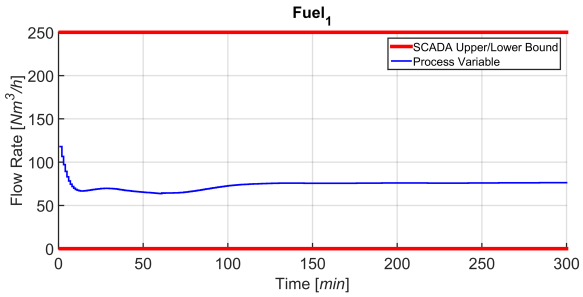


Figure 16. Simulation results: $Fuel_1$ trends.

The cooperative action between *TOCS* and *DO* modules, that exploits virtual sensor information and billets temperature LPV model ensures a coordinated management of the furnace zones temperature (Figure 10) that is directly tied to the manipulation of the fuel flow rates (Figures 11-16). Consequently, a more profitable plant configuration is guaranteed that, at the same time, respect all the defined specifications. For example, in Figure 10, note the increasing monotonicity of the temperatures along the furnace.

5. Field Results

The study and design phases of the project related to the considered process began in January 2015 and ended in May 2015. The APC system has been installed on the considered Italian steel plant in early June 2015, substituting the local PID temperature controllers managed by plant operators. This section shows a plant scenario under the control of "i.Process | Steel – RHF" APC system and some results about the obtained fuel specific consumption.

5.1. A Plant Scenario

Figures 17-26 show a field scenario where "i.Process | Steel – RHF" APC system is active. A four hours period is taken into account. Figure 17 shows the bCVs trends: virtual sensor estimation and optical pyrometer measurements are depicted, together with the defined constraints in the rolling mill area (1125 [°*C*] - 1065 [°*C*]). The furnace production rate is shown in Figure 18, while the billets furnace inlet temperature is shown in Figure 19. The inputs related to the bCVs model are depicted in Figure 20 and Figures 21-26 show the fuel flow rates. Note the fuel flow rates trends (Figures 21-26, blue line) and the defined constraints (Figures 21-26, red lines). All MVs and DVs are considered in the control problem, together with some zCVs (furnace zones temperatures, temperature differences between adjacent furnace zones, smoke-exchanger temperature) and the bCVs (some process variables have not been shown for brevity). Examples of the constraints defined for the furnace zones and smoke-exchanger temperatures have been reported in Table 8. The 136 billets that are already present in the furnace at the beginning of the considered plant scenario are characterized by temperatures in the range 55 [°*C*] – 1100 [°*C*] and the inlet temperature of the billets that will enter the furnace is in the range 40 [°*C*] – 105 [°*C*] (Figure 19). Besides the inlet temperature of the billets, also the furnace production rate is not constant (Figure 18): it assumes a

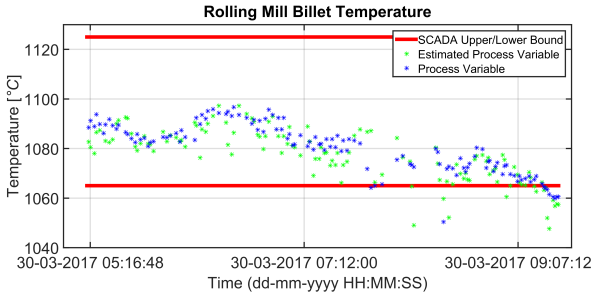


Figure 17. Field results: bCVs trends.

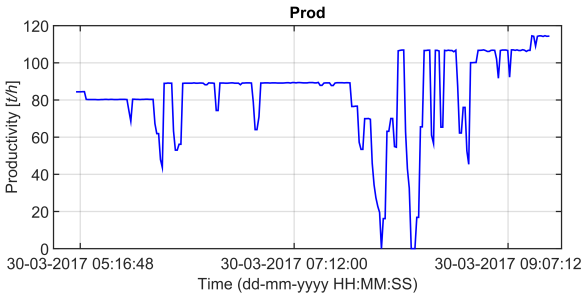


Figure 18. Field results: furnace production rate trends.

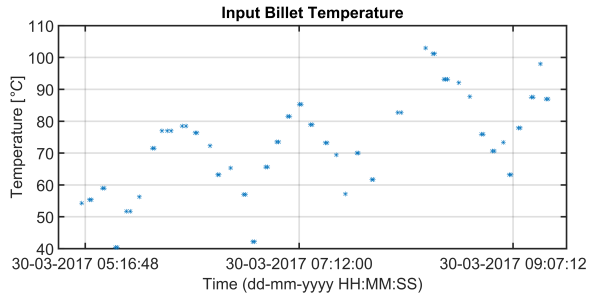


Figure 19. Field results: billets furnace inlet temperatures trends.

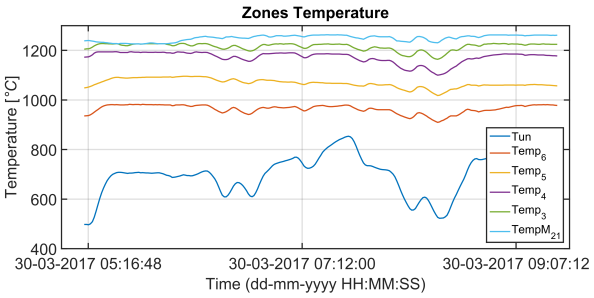


Figure 20. Field results: bCVs model inputs trends.

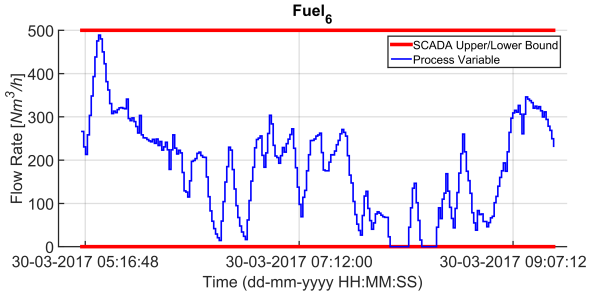


Figure 21. Field results: *Fuel*$_6$ trends.

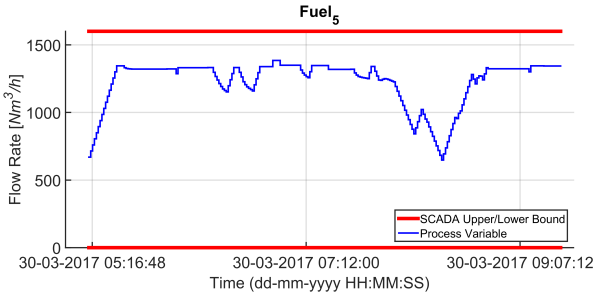


Figure 22. Field results: *Fuel*$_5$ trends.

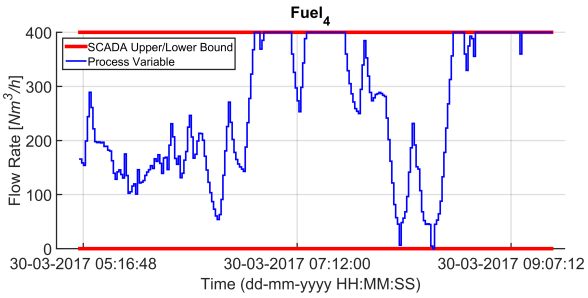


Figure 23. Field results: *Fuel*$_4$ trends.

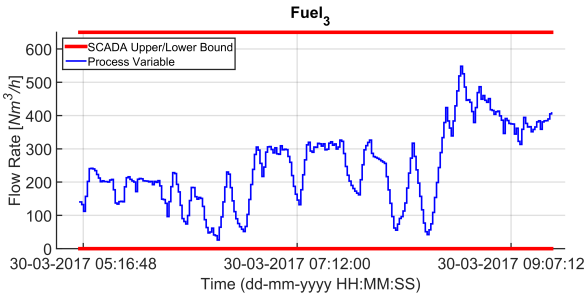


Figure 24. Field results: *Fuel*$_3$ trends.

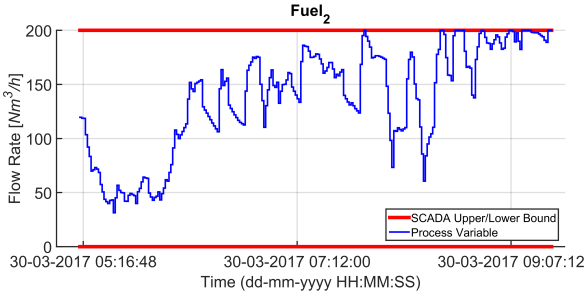


Figure 25. Field results: *Fuel*$_2$ trends.

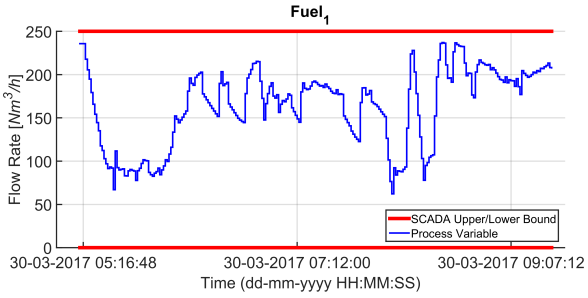


Figure 26. Field results: *Fuel*$_1$ trends.

maximum value that is equal to about 120 [*t/h*]. The furnace pressure and the air pressure (DVs) are about 0.9 [*mmH2O*] and 84 [*mbar*], respectively.

As can be observed in Figure 17, "i.Process | Steel – RHF" APC system ensures that the billets temperature detected by the optical pyrometer in the rolling mill area converges towards the minimum required temperature (1065 [°*C*]). The *TOCS-DO* cooperative action, exploiting virtual sensor information and billets temperature LPV model, ensures a profitable management of the furnace zones temperature (Figure 20) that is directly tied to the manipulation of the fuel flow rates (Figures 21-26). All the imposed constraints and specifications are fulfilled, despite a not constant furnace production rate (Figure 18) and a not constant billets furnace inlet temperature (Figure 19). The benefits deriving from the proposed multivariable predictive approach led the controller to conduct the plant to very profitable operating regions, but, at the same time, all the control specifications are satisfied. As it will be shown in the next subsection, an energy efficiency

Table 8: Field scenario: zCVs constraints

Acronym	Upper Constraint	Lower Constraint
Tun	950 [°C]	400 [°C]
$Temp_6$	980 [°C]	900 [°C]
$Temp_5$	1100 [°C]	1050 [°C]
$Temp_4$	1190 [°C]	1140 [°C]
$Temp_3$	1220 [°C]	1190 [°C]
$Temp_2$	1260 [°C]	1220 [°C]
$Temp_1$	1260 [°C]	1220 [°C]
T_{SE}	730 [°C]	300 [°C]

improvement has been obtained with the developed APC system, with respect to the previous furnace conduction.

5.2. Fuel Specific Consumption Results

The installation of the developed controller on the real industrial plant has guaranteed an improvement on the process control that has directly influenced the fuel specific consumption. The fuel specific consumption, that takes into account the fuel usage and the furnace production rate, represents a very significant indicator for the evaluation of the energy efficiency performances of "i.Process | Steel – RHF". A project baseline for the fuel specific consumption has been computed, that varies with the furnace hot charge.

Figure 27 shows a subpart of a synoptic of the developed GUI: the daily fuel specific consumption ([Sm^3/t]) is depicted. The "i.Process | Steel – RHF" APC system daily specific consumption

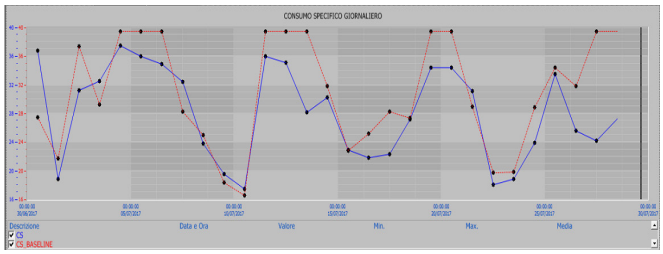


Figure 27. Field results: comparison between daily baseline and daily official specific consumption.

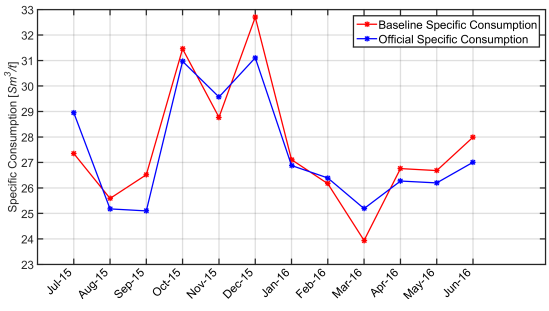


Figure 28. Field results: comparison between monthly baseline and monthly official specific consumption.

related to July 2017 is represented though a blue line, while the defined daily project baseline is shown through a red line. This page can be online monitored by plant operators; in this way, they can practically evaluate the controller performances from an energy efficiency point of view.

Figure 28 shows the monthly fuel specific consumption ([Sm^3/t]) related to the first year of "i.Process | Steel – RHF" APC system performances. The specific consumption is represented though a blue line, while the defined project baseline is shown through a red line. After about two years from the installation of "i.Process | Steel – RHF" APC system on the described pusher type billets reheating furnace, about 2 [%] reduction of the fuel specific consumption with respect to the defined project baseline has been achieved. A controller service factor about equal to 95 [%] has been observed.

6. Conclusions

In this paper, an Advanced Process Control system aimed at optimizing a pusher type billets reheating furnace located in an Italian steel plant has been proposed. The control system, denoted "i.Process | Steel – RHF", has been based on two-layer Model Predictive Control strategy formulated with linear models. The two-layer predictive controller also interacts with additional functional blocks.

Simulation and field results have demonstrated significant performances improvements guaranteed by the developed controller with respect to the previous control system based on Proportional Integral Derivative (PID) temperature controllers managed by plant operators. Thanks to the multivariable predictive approach, "i.Process | Steel – RHF" recognizes efficient operating zones and allows the process reaching them; in this way, process control and energy efficiency improvements have been guaranteed. Specifications related to the billets reheating are fulfilled and, at the same time, optimal configurations of the manipulated variables are reached. The developed control method has been patented [16].

After about two years from the installation of "i.Process | Steel – RHF" APC system on the described pusher type billets reheating furnace, about 2 [%] reduction of the fuel specific consumption with respect to the defined project baseline has been obtained, together with a controller service factor about equal to 95 [%].

References

[1] www.enea.it
[2] P. L. Latour, J. H. Sharpe, M. C. Delaney, "Estimating Benefits from Advanced Control" ISA Trans., **25**(4), 13–21, 1986.
[3] M. Bauer, I. K. Craig, "Economic assessment of advanced process control – A survey and framework" J. Proc. Contr., **18**(1), 2–18, 2008. https://doi.org/10.1016/j.jprocont.2007.05.007
[4] W. M. Canney, "Are you getting the full benefit from your advanced process control system?" Hydroc. Process., **84**(6), 55–58, 2005.
[5] W. Trinks, M. H. Mawhinney, R. A. Shannon, R. J. Reed, J. R. Garvey, Industrial Furnaces, John Wiley & Sons, 2004.
[6] A. Martensson, "Energy efficiency improvement by measurement and control: a case study of reheating furnaces in the steel industry" in 14th National Industrial Energy Technology Conference, Houston Texas USA, 1992. http://hdl.handle.net/1969.1/92210
[7] H. S. O. Santos, P. E. M. Almeida, R. T. N. Cardoso, "Fuel Costs Minimization on a Steel Billet Reheating Furnace Using Genetic Algorithms" Modelling and Simulation in Engineering, **2017**, Article ID 2731902, 2017. https://doi.org/10.1155/2017/2731902

[8] Z. Yi, Z. Su, G. Li, Q. Yang, W. Zhang, "Development of a double model slab tracking control system for the continuous reheating furnace" International Journal of Heat and Mass Transfer, **113**, 861–874, 2017. https://doi.org/10.1016/j.ijheatmasstransfer.2017.05.093

[9] Y. X. Liao, J. H. She, M. Wu, "Integrated Hybrid-PSO and Fuzzy-NN Decoupling Control for Temperature of Reheating Furnace" IEEE Trans. Ind. Electr., **56**(7), 2704–2714, 2009. https://doi.org/10.1109/TIE.2009.2019753

[10] A. Steinboeck, D. Wild, A. Kugi, "Nonlinear model predictive control of a continuous slab reheating furnace" Control Eng. Pract., **21**(4), 495–508, 2013. https://doi.org/10.1016/j.conengprac.2012.11.012

[11] A. Steinboeck, Model-based Control and Optimization of a Continuous Slab Reheating Furnace, Shaker Verlag GmbH, 2011.

[12] G. Astolfi, L. Barboni, F. Cocchioni, C. Pepe, S. M. Zanoli, "Optimization of a pusher type reheating furnace: an adaptive Model Predictive Control approach" in 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP), Taipei Taiwan, 2017. https://doi.org/10.1109/ADCONIP.2017.7983749

[13] J. Maciejowski, Predictive Control with Constraints, Prentice-Hall, 2002.

[14] E. F. Camacho, C. Bordons, Model Predictive Control, Springer-Verlag, 2007.

[15] J. B. Rawlings, D. Q. Mayne, Model Predictive Control: Theory and Design, Nob Hill Publishing, 2013.

[16] G. Astolfi, L. Barboni, F. Cocchioni, C. Pepe, "Metodo per il controllo di forni di riscaldo," Italian Patent n. 0001424136 awarded by Ufficio Italiano Brevetti e Marchi (UIBM), 2016. http://www.uibm.gov.it/uibm/dati/Avanzata.aspx?load=info_list_uno&id=2253683&table=Invention&

[17] Y. A. Cengel, Introduction to Thermodynamics and Heat Transfer, McGraw-Hill Companies, 2008.

[18] J. Mohammadpour, C. W. Scherer, Control of Linear Parameter Varying Systems with Applications, Springer, 2012.

[19] C. Pepe, "Model Predictive Control aimed at energy efficiency improvement in process industries," Ph.D. Thesis, Università Politecnica delle Marche, 2017.

[20] S. M. Zanoli, C. Pepe, "Two-Layer Linear MPC Approach Aimed at Walking Beam Billets Reheating Furnace Optimization" Journal of Control Science and Engineering, **2017**, Article ID 5401616, 2017. https://doi.org/10.1155/2017/5401616

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 76-81 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Design and Simulation of an RF-MEMS Switch and analysis of its Electromagnetic aspect in realtion to stress

Amna Riaz*,[1], Muhammad Umair Javed Ilam Sindhu[2], Tahir Zaidi[3]

[1]*University of Azad Jammu and Kashmir, Department of Electrical Engineering, Muzaffarabad, 13100, Pakistan*

[2]*Memorial University of Newfoundland, Department of Electrical and Computer Engineering, A1B 3X5, Canada*

[3]*National University of Science and Technology, Department of Electrical Engineering Department, EME College, 13100, Pakistan*

ARTICLE INFO

Article history:
Received: 31 May, 2017
Accepted: 05 September, 2017
Online: 30 January, 2018

Keywords :
RF-MEMS
Residual Stresses
Coplanar waveguide
Skin depth
Return loss
Insertion loss
Isolation
Reliability
Transmission coefficient
Reflection coefficient

ABSTRACT

Microelectromechanical Systems (MEMS) are devices made up of several electrical and mechanical components. They consist of mechanical functions (sensing, thermal, inertial) and electrical functions (switching, decision making) on a single chip made by microfabrication methods. These chips exhibit combined properties of the two functions. The size of system has characteristic dimensions less than 1mm but more than 1μm. The configuration of these components determine the final deliverables of the switch. MEMS can be designed to meet user requirements on any level from microbiological application such as biomedical transducers or tissue engineering, to mechanical systems such as microfluidic diagnoses or chemical fuel cells. The low cost, small mass and minimal power consumption of the MEMS makes it possible to readily integrate to any kind of system in any environment. MEMS are faster, better and cheaper. They offer excellent electrical performances. MEMS working at Radio frequencies are RF MEMS. RF-MEMS switches find huge market in the modern telecommunication networks, biological, automobiles, satellites and defense systems because of their lower power consumptions at relatively higher frequencies and better electrical performances. But the reliability is the major hurdle in the fate of RF MEMS switches. Reliability mainly arises due to the presence of residual stresses, charging current, fatigue and creep and contact degradation. The presence of residual stresses in switches the S-Parameters of the switches are affected badly and the residual stress affects the final planarity of the fabricated structure. Design and simulation of an RF-MEMS switch is proposed considering the residual stresses in both on and off state. The operating frequency band is being optimized and the best possible feasible fabrication technique for the proposed switch design is being analyzed. S-Parameters are calculated and a comparison for the switches with stress and without stress is performed in FEM based HFSS software.

1. Introduction

MEMS technologies are faster, better, and cheaper. They have reduced size and weight, means fewer power consumption and there are lower component counts. MEMS promises superior electrical performances and are built with low cost [1]. They can be produced at massive level. They have advanced functionality and superior reliability. They have new functionality and better system capability.

The ultimate goals of MEMS are to reduce size, cost, weight, power consumptions by increasing the performance of existing macroscopic devices [2]. To make devices never before possible at macroscopic scale. Easy to integrate into systems or modify. They have Small thermal constant and they can be highly resistant to vibration, shock and radiation. MEMS are Microelectro-mechanical system in which the device and structures are fabricated at micro levels using the micro level fabrication process [3]. It is basically an integrated micro devices or systems

*Corresponding Author: Amna Riaz, Email: amna_riaz99@yahoo.com

combining electrical and mechanical components fabricated using Integrated Circuit (IC) compatible processing technique and it range from micron to millimeter [4].

MEMS structures can be a simple system with all static elements or large complicated system with some parts moving and some static elements. Chief concept of MEMS is that there is at least one moving element [5]. MEMS most important elements are micro sensors and micro actuators. Basically these are the transducers that convert electrical to mechanical and vice versa. Micro actuators and micro sensors are widely used in the integrated circuits industry due to low per device cost and other benefits.

2. Radio Frequency MEMS

2.1. Maintaining the Integrity of the Specifications

MEMS can be classified into different types depending upon their mean of actuation. These include acoustic MEMS, optical MEMS and Radio Frequency MEMS (RF-MEMS) [6]. RF MEMS operate in the electromagnetic wave frequencies in the 3 kHz - 300 GHz range [7-8]. This frequency provides maximized efficiency in the circuitry and ability to flow through paths containing insulating material, such as the dielectric insulators of capacitors.

3. Microfabrication of RF-MEMS Switch and Thermal Induction of Residual Stress

In Trento Italy; at ITC-IRST research center; using gold electroplating a process was carried out on micro bridge and RF MEMS; which resulted in obtaining suspended microstructure by photoresist of 3micrometer thickness [2].

3.1. RF Switch Surface (RFS) Micromachining

The RF-MEMS STS whose design is considered in this study is realized through the gold electrodeposition microfabrication

1. At 975C in a wet atmosphere silicon wafer came in contact with thermal oxide of 1000nm thickness and then with nitrogen. Resistors and lines are formed by polysilicon while pattern is formed by dry ditching
2. At 718C silicon oxide is placed with wafer and after that applied to photolithography.
3. A metal is spitted for signal line generation. The general temperature is 400C but when LPCVD oxide is deposited temperature rises by 30C. The area which was free from photoresist; for electrical contact oxide is removed by dry etching then wafer is washed.
4. Wafer is now wet etched and floating metal (Cr/Au) layer is evaporated. 3micrometer thickness of Cr is placed between oxide layer and Au.
5. At 52C electroplating of gold on seed layer is done.
6. At 52C; gold layer is used for strengthen the structure. It's left for 30 minutes a 190C. Finally at 200C structure comes in contact with plasma oxygen for removing extra layers.

3.2. Mechanisms contributing to Residual Stress Induction

For residual stress following steps are made: -

1. Stress In this process; Cr-Au PVD is deposited on photoresist. Seed layer is used for ensuring adhesive property of Au. Although due to thermal instability chromium oxide is formed along with residual stress.
2. Gradient happens due to difference in thermal coefficient of gold and seed layer. Seed layer expands more rapidly and results in tensile residual stress. However micro beams get deformed due to release of Au.
3. Oxygen Plasma ashing helps in avoiding sticking but temperature changes too rapidly forming residual stress. Temperature is maintained at 200C so that Au reached even the narrow paths.

4. Design of RF-MEMS Symmetric Toggle Switches

4.1. On State without stress

The simply designed switch in HFSS is the extension of work presented in "Simulating the Electromagnetic Effects of Thermally Induced Residual Stresses in RF-MEMS Switches". This RF-MEMS switch is designed in FEM based HFSS software. In this switch, (Figure 1) simply a suspended type bridge was designed over the coplanar waveguide transmission line, the design of the switch is as follow. The gold material was assigned to the bridge, this switch was designed just to ensure the results validity with the already present standards in the literature.

Similarly, when the stress was induced in the switch, (Figure 2) the central bridge of the switch deflected upward, which ultimately affected the overall electromagnetic performance of the switch.

4.2. Identify the Headings Design of a simple RF-MEMS Switch in off-state without stress

When the switch is in the off state, the bridge deflects downwards at the center just over the dielectric layer of signal CPW (Figure 3). The side conductors of the CPW are grounded while the signal pass through the central conductor of CPW

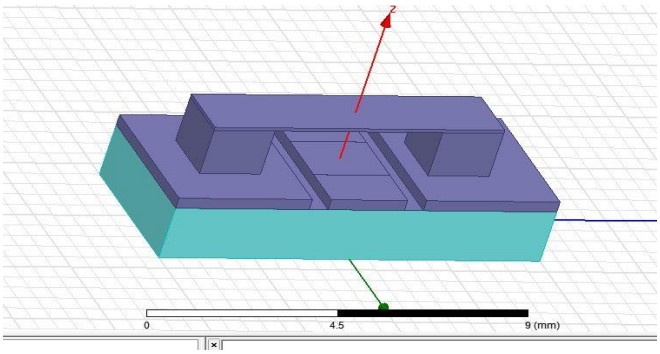


Figure 1: Design of a simple RF-MEMS Switch without stress and excitation

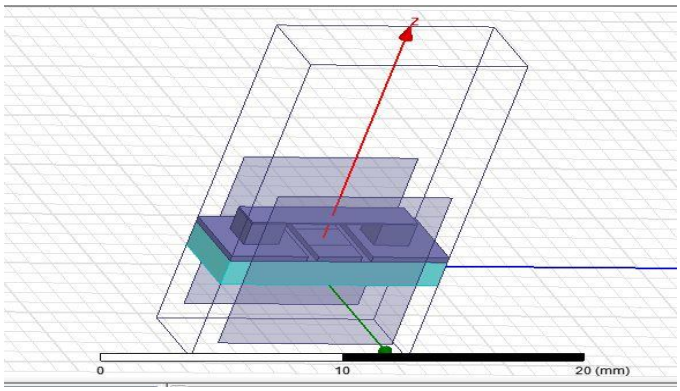


Figure 2: Design of a simple RF-MEMS Switch without stress and excitation and air box

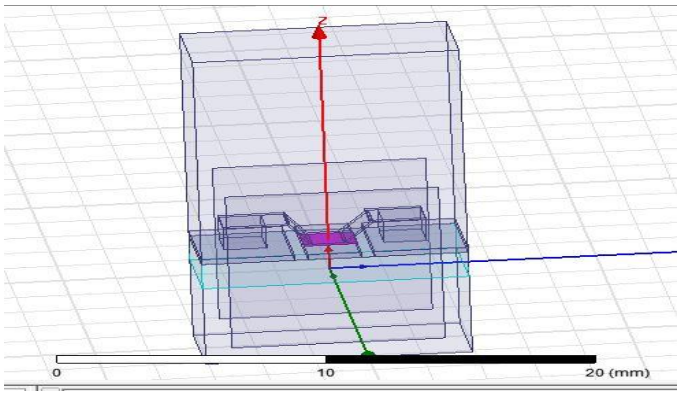


Figure 3: Design of a simple RF-MEMS Switch in off-state without stress

The excitation of the switch is applied through waveport excitations and it is shown in the figure as red shaded side conducting sheet. The waveport connects the central conductor of the CPW with the side conductors, and are used to supply the signal through the switch (Figure 4).

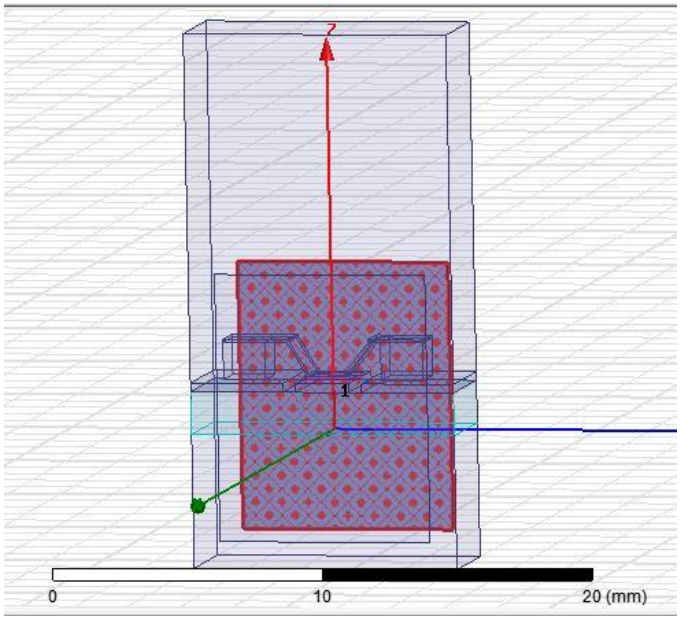


Figure 4: Design of a simple RF-MEMS Switch in off-state without stress with the excitation shown

5. Comparison of simply designed switch with the results already present in literature

5.1. Return Loss of On State

Per the Rangra at Kamal and Jacobi [9], the reflection coefficient of the switch during the on state must be always less than -10 dB, which ensures better switch performances. When we designed a simple switch RF-MEMS switch in HFSS software, the result of reflection coefficient was ascending curve along the X-axis. The result of this switch (Figure 5) is exactly in accordance with existing standards of literature [10, 11] which ensures that our approach was accurate.

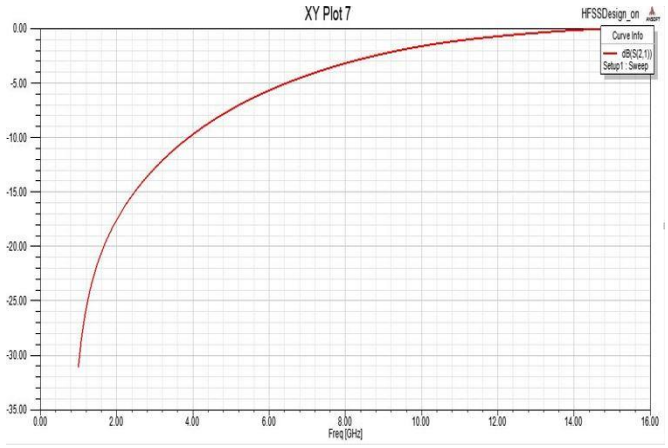


Figure 5: Simulated results of on state return loss

As shown in the graph above, at lower frequencies like between 1.5 GHz to 4 GHz, the reflection was good. Later as the operating frequency range increased, the switch moved to lesser matching or the reflection of the power from the switch increased. From the frequency range of 8GHz to 16 GHz, the switch was not in ideal conditions i.e. the refection increased at higher frequencies. The result already stated in literature (Figure 6) also showed the same trend as shown below [10, 12]. The reflection increased as we move on to higher frequencies.

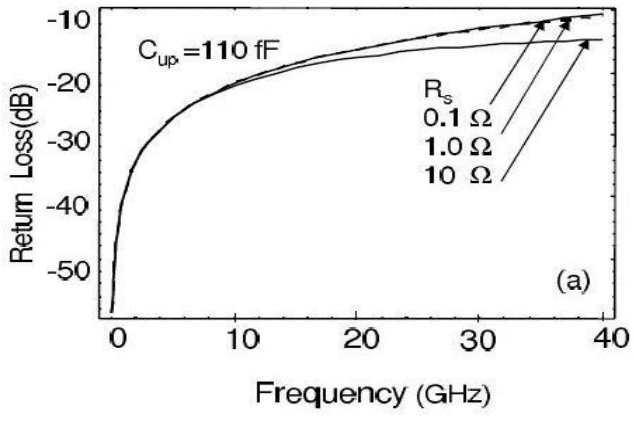


Figure 6: Literature results of on state return loss [2, 10]

5.2. Insertion loss of on state

Ideally, the switch during the on state should have minimum insertion loss, i.e. the value of transmission coefficient should be zero or close to zero. The results which we obtained for a simple RF-MEMS switch in the on state are shown in the graph below, here as the frequency increases, the value of transmission coefficient increases, which predicts lower or poor insertion loss (Figure 7). Ideally the transmission coefficient should be zero as state hypothetically in the literature (Figure 8), but in actual it is not possible to have perfectly zero transmission coefficient [2, 10]. But as the frequency increases, overall there is a decline in the value of transmission coefficient. Which means that at higher frequencies the transmission reduces [13].

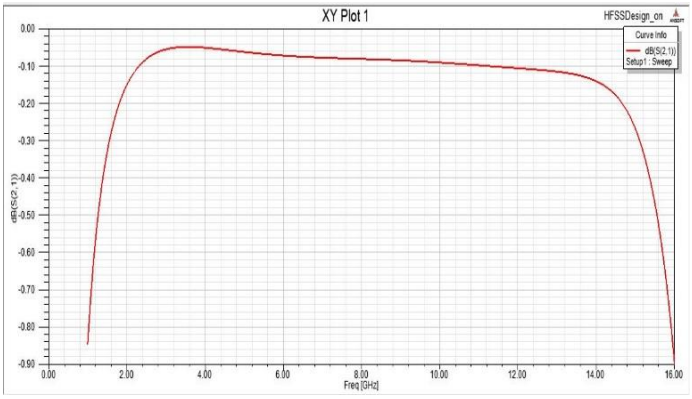


Figure 7: Simulated results of on state Insertion loss

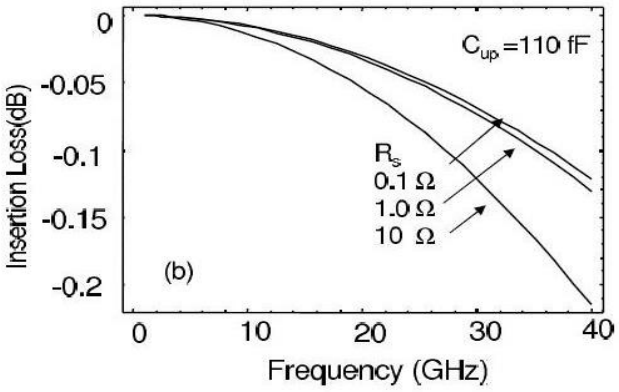


Figure 8: Literature results of on state Insertion loss [2, 10]

5.3. Return Loss of Off-State

Now when the switch is in off-state, the value of reflection coefficient should be greater than -10 dB because at the off-state, switch is not in the working conditions. Maximum power which is incident to the switch gets reflected back [14]. So as the literature predicts, if the operating frequency is taken on X-axis and the transmission coefficients along Y-axis, then the graph should be an ascending curve. So our results of a simple RF-MEMS switch (Figure 9) were found to be exactly similar to those, which were mentioned in the literature (Figure 10). The slight difference in the curve can be explained in terms of the operating frequency ranges i.e. in our graph the limit was between 0 GHz to 16 GHz, whereas in the standard the limit was between 0 GHz to 40 GHz.

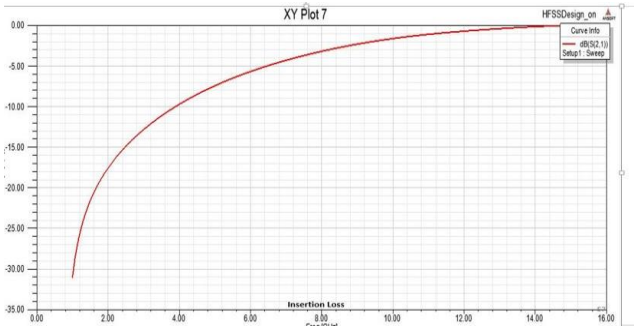


Figure 9: Simulated results of off state return Lost

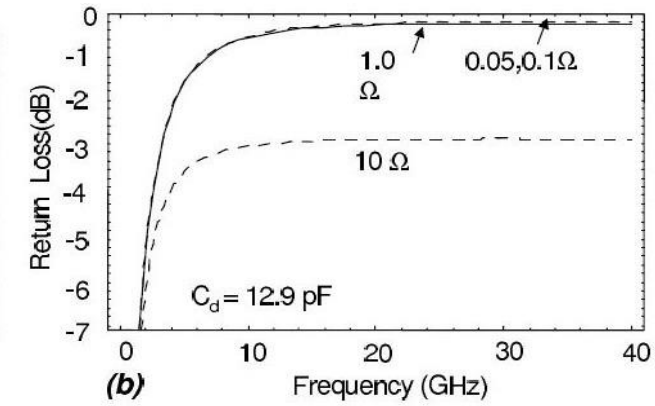


Figure 10: Literature results of off state return Loss. [2, 10]

5.4. Result of Isolation in Off State

According to the set standards, the isolation should be maximum in the off-state. Due to higher isolation, the leakage of power during off-state is minimum [15]. The S_{21} represents the isolation in the off-state. So, here the curve according to the set standard is descending curve, followed by a dip at nearly 15 GHz, and then a gradually ascending curve, after 15 GHz. So, our results of the simple RF-MEMS switch (Figure 11) are in accordance with the pre-defined standards (Figure 12). So, maximum isolation was observed at off-state nearly at 15 GHz.

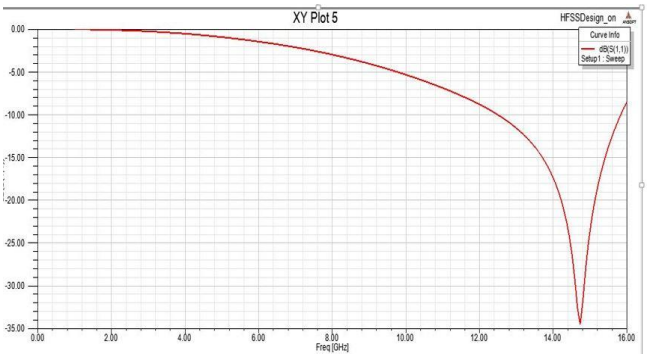


Figure 11: Simulated results of off state Isolation

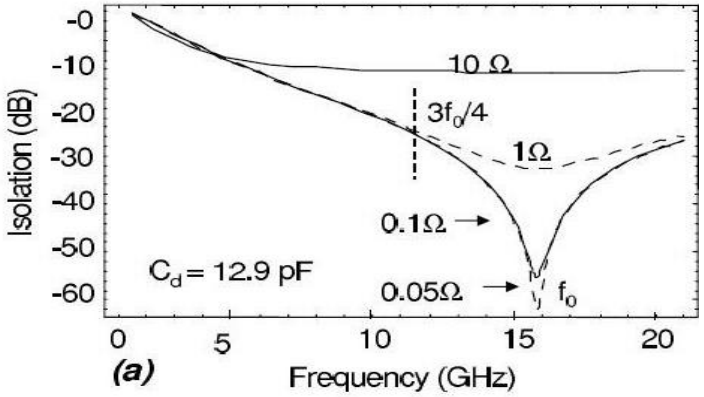


Figure 12: Literature results of off state Isolation. [2, 10]

6. ANALYSIS OF RESULTS

6.1. Without stress

The value of return loss S_{11} in the RF-MEMS switch without stress, shows excellent results. The value of reflection coefficients is much less than -10 dB, which indicates the best matching in the sense of transmitting power. The value of S_{11} lies between -58 dB to -42 dB, which indicates the very less power is reflected. As the value is much lesser than -10 dB, so we can say that switch is having excellent matching or very good reflection loss [16, 17]. Ideally any operating switch must have, reflection loss less than -10 dB, then it is assumed that the electrical performance of the switch is better. If any switch does not have this predefined reflection coefficient, then it means that maximum of the power is wasted or reflected back rather than transmitting through the switch [18]. So, the designed switch in HFSS software, which was already fabricated showed excellent electrical performance in the on state. Their electromagnetic performances are good for the frequency ranges between 2GHz to 16 GHz but, are best between the frequencies of 1.5 GHz to 8GHz. Similarly, the value of transmission coefficient is near 0dB, which means that overall transmission through the switch is excellent. The logic behind stating this is that as we know 0dB is actually equal to 1. It means when the value of transmission is or close to 0dB then the transmission through the switch will be excellent. Again, the operating frequency ranges between 1GHz to 8GHz. Because as the frequency increases, the transmission decreases.

6.2. With Stress

Due to the generation of residual stresses in the RF-MEMS switch, the reflection of the switch is badly affected. The reason is that due to generation of stresses, the bridge deflects upward at the center. Now the air gap between the top central bridge and the ground dielectric increases [19, 20, 21]. Due to rise in air gap the up-state capacitance of the switch is decreased now. As we already explained in the design aspect of RF-MEMS switch that S-parameters are dependent upon the up and down state capacitances. Reflection is dependent upon the up-state capacitances while the isolation is affected by down state capacitances [22, 23, 24]. The results of return loss with stress are shown in Figure 13.

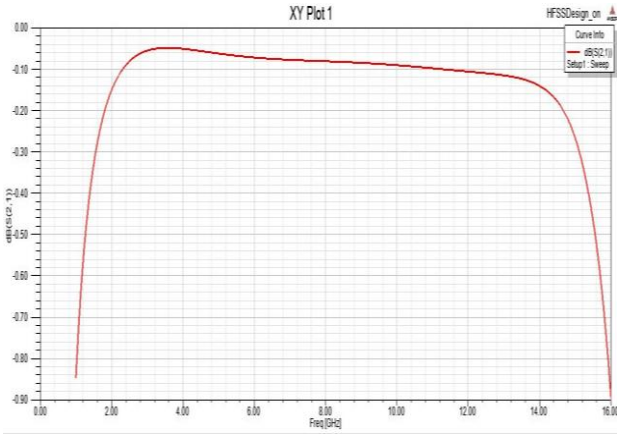


Figure 13: Simulated results of return loss with stress

Similarly, the results for insertion loss are shown in Figure 14.

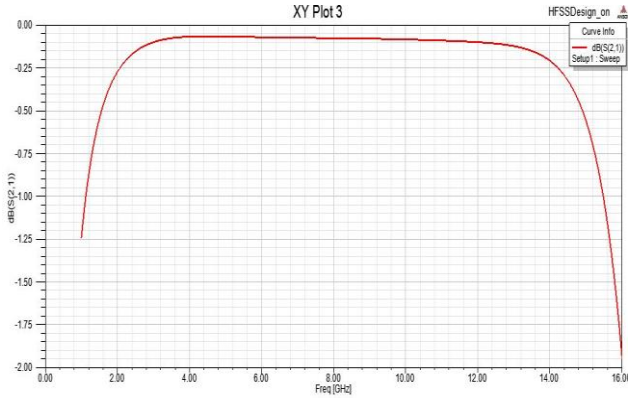


Figure 14: Simulated results of insertion loss with stress

7. Conclusion

So we can conclude that switch was operating efficiently when there were no residual stresses but due to the generation of residual stresses, not only the final planetary of the switch was affected but also the electrical performance of the switch was affected badly. Future work can be done on how to reduce the residual stresses in RF-MEMS switch during microfabrication when the temperature is changed abruptly. Some research can also be produced in the domain of improving reliability of RF-MEMS switches by reducing the effects of charging of dielectric, the effects of creep and fatigue along with the contact degradation. So that it may help to improve the overall efficiency of the RF-MEMS switch.

References

[1] G. Rebeiz, K. Entesari, I. Reines, S. J. Park, M. El-tanani, A. Grichener, and A. Brown, "Tuning in to RF MEMS", Microwave Magazine, IEEE, vol. 10, no. 6, pp. 55-72, 2009.

[2] Solazzi F, Margesin B and Faes A, "Novel Design Solutions for High Reliability RF MEMS Switches", 2010

[3] R. Osiander, M. A. G. Darrin, and J. L. Champion, MEMS and Microstructures in Aerospace Applications. New York, NY, USA: Taylor & Francis, 2006.

[4] G. M. Rebeiz, RF MEMS: Theory, Design, and Technology. John Wiley & Sons, Inc., 2003.

[5] N. Barker and G. Rebeiz, "Optimization of distributed MEMS transmission-line phase shifters-U-band and W-band designs," Microwave Theory and Techniques, IEEE Transactions on, vol. 48, no. 11, pp. 1957 - 1966, Nov. 2000.

[6] A. Verger, A. Pothier, C. Guines, A. Crunteanu, P. Blondy, J. Orlianges, J. Dhennin, F. Courtade, and O. Vendier, "Sub-hundred nanosecond reconfiguration capabilities of nanogap RF MEMS switched capacitor," in Microwave Symposium Digest (MTT), 2010 IEEE MTT-S International, May 2010, pp. 1238 -1241.

[7] H. Newman, J. Ebel, D. Judy, and J. Maciel, -Lifetime Measurements on a High-Reliability RF-MEMS Contact Switch," Microwave and Wireless Components Letters, IEEE, Feb 2008.

[8] Mulloni J, Giacomozzi V and Margesin B 2010 Controlling stress and stress gradient during the release process in gold suspended micro-structures Sensors Actuators A 162 93–9.

[10] B. Pillans, J. Kleber, C. Goldsmith, and M. Eberly, "RF power handling of capacitive RF MEMS devices," in Microwave Symposium Digest, 2002 IEEE MTT-S International, 2002.

[11]J. Muldavin, C. Bozler, S. Rabe, P. Wyatt, and C. Keast, "Wafer- Scale Packaged RF Microelectromechanical Switches," Microwave Theory and Techniques, IEEE Transactions on, vol. 56, no. 2, pp.522 -529, 2008.

[12]A. Grichener and G. Rebeiz, "High-Reliability RF-MEMS Switched Capacitors With Digital and Analog Tuning Characteristics," Mi-crowave Theory and Techniques, IEEE Transactions on, vol. 58, no. 10, pp. 2692 - 2701, 2010.

[13] B. Lakshminarayanan, D. Mercier, and G. Rebeiz, "High-Reliability Miniature RF-MEMS Switched Capacitors," Microwave Theory and Techniques, IEEE Transactions on, vol. 56, no. 4, pp. 971 -981, 2008.

[14] D. Mardivirin, A. Pothier, A. Crunteanu, B. Vialle, and P. Blondy, "Charging in Dielectricless Capacitive RF-MEMS Switches," Microwave Theory and Techniques, IEEE Transactions on, vol. 57,no. 1, pp. 231 -236, 2009.

[15] T. Rijks, J. van Beek, P. Steeneken, M. Ulenaers, J. De Coster, and R. Puers, "RF MEMS tunable capacitors with large tuning ratio," in17th IEEE International Conference on. MEMS, 2004.

[16] F. Casini, P. Farinelli, G. Mannocchi, S. DiNardo, B. Margesin, G. De Angelis, R. Marcelli, O. Vendier, and L. Vietzorreck, "High Performance RF-MEMS SP4T switches in CPW technology for space Applications", in Microwave Conference (EuMC), 2010 European, 2010, pp. 89 -92.

[17] I. Reines, S. J. Park, and G. M. Rebeiz, "Compact Low-Loss Tunable Band Bandstop Filter With Miniature RF-MEMS Switches," Microwave Theory and Techniques, IEEE Transactions on, vol. 58, no. 7, pp. 1887 -1895, 2010.

[18]A. Ocera, P. Farinelli, F. Cherubini, P. Mezzanotte, R. Sorrentino, B. Margesin, and F. Giacomozzi, "A MEMS-Recon_gurable Power Divider on High Resistivity Silicon Substrate," in Microwave Symposium, 2007. IEEE/MTT-S International, 2007, pp. 501 -504.

[19] M. Kim, J. Hacker, R. Mihailovich, and J. DeNatale, "A DC-to-40 GHz four-bit RF MEMS true-time delay network," Microwave and Wireless Components Letters, IEEE, vol. 11, no. 2, pp. 56 -58, Feb.2001.

[20] T. Vaha-Heikkila, J. Varis, J. Tuovinen, and G. Rebeiz, "A 20-50 GHz RF MEMS single-stub impedance tuner," Microwave and Wireless Components Letters, IEEE, vol. 15, no. 4, pp. 205 - 207, 2005.

[21] Rebeiz G M 2003 RF MEMS: Theory, Design and Technology (New York: Wiley)

[22] W. Auer, E. Hettlage, and G. Ru_, "RF-switches: Application and design," in ESA, Proceedings of the 3rd European Space Mechanisms and Tribology Symposium p 191 -195 (SEE N88-21191 14-18), Dec.1987, pp. 191-195.

[23] R.C. Hibbeler, "Mechanics of Materials", Edition 9, pp 153

[24] Fang W, Lee C H and Hu H H 1999 On the buckling behavior of micromachined beams J. Micromech. Microeng.9 236–44

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 76-86 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Real-Time Flux-weakening Control for an IPMSM Drive System Using a Predictive Controller

Tian-Hua Liu*, Shao-Kai Tseng, Yi Chen, Mao-Bin Lu

Department of Electrical Engineering, National Taiwan University of Science and Technology, 106, Taiwan

ARTICLE INFO

Received: 07 October, 2017
Accepted: 29 October, 2017
Online: 10 November, 2017

Keywords :
predictive controller
real-time tuning
flux-weakening control
IPMSM

ABSTRACT

This paper proposes extended-range high-speed control for an IPMSM drive system. A simple real-time tuning flux-weakening control algorithm is proposed and implemented to control an IPMSM drive system in a wide variable speed range, from 3 r/min up to 2700 r/min. This flux-weakening control algorithm does not require any motor parameters and only needs simple mathematical computations. The proposed drive system adjusts the angle between the d-axis and q-axis current to reach flux-weakening control. In addition, a multiple sampling predictive controller is implemented to enhance the dynamic responses of the proposed drive system, which yields improved overall transient responses, superior load responses, and good tracking responses. A detailed analysis of the proposed drive system's stability is discussed as well. A 32-bit digital signal processor, TMS-320F-28335, is used to execute the predictive controller and the flux-weakening control algorithm for the IPMSM drive system. Experimental results can validate the theoretical analysis.

1. Introduction

Use of the interior permanent magnet synchronous motor (IPMSM) has greatly increased due to its advantages over other motors, which include high efficiency, wide operating speed range, high power density, and robustness. The IPMSM has two different torque components: electromagnetic torque and reluctance torque. By strategically controlling the d-axis and q-axis currents, the maximum torque per ampere (MTPA) control or flux weakening control can be achieved at different operating speeds.

This paper is an extension of work originally presented in 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE) Conference [1]. Several researchers have proposed different flux-weakening control methods to increase the high-speed operating range of an IPMSM. For example, Bolognani et al. proposed an adaptive flux-weakening controller for an IPMSM drive system. The adaptation algorithm, however, is very complicated to implement [2]. Zhu et al. implemented a flux-weakening control for a PMSM drive system [3]. The proposed method accounted for resistance voltage drop and inverter nonlinearities. This method, however, is only suitable for a PMSM drive system, not an IPMSM drive system. Kwon et al. proposed a flux-weakening control for an IPMSM with quasi-six-step operation [4]. A feed-forward path, which consists of a one dimensional look-up table, is required. Pan et al. proposed a robust flux-weakening control strategy for a surface-mounted permanent-magnet motor drive [5]. A closed-form solution of the maximum available torque-producing current is generated to achieve both fast responses and real-time tuning flux-weakening control. This method, however, is suitable for a surface mounted PMSM but not an IPMSM. Uddin et al. proposed an on-line parameter-estimation-based high-speed control of an IPMSM drive. The method of estimating the d-axis inductance, q-axis inductance, and external load, however, is very complicated [6]. Pan et al. proposed a voltage-constraint-based flux-weakening control of an IPMSM drive system [7]. The method requires a d-axis current command and q-axis current command generator, which is rather complicated. Jung et al. proposed a hexagonal voltage limit to increase the voltage amplitude and to improve the DC-link voltage utilization in an IPMSM drive system. A torque control method based on the voltage angle was investigated. The idea is very good; however, the voltage vector selection algorithm is very complicated [8]. Chaithongsuk et al. proposed an optimal design of permanent magnet motors to improve flux-weakening performance in variable speed drives. It was shown that a salient-pole motor performed better at high speeds than a non-salient pole motor [9]. Kim proposed a novel magnetic flux-weakening method for permanent magnet synchronous motors in electric vehicles [10]. By adjusting the air gap between the stator and rotor, an increased maximum speed and maximum output power could be

*Corresponding Author: Tian-Hua Liu, Mobile: +886-22737-6108
Email: Liu@mail.ntust.edu.tw

obtained. This method, however, requires very complicated motor modifications. Jevremovic et al. proposed a simple robust integral controller to achieve field weakening control for a PMSM [11]. Harnefors et al. used a systematic analysis to determine the parameters of the current and speed combined controller to reduce the complicity [12]. To improve the performance of field weakening control, Wallmark et al. [13] and Bedetti et al. implemented a voltage regulator to achieve flux-weakening control for an electric vehicle and an IPMSM [14].

In this paper, a simple flux-weakening control method is proposed. This method has several advantages over previously proposed flux-weakening methods [1]-[14]. For one thing, the parameters of the IPMSM are not required. Also, the computations for the flux-weakening control are very simple. For example, only square root, multiplication, and addition are used. In addition, to achieve fast responses and good load disturbance rejection capability from the low-speed range to high-speed range, a predictive controller is employed. In contrast to previous research on predictive control [15]-[16], this paper proposes a multiple-loop predictive control motor drive system. The sampling interval of the current-loop is 100 , and the sampling interval of the speed-loop is 1 ms. Due to the quick current regulation, the proposed system has better performance than previous one-loop predictive control drive systems [15]-[16]. Unlike previous methods of flux-weakening control without using motor parameters [2], [16]-[17], the proposed method provides a simple control algorithm for voltage regulation without an adaptive controller or PI controller. To the authors' best knowledge, the proposed flux-weakening control method is a new idea. In addition, the idea of the predictive controller applied in the simplified flux-weakening control of an IPMSM is original.

2. Predictive Controller Design

The theory of predictive control was first proposed in the 1970s, and predictive controllers have been used in chemical and steel processing for more than 40 years. Recently, predictive control has been applied in motor drives due to the fast computation ability of the digital signal processor (DSP). The design objective of a predictive controller is to compute the control input u to optimize the future dynamic behavior for the output y of the plant. This optimization must be performed within a limited time window. Several researchers have investigated predictive control for motor drives. For example, Pacas et al. proposed a predictive direct torque control for PMSMs, which yielded a faster torque settling time than the classic field-oriented control with a PI controller [15]. Bolognani et al. investigated the design and implementation of model predictive control for electric motor drives [16]. Fuentes et al. implemented a predictive speed control for a two-mass system driven by a PMSM. Their proposed method allowed feedback of all the mechanical and electrical state variables into a single control input and provided a higher bandwidth closed-loop speed control [18]. Siami et al. proposed an improved predictive current control using prediction error correction for PMSMs [19]. Tarczewski et al. implemented a constrained state feedback speed control for PMSMs, which calculated the boundaries of control signals to provide permissible values of the future state variables [20].

In this paper, a systematic predictive controller design is proposed for an IPMSM drive system. The proposed predictive controllers, including a predictive current controller and a predictive speed controller, provide a multiple sampling rate control system and a closed-loop speed-control block diagram that are different from the predictive controllers proposed in previously published papers [15]-[16], [18]-[20].

The design of a predictive controller requires the following. First, a model of the uncontrolled drive system is developed. Next, a cost function that represents the desired behavior of the system is defined. Finally, a predictive controller for the motor drive system is derived.

2.1. Predictive Current Controller

The state-variable equation of an IPMSM can be expressed as

$$\frac{d}{dt}\begin{bmatrix} i_d \\ i_q \end{bmatrix} = \begin{bmatrix} \frac{-r_s}{L_d} & 0 \\ 0 & \frac{-r_s}{L_q} \end{bmatrix}\begin{bmatrix} i_d \\ i_q \end{bmatrix} + \begin{bmatrix} \frac{1}{L_d} & 0 \\ 0 & \frac{1}{L_q} \end{bmatrix}\begin{bmatrix} v_d \\ v_q \end{bmatrix} + \begin{bmatrix} \frac{1}{L_d} & 0 \\ 0 & \frac{1}{L_q} \end{bmatrix}\begin{bmatrix} \omega_{re} L_q i_q \\ -\omega_{re}(L_d i_d + \lambda_m) \end{bmatrix} \quad (1)$$

Where d/dt is the differential operator, i_d and i_q are the d-axis and q-axis stator currents, r_s is the stator resistance, v_d and v_q are the d-axis and q-axis stator voltages, L_d and L_q are the d-axis and q-axis self-inductances, λ_m is the permanent magnetic flux leakage, and ω_{re} is the electric motor speed.

After transferring the continuous-time domain into the discrete-time domain, (1) can be expressed as

$$i_d(z) = \frac{\beta_{cd} z^{-1}}{1-\alpha_{cd} z^{-1}}(v_d - \omega_{re} L_q i_q) \quad (2)$$

and

$$i_q(z) = \frac{\beta_{cq} z^{-1}}{1-\alpha_{cq} z^{-1}}(v_q - \omega_{re}(L_d i_d + \lambda_m)) \quad (3)$$

The related parameters in (2) and (3) are shown as follows

$$\alpha_{cd} = e^{\frac{-r_s T_{cur}}{L_d}} \quad (4a)$$

$$\beta_{cd} = \frac{1}{r_s}(1-\alpha_{cd}) \quad (4b)$$

$$\alpha_{cq} = e^{\frac{-r_s T_{cur}}{L_q}} \quad (4c)$$

$$\beta_{cq} = \frac{1}{r_s}(1-\alpha_{cq}) \quad (4d)$$

where T_{cur} is the sampling interval of current-loop control. Because the sampling time interval is shorter than the time constant of the stator current-loop, Euler approximation can be used here. As a result, the parameters shown in (4a)-(4d) can be described as follows

$$\alpha_{cd} \approx 1-\frac{r_s T_{cur}}{L_d} \tag{5a}$$

$$\beta_{cd} \approx \frac{r_s T_{cur}}{L_d} \tag{5b}$$

$$\alpha_{cq} \approx 1-\frac{r_s T_{cur}}{L_q} \tag{5c}$$

and

$$\beta_{cq} \approx \frac{r_s T_{cur}}{L_q} \tag{5d}$$

Then (2) and (3) can be rewritten as follows

$$X_{cur}(k+1) = A_{cur}X_{cur}(k) + B_{cur}U_{cur}(k) + B_{cur}F_{cur}(k) \tag{6}$$

and

$$X_{cur}(k) = \begin{bmatrix} i_d(k) & i_q(k) \end{bmatrix}^T \tag{7a}$$

$$F_{cur}(k) = \begin{bmatrix} \omega_{re}(k)L_q i_q(k) & -\omega_{re}(k)\left(L_d i_d(k) + \lambda_m\right) \end{bmatrix}^T \tag{7b}$$

$$U_{cur}(k) = \begin{bmatrix} v_d(k) & v_q(k) \end{bmatrix}^T \tag{7c}$$

$$A_{cur} = \begin{bmatrix} 1-\frac{r_s T_{cur}}{L_d} & 0 \\ 0 & 1-\frac{r_s T_{cur}}{L_q} \end{bmatrix} \tag{7d}$$

$$B_{cur} = \begin{bmatrix} \frac{T_{cur}}{L_d} & 0 \\ 0 & \frac{T_{cur}}{L_q} \end{bmatrix} \tag{7e}$$

According to (6), the (k+1)-th predictive current can be expressed as

$$\hat{X}_{cur}(k+1) = A_{cur}X_{cur}(k) + B_{cur}\left(U_{cur}(k) + F_{cur}(k)\right) \tag{8}$$

The performance index of the d-axis current control is defined as

$$J_{cur} = \left[\hat{i}_d(k+1) - i_d^*(k+1)\right]^2 \tag{9}$$

The following equation is the partial differential of (9) manipulated to equal zero.

$$\frac{\partial J_{cur_d}}{\partial U_d} = \frac{\partial\left(\hat{i}_d(k+1) - i_d^*(k+1)\right)^2}{\partial v_d(k)} = 0 \tag{10}$$

Substituting (8) into (10), we can obtain

$$\frac{\partial J_{cur_d}}{\partial v_d(k)} = \frac{\partial\left((1-\frac{T_{cur}r_s}{L_d})i_d(k) + (\frac{T_{cur}}{L_d})\left(v_d(k) + \omega_{re}(k)L_q i_q(k)\right) - i_d^*(k+1)\right)^2}{\partial v_d(k)}$$
$$= 2\left((1-\frac{T_{cur}r_s}{L_d})i_d(k) + (\frac{T_{cur}}{L_d})\left(v_d(k) + \omega_{re}(k)L_q i_q(k)\right) - i_d^*(k+1)\right)\left(\frac{T_{cur}}{L_d}\right) = 0 \tag{11}$$

After arranging the above, it is not difficult to obtain

$$\left((1-\frac{T_{cur}r_s}{L_d})i_d(k) + (\frac{T_{cur}}{L_d})\left(v_d(k) + \omega_{re}(k)L_q i_q(k)\right) - i_d^*(k+1)\right)$$
$$= \left((\frac{L_d}{T_{cur}} - r_s)i_d(k) + \left(v_d(k) + \omega_{re}(k)L_q i_q(k)\right) - \frac{L_d}{T_{cur}}i_d^*(k+1)\right) = 0 \tag{12}$$

Finally, from (12), we can derive the d-axis voltage command as

$$v_d^*(k) = r_s i_d(k) + \frac{L_d}{T_{cur}}\left(i_d^*(k+1) - i_d(k)\right) - \omega_{re}(k)L_q i_q(k) \tag{13}$$

By using the same principle, we can derive the q-axis voltage command as follows

$$v_q^*(k) = r_s i_q(k) + \frac{L_q}{T_{cur}}\left(i_q^*(k+1) - i_q(k)\right) + \omega_{re}(k)\left(L_d i_d(k) + \lambda_m\right) \tag{14}$$

2.2. Predictive Speed Controller

The differential equation of an IPMSM is

$$\frac{d}{dt}\omega_{rm} = \frac{1}{J_m}(T_e - T_L - B_m\omega_{rm}) \tag{15}$$

Where J_m ω_{rm} is the mechanical speed, $\frac{d}{dt}$ is the differential operator, T_e is the total output torque, T_L is the external load, and B_m is the viscous frictional coefficient.

According to (15) and assuming $T_L = 0$, the speed can be expressed as

$$\omega_{rm}(s) = \frac{T_e}{J_m s + B_m} \tag{16}$$

By converting (16) into discrete form, one can obtain

$$\begin{aligned}\omega_{rm}(k+1) &= e^{-(B_m/J_m)T_{sp}}\omega_{rm}(k) + \frac{1}{B_m}(1-e^{-(B_m/J_m)T_{sp}})T_e(k) \\ &= a_1\omega_{rm}(k) + b_1 u(k)\end{aligned} \tag{17}$$

and

$$a_1 = e^{-(B_m/J_m)T_{sp}} \tag{18}$$

$$b_1 = \frac{1}{B_m}(1-e^{-(B_m/J_m)T_{sp}}) \tag{19}$$

$$u(k) = T_e(k) \tag{20}$$

where T_{sp} is the sampling interval of the speed-loop control. Next, by setting the predictive window as 1 and the predictive control as 1, the performance index J_{sp} is expressed as

$$\begin{aligned}J_{sp} = \sum_{i=1}^{1}\left[P(z)\hat{\omega}_{rm}^{pre}(k+i) - P(1)\omega_{rm}^{*}(k+i)\right]^2 \\ + \rho\sum_{i=1}^{1}\left[Q(z)u(k+i-1)\right]^2\end{aligned} \tag{21}$$

where $\hat{\omega}_{rm}^{pre}$ is the predictive speed, $P(z)$ and $Q(z)$ are polynomials of z, and ρ is the weighting factor. By computing $\frac{\partial J_{sp}}{\partial u(k)}$, one can obtain

$$\begin{aligned}\frac{\partial J_{sp}}{\partial u(k)} = 2\frac{P(z)\partial\hat{\omega}_{rm}^{pre}(k+1)}{\partial u(k)}[P(z)\hat{\omega}_{rm}^{pre}(k+1) \\ - P(1)\omega_{rm}^{*}(k+1)] + 2\rho Q^2(z)u(k) = 0\end{aligned} \tag{22}$$

Substituting (17) into (22), one can derive

$$\begin{aligned}P(z)b_1\left(P(z)\left(b_1u(k) + a_1\omega_{rm}(k)\right) - P(1)\omega_{rm}^{*}(k+1)\right) \\ + \rho Q^2(z)u(k) = 0\end{aligned} \tag{23}$$

The speed command at (k+1)th sampling interval can be expressed as a fixed command Sp and then can be defined as

$$\omega_{rm}^{*}(k+1) = S_p \tag{24}$$

By substituting (24) into (23), we can obtain

$$(P(z)b_1)^2u(k) + \rho Q^2(z)u(k) = P(z)b_1P(1)S_p - P^2(z)b_1a_1\omega_{rm}(k) \tag{25}$$

The control input can be expressed as

$$u(k) = R(MS_p - N\omega_{rm}(k)) \tag{26}$$

and

$$M = \frac{P(1)}{b_1P(z)} \tag{27}$$

$$N = \frac{a_1}{b_1} \tag{28}$$

$$R = \frac{1}{1 + \frac{\rho}{b_1^2}(\frac{Q(z)}{P(z)})^2} \tag{29}$$

The predictive controller is shown in Figure 1. The block diagram of the predictive control includes a command gradation M(z), a controller R(z), a plant P(z), and a feedback gain N. The structure is very similar to a standard control system; however, the parameters can be systematically derived. Unlike [16], this paper outlines a systematic closed-loop block diagram of the predictive controller, which provides a clear feedback control structure for an IPMSM drive system. This is an unprecedented method of predictive control and also a major aspect of this paper.

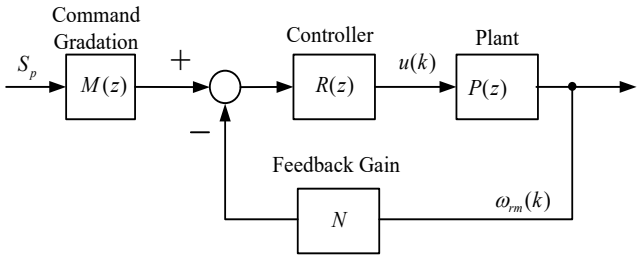


Figure 1 Block diagram of predictive speed-loop control.

2.3. Stability Analysis of the Proposed Drive System

The desired torque command of an IPMSM can be generated by the following equation

$$\frac{d}{dt}\omega_{rm}^{*} = \frac{1}{J_m}(T_e^{*} - T_L - B_m\omega_{rm}^{*}) \tag{30}$$

In the real world, there are parameter variations of the IPMSM and the external load measuring error. To compensate for that, the total equivalent load with uncertainty can be expressed as T_D, which includes the summation of the external load, the load influenced

by the parameter variations, torque tracking errors of both MTPA and flux weakening, and external load measuring error. Then (30) can be rewritten as

$$\frac{d}{dt}\omega_{rm}^{*} = \frac{-B_m}{J_m}\omega_{rm}^{*} + \frac{1}{J_m}T_e - T_{\mathrm{D}} \tag{31}$$

and

$$T_D = \frac{1}{J_m}(\Delta J_m \frac{d}{dt}\omega_{rm}^{*} + \Delta B_m \omega_{rm}^{*} + T_L) \tag{32}$$

where T_D is the total load disturbance, ΔJ_m is the variation of inertia, and ΔB_m is the variation of the viscous coefficient. Next, the dynamic speed equation of the IPMSM is

$$\frac{d}{dt}\omega_{rm} = \frac{-B_m}{J_m}\omega_{rm} + \frac{1}{J_m}T_e - \hat{T}_D \tag{33}$$

Where $\hat{T}_D$ is the estimated total load disturbance, which includes the influence of the motor parameters' uncertainty. The tuning step of the θ_{adv}, which is 0.18°, creates a varying torque that is less than 1% of the output torque T_e. As a result, the influence of the uncertainty is bound. The speed error, which is the difference between the speed command and the real speed, can be expressed as

$$e = \omega_{rm}^{*} - \omega_{rm} \tag{34}$$

Taking the differential of the speed error, one can obtain

$$\dot{e} = \frac{d}{dt}(\omega_{rm}^{*} - \omega_{rm}) = \frac{d}{dt}\omega_{rm}^{*} - \frac{d}{dt}\omega_{rm} \tag{35}$$

Subtituting (31) and (33) into (35), one can obtain

$$\begin{aligned}\dot{e} &= (\frac{-B_m}{J_m}\omega_{rm}^{*} + \frac{1}{J_m}T_e - T_{\mathrm{D}}) - (\frac{-B_m}{J_m}\omega_{rm} + \frac{1}{J_m}T_e - \hat{T}_D) \\ &= \frac{-B_m}{J_m}(\omega_{rm}^{*} - \omega_{rm}) - \tilde{T}_D\end{aligned} \tag{36}$$

We can define the load estimation error as $\tilde{T}_D = T_D - \hat{T}_D$. Assuming the external load changes slowly when compared to the quick responses of the speed control loop, it is possible to let $\dot{T}_D = 0$. The differential of the estimated load, therefore, can be expressed as follows

$$\dot{\tilde{T}}_D = -\dot{\hat{T}}_D \tag{37}$$

Define a lyapunov function as

$$V = e^2 + \frac{1}{k_1}(\tilde{T}_D)^2 \tag{38}$$

where k_1 is a weighting factor. By taking the derivative of the Lyapunov function, one can derive

$$\dot{V} = 2e\dot{e} + \frac{2}{k_1}\tilde{T}_D\dot{\tilde{T}}_D \tag{39}$$

Substituting (36) into (39), one can obtain

$$\dot{V} = 2e(\frac{-B_m}{J_m}e - \tilde{T}_D) + \frac{2}{k_1}\tilde{T}_D\dot{\tilde{T}}_D = 2\frac{-B_m}{J_m}e^2 + 2\tilde{T}_D(\frac{1}{k_1}\dot{\tilde{T}}_D - e) \tag{40}$$

According to (40), it is possible to choose an adaption law as follows

$$\dot{\hat{T}}_D = k_1 e \tag{41}$$

Substituting (41) into (40), we can obtain the differential of Lyapunov function as

$$\dot{V} = 2\frac{-B_m}{J_m}e^2 \le 0 \tag{42}$$

From (42), the differential of Lyapunov function is negative semi-definite. Then, Barbalet lemma can be applied. First, a function y_1 is set, which is equal to $\dot{V}$. Next, by integrating the y_1, we can obtain

$$\int_0^{\infty} y_1(\tau)d\tau = V(0) - V(\infty) < \infty \tag{43}$$

From (43), we can conclude

$$\lim_{t\to\infty} y_1(t) = 0 \tag{44}$$

According to (44), the speed error converges to zero as time approaches infinity.

3. MTPA Control and Flux-weakening Control

3.1. Basic Principle

The dynamic d-q axis voltages of an IPMSM can be expressed as

$$v_d = r_s i_d + L_d\frac{di_d}{dt} - \omega_{re}L_q i_q \tag{45}$$

and

$$v_q = r_s i_q + L_q \frac{di_q}{dt} - \omega_{re} L_d i_d + \omega_{re} \lambda_m \quad (46)$$

The total output torque of an IPMSM is

$$T_e = \frac{3}{2}\frac{P}{2}[\lambda_m + (L_d - L_q)\ i_d]\ i_q$$
$$= \frac{3}{2}\frac{P}{2}[\lambda_m i_q - (L_d - L_q)\sqrt{I_s^{\ 2} - i_q^{\ 2}}]i_q \quad (47)$$

Where I_s is the current amplitude. Two major control algorithms can be derived from (47) and can be used to adjust the torque of an IPMSM: field-oriented control and MTPA control. The field-oriented control sets a fixed d-axis current and adjusts the q-axis current to obtain linear proportional torque. The MTPA control sets a variable d-axis current, which is related to the q-axis current, to obtain maximum torque/ampere. The method can generate more torque than field-oriented control. However, the method creates a nonlinear relationship between the torque and q-axis current [3], which is a disadvantage of MTPA control. While it is a fact that MTPA's influence on torque linearity is a drawback, achieving maximum torque is a greater benefit that outweighs that disadvantage. The issue will be researched further by the authors of this paper.

Taking $\frac{d}{di_q} T_e = 0$ and substituting it into (47), we can obtain

$$\frac{3}{2}\frac{P}{2}[\lambda_m + (L_d - L_q)i_d - (L_d - L_q)i_q^2 \frac{1}{i_d}] = 0 \quad (48)$$

From (48), the d-axis current under MTPA control can be expressed as

$$i_d|_{MTPA} = \frac{-\lambda_m}{2(L_d - L_q)} - \sqrt{\frac{\lambda_m^2}{4(L_d - L_q)^2} + i_q^2} \quad (49)$$

The advance angle under MTPA control is

$$\theta_{MTPA} = \sin^{-1}\frac{i_d\ |_{MTPA}}{I_s} \quad (50)$$

and the current amplitude I_s is expressed as

$$I_s = \sqrt{i_d^2 + i_q^2} \quad (51)$$

In steady-state and neglecting the voltage drops of the resistance and inductance, it is not difficult to obtain

$$v_d = -\omega_{re} L_q i_q \quad (52)$$

and

$$v_q = \omega_{re} L_d i_d + \omega_{re} \lambda_m \quad (53)$$

The voltage amplitude is

$$v_s = \sqrt{(v_d^2 + v_q^2)} \quad (54)$$

If V_{om} is the maximum voltage amplitude, the voltage constraint can be shown as

$$(L_q i_q)^2 + (L_d i_d + \lambda_m)^2 = (\frac{V_{om}}{\omega_{re}})^2 \quad (55)$$

The maximum current amplitude I_{sm} can be expressed as

$$i_d^2 + i_q^2 = I_{sm}^2 \quad (56)$$

From (56), it is easy to obtain

$$i_q = \sqrt{I_{sm}^2 - i_d^2} \quad (57)$$

Combining (49), (54), and (57), the constraints of the IPMSM drive system can be obtained. The drive system operates under MTPA control when the motor speed is below the rated base-speed. After the drive system reaches its base speed, there are two constraints: the current constraint of the maximum amplitude, I_{sm}, and the voltage constraint of the maximum voltage amplitude, V_{om}. However, the $\frac{V_{om}}{\omega_{re}}$ is decreased as the motor speed increases, which is shown in Figure 2. As a result, the axes of the ellipse are reduced when the motor speed increases. When $\omega_{re} < \omega_{base}$, the motor operates under MTPA control, which is indicated as OCBA in Figure 2. When $\omega_{base} < \omega_{re} < \omega_c$, the motor operates in region II, in which the motor can operate either under MTPA control or flux-weakening control. If $i_d < i_d\ |_{MTPA}$, MTPA control is used; on the other hand, if $i_d > i_d\ |_{MTPA}$, flux-weakening control is used. When $\omega_c < \omega_{re}$, the motor operates in region III, in which only flux-weakening control can be used.

3.2. Proposed Real-Time Flux-weakening Control

In the real world, flux-weakening control is very complicated because it requires λ_m, L_d, and L_q, which are very difficult to measure [15].

To solve this difficulty, in this paper, a real-time tuning flux-weakening control is proposed. First, the voltage amplitude of the IPMSM, v_s, which is expressed in (54) is compared to the maximum voltage amplitude, V_{om}. The θ_{FW}, which is shown in Figure 3(a), is tuned as follows

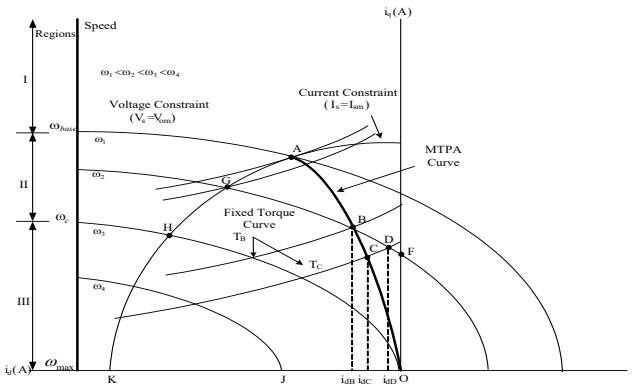


Figure 2 Operating regions of an IPMSM.

$$\theta_{FW}(\mathrm{k}) = \theta_{FW}(\mathrm{k}-1) + \alpha \cdot \mathrm{sgn}(v_s - V_{om}) \tag{58}$$

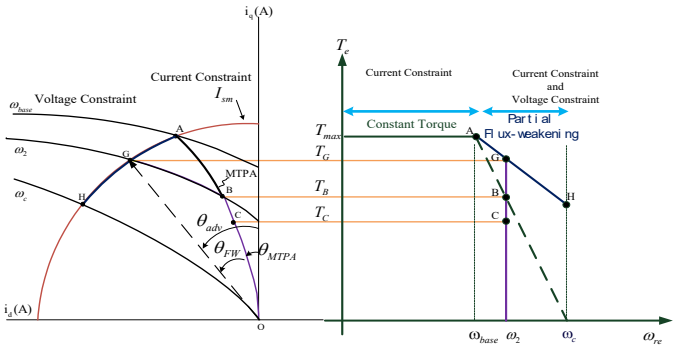
and

$$\mathrm{sgn}\ (v_s - V_{om}) = -1 \qquad \text{when } (v_s - V_{om}) < 0 \tag{59a}$$

and

$$\mathrm{sgn}\ (v_s - V_{om}) = 1 \qquad \text{when } (v_s - V_{om}) \geq 0 \tag{59b}$$

where α is the step size of the real-time tuning control. By using (58), (59a), and (59b), the θ_{FW} can be real-time tuned, as shown in Figure 3(a)(b). Finally, a block diagram of the proposed real-time tuning flux-weakening control is shown in Figure 4. First, the amplitude of the desired output voltage v_s is calculated. Next, the desired output voltage, v_s, is compared with the measured output voltage, V_{om}. After that, (58) is used to compute the $\theta_{FW}(k)$. Similar to previous research [2]-[5], [21], the proposed flux-weakening control method reduces the torque at high-speed operating range, and is not guaranteed the torque linearity. This is a common characteristic of flux-weakening control algorithms. However, the stability analysis of the whole drive system is discussed in the previous section of this paper.



(a) (b)

Figure 3 Operating curves and torque-speed curves of an IPMSM
(a) operating curves (b) torque-speed curves.

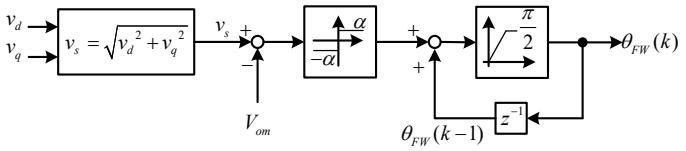


Figure 4 Block diagram of the proposed flux-weakening control.

4. Implementation

The implemented IPMSM drive system is shown in Figure 5(a) and (b). Figure 5(a) is a block diagram of the proposed IPMSM. The system consists of two main parts: the DSP system and the hardware circuit. The DSP system uses a TMS-320F-28335, which is a 32-bit, digital signal processor manufactured by Texas Instruments. The sampling interval of the current-loop control is 100 μs, while that of the speed-loop control is 1 ms.

Taking the MTPA operating region as an example, the controller works as follows. First, as shown in Figure 5(a) and based on the predictive speed controller, after the speed command is input and the motor speed is feedback, the q-axis current control i_{qp} is computed. Second, the load compensator compensates for the load disturbance $\hat{T}_D$ and then generates the compensation current $i_{q_{\hat{T}_D}}$. Then the total q-axis current command, which is the summation of the i_{qp} and $i_{q_{\hat{T}_D}}$, is computed and expressed as i_q^*. Next, the d-axis current is computed through (49) when the motor is operating in the MTPA region. After that, the advance phase angle θ_{MTPA} and stator current amplitude I_s^* are computed by using (50) and (51). Finally, the flux-weakening is obtained by adding θ_{FW}.

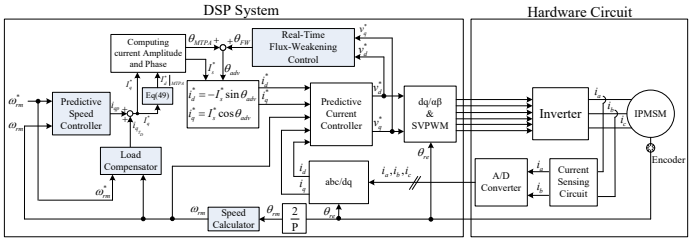
By using the MTPA table and the real-time tuning flux-weakening control, the advance angle of the current can be obtained. After that, the d-axis current command i_d^* and the q-axis current command i_q^* are calculated. A current regulated control and a space vector PWM modulation are executed. Finally, the switching states of the inverter are determined and output. A closed-loop system can thus be obtained. A DC motor is coupled with the IPMSM. The motor parameters are: $L_q = 31\ mH$, $L_d = 15\ mH$, $\lambda_m = 0.227\ v.\mathrm{sec}/rad$, and $R_s = 1.9\ \Omega$. The motor is 4-pole, 1-HP, and rated at 1500 r/min.

Figure 5(b) shows a photograph of the entire hardware circuit, which includes a DSP, gate drivers, peripheral devices, current sensing circuits, and an inverter. Figure 5(c) shows a photograph of the implemented IPMSM drive system, including an IPMSM and a dynamometer.

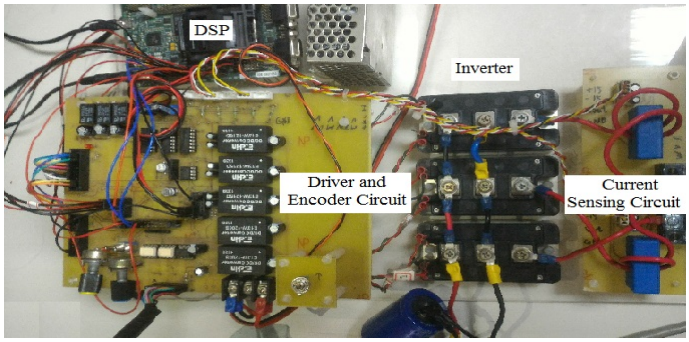
5. Experimental Results

The performance of the proposed flux-weakening control for an IPMSM drive system was evaluated according to experimental results. Figure 6(a) shows the measured steady-state torque-speed curves using the zero d-axis current control and the MTPA control. Figure 6(b) shows measured steady-state torque-speed curves

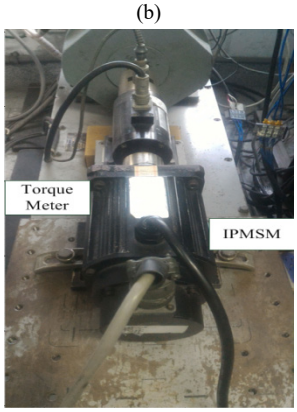
using the MTPA control and the hybrid control, which is a combination of the MTPA control and the flux-weakening control.



(a)



(b)



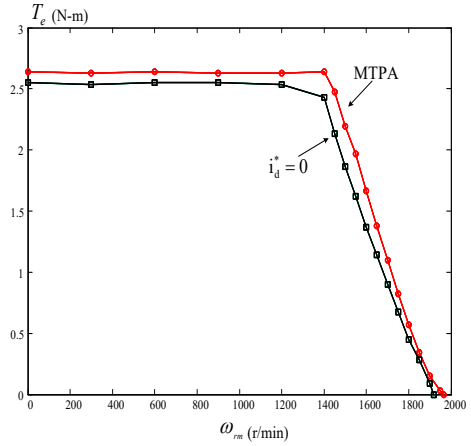
(c)

Figure 5 The implemented IPMSM drive system
(a) block diagram (b) hardware circuit (c) motor and dynamometer.

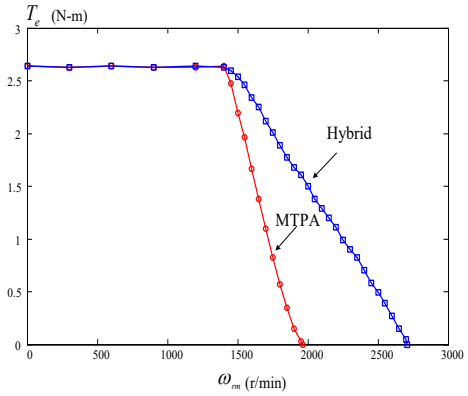
As you can observe, the hybrid control has better performance than the MTPA control. Figure 6 (c) shows the q-axis current to the d-axis current trajectory when the IPMSM is operating from the MTPA control region to the flux-weakening control region. The d-axis current increases slightly in the MTPA region; however, the d-axis current increases significantly in the flux-weakening region.

Figure 7(a)(b) and (c) show the measured waveforms of the IPMSM operating at 1800 r/min under loads of 0.57 N-m and 1.89 N-m. Figure 7(a) is the amplitude of the stator current I_s . Figure 7(b) shows the relative measured d-axis and q-axis currents. Figure 7(c) shows the measured advance angle θ_{adv} that is the sum of the θ_{FW} and θ_{MTPA}, which are the measured flux-weakening angle and the measured MTPA control angle. The delay of the measured responses is due to the lag between the load command and the external load that is generated by the dynamometer. Figure 8(a) shows the load disturbance responses when using the proposed predictive controller, the previous predictive controller that was proposed by [21], and the PI controller. The parameters of the PI controller are obtained by using the pole assignment technique. The locations of the desired poles for the closed-loop drive system are assigned as -4.06+j0.13 and -4.06-j0.13, respectively. The proposed predictive controller performs better than the previous predictive controller and the PI controller. At this 2000 r/min high speed, the field weakening algorithm is applied. The maximum allowed external load at this operating speed is 1.6 N.m. Experimental results show the proposed predictive controller performs the best. Figure 8(b) shows the repetitive 1.6 N.m load disturbance responses. The proposed predictive controller performs well again.

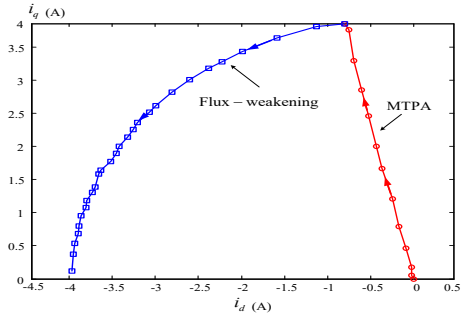
Figure 9(a) shows the measured step-input transient speed responses at the maximum speed, 2700 r/min. In this paper, the rated speed is 1500 r/min. The measured responses show the proposed predictive controller has a quicker rise time than the previous predictive controller proposed by [16] and the PI controller. Figure 9(b) shows the measured step-input transient speed responses at 3 r/min, which is the minimum operating speed. According to Figure 9(a) and Figure 9(b), we can conclude that the adjustable speed ratio of the maximum speed to minimum speed is 900:1. Figure 10 (a)(b)(c)(d) shows the current command and measured current errors using different controllers. Again, the proposed predictive controller has the smallest tracking errors. Figure 11(a)(b)(c) show the measured dynamic responses of the proposed flux-weakening control at 1800 r/min. A 5V step input at V_{om} is injected for every 1 second. The real voltage amplitude of the motor, which is expressed as v_s can track the voltage command V_{om} well. The rise time of the voltage control is about 50 ms. Figure 11(a), Figure 11(b) and Figure 11(c) show the responses of speed, voltage amplitude, and controlling angle θ_{FW} respectively. Figure 12(a) shows the measured efficiency of the whole drive system. The proposed method has higher efficiency than the linear flux-weakening method that was proposed by [21]. The major reason is that the proposed method provides greater output torque at the same motor speed. Figure 12(b) and (c) show the measured speed responses as the inertia of the drive system is increased to 5 times. The proposed predictive controller can provide a lower overshoot and a shorter settling time as the inertia increases.



(a)

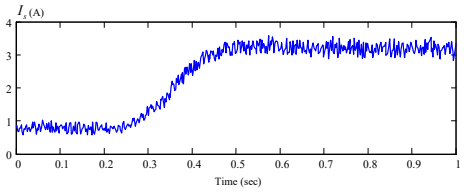


(b)

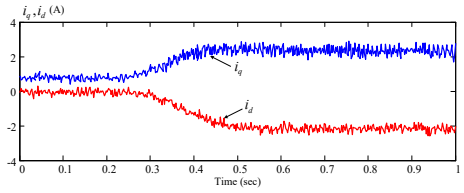


(c)

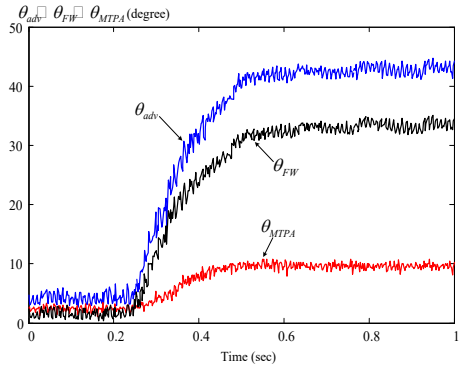
Figure 6 Measured steady-state characteristic curves
(a)MTPA and zero d-axis current (b) MTPA and hybrid control
(c) q-axis current to d-axis current curve.



(a)

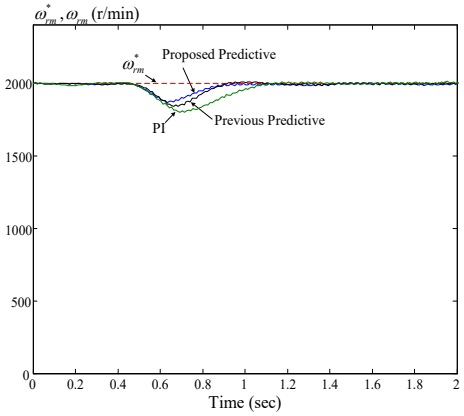


(b)

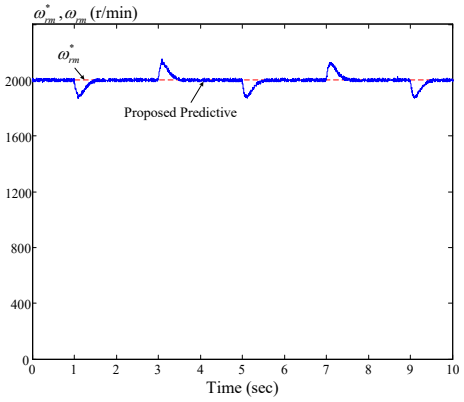


(c)

Figure 7 Measured 1800 r/min with loads of 0.57 N-m and 1.89 N-m
(a) I_s (b) measured d-q axis currents (c) advance angle.

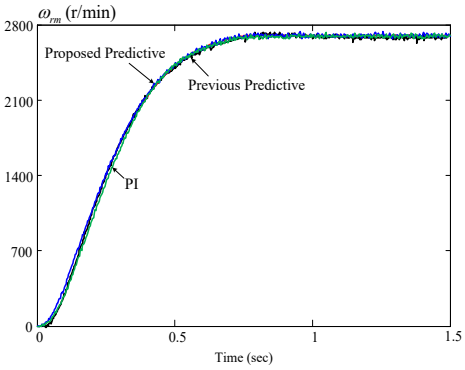


(a)

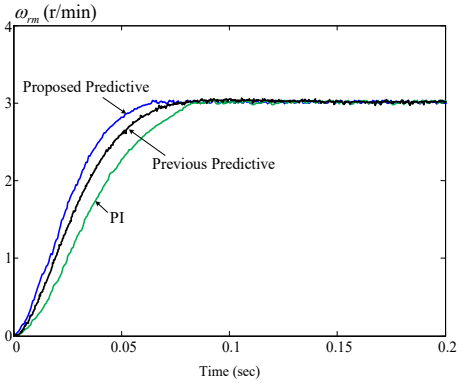


(b)

Figure 8 Measured field weakening operating speed with a large 1.6 N.m external load.
(a) step-input load (b) repetitive loads.

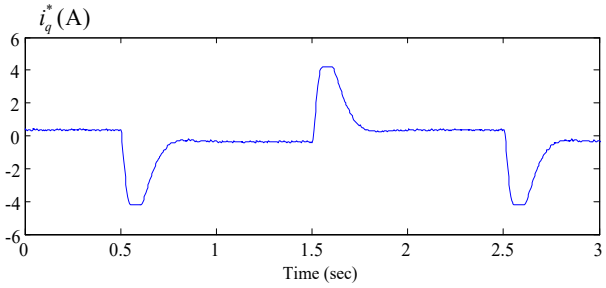


(a)

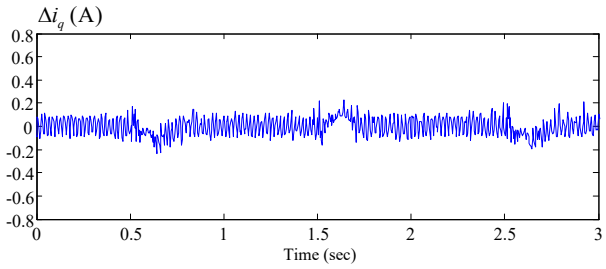


(b)

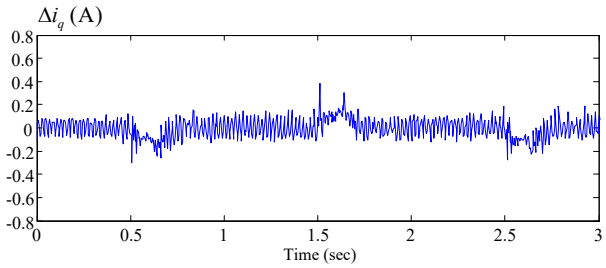
Figure 9 Measured step-input transient speed responses
(a) highest speed (b) lowest speed.



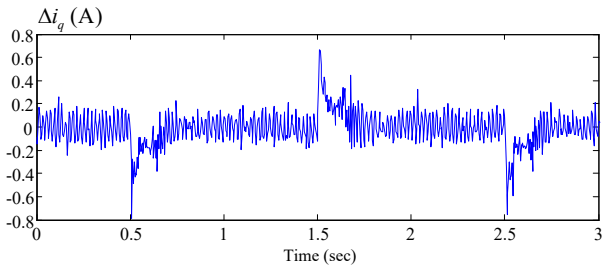
(a)



(b)

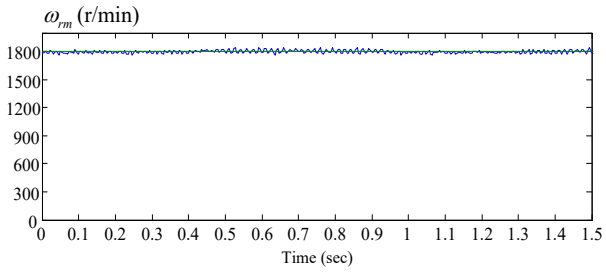


(c)

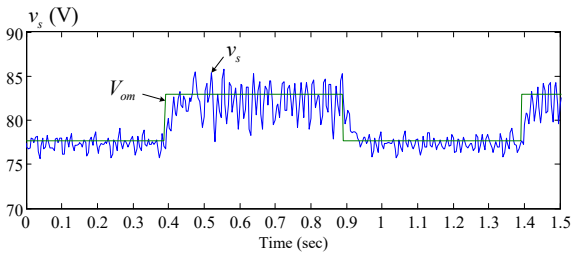


(d)

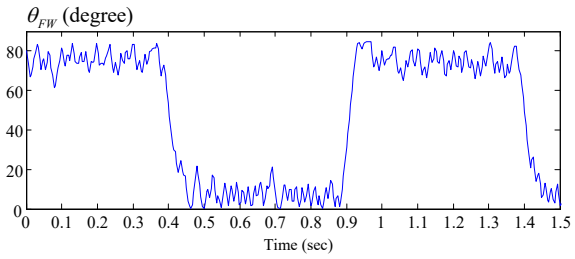
Figure 10 Comparison of measured current responses using different controllers (a) current command (b) proposed predictive (c) previous predictive (d) PI.



(a)

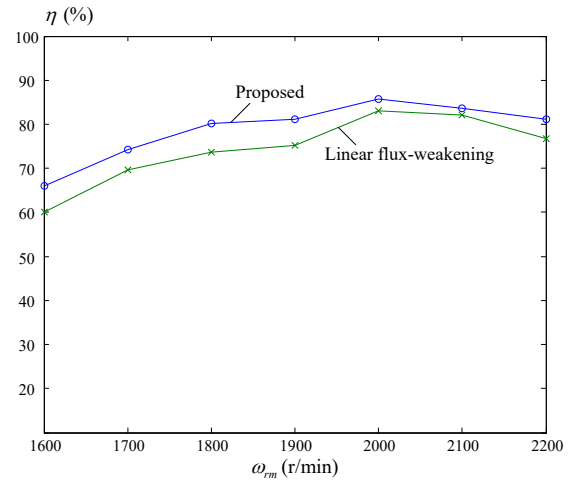


(b)

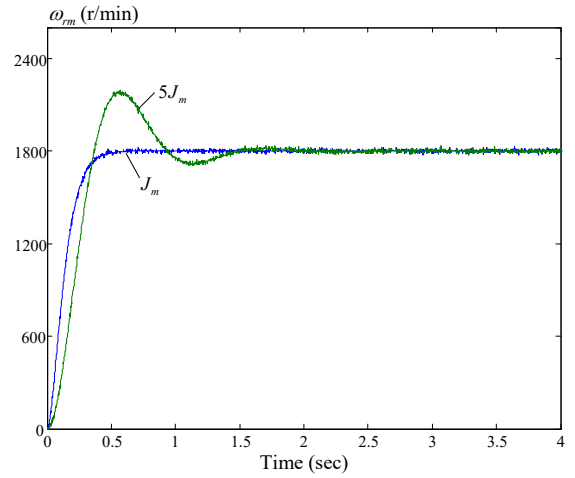


(c)

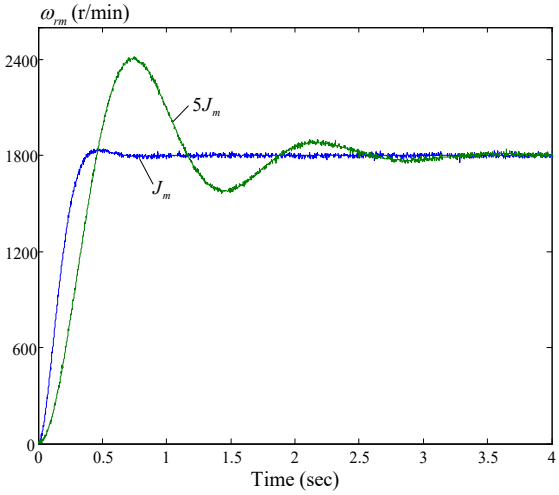
Figure 11 Measured dynamic responses of the flux-weakening control at 1800 r/min
(a) speed (b) voltage magnitude regulation (c) θ_{FW} .



(a)



(b)



(c)

Figure 12 Measured efficiency and varied inertia speed responses.
(a) efficiency (b) proposed predictive controller (c) PI controller.

6. Conclusions

In this paper, an MTPA control and a real-time flux-weakening control drive system using a predictive controller is proposed to improve the adjustable speed range and dynamic responses of an IPMSM. Experimental results show that the proposed drive system has good performance, including fast transient responses, good load disturbance rejection capability, and good tracking responses. Furthermore, the proposed method is easily implemented by using a DSP. Experimental results validate the theoretical analysis.

References

[1] T. H. Liu, S. K. Tseng, and M. B. Lu, "Auto-tunning flux-weakening control for an IPMSM drive system using a predictive controller," IEEE 26th International Symposium on Industrial Electronics (ISIE), 238-243, 2017.

[2] S. Bolognani, S. Calligaro, and R. Petrella, "Adaptive flux-weakening controller for interior permanent magnet synchronous motor drives," IEEE J. of Emerg. and Sel. Top. in Pow. Electron., **2**(2), 236-248, June 2014.

[3] H. Liu, Z. Q. Zhu, E. Mohamed, Y. Fu, and X. Qi, "Flux-weakening control of nonsalient pole PMSM having large winding inductance, accounting for resistive voltage drop and inverter nonlinearities," IEEE Trans. on Pow. Electron., **27**(2), 942-952, Feb. 2012.

[4] T. S. Kwon, G. Y. Choi, M. S. Kwak, and S. K. Sul, "Novel flux-weakening control of an IPMSM for quasi-six-step operation," IEEE Trans. on Ind. Appl., **44**(6), 1722-1731, Nov./Dec. 2008.

[5] C. T. Pan and J. H. Liaw, "A robust field-weakening control strategy for surface-mounted permanent-magnet motor drives," IEEE Trans. on Energy Conver., **20(**4), 701-709, Dec. 2005.

[6] M. N. Uddin and M. M. I. Chy, "Online parameter-estimation-based speed control of PM AC motor drive in flux-weakening region," IEEE Trans. on Ind. Appl., **44(**5), 1486-1494, Sep./Oct. 2008.

[7] S. M. Sue and C. T. Pan, "Voltage-constraint-tracking-based field-weakening control of IPM synchronous motor drives," IEEE Trans. on Ind. Electron., **55(**1), 340-347, Jan. 2008.

[8] S. Y. Jung, C. C. Mi, and K. Nam, "Torque control of IPMSM in the field-weakening region with improved DC-link voltage utilization,' IEEE Trans. on Ind. Electron., **62**(6), 3380-3387, June 2015.

[9] S. Chatithongsuk, B. Nahid-Mobarakeh, J. P. Caron, N. Takorabet, and F. Meibody-Tabar, "Optimal design of permanent magnet motors to improve field-weakening performances in variable speed drives," IEEE Trans. on Ind. Electron., **59**(6), 2484-2494, June 2012.

[10] K. C. Kim, "A novel magnetic flux weakening method of permanent magnet synchronous motor for electric vehicles," IEEE Trans. on Magn., **48**(11), 4042-4045, Nov. 2012.

[11] V. R. Jevremovic and D. P. Marcetic, "Closed-loop flux-weakening for permanent magnet synchronous motors," Conference on Power Electronics, Machines and Drives, PEMD 2008, 4th IET, 717-721, Apr. 2008.

[12] L. Harnefors, K. Pietilainen and L. Gertmar, "Torque-maximizing field-weakening control: design, analysis, and parameter selection," IEEE Trans. on Ind. Electron., **48**(1), 161-168, Feb. 2001.

[13] O. Wallmark, "On control of permanent-magnet synchronous motors in hybrid-electric vehicle applications," Technical Reports at the School of Electric Engineering, Technical Report, no. 495L, Chalmers University of Technology, Master Thesis, 2004

[14] N. Bedetti, S. Calligaro and R. Petrella, "Analytical design of flux-weakening voltage regulation loop in IPMSM drives," IEEE Energy Conversion Congress and Exposition (ECCE), ECCE 2015, 6145-6152, Sep. 2015.

[15] M. Pacas and J. Weber, "Predictive direct torque control for the PM synchronous machine," IEEE Trans. on Ind. Electron., **52**(5), 1350-1356, Sep. 2005.

[16] S. Bolognani, S. Bolognani, L. Peretti, and M. Zigliotto, "Design and implementation of model predictive control for electrical motor drives," IEEE Trans. on Ind. Electron., **56**(6), 1925-1936, June 2009.

[17] J. M. Kim and S. K. Sul, "Speed control of interior permanent magnet synchronous motor drive for the flux weakening operation," IEEE Trans. on Ind. Appl., **33**(1), 43-48, Jan./Feb. 1997.

[18] E. J. Fuentes and C. A. Silva, "Predictive speed control of a two-mass system driven by a permanent magnet synchronous motor," IEEE Trans. on Ind. Electron., **59**(7), 2840-2848, July 2012.

[19] M. Siami, D. A. Khaburi, A. Abbaszadeh, and J. Rodriguez, "Robustness improvement of predictive current control using prediction error correction for permanent-magnet synchronous machines," IEEE Trans. on Ind. Electron., **63**(6), 3458-3466, June 2016.

[20] T. Tarczewski and L. M. Grzesiak, "Constrained state feedback speed control of PMSM based on model predictive approach," IEEE Trans. on Ind. Electron., **63**(6), 3867-3875, June 2016.

[21] K. Chen, Y. Sun, and B. Liu, "Interior permanent magnet synchronous motor linear field-weakening control," IEEE Trans. on Energy Conver., **31**(1), 159-164, Mar. 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 82-93 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Perfect Ecosystem for Learning? Modern Thoughts for Organizing Higher Education

Pasi Juvonen*, Anu Kurvinen

Business and Culture, Saimaa University of Applied Sciences, 53850, Lappeenranta, Finland

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
experimental development ecosystem, team entrepreneurship, team learning

ABSTRACT

The background of the article lies in South Eastern Finland, in Lappeenranta, where an active University campus has attracted a group of ICT startups as well as SME's in the field to collaboration. The novel education approach also has another remarkable role as a developer and source of innovation. An experimental development ecosystem (EDE), where learning of knowledge, skills and character are combined, is presented in the article. Also future paths of the EDE are discussed. Companies in the ICT field worldwide are in constant need of competent experts who are ready to adopt the new tools and, at the same time, have an entrepreneurial mindset. We argue that inspiring students to learn through appropriate learning methods and providing them with a modern learning environment comes first. ICT tools, applications and systems to support learning objectives come second. The model presented in the article has been studied for some years as action research. The learning methods that have been found beneficial in IT and marketing bachelor education have been spread to other bachelor and master study programs as well. Results from the data show that students who study as team entrepreneurs have learned content knowledge, meta-skills and reflection skills via the learner-centric methods used in EDE. They have also been inspired to employ new ICT tools and applications to support their learning and project work.

1. Introduction

This article is an extension of the work originally presented in Educon 2017 in Athens in April 2017 [1].

The change in digitalization has been rapid and the change speed has been increasing during the recent decade. Since Facebook, we have seen plenty of new platform business models. Only some of them have succeeded and have been able to capture value for themselves. However, those who have succeeded have been able to capture themselves almost the whole market.

At the same time, many conventional businesses have run into severe challenges caused by rapid digitalization. To give an example, The Finnish Posti – a post delivery company owned by the government – has in recent years laid off employees due the diminishing amount of traditional mail deliveries. At the same time, agile delivery companies have challenged the Finnish Posti by offering customers faster delivery and several add-on services. The Finnish Posti has tried to find new business elsewhere, e.g. from lawn moving services to offering basic health care services.

Another example is the Finnish national railway company VR Group, which has held a monopoly for a long time (over 150 years). During the recent years, discussion about free competition in railway traffic has increased, and current plans may be deployed in the 2020s. Already, these plans have caused major reorganizations and layoffs within the VR Group.

There are more such examples found elsewhere in the world, e.g. Eastman/Kodak. When we consider what will happen in higher education, we can foresee what is going to happen in the future. How will platforms such as Khan Academy, Coursera and Udemy, and other more sophisticated platforms developed after them, change the current education systems? How many teachers will be needed in future to teach basic principles of programming, for example, when high-quality content can be freely loaded via learning platforms?

For us as educators and researchers an important question is: Are we going to build a shelter and try to resist the change, or would it be wiser to build a windmill and try to utilize the winds of change as well as possible?

*Corresponding Author: Pasi Juvonen, Email: pasi.juvonen@saimia.fi

In the world of today, there is plenty of information available. Thus, one has to be able to think critically, have skills to synthesize and put the information into action in a wise way. According to Fadel et al. [2], in order to deepen the learning process in the three essential dimensions – knowledge, skills and character qualities – an important dimension is needed – meta-learning. This means that there are some internal processes required for our learning, namely reflection and adaptation of our learning. Figure 1 presents the framework for the 21st century learner and for the curriculum redesign that tries to answer to these needs.

Our practical experiences with the same philosophy started several years ago. We – a group of lecturers - decided to start building a learning environment which would enable utilizing knowledge gained from different sources, combine theory, learning by doing and reflection, and make rapid changes possible when learning needs a change, without rethinking and redesigning the whole curriculum. This learning environment also provides practical measures to show students' progress.

The current curriculum (presented in section 2.1) is a result from coaching altogether 11 student teams comprising altogether more than 150 students from two areas of specialization - information technology and marketing. The development of the curriculum took place in 2009 - 2016. Since its early steps in 2009, this learning environment has expanded into a learning ecosystem that we call Experimental Development Ecosystem (EDE). This article describes the ecosystem and its pedagogical background, how the learning is organized within it, and the operational level practices that enable it. In this article, we present and discuss our current state of the art with the EDE, ongoing development activities as well as future development paths.

The article is organized as follows. Chapter 2 presents the results from a literature study about the requirements in bachelor education in Finland and in the OECD context. It also presents the current state of the art with the Experimental Development Ecosystem. Chapter 3 discusses data collection and data analysis. Chapter 4 lists some observations based on the data, and finally, Chapter 5 summarizes the findings and discusses our probable future paths on the subject.

2. Rapid changes will require rearrangements in organizing learning environments

Discussion on a right balance between studying explicit content that will, depending on the subject, easily become out-of-date, and meta-skills that are useful but, at the same time, may leave learners with an experience of not learning anything specific has been going on for a long time. Skills that are easy to teach and learning that is easily measured involve skills that are easily automated [3]. We are, to a certain extent, educating young people for future professions that do not exist when decisions about students' curricula are made. These new professions emerge (and some others disappear) while students are studying for their diploma.

Learning is less about reproducing content knowledge. It is more about extrapolating what we know in novel situations [3]. In future, more employees with versatile skills are needed and, at the same time, fewer specialists with deep expertise in one subject are needed. Communal learning skills and team working in multidisciplinary working groups or teams are important [3].

The demands presented for undergraduate education are versatile. Based on a study of literature [3-10], an undergraduate student needs at least the following skills:

- team working
- communal learning
- problem solving
- creativity and innovativeness
- critical thinking
- decision making
- leadership and self-leadership
- shared expertise
- reflection on one's values, and social and emotional skills.

At the same time, learning environments built for supporting learning should

- offer versatile methods for learning
- diminish teacher-led methods
- provide coaching for the constantly changing world
- foster entrepreneurship
- enable running pilots (e.g. establishing cooperatives).

[3-10].

To get an overall picture of what should be considered when organizing higher education in 2017, we looked for a framework for identifying knowledge, skills, meta-level skills and methods. We also wanted to make a cross-section of the EDE compared to other frameworks. The framework we chose is presented in Figure 1.

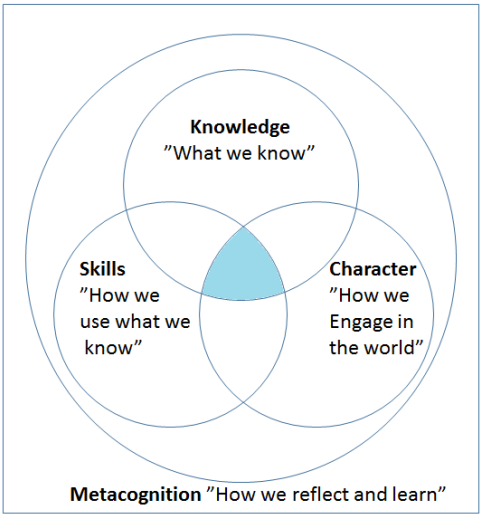


Figure 1. The framework for 21st century education by Centre for Curriculum Redesign [2].

According to [8], there are four forces that lead the learners towards new ways of learning for life in the 21st century. These forces are:

1. Knowledge work – employees who use brain power as well as digital tools for creating new solutions collaboratively in teams

2. Thinking tools – knowledge workers use a set of digital tools, devices and services
3. Digital lifestyles – today's student generations are born into the digital society, and grow up with the digital devices
4. Learning research – the latest research on learning during the last three decades has deepened the understanding of learning processes (See [34 – 37], [39-40], and [51-60] for more)

2.1. State of the art at our UAS – Experimental Development Ecosystem (EDE)

Since 2009 we have been developing a new learning environment, combining studying content knowledge (theory), learning by doing (practice), and employing dialogue [11, 12] in knowledge sharing, knowledge creation and reflection. Together with several local companies and municipalities, we have been able to build the Experimental Development Ecosystem. The current ecosystem offers our undergraduate students an excellent platform for both studying content knowledge, applying this knowledge in real customer projects, and reflecting what has been learned by doing with other team members and the team coach.

Since 2013 we have been developing the EDE in parallel with several RDI projects supported by the Finnish Funding Agency for Innovation (TEKES), Saimaa UAS and LUT University. An "easy-to-start" cooperation between stakeholder groups with little or no thresholds has been a cornerstone for the further development of the EDE.

The EDE was originally created and developed as a social innovation; it is a novel way to organize bachelor education for IT and Business Administration students specializing in marketing at Saimaa UAS. Tiimiakatemia [13] in Jyväskylä and Proakatemia [14] in Tampere were studied as models of how the core structures of team learning can be established. Both of the above academies are specialized in entrepreneurship. An extensive study of how team learning and team entrepreneurship has been organized at Saimaa UAS with IT students to support entrepreneurship education was carried out by Juvonen in 2014 [15].

In 2014 there was no other learning environment designed to support entrepreneurship education in IT Bachelor education [15]. The current employment of the EDE seems to be the only implementation of team entrepreneurship where a student can specialize in marketing (other deployments focus on entrepreneurship overall) and study as a team entrepreneur combining theory, practice and reflections within the EDE.

Studies on learning environments designed for supporting entrepreneurship education in higher education [16- 31], focus on promoting entrepreneurship by fostering either the mindset or skills needed in entrepreneurship, or focus on increasing the status of entrepreneurship as a career choice. Most of the studies we found from Scopus and Science Direct databases (from year 2010 until now) had a narrow focus either on a single course or group of students or a certain technique to foster entrepreneurial skills or thinking. We summarized these studies as follows:

- Applications, tools and methods for fostering entrepreneurial mindset, skills or intentions [19 – 25]
- Evaluation of entrepreneurship education programs overall [26, 27], perceived of value of entrepreneurship education program [27], evaluation of methods used in entrepreneurship education [28]
- How demand and supply meet on entrepreneurship education [30 - 32]
- Student entrepreneurship [33]

Applications with quite a similar approach than the EDE described in this study were found only at Lund University in Sweden and at University of Southern Denmark [18]. At Lund University, practice-based courses and projects are offered, and best practices are spread within an entrepreneurial ecosystem. At the University of Southern Denmark, there are many entrepreneurship education related courses, where different learning methods are used. Most of them are extra-curricular courses, which complicates student participation. Furthermore, basic concepts are taught in a conservative way and practical issues are learned via intensive courses, or via "real entrepreneurs" as visiting lecturers. An active reflection process has been found effective to unleash creative and innovative thinking potential [18].

Students in these two applications are not studying as team entrepreneurs, so they are not running and developing their own enterprise while they are studying. Rather they are participating in a series of teacher-led courses. Based on these differences, the current deployment of the EDE described in this study is considered as a novel approach to entrepreneurship education.

Also concepts of transformational learning [34] have been applied to service-learning while performing service work in [35]. The EDE is more than a team learning environment (Figure 2).

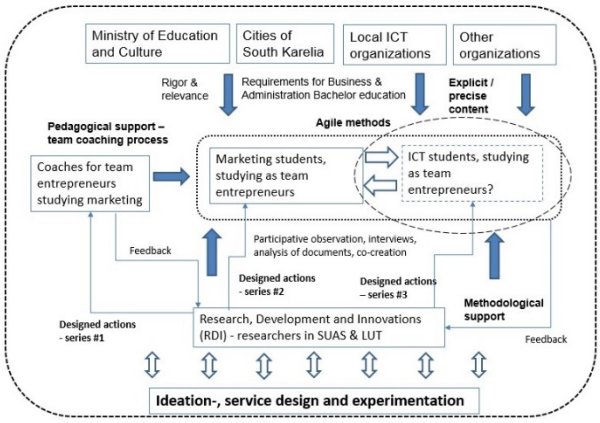


Figure 2. The Experimental Development Ecosystem. Adapted from [36].

The novelty of our approach (the EDE) is in combination of several factors. At the same time this combinations derives the EDE from the approaches found in literature. These factors are as follows:

- Students learn as team entrepreneurs by running a company (a cooperative) they own by themselves
- Focus for students is on specializing in learning marketing in versatile ways
- Experimental learning is emphasized - theory, practice and reflection is involved all the time

- A team coaching process where team development, team performance, and team leadership body of knowledge is applied to support learning
- Several stakeholder groups are involved in the implementation of the EDE
- Continuous cooperation between team entrepreneurs, team coaches, and researchers
- The curriculum has been rebuild to support the team learning and team entrepreneurship and it is further developed based on the experiences gained

Currently, Business Administration students who choose to specialize in marketing study their first year in a conventional way, enrolling on conventional study courses. After choosing the specialization in marketing, they continue their studies for two and half years as team entrepreneurs (Figure 3).

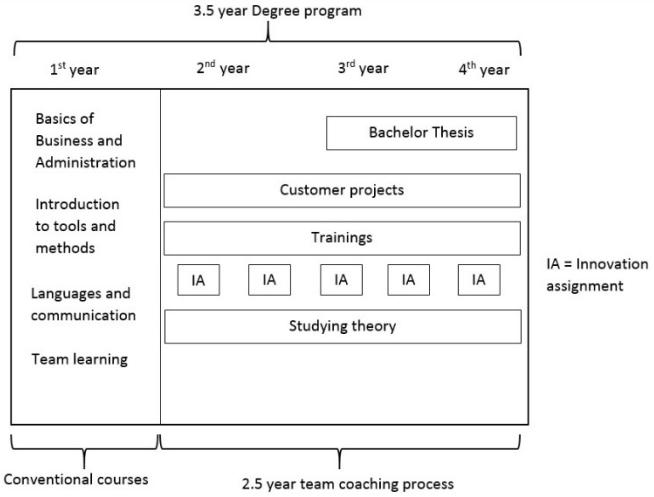


Figure 3. Overview of the curriculum for marketing students studying as team entrepreneurs.

The students who choose to specialize in marketing establish a team enterprise (in the form of a cooperative) and run its operations together during their studies. Their studies consist of studying theory (reading books and articles), carrying out customer projects for real customers for real money, and trainings, where a team coach (personnel of Saimaa UAS) coaches the team members and the team enterprise for team development and performance.

Progress of the team entrepreneurs' studies is measured by five practical measures. Four of these measures show the balance between theory, learning by doing, and sharing knowledge, and reflection. These measures are:

- amount of book points (3 book points equal to 1 ECTS point)
- amount of trainings (in hours, 133 hours equal to 5 ECTS points)
- amount of projects (in hours, 80 hours equal to 3 ECTS points)
- amount of innovation assignments (in hours, 80 hours equal to 3 ECTS points).

To give an example, when a team entrepreneur participates in every training session, which take place twice a week, she will get 5 ECTS points for trainings in a half year. Furthermore, when a team entrepreneur works for 16 hours per week in projects, she gets about 6 ECTS points for projects in a half year. Innovation sessions are held at least twice a year, and participating in them increased the ECTS points. The amount of book points is the only individual measure for the team entrepreneurs. Every team entrepreneur has to complete at least 102 book points, which equals 34 ECTS points. This means that every team entrepreneur has to read at least two business books per month. Book trainings are held twice a month, and those who have read a book and returned a book essay before a book training session are allowed to participate.

The measures described above produce the overall results that are monitored at Saimaa UAS, i.e. the numbers of students who complete at least 55 ECTS points per academic year (1.8 – 31.7). This measure is used by the Finnish Ministry of Education and Culture to monitor study progress in every university of applied sciences in Finland.

The team entrepreneurs are an important element of the ecosystem as collaboration partners and as a scalable source of creativity and innovativeness. Several team entrepreneurs have been recruited already during their studies by local companies. When a company and a team entrepreneur start cooperation on a customer project and continue cooperating on an internship scheme and/or Bachelor's thesis project, it is common, based on our experiences, to continue cooperation after the student has graduated.

The curriculum has been adjusted to make close cooperation with local companies and other organizations possible and fluent. When local companies and other organizations are not able to produce the knowledge they need on their own, they can ask team enterprises for help. A usual method of helping the local organizations is an innovation assignment (IA in Figure 2), where new knowledge is produced within a 12- or 24-hour time limit. Methods of experimental development are used in these assignments in order to create fast, concrete and applicable development ideas based on companies' current needs.

Innovation assignments serve several purposes. For local organizations, they function as a method of rapidly testing their assumptions on a certain topic. For team entrepreneurs, they provide an environment to develop problem-solving skills. Moreover, the assignments function as a measure of team development and substance skills. For all parties, the innovation assignments make it possible to find further cooperation opportunities. In most cases, the organizations participating in an IA will become cooperation partners in the EDE. Sometimes they also take part in RDI projects. Prototypes and/or concepts are usual outcomes of IA's, and prototyping is an important step in an experimental development process. This is the phase that can make a difference in comparison to the traditional workshops where the ideas are easily forgotten after the workshop. It is also easier for the client company to understand the idea and its possibilities for the company after seeing the prototype.

Also research, development, and innovation (RDI) projects support the current EDE model by offering team entrepreneurs

opportunities to work as research assistants. As research assistants, they learn about research methods, gain an understanding of organizing research processes and ways of producing value for local companies [15]. Acting as research assistants in an RDI project has led to employment via internships and/or Bachelors' thesis projects.

2.2. Pedagogics used in the Experimental Development Ecosystem (EDE)

The pedagogics used in the EDE follow ideally the cycle of learning as described by Kolb [37], where learning is described as a process: Firstly, concrete experience is gained, and as the next step in the process, the experience is reflected on. After that comes learning from the experience through abstract conceptualization, and finally testing the newly adopted knowledge and skills through active experimentation.

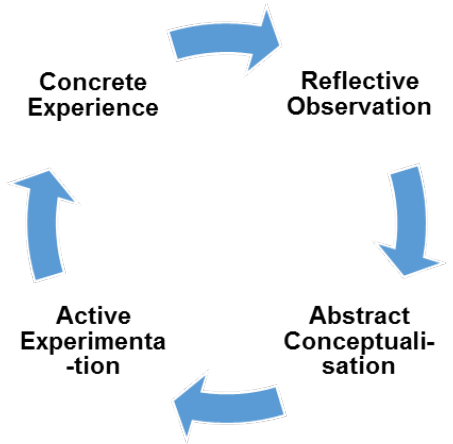


Figure 4. Cycle of learning, adapted from [35].

In constructivism, the learning of a human being is understood as a constant process where individuals are learning or creating their own understanding based on interaction between what they already know and believe, and ideas and knowledge with which they come into contact [38]. Constructivist learning involves at least the following five areas: 1) the educator's attention to the learners, the students and their backgrounds, 2) dialogue facilitation with the group with the purpose of creating a shared understanding of the topic, 3) planned or unplanned introduction of formal theory into the discussion, 4) creating opportunities for the students to challenge or change the existing beliefs and conceptions by using tasks that are structured in a way that makes this possible, and 5) developing students' awareness of their level of understanding and the learning process [39]. In addition to constructivism, team learning, as well as open and honest dialogue, are proven to support the learning objectives.

In the EDE pedagogics, the complexity of companies' operating environments can be learned via concrete experience. At the same time, the theories learned are linked to the reality. Here the principles of building an innovative learning organization are utilized [40]. When considering the current needs of the ICT field, ICT students require more insight into business operations, and practicing business. One way of supporting this is working together in shared projects with team entrepreneur students. Working in multi-disciplinary and heterogeneous project groups is analogous to working life experience, as project teams consist of experts from different fields. Reflection is an elemental part of the learning process and reflective dialogue is used as a pedagogical tool for deepening the learning experience. The researchers working in the ecosystem support the learning process, as they help the students in abstract conceptualization of the phenomena learned. The persons working in the ecosystem need coaching skills that can be applied to the EDE's needs as is required.

These modern thoughts about higher education will also require new skills from those who are employing the system with students. Instead of transferring information to the team entrepreneurs, the emphasis in the team coach's work is rather on helping the process of team development and facilitation of the learning of the information and skills that the team members need for their collaboration and learning (see [41] for more). These teamwork skills and competencies include adaptability, shared situational awareness, performance monitoring and feedback, leadership and team management, interpersonal relations, coordination, communication, and decision making. When these are managed with success, high commitment to learning can be achieved. High commitment usually leads to high performance [42].

A team coach should also make oneself familiar with different styles of consultation. These styles include acceptant, catalytic, confrontational, and prescriptive styles [43]. In the acceptant style, feelings are involved, and this style can be described as emphatic listening. The catalytic style helps the coachee to make decisions. In the confrontational style, the team coach points out what will follow if the coachee continues with her current behavior. The prescriptive style is common in conventional pedagogics. It gives direct advice; however, it does not offer the coachee any opportunity for growth. All these styles are needed in team coaching.

3. Data collection and analysis

The research framework for investigating all the development activities concerning the Experimental Development Ecosystem has been action research [44, 45]. During September 2015 – September 2017 there were several mini-cycles where designed development activities were carried out in different parts of the EDE, and the team entrepreneurs were active participants in many of these development activities.

The outcome of these development activities was monitored through participative observation. The team coaches are active agents for change when they act with student team entrepreneurs. By choosing to use qualitative methods of inquiry, the authors have, at the same time, committed themselves to continuous reflection of their own values and how they affect the research.

The field notes have provided valuable qualitative data, which has been analyzed with other researchers. Two other team coaches have been involved in the sense-making process of how to utilize

the current EDE in ICT education and how to foster cooperation between team entrepreneurs specialized in marketing and ICT bachelor students.

Survey and theme-based interviews during a one-year period between 10/2016 - 9/2017, followed by participative observation, have served as the main data collection methods for the study. An overview of the collected data for the study is presented in Table 1.

Table 1: Overview of the data.

Data collection method	Amount	Timeline	More information
Theme-based interviews	12	10/2016	13 specialists in 12 interviews, from different stakeholder groups involved with the EDE, were interviewed to find out development targets for the EDE.
Participative observation sessions	Over 50	Training sessions, four hours twice a week between 9/2016 – 9/2017	Team coach acts as an active agent for development and provides examples of digital tools for team entrepreneurs.
Theme-based interviews	34	3/2016 – 9/2017, lasted 15 – 40 minutes	Evaluation of the role of the ICT / Digital tools and reflections on why and how team entrepreneurs have been deploying them.
Survey	14 replies	8/2017 – 9/2017	E-mail survey followed by interviews as a part of development discussions.

Methodologically, this article is a partly descriptive and partly explorative case study [46, 47]. It presents the current implementation of the EDE, explores its possible development paths, and finally describes how the ICT Bachelor curriculum could benefit from it.

The basic assumption of the authors is that every research is value-laden and biased. By choosing to use qualitative methods for the inquiry, the authors have, at the same time, committed themselves to continuous reflection on their own values and how they affect the research. In this study, the objective has been the further development of the current EDE and integration of the ICT Bachelor education into it at some level. Therefore, there is an inbuilt bias in the observations and interventions made. However, the authors present the interpretation based on the analysis and the process of conducting research transparently and leave judgement of the validity of study to the reader.

The results of the study have been discussed with three team coaches and two researchers. Luckily, the team coaches share an office at the campus. This has helped to test inner validation of the observations made based on the interviews. The data for the article consists of qualitative interview material (12 specialist interviews with open-ended questions, notes on direct and participative observations, 10+ steering group meetings, researcher workshops, other workshops and meetings), and several unofficial discussions with colleagues and administrative staff at the Saimaa UAS campus and elsewhere where the authors have been actively involved in development activities.

Multiple sources of data and close cooperation between the two authors made it possible to utilize both investigator triangulation and data triangulation [48]. The triangulation of data and researchers has helped to test inner validation of the observations made based on the interviews and participative observation sessions. The data was analyzed applying the principles of grounded theory [49, 50]. The grounded theory analysis includes three main phases: open coding, axial coding, and selective coding [49], and the method requires the researcher theoretical sensitivity [51]. The researcher cannot force the data, but instead she has to let the data "speak". Naturally, this phase is extremely hard in cases where interviewers have a lot of pre-existing knowledge about the subject studied. The process can be made easier by asking the same open-ended questions about the subject studied from all the interviewees and carefully listening to and reporting their expressions.

In the open coding phase, interesting phenomena in the data were highlighted. In this study, the interview notes were first gathered into one text file and then analyzed by the two researchers. In the axial coding phase, the interesting phenomena marked in the open coding phase were grouped and their relations (causal and other) were analyzed. In the selective coding phase, a lot of data was abandoned, the core of the results - "What is going on here?" - was taken, and the research reports were written. An example of axial coding phase where i.e. associations and causal relations were searched by visualizing interesting phenomena with Atlas.ti software is presented in figure 5.

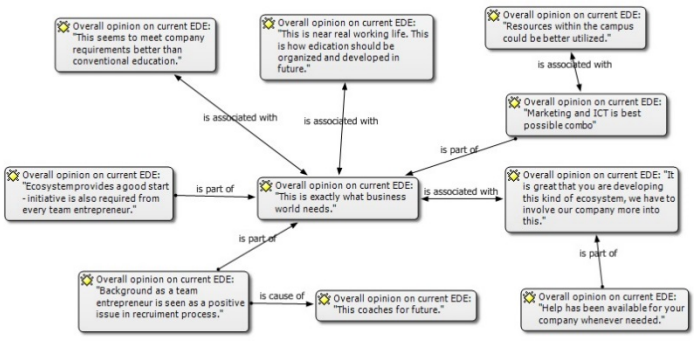


Figure 5. Axial coding with Atlas.ti code network view.

As mentioned above, the grounded theory analysis lets the data speak, and therefore no pre-existing theory is needed. In an ideal case, the grounded theory analysis is purely inductive. In practice, there always exists little or more pre-existing knowledge and bias related to the research subject. To be exact, a target to develop something is already a strong bias. Who defines development? Development for one stakeholder group may be stagnation for another group. When discussing values and biases, the best we can do as researchers is to be as open as possible about the motives we

recognize in ourselves. In this study, the researchers had a strong vision of how ICT Bachelor education should be organized in future. However, the interviewees were asked open-ended questions about the subject, and the results are presented as they are with no value-laden interpretations added.

Open coding and axial coding phases went on in parallel. During the open coding and axial coding phases, a constant comparative method and theoretical sampling [51, 52] were used. The use of the constant comparative method meant that when an interesting phenomenon was found later during data collection, all the earlier data was screened again to see if the phenomenon was found there, too. Theoretical sampling was used to collect more data on phenomena where more explanation was needed. The data was collected until a theoretical saturation was achieved. The use of the constant comparative method and theoretical sampling increased the researchers' interaction with the data. The selective coding phase also started early and parallel with the axial coding phase. The saturation of the new data [51] took place early, which meant that there was no need for extra interviews on that subject. A new tool for making word clouds was used to visualize part of the data (see Figure 4 for more details).

During the first round of interviews made with stakeholder groups, two seed categories [52] emerged during the open coding phase: 1) creating interdisciplinary interaction during ICT education, and 2) entrepreneurial mindset and its value. Based on an analysis of these seed categories and combined with the analysis of the new field notes (including memos, emails, book essays, and observation data) between January 2016 and mid November 2016, a pattern of "Where should we be heading?" was written. The results were published in a conference article in the Educon 2017 conference in Athens, Greece on 27th April 2017.

At the same time when employing designed actions to the Experimental Development Ecosystem learned from the previous study, two new research questions were chosen for this study:

1. *What are the learning experiences when dialogue is used as learning method?*
2. *How new ICT tools and applications have been utilized by team entrepreneurs during the first year?*

Between mid of November 2016 and mid November 2017, plentiful new data was gathered and analyzed. This data collection comprised two interview rounds as part of the team entrepreneurs' development discussions, over 50 training sessions (where team entrepreneurs and the team coach were present), and a survey. The survey was presented to the team entrepreneurs face to face and they were able to ask further questions about the purpose of reflecting on these issues. The team entrepreneurs were asked to evaluate their dialogue skills, make observations on dialogue in other groups they had visited, and describe how they had employed ICT tools in their project work. The expressions made by 14 interviewees were later observed in training sessions, project reflection sessions, and finally by end products made with the ICT tools that had been utilized.

4. Observations based on the data

In general, the majority of the interviewees within our stakeholder groups shared the opinion that the ecosystem model (EDE) in modern education is a response to many challenges in the employment of the newly graduated, and at the same time it is a way to pave the way to embarking the working career smoothly after studies. The students' early adoption of an open view of the real business life was generally seen as a positive and novel way of educating experts for the business world, where the described individual capabilities are appreciated by employees.

Entrepreneurial mindset and thinking as a driving force for students was seen as a positive feature to be maintained in education. The necessity of establishing a cooperative and working as team entrepreneurs was seen differently among the interviewees. Some of the interviewees from Saimaa UAS did not see that learning to run an enterprise would bring much value to ICT students. The interviewees from industry, however, saw that learning to think and operate as entrepreneurs would be one of the most important topics to learn overall.

As the most promising pathways, the interviewees saw the possibilities of interdisciplinary interaction between marketing and ICT students, who all are familiarized with the ICT business and the ICT customers' businesses. During the studies, the two student groups could make use of each other's specialty areas by participating in shared projects either in RDI or in business collaboration projects. As the study aims at understanding the future requirements for a competent ICT education, several ICT company representatives were interviewed. As a way to prepare competent future employees for the ICT field, the interviewed ICT companies mentioned a possibility to participate directly in the education of ICT students via both direct education activities and placement opportunities or shared customer projects in RDI, for example. This would create a firm an efficient recruitment process with less need for training the newly recruited personnel.

The interviewees that have graduated from the team entrepreneurship ecosystem or are currently studying marketing in that environment had positive experiences of this type of an education system. The most important and positive feature mentioned was learning through real business cases. The most valuable part in the model was the early networking with businesses. After graduation, it has been easy to find a job relating to the field of studies. At least, this different form of studies in a cooperative has proven a positively differentiating factor in job applications.

ICT business is mostly a project-based business where experience and an entrepreneurial mindset are a benefit as such. The current marketing team entrepreneurs have started to build also cooperation with students from different sectors of education within Saimaa UAS to be able to expand their domain knowledge. Making oneself familiar with different contexts where marketing activities (and ICT) are carried out is another example of a positively differentiating factor for the job applicant.

The ICT education ecosystem and ICT education in general need to respond to the changing skill requirements in the industry. In the EDE this can be promoted by involving the ICT sector in the constant development of the study programme and in the education activities in special projects. This would ideally result in long-term interaction between the ICT industry and the Saimaa University of Applied Sciences (SUAS) and the students. In some cases, even weekly co-operation between the ICT companies and the coach

and students would form a mutually beneficial cooperation model supporting a high level of motivation among both the students and the ICT companies. When mutual benefits exist and the captured value will be fairly shared between the parties, the cooperation will sustain itself.

There were some concerns towards the coached team learning and team entrepreneurship model raised in the interviews. In general, the team coaches have a key role and they are responsible for ensuring that the students are provided with the basic capabilities required in the education programme at a University of Applied Sciences. It was also stated that this type of a learning method is not ideal for students who still have to grow in order to mature for the responsibility and self-leadership skills that are required by the learning style. The entrance examination for the students was seen as one important step in this process. Also, the new ICT education should be marketed as a more interdisciplinary education program that is not too much technically oriented in order to lure the business oriented people with a high motivation for fast career building.

As the student teams operate in groups or as working teams, it would be beneficial if the different team roles could be consciously tested by each student. This would help the young students safely test their own personal strengths, which would be a supporting factor in building the students' professional self-esteem. As all companies, also the student teams have to set goals for their actions and all the steps during the studies should be taking the team towards the goals. It is not enough to act as active since that does not suffice to develop the situation. Supporting the building of students' versatile knowledge base requires a broad scale of practical learning projects, which has to be ensured during the studies. Here, the role of the team coach is emphasized again. A frequent presentation discussion of the learning goals and how they have (or have not) been achieved is crucial.

In the Business Administration degree program, the number of students who complete 55 ECTS points per year has been monitored since 2014 when criteria for funding University of Applied Sciences was changed. However, the measuring criteria were changed again in 2016, so we do not yet have enough new data to be able to make valid arguments about overall results. In the academic year 2015 – 2016, the overall result in achieving 55 ECTS points in the Business Administration and International Business degree programs (measured together) was 57.1 % (68.1 % with team entrepreneurs). In the academic year 2016-2017, the figures were 61.3 % (58.7%) respectively. The measuring system is now being updated to be able to provide more accurate data in future.

Even though we do not have enough performance data yet from team entrepreneurs' studying within the EDE, we are able to argue something. One of the biggest worries at the beginning of team learning and team entrepreneurship was how the increased freedom of students in choosing what they study and when will affect the performance measured in ECTS points. With this small amount of data we are able to argue that the pedagogics used with the EDE is working and no one has to worry about the performance metrics.

Two of the interviewees strongly emphasized that in future the EDE should emphasize the role of internationalization of students and companies cooperating in the EDE. This kind of organization of learning, development, and research activities is not familiar elsewhere and therefore the EDE ideas should be exported to other countries as well. The main strengths that this kind of ICT education would offer include the digital thinking of "Digi-native" generations together with marketing capabilities. These make a combination that a successful and competitive international business requires. The students would also, right from the beginning, start to understand that their future work will be part of someone's business – if not their own.

4.1. Towards a perfect ecosystem of learning

The key findings, as also described in pedagogic research literature that have proved to be successful in the researched EDE model are the following: authentic learning, mental model building, internal motivation, multiple intelligences as well as social learning. The curriculum applied in Saimaa UAS, as well as the piloted methodology applied on certain study courses, aims at developing skills needed when coping with the complexities of the 21st century: team working skills, communal learning skills, problem solving skills, leadership and self-leadership skills as well as innovativeness, shared expertise, and ability to reflect on one's own values and attitude.

These suggestions follow the advice from studies on how to build great teams. They suggest that energy (how team members contribute to a team as a whole), engagement (how team members communicate with one another), and exploration (how teams communicate with one another) are crucial factors for teams' success [53]. Furthermore, when quality of communication includes balance of advocacy and inquiry, a lot of positive feedback and true interest in others' opinions [54], there are many of the required elements available to develop great teams. Applications selected should support the objectives listed here, and this means they cannot be selected from a teacher-centric viewpoint.

Students' experiences from learning dialogue skills are proving that practicing dialogue in a safe environment gives them valuable communication exercise. It also supports the readiness for dialogue of the future company developers. Within one year's time every respondent said that there is development gained in one's own dialogue skills. Several respondents stated that they already are good at listening to other students. Several stated that they have learned to express their own opinions in the group. Based on external working life experiences, some students also had noticed that in many companies the leadership culture does not encourage dialogue at all. These future employees will be motivated as well as committed when they are part of a team where dialogue exists.

4.2. Where there's a will there's a way – A practical view of communication technologies and platforms.

A framework for making interpretations based on the qualitative data analysis for research question 2 had four phases:

1. Notes from interview data (What tools were said to be employed?)
2. Direct and participative observation in training sessions (What tools were actually employed in practice and by who?)

3. Reflection of what was learned (Who actually participated to implementation of new tools? How tools were selected? Why tools were abandoned?)
4. What outputs were done for customers (Videos, web-sites, etc.) with the ICT tools employed?

These phases helped to validate our findings. When a team entrepreneur expressed in an interview (1) that she was interested to learn a new tool for making videos, later discussed in training session (2) about utilizing the tool in a customer project, reflected the use of same tool when a project was completed (3), and finally the a video was publicly available (4) in social media platform – we were able to validate integrity of our chain of making interpretations. An example of how these 14 cases were analyzed is presented in figure 6.

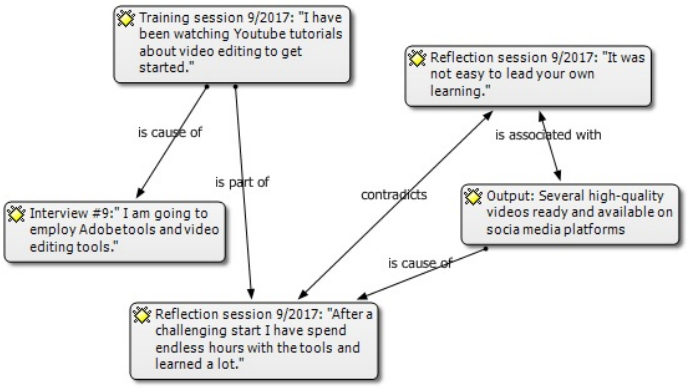


Figure 6. An example of qualitative data analysis, case #9

Based on our current experience, better utilization of tools and applications is not limited by availability of ICT tools and applications. Rather, it is limited by individual differences in the will to share information about one's working. Sharing work-related information is a much more sensitive issue than it is usually expected to be. This very same phenomenon limits the use of many other ICT systems as well. Every time when there is competition of projects within teams or departments of the same organization, openness in sharing information is limited. Only part of the employees are willing to share their business contacts via the company-wide CRM system, for example. A better understanding of the benefits of balance between individual work and team work (see [55] for more details) is needed to better utilize ICT systems. Cui bono (Who benefits?) question has to be answered clearly when new systems are employed.

Adopting the use of some practical IT tools while learning in projects is important. To facilitate communications between learners and to support project management, a broad selection of "free to use" tools have been applied within the EDE. The student respondents in the study in August 2017 stated that they had learned to use 4 - 5 new IT tools, applications or platforms on average during one year's time. Some of these tools had been presented by the team coach or a visiting company representative. Currently, also learners find and introduce new tools to the community on a regular basis. The list of the tools, applications and platforms used is long, including the following ones that were mentioned the most frequently: Trello, Slack, Skype, Doodle, WhatsApp, Moodle, Dropbox, Google Drive, Canva, Mention, Facebook, Twitter, and Instagram. Naturally, email, SMS messages and other conventional communications tools are used as well.

The most frequently used tools in project management were Google Drive, Google Sheets and Trello. In a brief unofficial discussion with the team entrepreneurs from other teams, the same tools were mentioned as the most popular ones in other teams as well.

According to our experiences, it is rather simple for the students to start using new digital tools when working on projects. This also helps students to quickly gain valuable experiences with different tools and platforms and their usability in different tasks. Sometimes students want to challenge themselves by learning to use a new tool, and harness it into the use of their cooperative. There is no need to teach the system itself, but rather act as an example and invite students to use the system. That way they will learn how the real work processes can make use of different technologies.

Though learners are using several tools and applications to support their learning in projects and to facilitate their communication, we are still in the early stages of truly benefitting from digitalization. Technology is not a limitation anymore. The real limitations are found in normal human behavior. To pick an example, a project manager and team members would benefit from knowing what tasks other team members currently have, who might need more tasks do to or who would need help. This data could easily be entered in a Trello table and shared to every team member.

4.3. Objectives first, systems to support productivity second

Making experiments and building prototypes [56, 57], associating different types of knowledge [58 - 60], and sharing knowledge of learning experiences [60 - 62] are part of the innovation management concept of the EDE. The team entrepreneurs experiment with different types of tools, applications, systems and platforms to find the ones that are appropriate and useful for their purposes.

Our target with the EDE is to help a new culture of agile methods, applications, and practices to emerge. The easiest way so far has been to start with the team entrepreneurs by introducing new applications to them and just starting using them in project management and communications, when students can directly see the benefit of the systems. So far, the results have been promising.

Experimental development is inbuilt in the EDE, in the form of innovation assignments mentioned before. Also the learning by doing approach, supported by reflection on what was learned, creates confidence in the cooperatives and students' own skills. Dedicating time for reflections is an important learning outcome on the way of becoming a professional in one's own area [63].

5. Discussion and summary

Combining and analyzing all data from the stakeholder groups of the EDE and from the team entrepreneurs studying within the ecosystem, we can clearly identify three steps we have to take next:

1) to leverage lessons learned so far in the ICT bachelor curriculum and master study curricula 2) to re-evaluate how extensively the use of new ICT tools has to be taught, and 3) to set objectives for further studies on the topic.

The current Experimental Development Ecosystem offers a pedagogically solid, tested, experimentally developed, and continuously evolving basis for supporting bachelor education in higher education. Also, the use of digital tools and applications to support learning by doing in projects and to facilitate communications within and between teams has been increasing rapidly.

5.1. Probable future for the EDE

Professional development should not focus on educating students about how current professions are executed [64]. As suggested, professional education should rather be focused on developing professions in cooperation between educational institutes and companies. Underneath these suggestions lies the activity theory [65, 66]. The ultimate goal is to help school and work life to collaborate in a better way. In times of rapid changes, educational arrangements including theory, practice and reflection repeat themselves in a continuing manner.

- International partners are needed to serve both education and business requirements. Companies who are interested in internationalizing their operations will benefit from opportunities for rapid experimentation in two or more locations.
- ICT students need a balanced amount of core ICT skills and business skills. This could be achieved by working with team entrepreneurs studying marketing with the EDE.
- ICT and marketing students should be put in mixed teams for sharing knowledge, and especially for creating new knowledge together. More cross-fertilization is needed for more learning and more versatile ideas.

There is also a lot of work to be done to better integrate the competencies of Saimaa UAS staff into the coaching process of marketing team entrepreneurs and/or ICT students, whether they study as team entrepreneurs or conventional students. This will require a more adaptive attitude towards how learning is organized from all the parties involved: teachers, administration, and team entrepreneurs.

In the beginning, this will require a lot of encouragement because even though higher education teachers may have extensive experience in teaching, they might lack competencies in learner-centric pedagogy and its methods, the mindset required for experimental development, or even both. Changing methods will also require teachers to accept that employing these methods means that they will frequently face situations that are novel to everyone. If someone has a strong routine of teaching, there might not be much interest or will to change it.

When the courage to experiment new methods to support learning is found (usually with support of colleagues or appropriate further education), at the beginning the change of habits will raise anxiety among teachers and learners. Teachers are used to acting as specialists providing answers and at the same time learners have been passive listeners. When tables are turned, both parties feel uncomfortable and the temptation to go back to old habits is strong. If this happens, it will return the potential of learners taking responsibility and teachers role of not giving easy answers to learners directly to zero. During the process of change we have gone through we have seen this phenomenon take place several times.

Very soon teachers, however, usually find out that with longer experience in life in general and operating in different types of situations helps to coach others to find means of how to solve them. Teachers are usually very good at abstracting and conceptualizing what has been learned. By reflecting on the learned skills and discussing how they can be applied in other contexts will help transfer the learning outcomes.

5.2. How teachers, lecturers and team coaches are able to benefit from the changes

At the beginning, creating a learning environment required in the 21st century will take more time than is usually used when organizing courses in higher education. After a two or three years' time, a new culture of students' responsibility for their own learning process will start to emerge and strengthen. At this tipping point, a team coach will recognize that there will be time for tactical and strategic thinking again.

5.3. Summary

By combining and analyzing the versatile data we collected, we were able to answer the research questions:

1. *What are the learning experiences when dialogue is used as learning method?*

2. *How new ICT tools and applications have been utilized by team entrepreneurs during the first year?*

This data shows that dialogue as a learning method was appropriate to support the professional development of the team entrepreneurs. Various new tools for project management, communication, and marketing and sales were employed in practice. It is noteworthy that these ICT tools, applications or systems were not taught. The dialogue and presentation of the tools was enough to inspire action.

The higher education institutes should foster creation of such ecosystems where knowledge, skills and character are combined. The learning environments should provide learner-centric methods, and acknowledge the importance of metacognition. Focusing mainly on content is not enough.

The systems (tools, applications, platforms) needed to support the ecosystem are nowadays mostly free to use and easy to use, meaning that most of them can be applied to practice at a fast pace. These are used in several visionary companies and they should be applied also in higher education institutes if they are not in use yet. The authors warmly welcome new partners to write and discuss the topics covered in this article.

Acknowledgments

This article has been written as a part of the Digikaappaus project, which is funded by Saimaa University of Applied Sciences,

Lappeenranta University of Technology, Tekes – the Finnish Funding Agency for Innovation, and by 11 Finnish organizations.

References

[1] P. Juvonen, A. Kurvinen, "Developing Experimental Development Ecosystem to serve ICT Education – A follow-up Study of Collaboration possibilites between Stakeholder Groups" in IEEE EDUCON 2017 Conference on 26th April 2017, Athens, Greece. http://ieeexplore.ieee.org/document/7943052/?reload=true

[2] C. Fadel, M. Bialik, B. Trilling, Four-dimensional-education – The competencies learners need to succeed. The Center for Curriculum Redesign, 2015.

[3] OECD, Schooling Redesigned: Towards Innovative Learning Systems, Educational Research and Innovation. OECD Publishing, 2015. http://www.keepeek.com/Digital-Asset-Management/oecd/education/schooling-redesigned_9789264245914-en#.WYvvnMuwfbg#page3

[4] S. Leppimäki, T. Meristö, H. Tuohimaa, J. Laitinen, "Future needs for competencies in ICT companies. Tulevaisuuden osaamistarpeet tietotekniikkayrityksissä", Corporate Foresight Group, CoFi / Åbo Akademi, 2007, (In Finnish).

[5] OECD, "Moving Up the Value Chain: Staying Competitive in the Global Economy", OECD, 2007.

[6] OECD (2012). Education Today 2013: The OECD Perspective, OECD Publishing.

[7] T. Meristö, S. Leppimäki, J. Laitinen, H. Tuohimaa, "The skill foresight of The Finnish Technology Industries. Tulevaisuuden osaamistarpeet teknologiateollisuudessa Yhteenvetoraportti toimialakohtaisista yrityskyselyistä.", Teknologiateollisuus ry, 2008. (In Finnish).

[8] B. Trilling, C Fadel, 21st Century Skills – Learning for Life in our Times, Jossey-Bass, 2009.

[9] P. Ylä-Anttila, "Dispersion of value chains, future skills requirements change. Sähkö-, elektroniikka- ja tietotekniikka-ala. Tuotantoketjut hajautuvat, osaamistarpeet muuttuvat." ETLA, The Research Institute of the Finnish Economy. 2012, (In Finnish).

[10] R. Rajander-Juusti, "Competency needs in Business and Administration. Liiketalouden osaamistarpeet. Ennakkotietoa koulutuksen suunnittelun tueksi." Raportit ja selvitykset 2013:1. Opetushallitus, 2013, (In Finnish).

[11] D. Bohm, On dialogue, Routledge Classics, 1996.

[12] W. Isaacs, Dialogue: The art of thinking together. Doubleday, Randomhouse Inc, 1999.

[13] Tiimiakatemia. Visited on 11/2017, http://tiimiakatemia.fi/en

[14] Proakatemia. Visited on 11/2017, http://proakatemia.fi/en/

[15] P. Juvonen, "LEARNING INFORMATION TECHNOLOGY BUSINESS IN A CHANGING INDUSTRY LANDSCAPE. The Case of Introducing Team Entrepreneurship in Renewing Bachelor Education in Information Technology in a University of Applied Sciences." Doctoral Thesis, Acta Universitatis Lappeenrantaensis 606, 2014.

[16] V. A. Alexandria, P. Parton, A. Robb, "Entrepreneurship Education and Training Programs around the World. Dimensions for Success.", International Bank for Reconstruction and Development / The World Bank, 2014.

[17] R.A. Baron, "Behavioural and Cognitive Factors in Entrepreneurship: Entrepreneurs as the Active Element in New Venture Creation", Strategic Entrepreneurship Journal **1**, 167–82, 2007.

[18] C. K. Volkmann, D. B. Audretsch (Eds.) "Entrepreneurship Education at Universities. Learning from Twenty European Cases", Springer International Publishing 2017.

[19] D. La Guardia, M. Gentile, V. Dal Grande, S. Ottaviano, M. Allegra, "A Game based Learning Model for Entrepreneurship Education", Procedia - Social and Behavioral Sciences, **141**, 195-199, 2014.

[20] A. Daniel, R. Costa, M. Pita, C. Costa, "Tourism Education: What about entrepreneurial skills?", Journal of Hospitality and Tourism Management, **30**, 65-72, 2017.

[21] R.G. Klapper, V.A. Farber, "In Alain Gibb's footsteps: Evaluating alternative approaches to sustainable enterprise education (SEE)", The International Journal of Management Education, **14**(3), 422-439, 2016.

[22] R. Bell, "Developing the next generation of entrepreneurs: Giving students the opportunity to gain experience and thrive", The International Journal of Management Education, **13**(1), 37-47, 2015

[23] S. Mat, S. Maat, N. Mohd, "Identifying Factors that Affecting the Entrepreneurial Intention among Engineering Technology Students", Procedia - Social and Behavioral Sciences, **211**, 1016-1022, 2015.

[24] C. Shlaegel, M Koenig, "Determinants of Entrepreneurial Intent: A Meta-Analytic Test and Integration of Competing Models", Entrepreneurship Theory and Practice, **38**(2), 291-332, 2013.

[25] C. Luthje, N. Franke, "The "making" of an entrepreneur: Testing a model of entrepreneurial intent among engineering students at MIT", R&D Management, **33**(2), 135 – 147, 2003.

[26] B. Hj Din, A. Rahim, R. Anuar, M. Usman, "The Effectiveness of the Entrepreneurship Education Program in Upgrading Entrepreneurial Skills among Public University Students", Procedia - Social and Behavioral Sciences, **224**, 117-123, 2016.

[27] B, Askun, N. Yıldırım, "Insights on Entrepreneurship Education In Public Universities In Turkey: Creating Entrepreneurs Or Not?", Procedia - Social and Behavioral Sciences, **24**, 663-676, 2011.

[28] J. Kirkwood, K. Dwyer, B. Gray, "Students' reflections on the value of an entrepreneurship education", The International Journal of Management Education, **12**(3), 307-316, 2014.

[29] A.S. Zamberi, A. Bakar, N. Ahmad, "An evaluation of teaching methods of entrepreneurship in hospitality and tourism programs", The International Journal of Management Education, **16**(1), 14-25. Forthcoming, 2018.

[30] M. Küttim, M. Kallaste, U. Venesaar, A. Kiis, "Entrepreneurship Education at University Level and Students' ", Entrepreneurial Intentions, Procedia - Social and Behavioral Sciences, **110**, 658-668, 2014.

[31] T. Shih, Y-Y. Huang, "A case study on technology entrepreneurship education at a Taiwanese research university." Asia Pacific Management Review, **22**(4), 202-211, 2017. https://doi.org/10.1016/j.apmrv.2017.07.009

[32] I, S, Ahmad, M, F. R. Buchanan, "Examining the entrepreneurship curriculum in Malaysian polytechnics." The International Journal of Management Education, **12**(3), 397-406, 2014.

[33] S , Jansen, T. De Zande, S. Brinkkemper, E. Stam, V. Varma, "How education, stimulation, and incubation encourage student entrepreneurship: Observations from MIT, IIIT, and Utrecht University", The International Journal of Management Education, **13**(2) 170-181, 2015.

[34] J. Mezirow, Transformative dimensions of adult learning, Jossey-Bass, 1991.

[35] R. Kiely, "A Transformative Learning Model for Service-Learning: A Longitudinal Case Study", Michigan Journal of Community Service Learning. **12**(1), 5-22, 2005.

[36] P. Juvonen, "Comparison of two learning and team entrepreneurship models at a Finnish University of Applied Sciences. Setting the scene for future development", **7**(1). International Journal of Engineering Pedagogy, 2017. http://online-journals.org/index.php/i-jep/article/view/6517

[37] D. Kolb, D, Experiental Learning, Prentice-Hall, Englewood Cliffs, NJ, 1984

[38] L. B. Resnick, Introduction. In L. B. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser, 1-24, Hillsdale, NJ: Erlbaum, 1989.

[39] V. Richardson, "Constructivist Pedagogy." University of Michigan. Teachers College Record, 105(9) 1623-1640 , 2003. Teachers College, Columbia University, 0161-4681.

[40] R. Sarder, Building an Innovative Learning Organization. A framework to build a smarter workforce, adapt to change and drive growth, Wiley, 2006.

[41] D. Levi, Group Dynamics for teams, 2nd edition, Sage Publications, 2007.

[42] M. Beer, High Commitment High Performance. How to Build a Resilient Organization for Sustained Advantage, Jossey-Bass, 2009.

[43] P. Cockman, B. Evans, B. P. Reynolds, Consulting for Real People. Second Edition. McGraw-Hill Publishing Company, 1999.

[43] K. Herr, G.L.Anderson, The Action Research Dissertation. A guide for Students and Faculty, Sage Publications, 2005.

[45] E.T. Stringer, Action Research 3rd ed, Sage Publications, 2007.

[46] R.K. Yin, Case Study Research: Design and Methods, Sage, 2003.

[47] K.M. Eisenhardt, M.E. Graebner, Theory building from cases: Opportunities and challenges, Academy of Management Journal, **50**(1) 25 – 32, 2007.

[48] N.K. Denzin, Y.S. Lincoln, (eds.), Collecting and Interpreting Qualitative Materials, Sage Publications 2003.

[49] A. Strauss, J. Corbin, Grounded Theory in Practice, Sage Publications, 1997.

[50] U. Flick, An Introduction to Qualitative Research. 4th ed. Sage Publications, 2009.

[51] B.G. Glaser, Theoretical sensitivity. Advances in the Methodology of Grounded Theory. Sociology Press, 1978.

[52] B.G. Glaser, Emergence vs. Forcing: basics of grounded theory analysis. Sociology Press, 1992.

[53] A. Pentland, The New Science of Building Great Teams. Harvard Business Review, April 2012.

[54] M. Losada, E. Heaphy, The Role of Positivity and Connectivity in the Performance of Business Teams. American Behavioral Scientist, **47**(6), 740 – 765, 2004.

[55] P. Lencioni, The five dysfunctions of a Team. A workshop for teams, Jossey-Bass, 2012.

[56] M. Schrage, Serious Play. How the World's Best Companies Simulate to Innovate, Harvard Business School Press, 2000.

[57] N. Furr, Dyer, J, The Innovator's Method. Bringing the Lean Startup into your Organization, Harvard Business Review Press, 2014.

[58] J. Dyer, H. Gregersen, C.L. Christensen, The Innovator's DNA. Mastering the Five Skills of Disruptive Innovators. Harvard Business Review Press, 2011.

[59] E. Wenger, R. McDermott, W.M. Snyder, Cultivating Communities of Practice. A guide to managing knowledge, Harvard Business School Press 2002.

[60] J.R. Katzenbach, D.K. Smith, The Discipline of Teams. A Mindbook-Workbook for Delivering Small Group Performance, John Wiley & Sons, 2001.

[61] G. Von Krogh, K. Ichijo, I. Nonaka, Enabling Knowledge Creation. How to Unlock the Mystery of Tacit Knowledge and Release the Power of Innovation, Oxford University Press, 2000.

[62] P.J. Denning, R. Dunham, The Innovators Way. Essential practices for successful Innovation, The MIT Press, 2010.

[63] D. Boyd, R. Keogh, D. Walker, Reflection: Turning Experience into Learning, Kogan Page, 1985.

[64] J. Virkkunen, H. Ahonen, M. Schaupp, L. Lintula, Toimintakonseptin yhteisen kehittämisen mahdollisuus. TYKES, raportteja 70, 2010, (In Finnish).

[65] L.S. Vygotsky, Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, 1978.

[66] Y. Engeström, Developmental work research: Expanding activity theory in practice, Lehmanns Media, 2005.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 85-89 (2018)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Estimating short time interval densities in a CTM-KF model

Arlinda Alimehaj Rrecaj[*,1], Marija Malenkovska Todorova[2]

[1]*University of Prishtina, Mechanical Engineering Faculty, Traffic and Transportation Department, 10000 Prishtina, Kosovo*

[2]*University of St.Klement Ohrid, Faculty of Technical Sciences, Transport Engineering, 7000 Bitola, Republic of Macedonia*

ARTICLE INFO

Article history:
Received: 27 January, 2018
Accepted: 21 February, 2018
Online: 08 March, 2018

Keywords:
Cell
Density
Kalman Filter

ABSTRACT

On-ramping is being widely used as e method to increase the freeway operational efficiency. The main traffic parameter that must be taken in consideration for the implementation of the feedback control strategies for the on-ramp metering is density on the main section of road. In this paper is given discretized model of traffic which is then improved by a recursive technique called Kalman-Filter with the aid of which is possible to predict the density, by only having the traffic flow measured on the start and end road section. Kalman Filter is based on linear relationship of flow and density. By minimizing the square of error between of the measurements and the estimated values of flows, a gain is derived which then is applied to the densities of the model in order to obtain the greatest accuracy of these values.

1. Introduction

The increasing demand of motorized vehicles is becoming one of the major problems that face the developed world places in nowadays. Based on some statistics in Great Britain, 90% of the journeys are by road, during the last decade, and for more the road distances travelled, have increased by over 1000% on the last sixty years. [1]

On-ramp metering [2] is being widely used as e method to increase the freeway operational efficiency by regulating the traffic interruption by minor roads, while maintaining the right of way on the major section until it reaches me critical densities, which assure the maximal flow [3]. Measure of the density on the major road is difficult to measure. In this paper is proposed a discretized traffic model (CTM) to obtain traffic densities which then will be accurate through Kalman Filter [4, 5 and 6].

2. CTM model of a Highway with three cells

If we denote with ρ_i (k) the density of a cell (uniform or non-uniform length), instead of the number of vehicles n_i on the a unit length cell, then we can bring equation *(1.1)*[6].

for density $\rho_i(k+1)$ of the cell i updated time step in *(k+1),* where T_s is the discrete time unit in seconds.

$$\rho_i(k+1) = \rho_i(k) + \frac{T_s}{L}(q_i(k) - q_{i+1}(k)) \quad (1.1)$$

Analyzing a highway partitioned in three cells (for the sake of simply illustration) with an on ramp and an off ramp, by assuming that the belonging cells can be in the free flow either in congested mode the densities on each cell can be written as in equations (*1.2, 1.3, 1.4*).

The densities on each cell are:

$$\rho_{i-1}(k+1) = \rho_{i-1}(k) + \frac{T_s}{L}(q_{i-1}(k) - q_i(k) + r(k)) \text{ or}$$
$$\rho_1(k+1) = \rho_1(k) + \frac{T_s}{L}(q_1(k) - q_2(k) + r(k)) \quad (1.2)$$

$$\rho_i(k+1) = \rho_i(k) + \frac{T_s}{L}(q_i(k) - q_{i+1}(k)) \text{ or}$$
$$\rho_2(k+1) = \rho_2(k) + \frac{T_s}{L}(q_2(k) - q_3(k)) \quad (1.3)$$

$$\rho_{i+1}(k+1) = \rho_{i+1}(k) + \frac{T_s}{L}(q_{i+1}(k) - q_4(k) - f(k)) \text{ or}$$
$$\rho_3(k+1) = \rho_3(k) + \frac{T_s}{L}(q_3(k) - q_4(k) - f(k)) \quad (1.4)$$

With the elaboration of the inter-cell flow law can be defined the expressions for the inter cell flows q_1, q_2 and q_3 in the above equations (*1.1, 1.2 and 1.3)*. Before the inter cell flows elaboration is given, a reasonable description of the congestion must be given further, since as we assumed above, the cells can be in either free

[*]Arlinda Alimehaj Rrecaj, Email: arlinda.alimehaj@uni-pr.edu

or congested mode. Congestion is defined as the state of the traffic with high density rates, or with other words the density of that part of the highway expressed in cell is equal or higher than the critical density based on the fundamental diagram of relationship of flow and density. Referred to the mentioned diagram, can be noticed that the congested flow belongs to higher values of the density, above the critical density values where the flow drops down. That can be described with enormous number of vehicles travelling at low speeds and with short distance spaces between each other.

The common modes, used in analysis of researchers are the fully congested mode when the three cells are congested, denoted by *CCC* mode, and free flow mode when the three cells are in free flow mode, denoted by *FFF* mode. The other middle modes that are out of the scope of this paper are those with last one and two cells in congested mode, written by *FCC* and *FFC*, respectively. To emphasize the modes, the congested cells are highlighted further.

Now, for the *FFF* mode, the densities of the cells are lower than the critical density and the inter cell flows are as follows:

$$q_i(k) = \min\left(v_{fi-1}.\rho_{i-1},\ Q_{i-1}\ w_i(\rho_J - \rho_i)\right) or$$

$$q_2(k) = \min\left(v_{f1}.\rho_1,\ Q_1\ w_2(\rho_J - \rho_2)\right) = v_{f1}.\rho_1 \quad (1.5)$$

$$\left| q_{i+1}(k) = \min\left(v_{fi}.\rho_i,\ Q_i\ w_{i+1}(\rho_J - \rho_{i+1}),\right)\right.$$

$$q_3(k) = \min\left(v_{f2}.\rho_2,\ Q_2\ w_3(\rho_J - \rho_3)\right) = v_{f2}.\rho_2 \quad (1.6)$$

In *CCC* mode, the densities of the cells are higher that the critical density, and the inter cell flows are:

$$q_i(k) = \min\left(v_{fi-1}.\rho_{i-1},\ Q_{i-1}\ w_i(\rho_J - \rho_i)\right) or$$

$$q_2(k) = \min\left(v_{f1}.\rho_1,\ Q_1\ w_2(\rho_J - \rho_2)\right) = w_2(\rho_J - \rho_2) \quad (1.7)$$

$$q_{i+1}(k) = \min\left(v_{fi}.\rho_i,\ Q_i\ w_{i+1}(\rho_J - \rho_{i+1}),\right)$$

$$q_3(k) = \min\left(v_{f2}.\rho_2,\ Q_2\ w_3(\rho_J - \rho_3)\right) = w_3(\rho_J - \rho_3) \quad (1.8)$$

After subtracting the expressions for inter cell flows (1.5 *and 1.6*) in the equations of densities *(1.2, 1.3* and *1.4),* for the *FFF* mode, we have:

$$\rho_1(k+1) = \rho_1(k) + \frac{T_s}{L}(q_1(k) - v_{f1}.\rho_1(k) + r(k)) \quad (1.9)$$

$$\rho_2(k+1) = \rho_2(k) + \frac{T_s}{L}(v_{f1}.\rho_1(k) - v_{f2}.\rho_2(k)) \quad (1.10)$$

$$\rho_3(k+1) = \rho_3(k) + \frac{T_s}{L}(v_{f2}.\rho_2(k) - q_4(k) - f(k)) \quad (1.11)$$

And after subtracting the expressions for inter cell flows (*1.7* and *1.8*) in the equations of densities (*1.2, 1.3* and *1.4*), for the *CCC* mode, we have:

$$\rho_1(k+1) = \rho_1(k) + \frac{T_s}{L}(q_1(k) - w_2\left(\rho_J - \rho_2(k)\right) + r(k)) \quad (1.12)$$

$$\rho_2(k+1) = \rho_2(k) + \frac{T_s}{L}(w_2(\rho_J - \rho_2(k)) - w_3(\rho_J - \rho_3(k))) \quad (1.13)$$

$$\rho_3(k+1) = \rho_3(k) + \frac{T_s}{L}(w_3(\rho_J - \rho_3(k)) - q_4(k) - f(k)) \quad (1.14)$$

3. State-Space presentation of CTM Model

The state space presentation (particularly the state space, eq.*1.17*) of *CTM* based traffic densities of a highway segment in *FFF* mode differs from that of *CCC* mode. [7].

What it characterizes the state space model of the traffic density based on *CTM* model, is implication of some other extension parts of the state space, B_q which is the input matrix of upstream and downstream flows q_1 and q_4 respectively, B_r the input matrix for on ramp and off ramp flows, r and f respectively that are applicable on the *FFF* mode and input matrices, B_w which takes into consideration the backward waves w_2 and w_3 and B_J the input matrix of the jam density that are applicable on the state space model of the *CCC* mode. (1.17) and (1.18)

$$\rho_{(k+1)FFF} = A\rho_{(k)} + B_u q_{(k)} + B_{r,f}(r,f)_{(k)} + Bw_{(k)} \quad (1.17)$$

$$\rho_{(k+1)CCC} = A\rho_{(k)} + B_u u_{(k)} + B_{r,f}(r,f)_{(k)} + B_J\rho_{J(k)} + Bw_{(k)} \quad (1.18)$$

Where: *x (k+1)* is the system state vector and in this paper, according to the *CTM* model, corresponds to the density in cell of cell *i.*

A is the state matrix, *B* is the input matrix, *u(k)* is the input or control and w_k represents the process noise.

They are assumed to be independent (of each other), white, and with normal probability distributions (Gaussian) as:

$p(w) \tilde{} N(0, Q)$ *and* $p(v) \tilde{} N(0, R)$.

From the system of equations in (*3.12, 3.13* and *3.14*) can be drawn (after some regulations finding partial derivatives of the functions of densities to the parameters previous densities *ρ(k),* flows q_1 and q_4 and backward waves w_2 and w_3 which provide the elements of the respective matrices of the i^{th} row that correspond to density of i^{th} cell) the belonging matrices of the system matrices.

After bringing up together (eq. in *1.19* and *1.20* can be expanded to the form of state space: For *FFF* mode:

$$\begin{pmatrix}\rho_1\\ \rho_2\\ \rho_3\end{pmatrix}_{FFF}(k+1) = \begin{pmatrix}1 - v_{f1}\cdot\frac{T_s}{l_1} & 0 & 0\\ v_{f1}\cdot\frac{T_s}{l_2} & 1 - v_{f2}\cdot\frac{T_s}{l_2} & 0\\ 0 & v_{f2}\cdot\frac{T_s}{l_3} & 1\end{pmatrix}\cdot\begin{pmatrix}\rho_1\\ \rho_2\\ \rho_3\end{pmatrix}(k) +$$

$$\begin{pmatrix}\frac{T_s}{l_1} & 0 & 0\\ 0 & \frac{T_s}{l_2} & 0\\ 0 & 0 & \frac{T_s}{l_3}\end{pmatrix}\cdot\begin{pmatrix}q_1\\ 0\\ -q_4\end{pmatrix}(k) + \begin{pmatrix}\frac{T_s}{l_1} & 0 & 0\\ 0 & \frac{T_s}{l_2} & 0\\ 0 & 0 & \frac{T_s}{l_3}\end{pmatrix}\cdot\begin{pmatrix}r\\ 0\\ 0\end{pmatrix}(k) \quad (1.19)$$

By recalling the standard state space models in (*1.19*) and *(1.20*), for completely free flow mode FFF and congested mode CCC respectively, there can be derived other variants, by changing the outflow from which is considered to have congestion the density formula, that is dictated by the backward speed and jam density of the downstream cell.

For *CCC* mode:

$$\begin{pmatrix}\rho_1\\ \rho_2\\ \rho_3\end{pmatrix}_{CCC}(k+1)=\begin{pmatrix}1 & w_2\cdot\frac{T_s}{l_1} & 0 & 0\\ 0 & 1-w_2\frac{T_s}{l_2} & w_3\cdot\frac{T_s}{l_2} & \\ 0 & 0 & 1-w_3\cdot\frac{T_s}{l_3} & \end{pmatrix}\cdot$$
$$\begin{pmatrix}\rho_1\\ \rho_2\\ \rho_3\end{pmatrix}(k)+\begin{pmatrix}\frac{T_s}{l_1} & 0 & 0\\ 0 & \frac{T_s}{l_2} & 0\\ 0 & 0 & -\frac{T_s}{l_3}\end{pmatrix}\cdot\begin{pmatrix}q_1\\ 0\\ q4\end{pmatrix}(k)+\begin{pmatrix}\frac{T_s}{l_1} & 0 & 0\\ 0 & \frac{T_s}{l_2} & 0\\ 0 & 0 & \frac{T_s}{l_3}\end{pmatrix}\cdot\begin{pmatrix}r\\ 0\\ 0\end{pmatrix}(k)+$$
$$\begin{pmatrix}0 & -w_2\cdot\frac{T_s}{l_1} & 0\\ 0 & w_2\cdot\frac{T_s}{l_2}0 & 0\\ 0 & 0 & w_3\cdot\frac{T_s}{l_3}\end{pmatrix}\cdot\begin{pmatrix}\rho_{1J}\\ \rho_{2J}\\ \rho_{3J}\end{pmatrix}(k) \qquad (1.20)$$

4. A numerical example of the CTM model-Traffic density1.1.

For the purpose of the demonstration of the CTM model, in this paper is performed a numerical example which is described below. For the sake of simplicity, are chosen the approximately the same freeway segment partitioning characteristics as that in earlier sections in order to do an interconnection with the laid state space model of traffic density. The system of performance measurements of the traffic road networks of the Californian state *PeMs* is used for traffic collection data and is considered a freeway link for on the street "Broadway Avenue", Stockton/San Francisco. The freeway is consisted from three cells with different lengths with one on-ramp on the first cell. (*fig.1).*

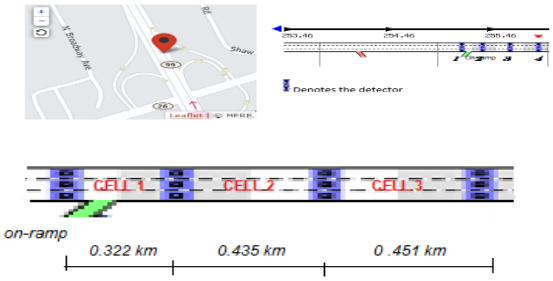


Fig. 1. Freeway segment with three cells

Calibrated parameters are given below [8].

Table 1. Summary of traffic parameters for three cells

	Q_M	*vf*	*ρcr*	*ρJ*	*w*	**FF/ /CC**
Cell 1	5580	84.8	65.8	248	30.6	FF
Cell 2	4176	96.8	43.1	248	20.3	FF
Cell 3	4268	106.7	40.0	248	20.5	FF

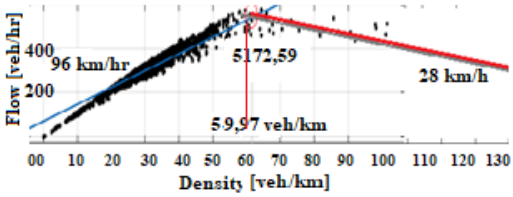


Fig.2. Calibrated parameters

5. Kalman Filter

Kalman filter-*KF* (*Kalman, 1960; Welch and Bishop 2001)* is a recursive data processing algorithm that uses only the previous time-step's prediction with the current measurement in order to make an estimate for the current state [4]. This means the *KF* does not require previous data to be stored or reprocessed with new measurements.

The building structure of the KF

The Kalman Filter consists of a set of mathematical equations that provides an efficient recursive computation to estimate the state of a process by minimizing the mean of the squared error. [4] The KF estimates the value of the variable x at any time (k+1), represented by a linear stochastic equation.

$$x_{k+1}=Ax_k+BX_k+w_k \qquad (1.22)$$

Where: A (k) is matrix which relates the state a time interval *k* with the state at current time interval *k+1. B (k)* is matrix which relates the current state to the control inputX_k.

The random variable *w* represents noise in modelling process. It is assumed to be within normal probability distributions with zero mean and variance Q (Gaussian) as: $p(w)\tilde{}\ N(0,Q)$

The system measurement equation describes the relationship between system states and measurements. Acknowledging that measurements inevitably contain noise, the output equation is expressed as follows:

$$\mathrm{Z(k+1)=H\cdot x(k)}, \qquad (1.23)$$

$\tilde{Z}(k)$ is the measurement variable (outflow of vehicles from cell 3-measured by loop detector 2) *H (k)* is the output matrix, and *v (k)* is the measurement noise variable. The errors in estimating *a priori* and *a posteriori* states are defined as follows:

$$P_k^-=x_k-\hat{x}_k^- \qquad (1.24)$$
$$\mathrm{P_k=x_k-\hat{x}_k} \qquad (1.25)$$

The *a priori* and a posteriori estimate covariance is given by:

$$P_k^-=E[e_k^-,e_k^{-T}]=AP_{k-1}\cdot A^T+Q \qquad (1.26)$$

$$P_k=E[e_k,e_k{}^T)=(I-K_kH)\cdot P_k^- \qquad (1.27)$$

The KF estimates a posteriori state of the process using a linear combination of a priori state and a weighted difference between the actual measurement and the model measurement of the state.

$$\hat{x}_k = \hat{x}^-_k + K(z_k - H \cdot \hat{x}^-_k) \quad (1.28)$$

$$K_k = P^-_k H^T (H P^-_k H^T + R)^{-1} \quad (1.29)$$

Based on the above equation, especially on (1.29), the *KF* process can be divided in two steps ore phases. The first step is the prediction step and the second step is the correction step.

6. Estimation with Kalman Filter

In this section are described in detail the applied matrices to the algorithm of the *CTM-KF* model. It is necessary to recall the equations state space of *CTM* model (Section V.1) first and then to do an interconnection of it with the *KF* algorithm equations. Since in our model, we are estimating the traffic densities of the three cells of the mentioned link, by the usage of the inputs values of the inflow q_1, output values q_4 and the flow from on ramp, then the state space vector of our algorithm are the densities $x=[x_1,x_2,x_3]= [\rho_1, \rho_2, \rho_3]^T$,the input vectors are

$$x_u= [q_1,0, q_4]^T; x_r=[qr,0,0]^T$$

$$x_{k+1} = Ax_k + BX_{Uk} + w_k$$

$$\begin{pmatrix}\rho_1\\ \rho_2\\ \rho_3\end{pmatrix} = A \cdot \begin{pmatrix}\rho_1\\ \rho_2\\ \rho_3\end{pmatrix} + Bu \cdot \begin{pmatrix}q_1\\ 0\\ -q_4\end{pmatrix} Br \cdot \begin{pmatrix}r\\ 0\\ 0\end{pmatrix} \quad (1.30)$$

On this paper we are going to use the measurement of the outflow from the cell three, that corresponds to the flow q_4 in the figure (3.1). Based on the fundamental diagram we model the traffic flow measurement through the densities on the last cell (ρ_3) and the free flow speed on cell 3 v_{f3} we will have $q_{out} = v_{f3} \cdot \rho_3$ that is consistent with the equation (6.2)

$$Z(k+1) = H \cdot x(k) \quad (1.31)$$

$$H = [0\ 0\ v_{f3}]\ and\ x = \begin{pmatrix}\rho_1\\ \rho_2\\ \rho_3\end{pmatrix}, q_{out} - [0\ 0\ v_{f3}] \begin{pmatrix}\rho_1\\ \rho_2\\ \rho_3\end{pmatrix}$$

Where*:* Q -the model error covariance matrix which elements standard deviations of the density variables. The off diagonal elements are equal to zero while R is the measurement or output error covariance. In this seminar paper, the matrices Q and R are assumed to be constant.[9]

$$Q = \begin{vmatrix} w & 0 & 0 \\ 0 & w & 0 \\ 0 & 0 & w \end{vmatrix}; R = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & v \end{vmatrix}$$

7. Results and Conclusion

Evaluation of the density values of cell is performed with discrete time intervals of $T_s=10$ seconds, where the initial values of the densities $\rho_0 = [\rho_{10}, \rho_{20}, \rho_{30}]^T$ and estimated covariance matrix P_o are assumed. T_s is chosen to be 10 second in order to fill the conditions $T<L/v_f$, for proper work with system matrices, otherwise there will be obtained negative values of density parameters. For the purpose of the results evaluation, measured traffic densities for five minute intervals are used for comparison with the estimated densities with *CTM* model. The performance of the model was quantified by calculating the Mean Absolute Percentage Error (*MAPE)* given in (*1.32).*

$$MAPE = [\frac{1}{n} \cdot \sum_{k=1}^{n} \left| 100 \frac{\rho_{mod}(k) - \rho_{meas}(k)}{\rho_{meas}(k)} \right] \cdot 100 \quad (1.32)$$

The MAPE results for CTM model for modes FFF -KF,

for Cell 1, Cell 2 and Cell3, 2 %, 0.6 % and 1 % respectively, nd for CCC-KF, 14% on the three cells. The results of the estimated values by CTM model against the measured values of traffic density are also graphically presented on the below figure (3-5).

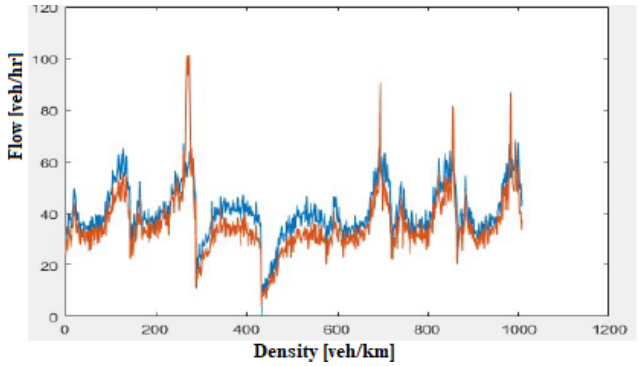


Fig.3. Graphic results densities of KF FFF and measurements- Cell 1

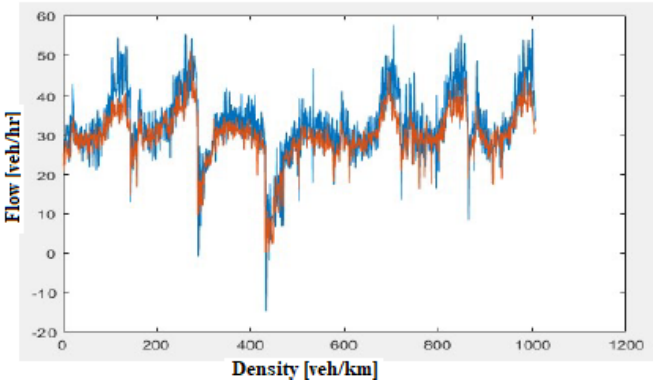


Fig.4. Graphic results densities of KF FFF and measurements- Cell 2

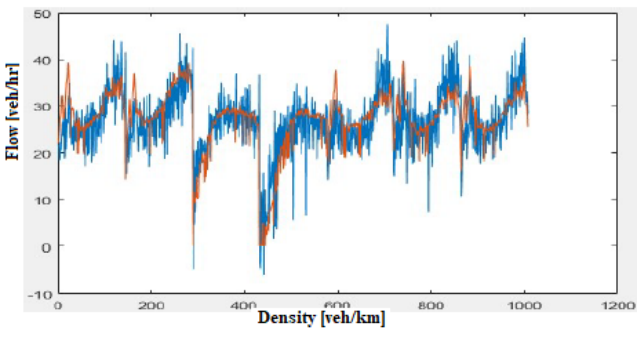


Fig.5. Graphic results densities of KF FFF and measurements- Cell3

References

[1] Nicola George - Kathryn Kershaw, Road Use Statistics, Great Britain, 2016
[2] Ramp Metering: A Review of the Literature, E.D. Arnold ,Jr (1998)
[3] Chen X.M, Li, L; Stochastic Evolutions of Dynamic Traffic Flow Modeling and Applications, Chapter 2, http://www.springer.com/978-3-662-44571-6 Springer (2015)
[4] M.J .Lighthill ; G.B. Whitham "On Kinematic Waves, II: A theory of traffic flow on long crowded roads, Proceedings of the Royal Society of London, Series A-Mathematical and Physical Science.
[5] X. Sun, L. Munoz. R. Horowitz, A Mixture Kalman Filter Highway Congestion Mode and Vehicle Density Estimator and its Application (2004)
[6] Carlos F. Daganzo, "The Cell Transmission Model: Network Traffic", (1996)
[7] L. Munoz, X. Sun, R. Horowitz, L. Alvarez; "Traffic Density Estimation with the Cell Transmission Model1" (2003)
[8] G.B. Witham "Linear and Nonlinear Waves", Pure and Applied Mathematics, John Wiley Sons, New York City, USA, (1974)
[9] L. Munoz, X. Sun, R. Horowitz, D. Sun, G. Gomes; "Methodological calibration of the cell transmission model" Proceedings of the American Control Conference, Massachusetts 2004.
[10] A. Katsivalis, M. Papa Georgiou, 'The importance of Traffic Flow Modeling for Motorway Traffic Control', 2001

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 87-93 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Discovering Interesting Biological Patterns in the Context of Human Protein-Protein Interaction Network and Gene Disease Profile Data

Rami Alroobi[*,1], William Perizzo[2]

[1]*Department of Math and Computer Science, Southern Arkansas University, 71753, USA*

[2]*Department of Computer Science, North Dakota State University, 58105, USA*

A R T I C L E I N F O

Article history:
Received: 11 October, 2017
Accepted: 08 November, 2017
Online :16 November, 2107

Keywords:
Organisms
Bioinformatics
Functional modules
Interaction networks
Disease
Expression profiles

A B S T R A C T

The current advances in proteomic and transcriptomic technologies produced huge amounts of high-throughput data that spans multiple biological processes and characteristics in different organisms. One of the important directions in today's bioinformatics research is to discover patterns of genes that have interesting properties. These groups of genes can be referred to as functional modules. Detecting functional modules can be accomplished by the deep analysis of protein-protein interaction (PPI) networks, gene expression profiles, or both. In this work the focus will be on Human protein-protein interaction network and genes expression data that represents genes behavior in a group of diseases. Two of the most well-established clustering methods that target the interaction networks and the expression data will be used in this analysis. In addition, and to have more insights, genes molecular functionality will be studied. Finally, I will introduce the relation of the extracted modules on biological pathways. This study mainly illustrates the importance of including protein interaction activities as part of any study that aims at discovering meaningful knowledge about the biological scene where many actors play different roles.

1. Introduction

The current advances in proteomic and transcriptomic technologies produced huge amounts of high-throughput data that spans multiple biological processes and characteristics in different organisms. One of the important directions in today's bioinformatics research is to discover patterns of genes or genes products that have interesting characteristics that can be related to specific activities or functions inside the living organism. Researchers believe that those groups of genes are can give deeper insights of what is happening inside a living cell than studying individual genes in isolation. These groups of genes are called functional modules (patterns). Different approaches can be employed to discovering functional modules. One of the most important approaches is to focus on protein-protein interaction (PPI) networks and try to find well connected sub-networks. Another approach is to consider gene expression data and try to find modules that show similar behavior according to their expression levels by calculating correlations. The last approach is to integrate both the PPI and the expression data in a complementary effort because both PPI and expression data suffer from incompleteness and inconsistency problems [1].

However, for the expression profiles that are related to diseases in specific, most of the available data is in the form of genome-wide expression profiles that extracted from microarray measurements. A number of methods have been proposed to clarify the biological mechanisms from this expression data [2,3]. Most of those methods are based on individual genes ability to be used as a strong indication of the disease under consideration. However, considering only individual genes will give limited insights about the molecular mechanisms and biology of an organism under different disease conditions. On the other hand, studying groups of genes that show specific behavior related to some disease will help in grasping more comprehensive views about the disease itself and what molecular functionality those genes may have and can be involved in this disease. The reason for that is the fact that any biological process, such as a disease, occurring inside the living organism is affected and affects multiple and different biological components and pathways of that organism. Patterns of genes,

*Corresponding Author: Rami Alroobi, Magnolia, Ar 71753, 870-235-4295, Email: rmalroobi@saumag.edu

or functional modules, can be discovered in different ways. One approach is by applying clustering techniques on PPI networks to find patterns of well-connected genes. One of the most established methods used in this field is the stochastic flow-based clustering (Markov Cluster Algorithm, MCL) [4, 5]. The Markov cluster algorithm (MCL) uses a graph's natural transition probability matrix to cluster a graph by combining random walks with two alternating operations (expansion and inflation). With these operations, the algorithm iteratively simulates the dissipation and elimination of stochastic flow across the graph structure to produce a non-negative matrix representation of a graph clustering. In other words, the nodes of the graph are clustered with each other based on the strength of connections between these nodes. The MCL method has been shown to be significantly robust and superior to other network-based methods when used to cluster benchmarking biological networks [6].

Another approach is to use gene expression profiles to cluster genes that show similar behavior based on their expression levels that are extracted from microarray measurements. In this case, any clustering method can be used like K-means [7]. K-means and its variations are among the most popular iterative methods used for clustering data and can be easily applied to biological data. The general idea of K-means is that a number of centers are specified and a distance metric from these centers is used as a dissimilarity measure. The points are grouped towards the closest center based on the criteria used. For example, one of the widely used metrics is the Euclidean Distance. The Euclidean Distance [8] between values x_i and $x_{i'}$ can be defined as follows:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2$$

Then each of the points, genes in this case, will be assigned to cluster so that the distance between these points and the specified center is minimized. After that, a new set of centers are calculated and the assignment step is repeated. The process continues iteratively until the assignments stabilize so that no further assignments changes are possible.

2. Experimental Analysis

2.1 Datasets

In order to elucidate the work in this paper, a group of data sets were obtained and preprocessed in order to suite the setting required for this work. The following is a listing of the datasets used in the analysis.

- **Interaction Networks**
  The experiments were performed using Homosapiens (Humans) protein-protein interaction network. The Human network was obtained from the Human Protein Reference Database (HPRD), release 9 [9]. The Human PPI network has 9465 nodes with 37039 interactions.

- **Disease Gene profiles**
  In order to perform the planned analysis, gene expression profiles for 16 diseases were obtained from the GEO database [10]. Table 1 shows the names and identification numbers of these diseases.

- **OMIM dataset**
  The OMIM, Online Mendelian Inheritance in Man, database [11] was used to extract genes-disease data. The data includes 4022 different diseases. Each disease is represented by the genes causing it or in relation to that disease.

- **Human protein complexes**
  The human protein complexes are groups of proteins that have strong evidence that they are interact with each other. The protein complexes were extracted from HPRD (Human Protein Reference Database), release 9 [9]. This data set contains 1521 of manually curated protein complexes.

- **Molecular Functionality**
  For studying Molecular functions of the Human genes, the Molecular Functions from the GO annotation database were used (release date of the data base: 11/15/2011, version: 1.216) [12].

2.2 Performing Experiments

After obtaining all the datasets needed for this work, both MCL and K-means were employed to discover the interesting patterns.

Firstly, the MCL algorithm was applied to the HPRD interaction network. MCL uses a parameter that controls its operation. That parameter is called inflation. I used an inflation of 1.75 which is considered optimal according to [6]. I only considered the resulting patterns of sizes greater than or equal to 4 genes. The MCL produces connected modules when applied to the interaction network. The result contains 776 modules.

Table 1. Disease Datasets for expression profiles

GEO Series ID	GEO DataSet ID	Disease Name
GSE2503	GDS2200	Actinic keratosis
GSE1420	GDS1321	Adenocarcinoma of esophagus
GSE1297	GDS810	Alzheimer's disease
GSE5388	GDS2190	Bipolar disorder
GSE475	GDS289	Chronic obstructive lung disease
GSE1462	GDS1065	Chronic progressive ophthalmoplegia
GSE1629	GDS1850	Complex dental cavity
GSE3585	GDS2205	Congestive cardiomyopathy
GSE3365	GDS1615	Crohn's disease
GSE5370	GDS2153	Dermatomyositis
GSE2006	GDS1376	Essential thrombocythemia
GSE1751	GDS1331	Huntington's disease
GSE2018	GDS999	Lung transplant rejection
GSE3189	GDS1375	Malignant melanoma
GSE3868	GDS1746	Malignant neoplasm of prostate
GSE2549	GDS1220	Malignant pleural mesothelioma

2.3 Performing Experiments

After obtaining all the datasets needed for this work, both MCL and K-means were employed to discover the interesting patterns.

Firstly, the MCL algorithm was applied to the HPRD interaction network. MCL uses a parameter that controls its

operation. That parameter is called inflation. I used an inflation of 1.75 which is considered optimal according to [6]. I only considered the resulting patterns of sizes greater than or equal to 4 genes. The MCL produces connected modules when applied to the interaction network. The result contains 776 modules.

Secondly, the K-means algorithm was applied to the gene disease expression profile data. In order to have comparable number of modules to the one produced by the MCL, the number of modules resulting from the MCL was used as the number of K-means initial centers (clusters). Then, the analysis was performed with clusters of sizes greater than 4 genes (similar to what was done in MCL). K-means produced 1039 module of size ≥ 4 genes. The genes in these modules show similar behavior according to the disease expression data used in this study.

2.4 Human protien complexes

The following analysis was performed to assess the quality of the MCL and K-means produced modules in predicting known protein complexes. Those complexes are proved to have biological significance. Therefore, more matched complexes by the produced patterns means higher quality patterns from the biological point of view.

For assessing complex matching, I used the matching criteria introduced by [13]. The overlap score was computed by the formula $w = i^2/a*b$, where i is the size of the intersection set between the discovered patterns and the known complex, a is the size of the discovered protein pattern and b is the size of known protein complex. The overlap threshold between produced patterns and the protein complexes ranged from 0.1 to 1.0 with 0.1 increments; 1.0 implies a 100% match between the module and the protein complex. The range was chosen to make the overlapping abilities of the discovered patterns a tangible amount illustrating the quality of these patterns. Furthermore, we have not considered any overlap to be considered as a match which would may be seen as an overstatement of these results. More overlap indicates a better match with protein complexes, however knowing that more protein complexes are yet to be identified, even a lower match percentage can highlight some promising results. Figure 1 illustrates the MCL performance. While Figure 2 shows the performance of K-means prediction capability. Clearly, MCL has superiority over the K-means and was able to predict protein complexes even with higher overlap thresholds. This can be referred to the fact that many of the complexes are known to be well connected sub-graphs originally, and MCL aims at producing connected modules. Furthermore, K-means does not care about patterns connectivity. It is only concerned of expression similarity.

Another note about the results that I found interesting is that the larger the produced pattern the more penalty it will face in the complex matching process. This is because the average size of protein complexes is ≈ 5 genes and the criteria used for matching penalizes large patterns.

2.5 OMIM disease modules

The OMIM dataset provides a comprehensive collection of diseases that relates a large number of known diseases to the genes that are causing them or have a strong involvement in the cause and the mechanisms of the disease, in other words the Etiology and Pathogenesis of a disease. To evaluate the produced patterns and to have more insights that they can provide in the disease domain, both MCL and K-means produced patterns were tested against the OMIM disease dataset. This analysis aimed at discovering modules that might be related to known diseases.

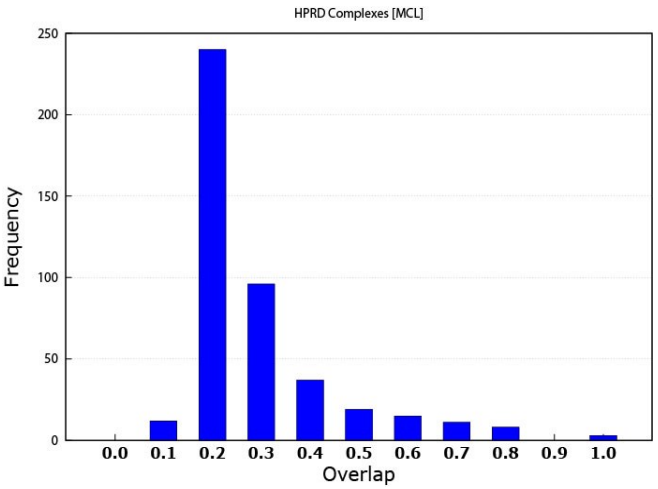


Figure 1. MCL protein complex prediction

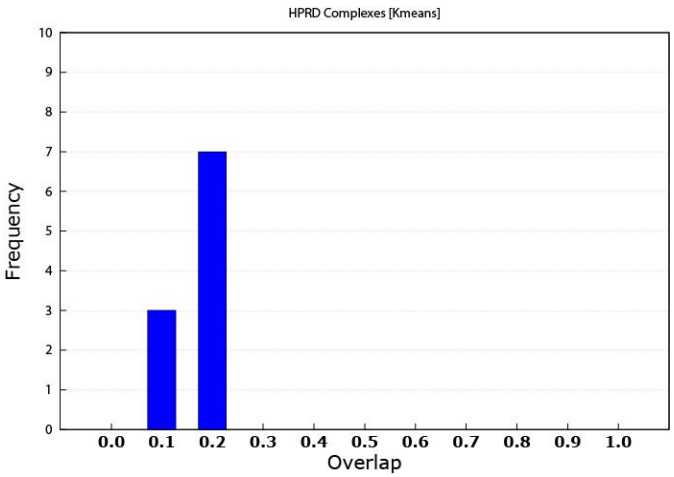


Figure 2. K-means protein complex prediction

As expected, the K-means slightly outperformed the MCL in this case. In this work, the patterns in the OMIM dataset are referred to as disease modules.

MCL produced patterns matched 68 disease modules while K-means matched 134 disease modules. The reason is that K-means patterns were originally based on data that is disease related. Some of the diseases that were used to create expression profile dataset were among the ones that were found from the OMIM dataset. As mentioned above, this study used only 16 diseases expression data, while the resultant matched diseases were multiples of that number. An explanation is that some genes can be involved in multiple diseases. This note means that some genes play the rule of a link between different diseases.

Thus, targeting these specific genes can illuminate hidden information that might lead interesting results; the cure for example. However, this statement needs more investigation by intensive research that is out of the scope of this work. In addition, a produced pattern can match more than one disease.

Figure 3 shows the performance of MCL when used with the OMIM dataset. Figure 4 is similar but for K-means. Table 2 and Table 3 show some of the disease matched by the MCL and K-means modules respectively. More specifically, Table 3 shows that some produced modules by K-means has matched multiple types of disease.

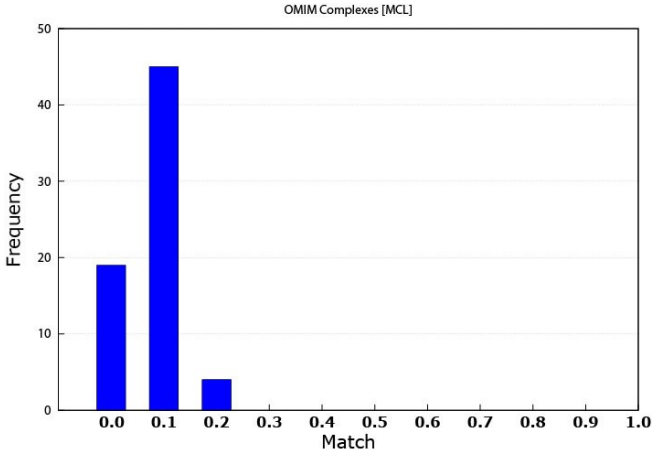


Figure 3. MCL matching performance against the OMIM data (0.0 means that the threshold is ≤ 0.1)

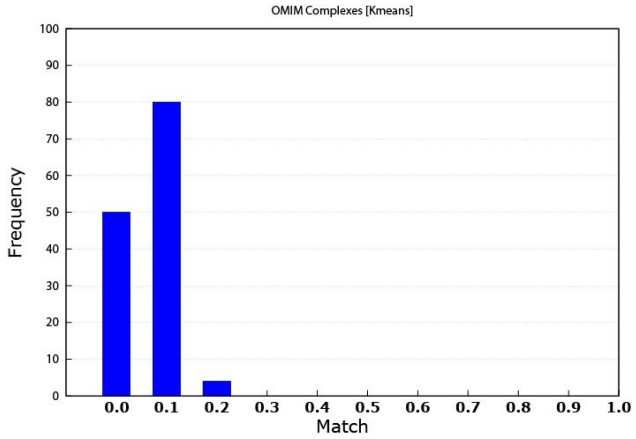


Figure 4. K-means matching performance against the OMIM data (0.0 means that the threshold is ≤ 0.1)

Table 2. Some of the MCL matched disease

Module's Genes	Related Disease
CDK2AP1 CDK2AP2 WARS PC GOT2 MDH2 CS ZDHHC6 FH NDUFS1	Thrombophilia due to protein C deficiency autosomal dominant
FANCC FANCA FANCE FANCG FANCF HES1 SAMD3 CYP19A1	Fanconi anemia complementation
GORASP2 RAB2A BLZF1 MIF NQO2 KCTD5 FAM71C	Persistent Mullerian duct syndrome type I
LYST SEMA4C DGCR14 MRPL17 NCDN DNAJC14 CNTROB	Chediak-Higashi syndrome
SHH PTCH1 PTCH2 IHH SMO DHH HHIP	Basal cell carcinoma somatic
NSF GABBR1 NAPG ATF5 GABBR2 PTPN9	Nicotine dependence susceptibility
CYP17A1 POR CYP2C19 CYP1A2 CYP2C9 CYP2E1	Mephenytoin poor metabolizer
NHEJ1 LIG4 XRCC4 APLF IRX5	Multiple myeloma resistance, Severe combined immunodeficiency with sensitivity to ionizing radiation
KDR ITGA9 FLT4 FIGF VEGFC	Hemangioma capillary infantile somatic
DHX9 SERPINB2 C6 MGEA5 PRPF8	Combined C6/C7 deficiency

3. Molecular functionality

As a further step, the Molecular Functionality (MF) of the genes can be considered as a useful source of information in many aspects. First, knowledge about the MF of the genes in the produced patterns is informative especially if some of the genes in the pattern are not well studied and their functions are not well known, but the association in the pattern may result in new data about these genes. Second, the combination between the MF of the patterns and the disease expression data can lead to meaningful indications about the disease mechanisms inside the living organism, the human in this case. Molecular Functions Gene Ontologies were extracted from the Gene Ontology project [12]. Studying the resulting modules showed that modules span multiple molecular functionalities and they did not show any tendency towards some particular biological processes. However, modules gene members have shown high similarities in what biological processes they share.

Table 3. Some of the K-means matched disease

Module's Genes	Related Disease
PKD2 IGBP1 TXNL4A NID1	Polycystic kidney disease
PKD2 IGBP1 TXNL4A NID1	Corpus callosum agenesis of with mental retardation ocular ,coloboma and micrognathia
SLC9A1 ADAM2 GMPS HIVEP2	Leukemia acute myelogenous
PCSK7 TES BSN COL10A1	Metaphyseal chondrodysplasia Schmid
PHEX INSL3 WDR61 APPL1	Cryptorchidism idiopathic
SI PDX1 BNIP2 DNAJA2	Sucraseisomaltase deficiency congenital
SI PDX1 BNIP2 DNAJA2	Lacticacidemia due to PDX1 deficiency
PTRF MPP3 TSPAN4 SERPIND1 GRIP1	Thrombophilia due to heparin cofactor II deficiency
AKAP4 MDK SH2D1A SLC12A3 RPP40	Mesomelic dysplasia Kantaputra
GADD45G TRO NPR1 BACH1	Breast cancer early-onset

Figure 5 has an example module that is enriched in a number of multiple molecular functions such as, GO:0008601: Modulation of the activity of the enzyme protein phosphatase type 2A. GO:0019888: Modulates the activity of a protein phosphatase, an enzyme which catalyzes of the removal of a phosphate group from a protein substrate molecule GO:0019208: Modulates the activity of a phosphatase, an enzyme which catalyzes of the removal of a phosphate group from a substrate molecule. GO:0005488: The selective, non-covalent, often stoichiometric, interaction of a molecule with one or more specific sites on another molecule. GO:0004512: Catalysis of the reaction: D-glucose 6- phosphate = 1D-myo-inositol 3-phosphate. This reaction requires NAD, which dehydrogenates the CHOH group to CO at C-5 of the glucose 6-phosphate, making C-6 into an active methylene, able to condense with the aldehyde at C- 1. Finally, the enzyme-bound NADH reconverts C-5 into the CHOH form. GO:0005515: Interacting selectively and non-covalently with any protein or protein complex. The above results were obtained using The Database for Annotation, Visualization and Integrated Discovery (DAVID) [14] [15].

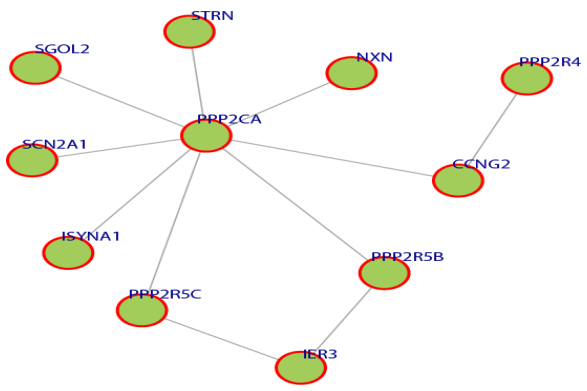


Figure 5. A module that was extracted from the network where all the genes have similar molecular functionality.

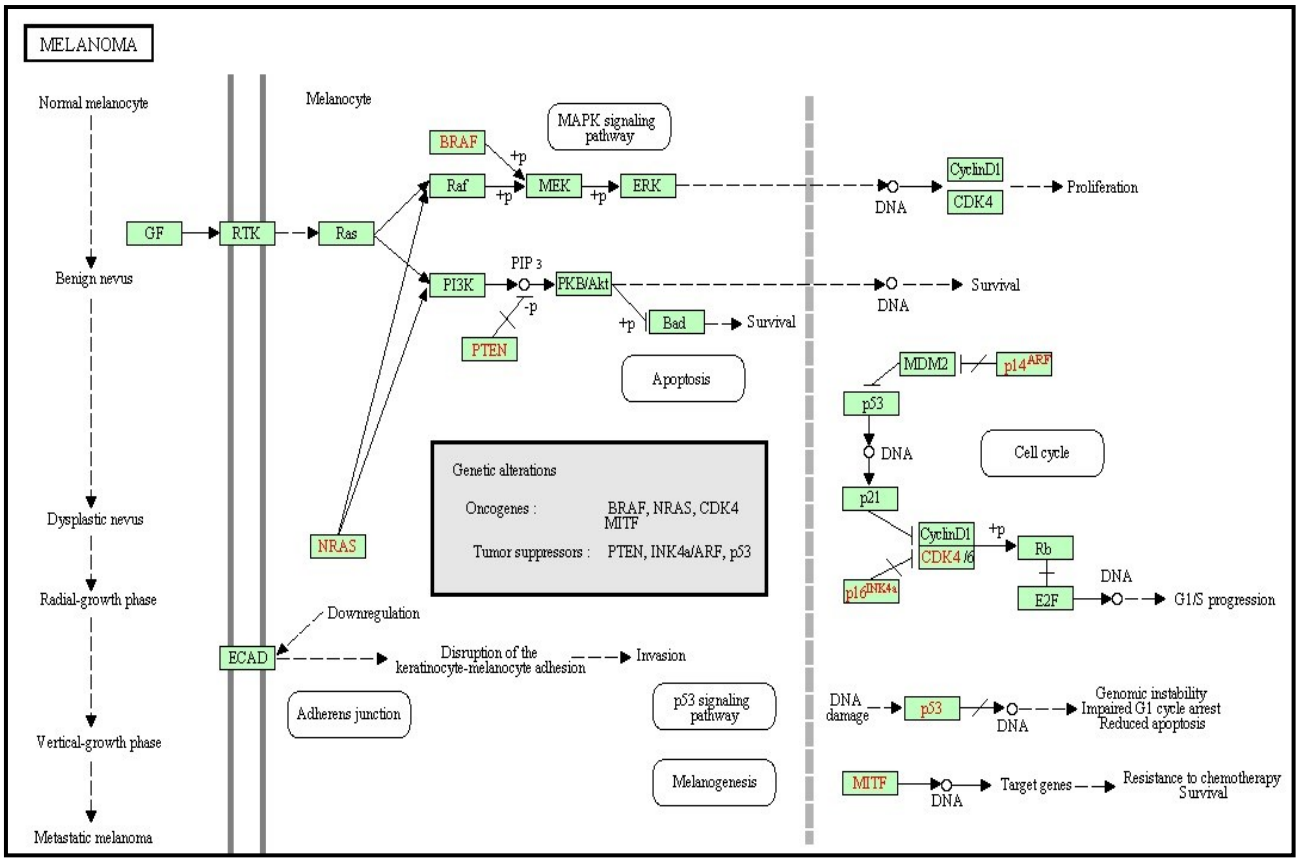


Figure 6. The prostate cancer pathway.

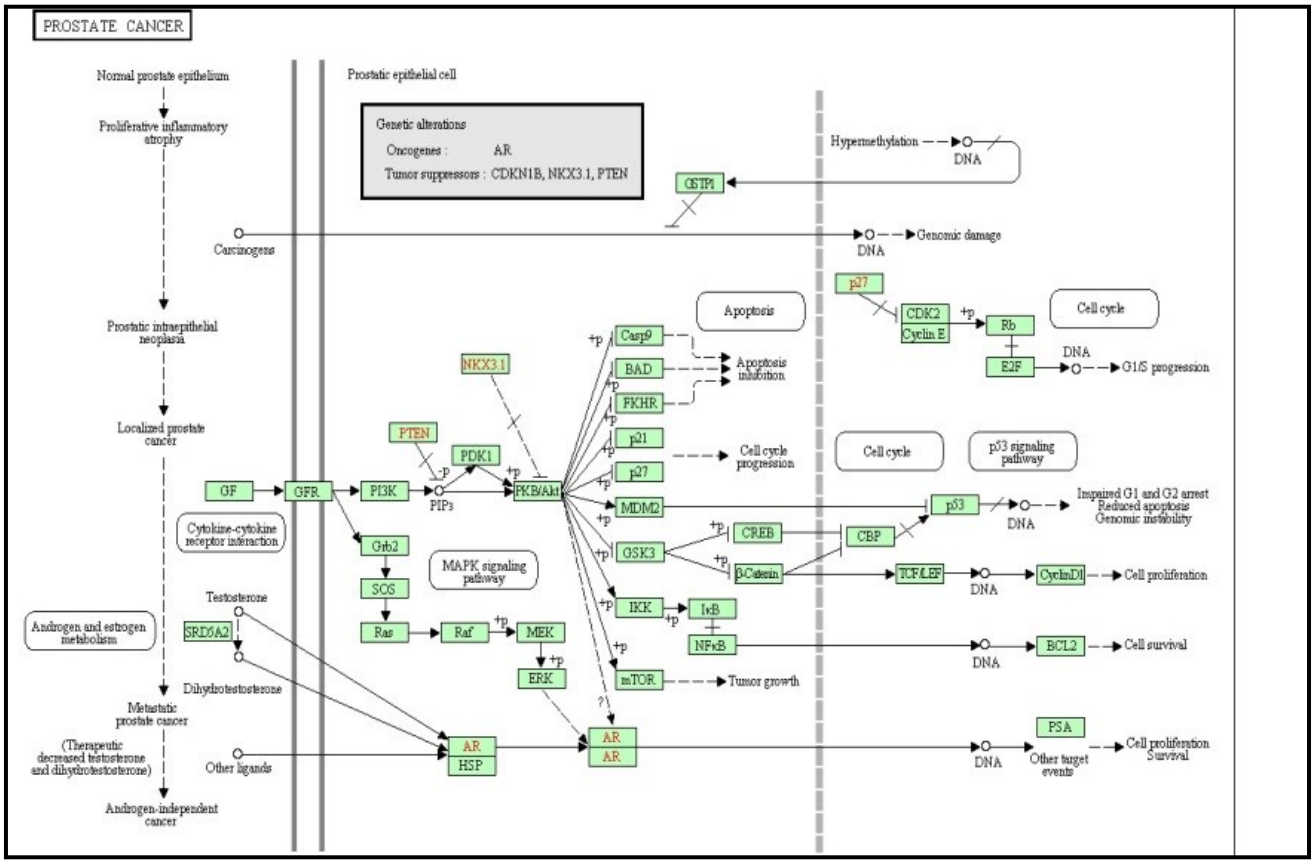


Figure 7. The Melanoma pathway.

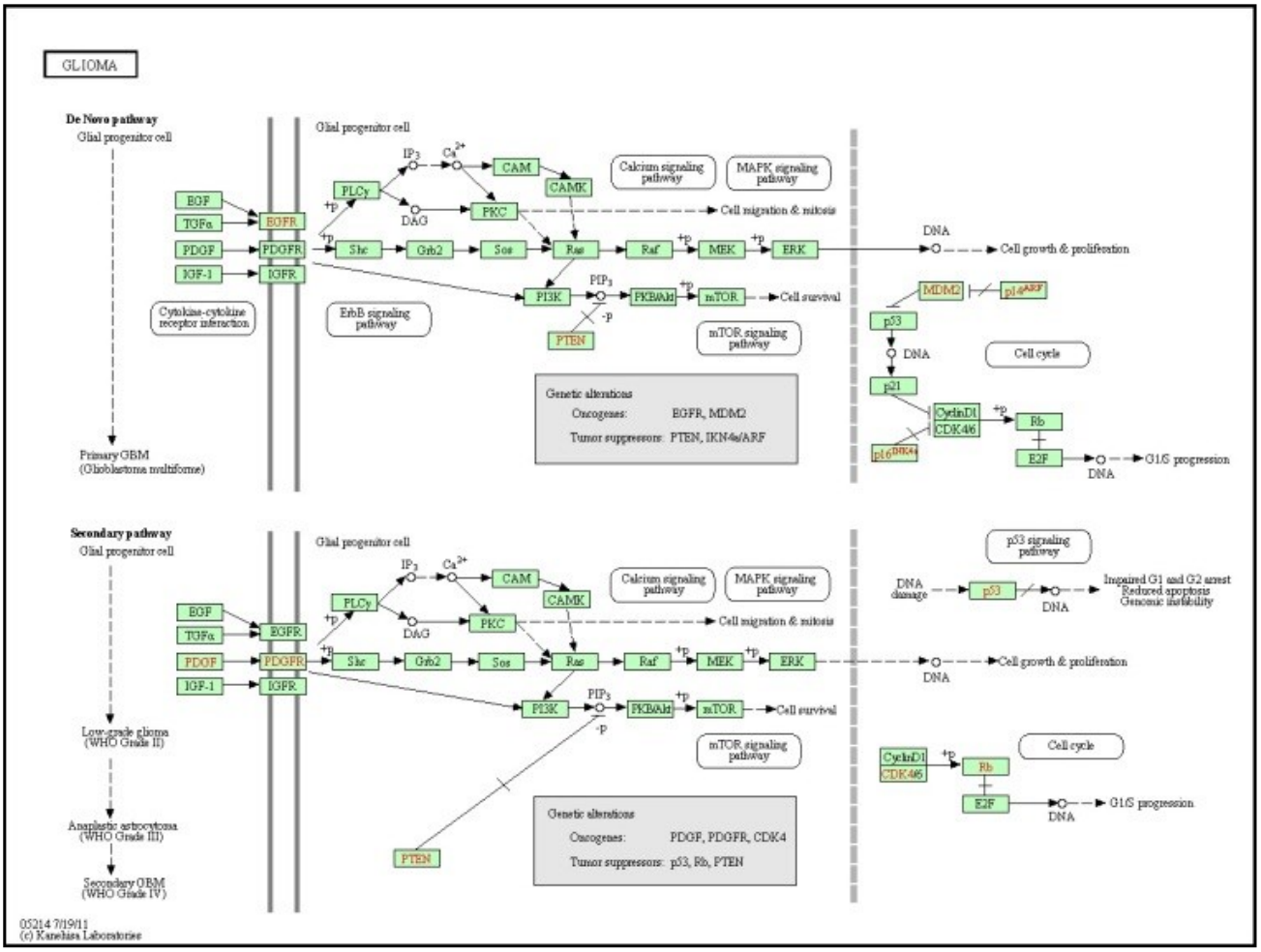


Figure 8. The Glioma pathway.

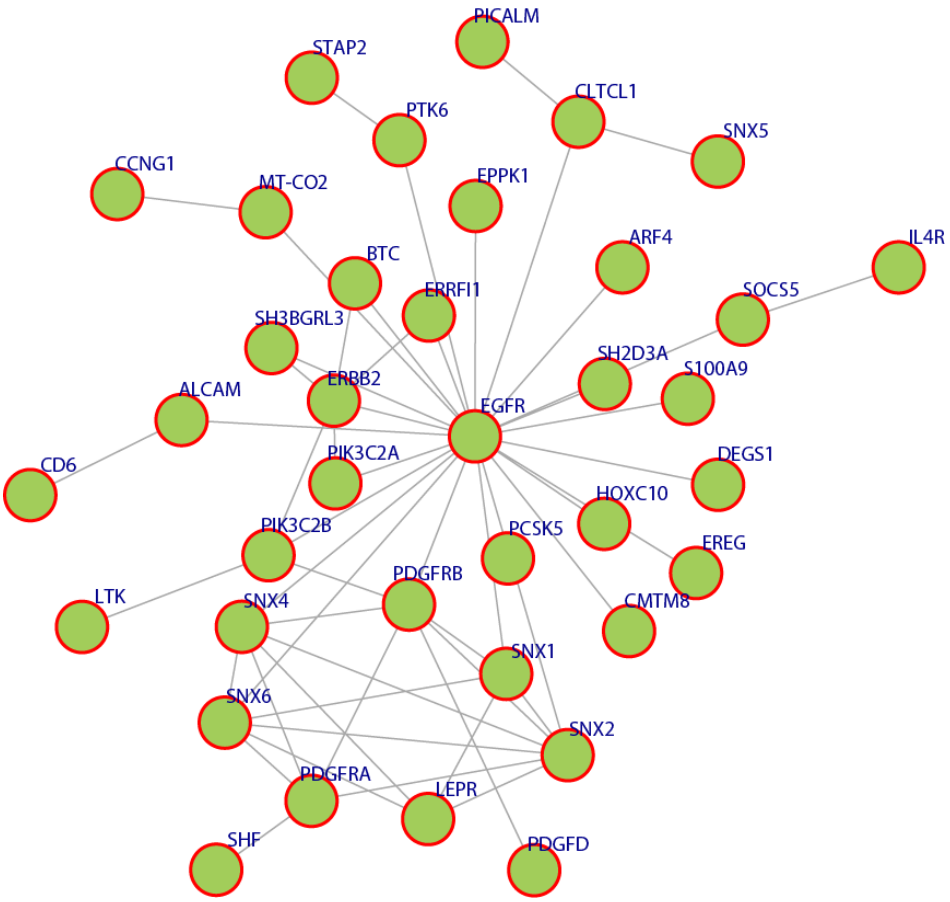


Figure 9. This discovered pattern above shows a module that is overlapped with multiple disease pathways

4. Pathways

Many of the discovered modules have biological pathways associations. Biological pathway is a group of actions that occur between the molecules in the cell that result in the formation of a new product or change in the cell. This association with pathways was checked against the Kyoto Encyclopedia of Genes and Genomes (KEGG) [16]. To illustrate the idea further, some example pathways chosen were very interesting in that they have a relation with different types of diseases. Figure 6, 7, 8, and 9 have some examples of disease-related pathways like: Prostate cancer pathway, Glioma pathway, and Melanoma pathway. In addition, another example (not shown here) matched the Notch Signaling Pathway. Notch signaling has been related to a number of malignancies including leukemia, lymphomas and carcinomas of the breast, skin, lung, cervix and kidneys [9]. It is clear from Figure 6, that one of the discovered modules has several genes that are highly incorporated in multiple disease pathways. Furthermore, we can see that not all of the module genes are present in those pathways; this raises an interesting question; can future research prove that those absent genes are related to the same pathway and therefore to the diseases because they were clustered with several of the pathway genes?

5. Conclusion

The availability of large amounts of microarray data in, both interaction networks and expression profiles, has been used to have deeper insights of what biological processes are taking place inside an organism. Much of this data is related to different aspects of the living cell activities. In addition, this expression data can be extracted from organism or tissues under a group of experimental, environmental, or any kind of stress conditions. Diseases expression data can be considered a very useful source of information regarding what genes are involved in or affected by them. In this study I tried to shed some lights about modules extracted from two separate sources of information and what relations they may have with known diseases and try to uncover some of new connections between genes and diseases. Both of the two methods employed here, MCL and K-means, have its strength sides and weakness sides. When it comes to interaction networks, MCL was superior. However, K-means has shown acceptable performance with expression profile data. The method has produced interesting patterns when they are tested against several biological concepts such as protein complexes, disease modules, molecular functionality, and biological pathways.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

I would like to thank the reviewers of this work for their help. This work was accomplished by the support of Southern Arkansas University, College of Science and Engineering.

References

[1] Y. Fang, M. Sun, G. Dai and K. Ramain, "The Intrinsic Geometric Structure of Protein-Protein Interaction Networks for Protein Interaction Prediction," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 1, pp. 76-85, Jan.-Feb. 1 2016. doi: 10.1109/TCBB.2015.2456876

[2] M . Bansal, V. Belcastro, A. Ambesi-Impiombato, D. di Bernardo." How to infer gene networks from expression profiles". Mol Syst Biol 3: 78, 2007. https://dx.doi.org/10.1038%2Fmsb4100120

[3] I. Ulitsky, I. Gat-Viks, R. Shamir. "MetaReg: a platform for modeling,analysis and visualization of biological systems using large-scale experimental data". Genome Biol 9: R1, 2008. https://dx.doi.org/10.1186%2Fgb-2008-9-1-r1

[4] S. van Dongen. "A new cluster algorithm for graphs". Report No. INS-R0012, Center for Mathematics and Computer Science (CWI), Amsterdam, 2000.

[5] S. van Dongen. "Graph clustering via a discrete uncoupling process". SIAM J Matrix Anal Appl, 30(1): 121–141, 2008. https://doi.org/10.1137/040608635

[6] S. Broh´ee and J. van Helden. "Evaluation of clustering algorithms for protein-protein interaction networks". BMC Bioinformatics, 7(488), 2006. https://doi.org/10.1186/s12859-014-0445-4

[7] J. B. MacQueen. "Some methods for classification and analysis of multivariate observations" Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. https://projecteuclid.org/euclid.bsmsp/1200512992

[8] L. Liberti, C. Lavor, N. Maculan, and A. Mucherino. "Euclidean Distance Geometry and Applications". SIAM Review. 56:1, 3-69. 2014. DOI10.1137/120875909.

[9] L. Baolin and H. Bo. "Hprd: a high performance rdf database". In Proceedings of the 2007 IFIP international conference on Network and parallel computing , NPC'07, pages 364–374, Berlin, Heidelberg, 2007. Springer-Verlag. https://doi.org/10.1007/978-3-540-74784-0_37

[10] T. Barrett , D.B. Troup , S. E. Wilhite , P. Ledoux , D Rudnev, et al. "NCBI GEO: archive for high-throughput functional genomic data". Nucleic Acids Res 37: D885–890. 2009. https://doi.org/10.1093/nar/gkn764

[11] "Online Mendelian Inheritance in Man, OMIM". McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). 2011. URL: http://omim.org/. https://doi.org/10.1093/nar/gki033

[12] The gene ontology consortium. "Gene ontology: tool for the unification of biology". Nat. Genet., 25(1):25–29, May 2000. http://dx.doi.org/10.1038/75556

[13] G. D. Bader and W.V. Christopher. "An automated method for finding molecular complexes in large protein interaction networks". BMC Bioinformatics, 4 (2), 2003. https://dx.doi.org/10.1186%2F1471-2105-4-2

[14] D. W. Huang, B. T. Sherman, R. A. Lempicki. "Systematic and integrative analysis of large gene lists using david bioinformatics resources". Nat. Protocols, 4(1754-2189), 2008/12//print. http://dx.doi.org/10.1038/nprot.2008.211

[15] D. W. Huang, B. T. Sherman, and R. A.Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". Nucleic Acids Research, 37(1):1–13, 2009. https://dx.doi.org/10.1093/nar/gkn923

[16] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. "Kegg for representation and analysis of molecular networks involving diseases and drugs". Nucleic Acids Research, 38(suppl 1):D355–D360, 2010. https://dx.doi.org/10.1093/nar/gkp896

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 2, 90-98 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Agent Based Fault Detection System for Chemical Processes using Negative Selection Algorithm

Naoki Kimura*, Yuya Takeda, Yoshifumi Tsuge

Department of Chemical Engineering, Faculty of Engineering, Kyushu University, 819-0395, Japan

ARTICLEINFO

Article history:
Received: 15 November, 2017
Accepted: 07 January, 2018
Online: 12 March, 2018

Keywords:
Negative selection algorithm
Plant alarm system
Fault detection

ABSTRACT

Recently, the number of industrial accidents of chemical plants has been increasing in Japan. The fault detection system is required to keep chemical plant safely. In this study, a fault detection system for a chemical plants using agent framework and negative selection algorithm was proposed. The negative selection algorithm is one of artificial immune systems. The artificial immune system is an imitative mechanism of vital actions to discriminate self/nonself to protect itself. The method was implemented and applied to a complicated chemical plant—which is a boiler plant virtually operated using a dynamic plant simulator. The simulations of fault detection were carried out. And also the results of simulations are presented in this paper.

1 Introduction

This paper is an extension of work originally presented in 2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP) [1].

Recently, it has been increasing the accidents of chemical plants. Figure 1 shows the annual numbers of the industrial accidents in Japan (the data is based on the summary of accidents in specified business facilities inside petrochemical complex in Japan, published by Fire and Disaster Management Agency, Ministry of Internal Affairs and Communications, Japan, posted on their official web site on May 2017, article in Japanese). The numbers of the industrial accidents in Japan—except accidents caused by earthquakes or tsunami—has been increasing from 45 in 1993 to over 250 in 2016. It is said that the remote causes of the rise of accidents in Japan are mass retirements of skilled engineers, insufficient technical tradition, labor-savings in production lines or plant operations, aged deterioration of productive facilities, or maintenance cost reduction in assertive ways. Therefore, effective fault detection system for chemical plants is required. In the general chemical plant operations, *plant alarm system* has been used to notify the process deviance to operators via warning lights or buzzers in the operation rooms, where the upper and lower thresholds of the measured values or the thresholds of their amounts of changes have been set to the sensors in the chemical plants. However, it is so difficult to determine the adequate values of thresholds (that is 'alarm setpoint') that if the alarm setpoints are too small, the alarm floods will be caused, if the setpoints are too large, missed detection of deviation will be caused. And also it is difficult to detect if the plant has normal and regular load fluctuations under both the normal and the abnormal situations. Therefore, a method is required that observes the relationship among several variables to detect faults in a complicated system. We focus on the Artificial Immune System.

Artificial Immune Systems—which are imitative mechanisms of vital actions of discrimination between self and nonself—have been proposed since 1990s. And a lot of methods have been proposed using various parts of artificial immune systems, for example, pattern recognition by B-cells for fault detection in gas lift oil well by Aguilar [2, 3], Natural Killer (NK) immune cells mechanisms by Laurentys [4], clonal selection algorithm for maintenance scheduling of power generators by El-sharkh [5], dendritic cell for failure detection of aircraft by Azzawi [6]. Also lots of applications have been proposed—Wada et al. [7] proposed an fault mode detection method for automotive exhaust gas treatment system, Inomo et al. [8] proposed an failure diagnosis method for water supply network by using immune system.

*Corresponding Author, nkimura@chem-eng.kyushu-u.ac.jp

https://dx.doi.org/10.25046/aj030211

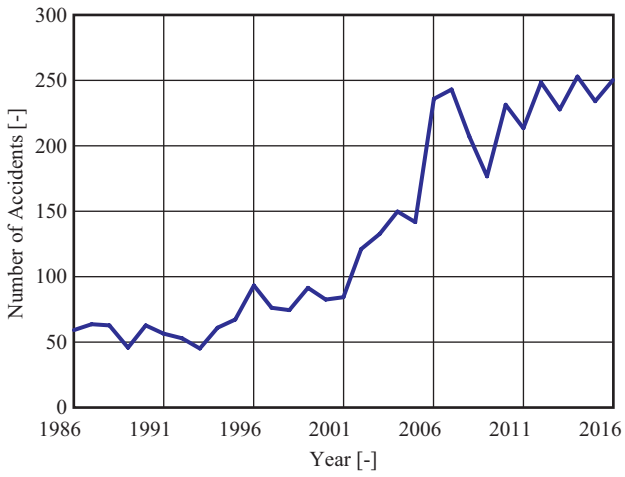


Figure 1: The annual numbers of the industrial accidents in Japan, data from FDMA, Japan.

In this study, we adopt a negative selection algorithm to detect faults in a chemical plant. Negative selection algorithm is an imitative method of mechanisms of differentiation and maturation, and discrimination of normal/abnormal. The algorithm was proposed by Forrest in 1994 [13] for the detection of computer virus. The negative selection algorithm has been applied to various domains—Dasgupta et al. [9] applied to aircraft fault detection, Gao et .al [10] applied to motor fault detection, Xiong et al. [11] and Prasad et al. [12] applied to fault detection in the Tennessee Eastman process.

In this study, we introduce detectors to detect faults based on the negative selection algorithm. The adopted method of the negative selection algorithm is mentioned in section 2. In order to utilize negative selection algorithm, we designed and implemented an agent based framework of fault detection system, mentioned in section 3. The target chemical process, the simulation conditions and the results are mentioned in section 4. And the conclusion in section 5.

2 Negative Selection Algorithm

Negative selection algorithm is one of the methods of the artificial immune systems inspired by the vital immune systems. Negative selection algorithm borrowed from the mechanism of T-cell generation in thymus. On T-cell generation in vital system, immature T-cells are randomly generated with various immunological types. And then some of T-cells are eliminated if they have high affinity with self-antigen to avoid response to "self". T-cells which are not self-affinitive can be matured to react with foreign antigen to protect itself.

In our system, detectors—correspond to mature T-cells in vital—can detect faults by recognizing the normal operational data of chemical processes which are assumed as *self* and the abnormal operational data which are assumed as *nonself*. In Figure 2(a)–(c), the steps of the generation phase of detectors are illustrated. Figure 2(a) illustrates that there are *self* regions in a 2-dimensional process variables space. And the rest area of the variables space are *nonself* regions. Figure 2(b) illustrates that detector candidates which indicated by plus(+) signs are generated with various position, where the radius is set as the minimum distance to the *self* region. In this study, we implement two steps of candidate generation. The first step is a grid-based generation. Place the detector candidates at every 0.1 of each axis—axis is normalized between [0.0, 1.0]—, therefore 121 (each axis divided into eleven) grid-based candidates are generated. The second step is a semi-randomly generation. Place a detector candidate at random position. Then check whether the new candidate is overlapped on the previously generated candidates. If it overlapped, the new candidate is eliminated in this step, if not, the new candidate is added to the candidates. And then, check the coverage of the detectors. If the coverage is over 90%, candidate generation will be terminated. From Figure 2(b) to (c), the elimination of some candidates which have affinity with "self" is carried out. Figure 2(c) illustrates that seven valid detectors are remained. Then the generation phase is finished.

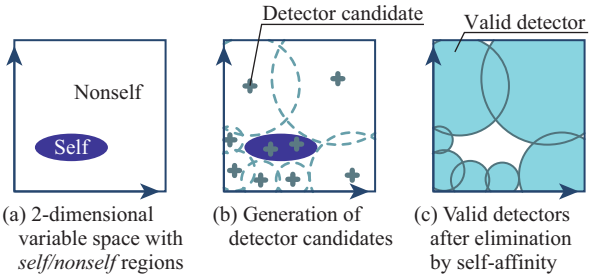


Figure 2: A schematic diagram of the detectors generation.

In the detection phase, a sample consists of the values of the current process variables—which are to be examined— is plotted into the variable space indicated by a 'star'. If the plotted sample is inside the detection area of at least one detectors, the sample is recognized as "abnormal" (Figure 3(a)). And if the sample is out of any detection area, the sample is recognized as "normal" (Figure 3(b)).

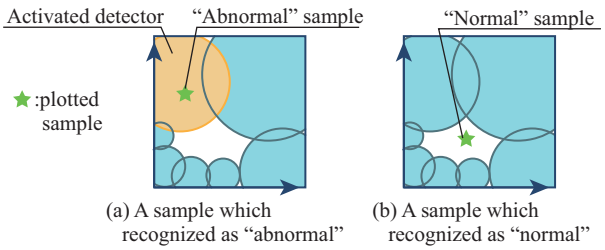


Figure 3: A schematic diagram of the detection by detectors.

3 Fault Detection System

A multiagent framework is adopted to implement a fault detection system using negative selection algorithm. In our system, there are *Detector Manager*, *Detector Leader(s)* and *Database Agent* illustrated in Figure 4. *Detector Manager* is an interface between human operator and the system. *Detector Leader(s)* have their own variable spaces to detect faults. These variable spaces represent relationships between two certain process variables. And there are a lot of detectors under the dominion of a *Detector Leader*. In order to avoid missing detection, some *Detector Leaders* with variety of combinations of the process variables are required. *Database Agent* have a database which stores operational database of the target process. All these agents can communicate with other agents via TCP/IP network connection.

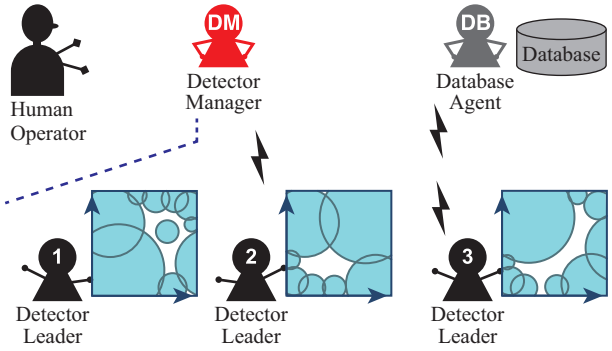


Figure 4: A schematic diagram of the detection by detectors.

4 Simulations and Results

4.1 Target Process

Figure 5 shows a boiler plant which is the target process of this study. It is an utility plant which supplies three steam headers—high pressure steam (HP), middle pressure steam (MP), and low pressure steam (LP) to the nearby plants by boiling pure water in a furnace. Due to the fluctuation of the steam demands from user plants, the boiler plant is always under unsteady operation. Therefore, it is difficult to detect faults by setting up the constant thresholds to some process variables of the boiler plant.

In this study, nine variables were selected from among 120 measured variables in the boiler plant. The selected nine variables are listed in Table 1 and also illustrated in Fig.5. The operational data was obtained by using dynamic plant simulator "Visual Modeler" (Omega Simulation Co., LTD).

4.2 Normal Operational Data

Data of a normal operation—in which the steam demands were stepwise changed without any malfunction—were obtained using the dynamic simulator. The data contain 7200 samples of the above-mentioned nine process variables and its sampling interval is one second. The trend graphs of PI1422 and TI1422 were indicated in Figure 6. These data were normal operational data which should be recognized as *self* in the artificial immune system even though the steam demands had increased at time 600 second.

4.3 Abnormal Operational Data

Abnormal operational data with three kinds of assumed malfunctions were obtained using the dynamic simulator. Table 2 shows the list of three assumed malfunctions. And Figures 8–10 illustrate the trend graphs of PI1422 and TI1422 when the one of malfunctions was occurred at time 600 second, but without any steam demand change during these 7200 seconds.

Table 1: Selected nine process variables of the boiler plant.

Variable name	Description
FI1422	Flow rate of the second extraction steam from the turbine
PI1422	Pressure of the second extraction steam from the turbine
TI1422	Temperature of the second extraction steam from the turbine
TC1423	Temperature of the low pressure steam from the desuperheater
PI1315	Pressure inside the furnace (upper)
GB401.pos	Valve position of the combustion air
PI1311	Pressure of the outlet of the forced draft fan
TI1310	Temperature of the exhaust gas at the gas air heater outlet
TI1308	Temperature of the combustion air at the gas air heater outlet

4.4 Generation of Detectors

Detectors were generated by using normal operational data for each of 36 combinations of 2-dimensional variable spaces—all the 2-dimensional variable spaces consists of two different process variables from the nine variables. Figure 7 illustrates the detectors in the variable space consisting of TI1422 vs. PI1422. The axes are normalized to the range of the normal operation occupies [0.05, 0.95] in the normalized axis respectively. In this figure, there are 261 detectors (118 grid detectors and 143 randomized detectors) which indicated as sky blue circles. In other words, it is recognized that there exist the normal operational data—which correspond "self" region—at

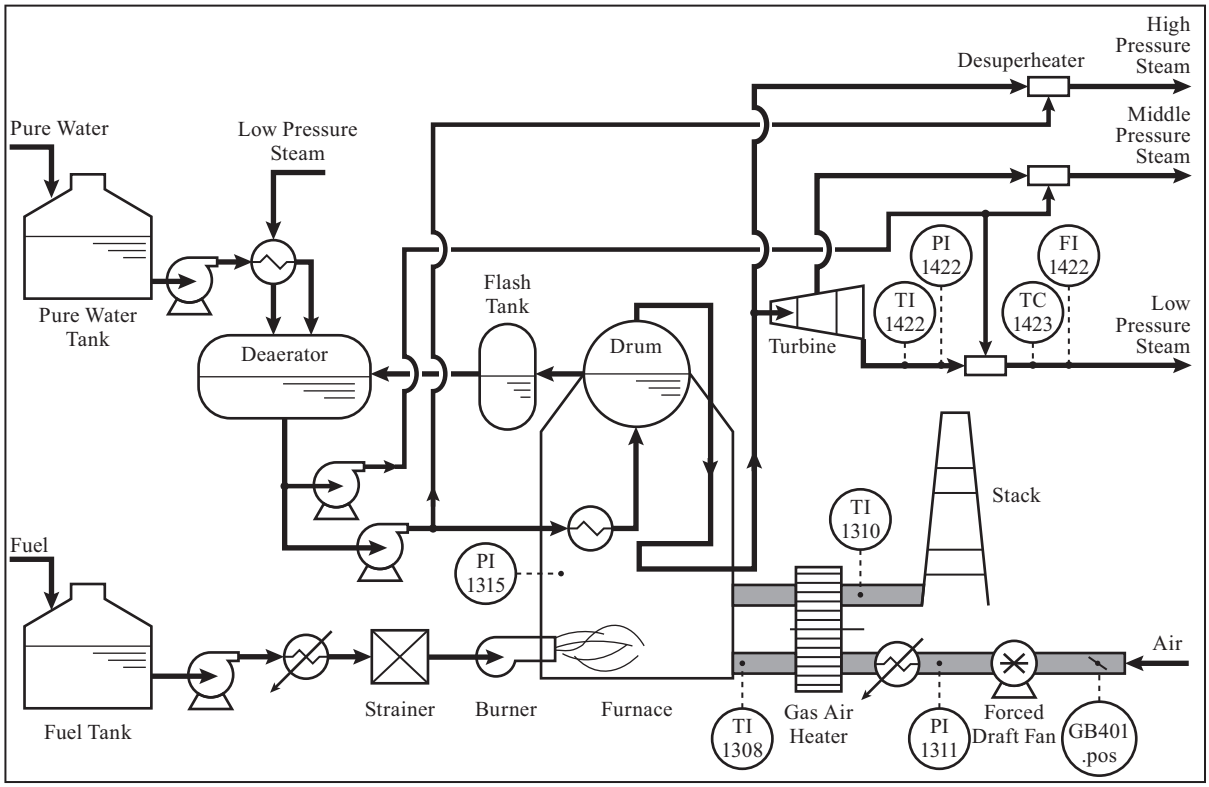


Figure 5: A schematic diagram of the boiler plant.

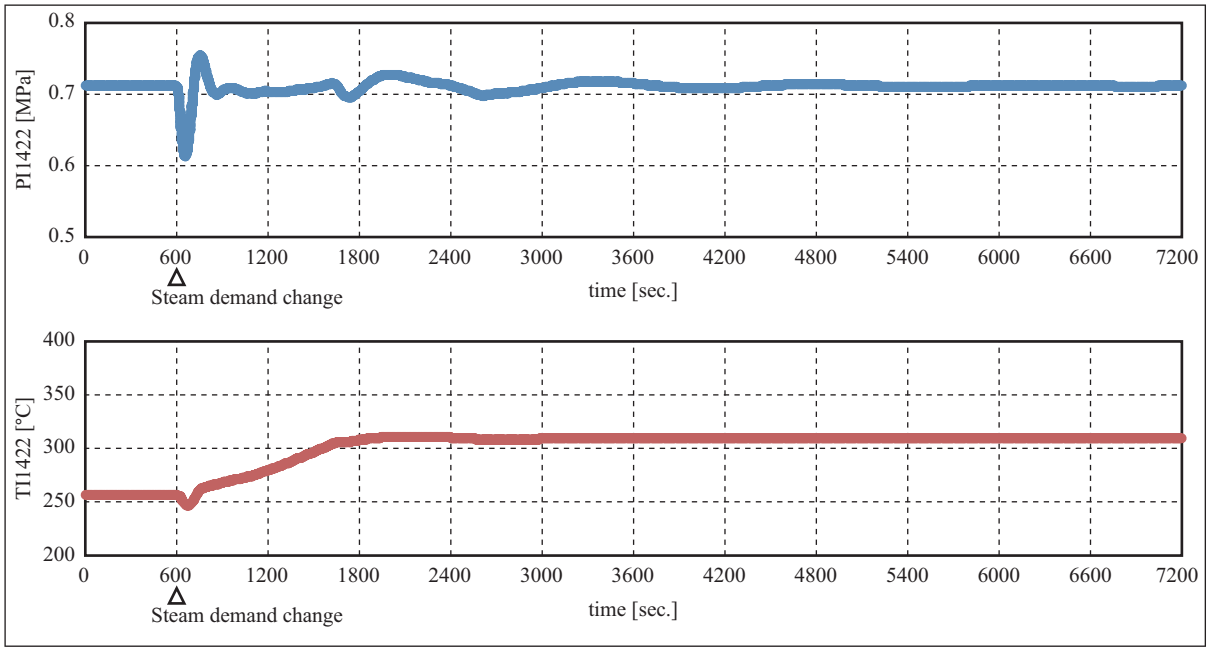


Figure 6: The trend graphs of PI1422 and TI1422 during normal operation with steam demands change at time 600 second.

the unpainted parts in Figure 7, and not exist the normal operational data—which correspond "nonself" region—at the painted sky blue parts. The sets of detectors were generated for the rest of variable spaces in the same way. The numbers of the detectors in each variable space were from 121 to 316 including both the grid detectors and randomized detectors.

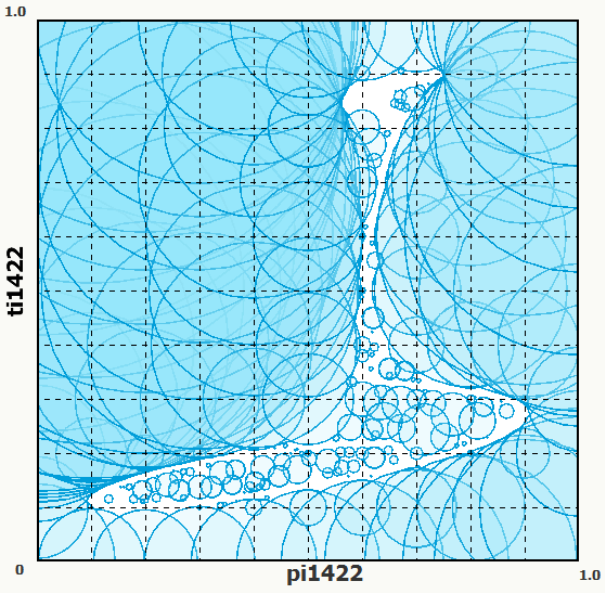


Figure 7: Generated detectors for a 2-dimensional variable space (TI1422 vs. PI1422).

The bold dashed lines in the trend graphs are the maximum and minimum values under the normal operation (Figure 6). We can find that it is impossible to detect fault using upper/lower thresholds of variable PI1422 and/or TI1422 when the malfunction BFO or GAH occurred, because the values of PI1422 and TI1422 have not exceeded the ranges under the normal operation. On the other hand, it may possible to detect at time 671 second by lower threshold of TI1422 when the malfunction FDFclose occured at time 600 second. TI1422 also exceeds upper threshold after time 1340 second, shown in Figure 10.

Table 2: Three assumed malfunctions in the boiler plant.

Malfunction ID	Description
BFO	Burner frame out
GAH	Gas air heater rotation failure
FDFclose	Forced draft fan inlet vane closure

4.5 Fault Detection using Detectors

Fault detections for the abnormal operational data were carried out. Figure 11 shows an outline of the fault detection in a 2-dimensional variable space consisting of TI1422 and PI1422 when malfunction BFO occurred. The axes and the detectors—indicated by sky blue circles— are the same as Figure 7. The sampling data were plotted by green dots on the 2-dimensional variable space after the normalization for every second serially. The blue dots are the normal operational data and the green dots are the sampling data to be examined by the detectors. These dots are moving in the variable space as time proceeds. If the green dots were placed on the unpainted region, they are recognized as "self"—where the values are similar to the normal operational data. If the green dots were placed over the sky blue region, they are recognized as "nonself"—in other words, a fault was detected in this variable space by detector(s).

In Figure 11, the painted orange circles indicate activated detectors—which have detected fault. Figure 12 shows the detection status through time. If the value is '1', at least one detector in this variable space detects fault, and if the value is '0', no detector detects fault at that time. The figure shows that the first detection was at time 630 second—which is 30 second after malfunction occurred—, and there are missing between time 897 and 904 second in this variable space. The detections using the sets of detectors in the rest of variable spaces were simultaneously carried out in the same way. Figure 13 shows the number of variable spaces whose detection status is '1'. The figure shows that there are 15 variable spaces detected at time 601 second, 30 variable spaces—which corresponds to 80% of 36 variable spaces—detected at time 604 second, and the number does not fall below the 80% after that time. Therefore it can be said that this system can detect malfunction BFO successfully.

Figure 14 shows the outline of the fault detection, Figure 12 shows the detection status through time when malfunction GAH occurred. The figures show that the fault was detected at time 666 second by the detectors. On the other hand, it have not been detected after time 1047 second in this variable space TI1422 vs. PI1422. However 16 shows that 21 variable spaces—which corresponds to 58% of 36 variable spaces—detected at time 602 second, and the number does not fall below 58% after that time. It can be said that this system can detect malfunction GAH successfully, although the variable space TI1422 vs. PI1422 could not detect after time 1047 second.

Figure 17 also shows the outline of the fault detection, Figure 18 shows the detection status through time when malfunction FDFclose occurred. The figures show that the fault was detected at time 630 second by the detectors. Although the detection status fluctuate between time 897–1102 second in the variable space TI1422 vs. PI1422, over 30 variable spaces detect after time 604 second and the number does not fall below the 80% of the 36 variable spaces after time 604 second. It can be also said that this system can detect malfunction FDFclose successfully.

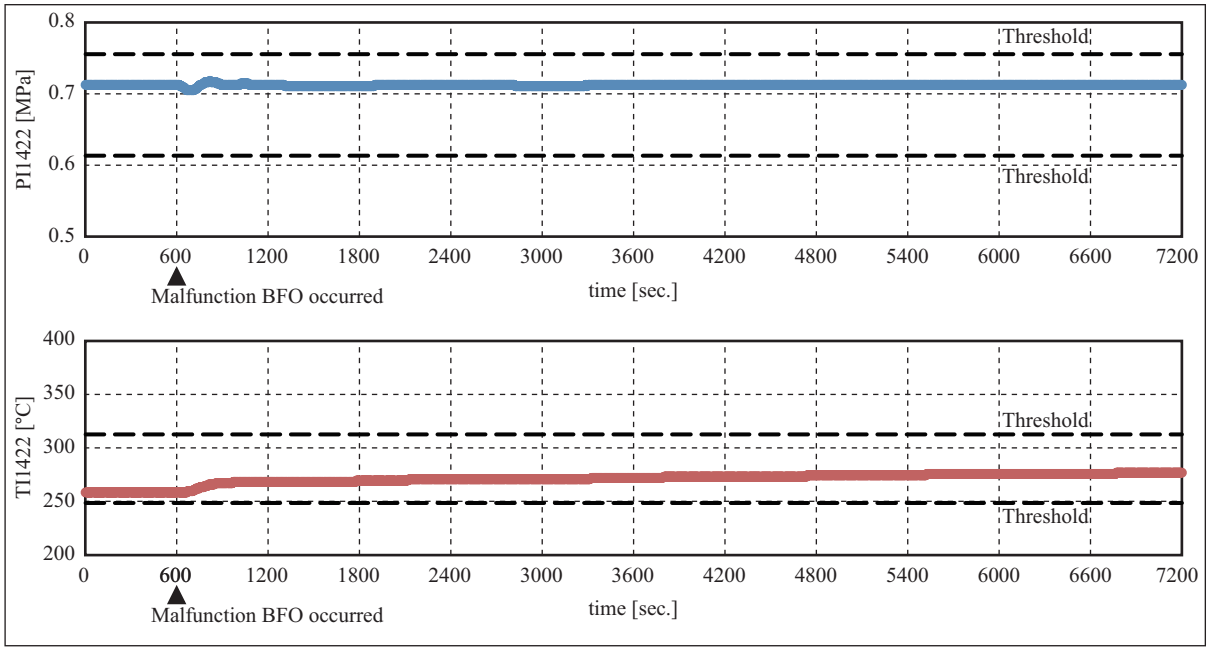


Figure 8: The trend graphs of PI1422 and TI1422 when malfunction BFO was occurred at time 600 second without steam demands change.

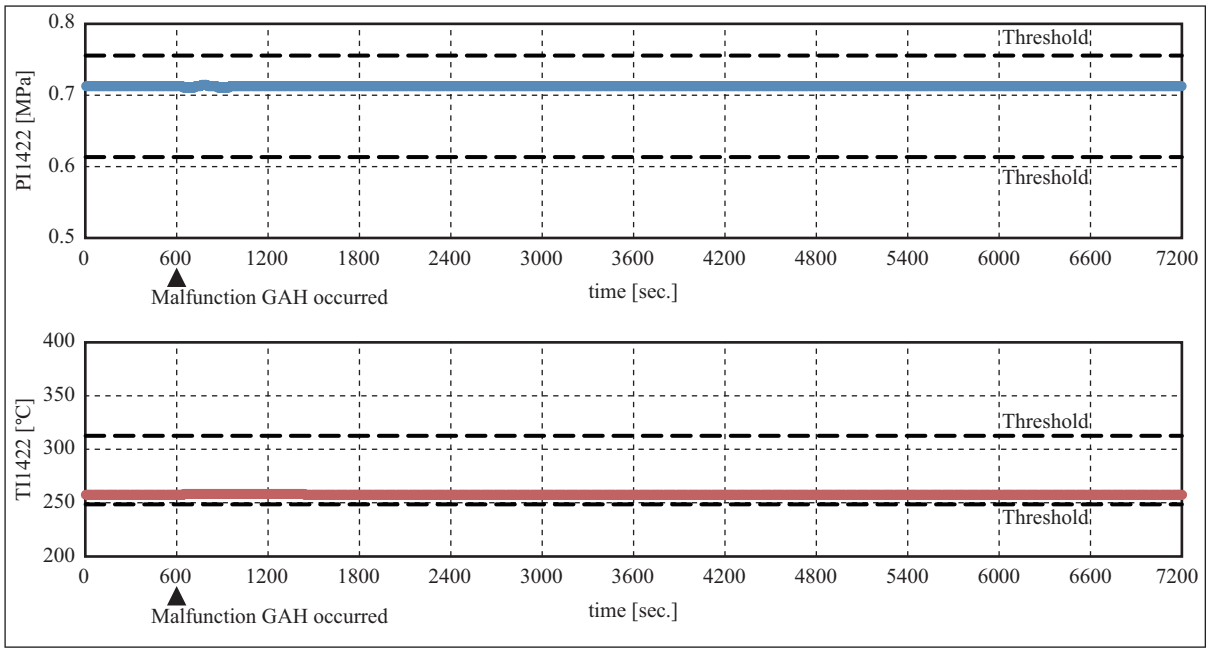


Figure 9: The trend graphs of PI1422 and TI1422 when malfunction GAH was occurred at time 600 second without steam demands change.

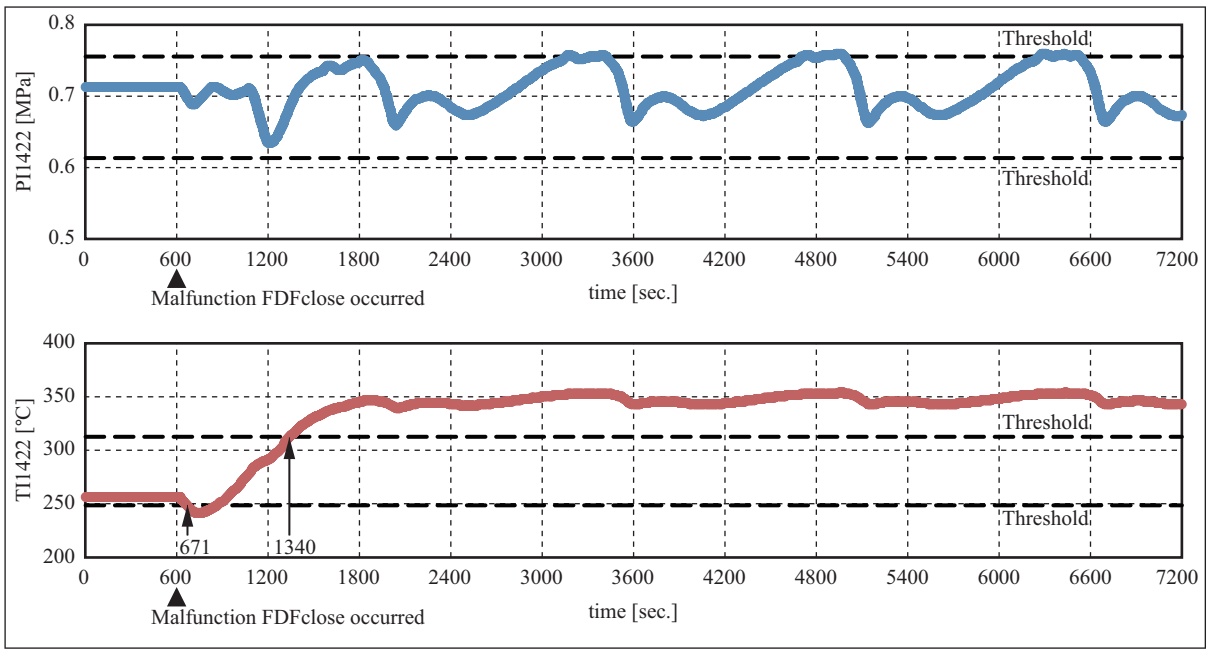


Figure 10: The trend graphs of PI1422 and TI1422 when malfunction FDFclose was occurred at time 600 second without steam demands change.

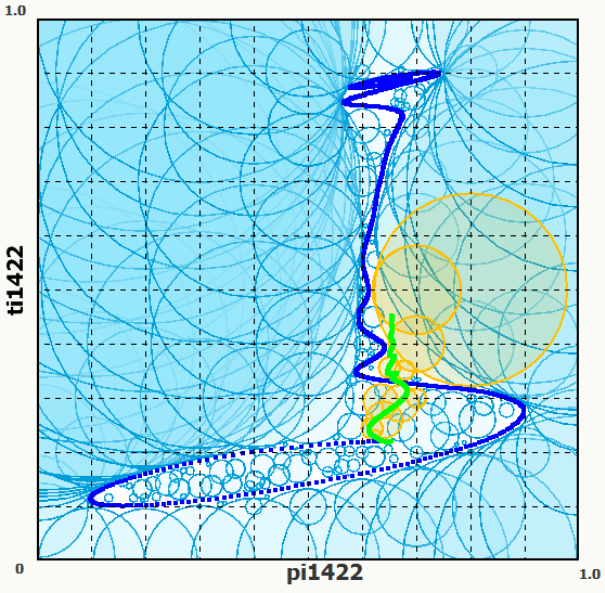


Figure 11: Fault detection by detectors in a 2-dimensional variable space when malfunction BFO was occurred.

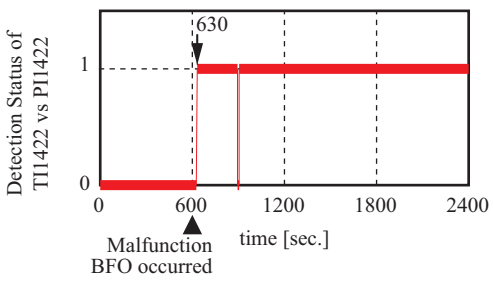


Figure 12: Detection status of the TI1422 vs. PI1422 variable space when malfunction BFO was occurred.

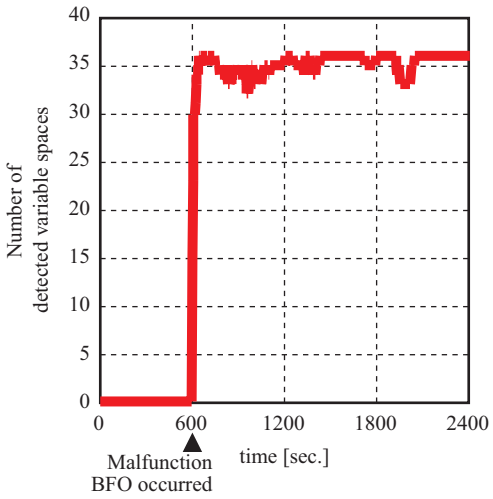


Figure 13: The number of detected variable spaces when malfunction BFO was occurred.

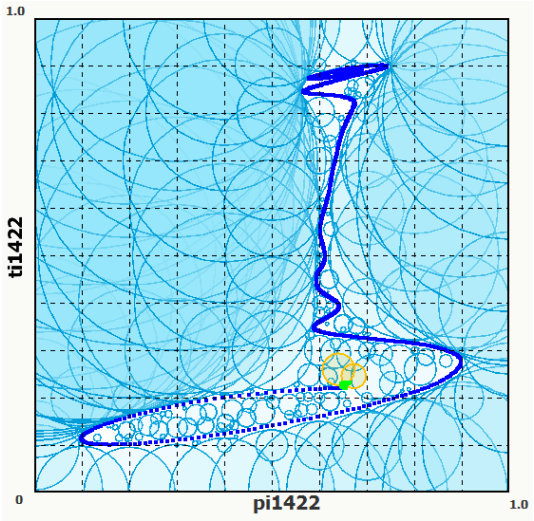


Figure 14: Fault detection by detectors in a 2-dimensional variable space when malfunction GAH was occurred.

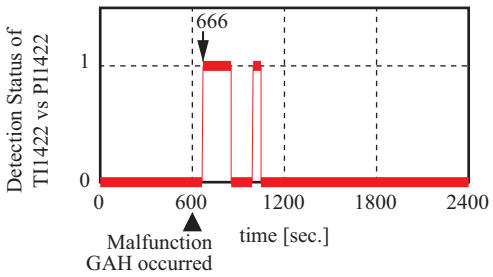


Figure 15: Detection status of the TI1422 vs. PI1422 variable space when malfunction GAH was occurred.

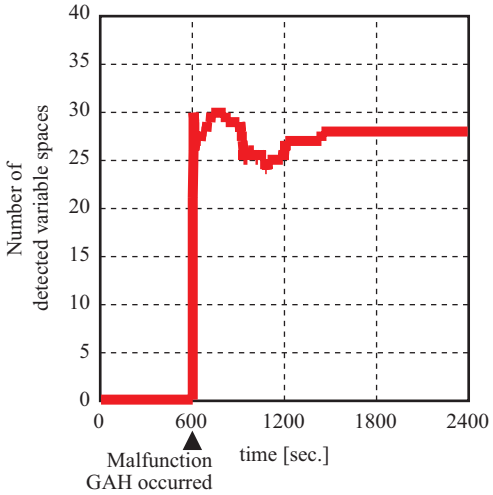


Figure 16: The number of detected variable spaces when malfunction GAH was occurred.

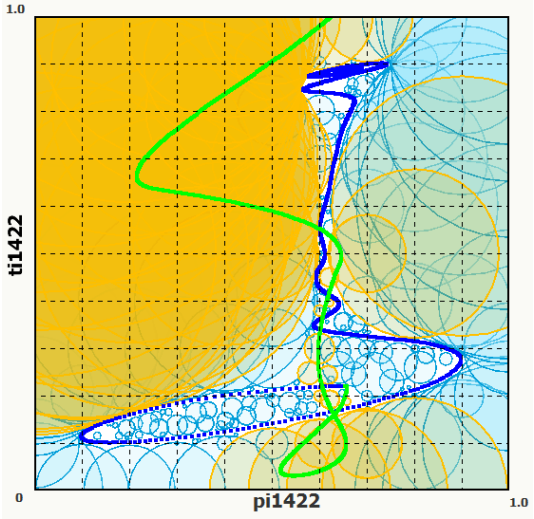


Figure 17: Fault detection by detectors in a 2-dimensional variable space when malfunction FDFclose was occurred.

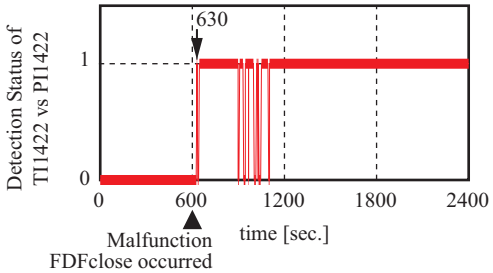


Figure 18: Detection status of the TI1422 vs. PI1422 variable space when malfunction FDFclose was occurred.

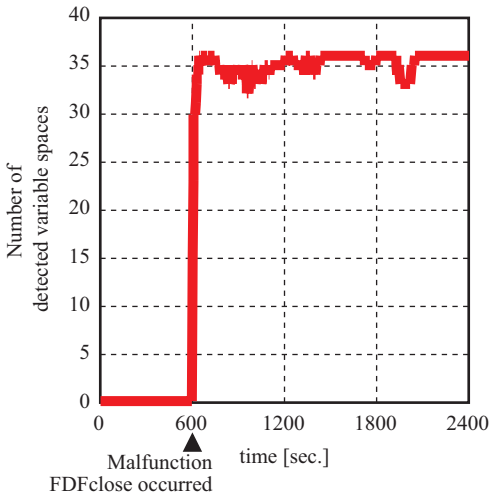


Figure 19: The number of detected variable spaces when malfunction FDFclose was occurred.

5 Conclusion

We built up a fault detection system using negative selection algorithms which can focus on the relationships between process variables.

References

1. N. Kimura, Y. Takeda, T. Hasegawa, Y. Tsuge, "Agent based fault detection using negative selection algorithm for chemical processes" in 2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP), Taipei, Taiwan, 2017. https://doi.org/10.1109/ADCONIP.2017.7983822
2. J. Aguilar, "An artificial immune system for fault detection" Innovations in Applied Artificial Intelligence (17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE 2004, Lecture Note in Artificial Intelligence), **3029**, 219–228, 2004. https://doi.org/10.1007/978-3-540-24677-0_24
3. M. Araujo, J. Aguilar, H. Aponte, "Fault detection system in gas lift well based on artificial immune systems" in the International Joint Conference on Neural Networks, 2003 , 1673–1677, 2003. https://doi.org/10.1109/IJCNN.2003.1223658
4. C.A. Laurentys R.M. Palhares, W.M. Caminhas, "A novel artificial immune system for fault behavior detection" Expert Systems with Applications, **38**, 6957–6966, 2011. https://doi.org/10.1016/j.eswa.2010.12.019
5. M.Y. El-Sharkh, "Clonal selection algorithm for power generators maintenance scheduling" International Journal of Electrical Power & Energy Systems, **57**, 73–78, 2014. https://doi.org/10.1016/j.ijepes.2013.11.051
6. Dia Al Azzawi, Mario G. Perhinschi, Hever Moncayo, Andres Perez, "A dendritic cell mechanism for detection, identification, and evaluation of aircraft failures" Control Engineering Practice, **41**, 134–148, 2015. https://doi.org/10.1016/j.conengprac.2015.04.010
7. K. Wada, T. Toriu, H. Hama, "Improving the efficiency of known fault mode detection for immunity-based diagnosis" Transactions of the Institute of Systems, Control and Information Engineers, **27**(2), 59–66, 2014 (article in Japanese with English abstract). https://doi.org/10.5687/iscie.27.59
8. H. Inomo, W. Shiraki, Y. Imai, H. Kanamaru, "Failure diagnosis of water supply network by immune system" J. Soc. of Mat. Sci., Japan, **52**(1), 40–45, 2003. (article in Japanese with English abstract). https://doi.org/10.2472/jsms.52.40

9. D. Dasgupta, K. KrishnaKumar, D. Wong, M. Berry, "Negative selection algorithm for aircraft fault detection" Artificial Immune Systems (Third International Conference, ICARIS 2004, Lecture Note in Computer Science), **3239**, 1–13, 2004. https://doi.org/10.1007/978-3-540-30220-9_1
10. X.Z. Gao, H. Xu, X. Wang, K. Zenger, "A study of negative selection algorithm-based motor fault detection and diagnosis" International Journal of Innovative Computing, Information and Control, **9**(2), 875–901, 2013.
11. C. Xiong, Y. Zhao, W. Liu, "Fault detection method based on artificial immune system for complicated process", Computational Intelligence (International Conference on Intelligent Computing, ICIC 2006, Lecture Note in Artificial Intelligence), **4114**, 625–630, 2006. https://doi.org/10.1007/11816171_77
12. J. V. Prasad, K. Ghosh, "Negative selection algorithm for monitoring processes with large number of variables", in 2014 IEEE Conference on Control Applications (CCA), **7**78–783, Antibes, France, 2014. https://doi.org/10.1109/CCA.2014.6981435
13. S. Forrest, A. S. Perelson, L. Allen, R. Cherukuri, "Self-nonself discrimination in a computer" in 1994 IEEE Computer Society Symposium on Research in Security and Privacy, **2**02–212, 1994 https://doi.org/10.1109/RISP.1994.296580

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 94-102 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Auto-Encoder based Deep Learning for Surface Electromyography Signal Processing

Marwa Farouk Ibrahim Ibrahim*, Adel Ali Al-Jumaily

School of Mechanical and Electrical Engineering, Faculty of Engineering and IT, University of Technology Sydney, NSW 2007, Australia

ARTICLE INFO

Article history:
Received: 01 November, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords:
Feature learning
Deep learning
Bio-signal processing
Sparse autoencoder
Analysis of variance
Confidence interval
Support vector machine
Extreme learning machine
Softmax layer
Classifier fusion

ABSTRACT

Feature extraction is taking a very vital and essential part of bio-signal processing. We need to choose one of two paths to identify and select features in any system. The most popular track is engineering handcrafted, which mainly depends on the user experience and the field of application. While the other path is feature learning, which depends on training the system on recognising and picking the best features that match the application. The main concept of feature learning is to create a model that is expected to be able to learn the best features without any human intervention instead of recourse the traditional methods for feature extraction or reduction and avoid dealing with feature extraction that depends on researcher experience. In this paper, Auto-Encoder will be utilised as a feature learning algorithm to practice the recommended model to excerpt the useful features from the surface electromyography signal. Deep learning method will be suggested by using Auto-Encoder to learn features. Wavelet Packet, Spectrogram, and Wavelet will be employed to represent the surface electromyography signal in our recommended model. Then, the newly represented bio-signal will be fed to stacked autoencoder (2 stages) to learn features and finally, the behaviour of the proposed algorithm will be estimated by hiring different classifiers such as Extreme Learning Machine, Support Vector Machine, and SoftMax Layer. The Rectified Linear Unit (ReLU) will be created as an activation function for extreme learning machine classifier besides existing functions such as sigmoid and radial basis function. ReLU will show a better classification ability than sigmoid and Radial basis function (RBF) for wavelet, Wavelet scale 5 and wavelet packet signal representations implemented techniques. ReLU will illustrate better classification ability, as an activation function, than sigmoid and poorer than RBF for spectrogram signal representation. Both confidence interval and Analysis of Variance will be estimated for different classifiers. Classifier fusion layer will be implemented to glean the classifier which will progress the best accuracies' values for both testing and training to develop the results. Classifier fusion layer brought an encouraging value for both accuracies either training or testing ones.

1. Introduction

Supervised learning is widely utilised in various applications. However, it is still quite limited method. The majority of applications need handcrafted engineering extraction of features by implementing different techniques. This means that the principal purpose is to represent the bio-signal by applying proper feature representation methods. Whenever significant features represent bio-signal, classification error should be anticipated to be lower than extracting features, which are not genuinely representing data. However, the general engineering handcrafted representation is still effortful and consumes a long time. Moreover, the standard feature extraction algorithm relies on researcher's experience. Many proposed feature learning methods may be implemented to improve feature representation automatically and save both effort and time. The primary evaluation of the behaviour of implemented feature learning method is the classification error. Deep learning is considered the

*Marwa Farouk Ibrahim Ibrahim, 15 Broadway, Ultimo, NSW 2007, Australia, 0061405565739 & marwafaroukibrahim.ibrahim@student.uts.edu.au

most common technique to implement feature learning. Rina Detcher was the first to introduce the fundamentals for both first and second order deep learning [1]. Deep learning is an essential division of machine learning that consists of a multilayer. The output of each layer is considered as features that will be introduced to the following cascaded layer [2]. Artificial neural networks use a hidden layer to implement each layer of multilayers that construct deep learning [3]. Fig.1 shows a simple architecture of deep learning steps. The learning technique is done in a hierarchal method starting from the lower layers to the upper ones [4]. Deep learning can be used for both supervised and unsupervised learning where it learns features from data and eliminates any redundancy that might be existing in the representation. Unsupervised learning recruitment brings more defy than supervised one. Unsupervised learning for deep learning was implemented by Neural history compressors [5]and deep belief networks[6].

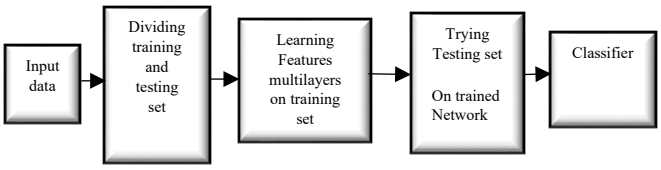


Fig.1. Traditional Deep Learning Steps

This paper is organised by presenting a brief study on previous work that has been done on classification finger movement and deep learning in different fields, then a review study on autoencoder including the main equation for Auto-Encoder will be introduced. The surface electromyography will be assimilated by Wavelet Packet, Spectrogram and Wavelet. We will compare our results by implementing three different classifiers, which will be Support vector machine, Extreme learning machine with three activation functions and Softmax layer.

The Analysis of Variance (ANOVA) will be calculated for different classifiers in Auto-Encoder deep learning method. Also, the confidence interval for Auto-Encoder will be implemented as well. At last, each of training and testing accuracy will be promoted by concatenating classifier fusion layer.

2. Previous Work

In this research, we will suggest a deep learning system that will be capable of providing essential features from the input signal without recourse to traditional feature extraction and reduction algorithms. The suggested system will be talented in assert the ten hand finger motions. The classification of different Finger motions was discussed earlier in many published scientific types of research. The early pattern recognition for finger movements was proposed in [7] where the researchers suggested using neural networks in analysing and classifying the introduced EMG pattern. They classified both finger movement and joint angle associated with moving finger. Later, in [8] the authors investigated and optimised configuration between electrode size and its arrangement to achieve high classification accuracy. Then, in [9] the researchers gave more attention to selecting the extremely discriminative features by employing Fuzzy Neighbourhood Preserving Analysis (FNPA) where the main purpose of this technique is to reduce the distance between the samples that belong to the same class and maximise it between samples of different classes. In the same year, other researchers explored the traditional machine learning well-known algorithm. Where, they used time domain features and implemented support vector machine, linear discriminate analysis and k-nearest neighbours as different classifiers then, they took advantage of Genetic Algorithm to search for redundancy in the used dataset and selected features as well [10]. In the same context, authors proposed an accurate finger movement classification system by extracting time domain-auto regression features, reducing features by using orthogonal fuzzy neighbourhood discriminant analysis technique and implementing linear discriminant analysis as classifier [11]. After that, other researchers suggested an accurate pattern recognition system for finger movement by extracting 16-time domain features to process the Electromyography signal and implementing two layers feed forward neural networks as classifiers [12]. In contrast, effort and time that are being wasted, as mentioned before, in feature extraction and reduction were the motivation behind introducing the concept of deep learning. Therefore, many researchers published valuable achievements in deep learning for the biomedical signal. An extensive review study was presented on different types of research that recalled deep learning in health field [13]. The common factor in each study was the recruitment of neural network to learn features from input bio-signal. In the same context, researchers proposed a model by using convolutional neural networks to convert the information which was given by wearable sensor into highly related discriminative features [14]. Another research presented a deep learning record system that predicted the future medical risk automatically after extracting essential features by implementing convolutional neural networks [15]. Also, researchers implemented a system that used to extract shallow features from wearable sensor devices then the features were introduced to convolutional neural networks and finally to the classifier layer [16]. Based on the above, we can conclude that deep learning is an initial step towards implementing self-learning system by using neural networks. In our proposed system, we will implement neural networks in the form of two stages autoencoder, which read represented bio-signal by either spectrogram, wavelet or wavelet packet. We will use different classifiers to evaluate our system behaviour. Finally, we will add classifier fusion layer, which will follow best local classifier methodology. Adding classifier fusion was a promising contribution to the accuracies. Moreover, both confidence interval and Analysis of Variance will be estimated for different classifiers.

3. Sparse Auto-Encoder

An Autoencoder is an extensively used technique to reduce dimensions [17]. Sparse autoencoder idea first started in [18]. Where it started to reduce the redundancy that may result from complex statistical dependencies. Building a neural network and train it by using sparse method penalty as mentioned in [19] and taking into account the number of hidden nodes in the developed neural network, is considered a straightforward factor but as crucial as choosing the learning algorithm [20].

Auto-Encoder is a feed-forward neural network that is used in unsupervised learning [21]. The implemented neural network is being trained to learn features and produce it as its output rather than generating classes in case of recalling the classification ability of the hired neural network [22]. The encoder input is the represented data while its output is the features learnt by autoencoder. The learnt features learnt from the autoencoder will be introduced to classifier to be used in the assort of the data into

predefined classes[23]. Lately, autoencoder is commonly employed to extract highly expressing features from data.

Unlabelled data can be used to train an autoencoder where training is mainly interested in optimising the cost function. The cost function is mainly responsible for estimating the miscalculation that may occur in calculating the reconstructed copy of input at the output and the input data.

Assume that we have an input vector $x \in R^{D_x}$. The autoencoder maps this input to a new vector $z \in R^{D^{(1)}}$.

$$z^{(1)} = h^{(1)}(W^{(1)}x + b^{(1)}) \quad (1)$$

Where the superscript (1) represents the first layer of the autoencoder. $h^{(1)}: R^{D^{(1)}} \rightarrow R^{D^{(1)}}$ represents the transfer function, $W^{(1)} \in R^{D^{(1)}}$ represents the weight matrix, and $b^{(1)} \in R^{D^{(1)}}$ represents the bias vector. Then the decoder transfers the encoded representation z as a reconstruction of the input x following the next equation

$$\hat{x} = h^{(2)}(W^{(2)}x + b^{(2)}) \quad (2)$$

Where the upper character (2) signifies the second layer of the autoencoder. $h^{(2)}: R^{D^{(2)}} \rightarrow R^{D^{(2)}}$ accounts for the transfer function, $W^{(2)} \in R^{D^{(2)}}$ represents the weight matrix, and $b^{(2)} \in R^{D^{(2)}}$ represents the bias vector.

The sparsity term can be introduced to autoencoder by adding an adapted cost function in the form of regularisation term. The regularisation function is estimated for each neuron i by averaging its activation function, which can be expressed as follows

$$\hat{\rho}_i = \frac{1}{n}\sum_{j=1}^{n} z_i^{(1)}(x_j) = \frac{1}{n}\sum_{j=1}^{n} h(W_i^{(1)T}x_j + b_i^{(1)}) \quad (3)$$

Where n is the number of training samples, x_j is the j^{th} training sample of input, $W_i^{(1)T}$ is the i^{th} row of the weight matrix transpose of the first layer, and $b_i^{(1)}$ is the i^{th} term of the bias vector for the neural network. The neurone is considered to be firing if its output activation function is high and in the case of having a low activation value, this means that the neurone is only responding to a small number of input samples, which in turn encourages the autoencoder to learn. Accordingly, adding a limitation term to activation function output $\hat{\rho}_i$ limits every neurone to learn from limited features. This motivates the other neurones to respond to only another small number of features, which initiates every neurone to be responsible for responding to individual features for each input.

Introducing a sparsity regularise value is considered as a measure of how far or close is the targeted activation value ρ from the actual activation output function $\hat{\rho}$. Kullback-Leibler divergence is a very well know the equation that describes the difference between two different distributions. This equation is shown as follows:

$$\Omega_{Sparsity} = \sum_{i=1}^{D^{(1)}} KL(\rho||\hat{\rho}_i) = \sum_{i=1}^{D^{(1)}} \rho \log\frac{\rho}{\hat{\rho}_i} + (1-\rho)\log(\frac{1-\rho}{1-\hat{\rho}_i}) \quad (4)$$

The cost function is decreased to initiate the two distributions $\hat{\rho}_i$ and ρ to be as close as possible. The cost function can be represented by a mean square error equation as follows:

$$E = \underbrace{\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}(X_{kn} + \hat{X}_{kn})^2}_{\text{Mean Square Error}} + \underbrace{\lambda * \Omega_{weights}}_{\text{L2 Regularisation}} + \underbrace{\beta * \Omega_{Sparsity}}_{\text{Sparsity Regularisation}} \quad (5)$$

Where L2 regularisation is a term to be added to the cost function to regulate and prevent the value of Sparsity Regularisation value of being small during the training due to the increase that may happen to the values of weights and decrease to the value of the mapped vector z

$$\Omega_{wights} = \frac{1}{2}\sum_{l}^{L}\sum_{j}^{n}\sum_{i}^{K}(w_{ji}^{(1)})^2 \quad (6)$$

L is the number of hidden layers, N is the number of input data samples, and K is the number of classes.

Autoencoder was hired in many research areas as a feature learning layer. Where its primary task, was to learn features from input data. A robust study was published to compare between many applications for autoencoder in deep learning field [24]. Autoencoder was implemented in [25] to learn incremental feature learning by introducing an extensive data set to denoising autoencoder. Denoising autoencoder provides an extremely robust performance against noisy data with a high classification accuracy [26, 27]. Another suggested autoencoder was a marginalised stacked one which showed a better performance, with high dimensional data, than the traditionally stacked autoencoder regarding accuracy and simulation time [28]. Denoising stacked autoencoder was hired to learn features from unlabeled data in a hierarchical behaviour [29] and was applied to filter spam by following greedy layer-wise to the implemented denoising stacked autoencoder [30].

In our proposed model, Auto-Encoder is a feed-forward neural network that is used in feature learning. The implemented neural network is being trained to learn features and produce it as its output rather than generating classes in case of recalling the classification ability of the hired neural network. Where, we implemented a stack autoencoder, which consists of two successive encoder stages. The input to the encoder is the data while the output is the features or representations. The classifier uses features produced from encoder as an input while; its output is the classes equivalent to input data [23]. Fig.2 demonstrates the steps that the surface electromyography signal moves through by using a sparse autoencoder.

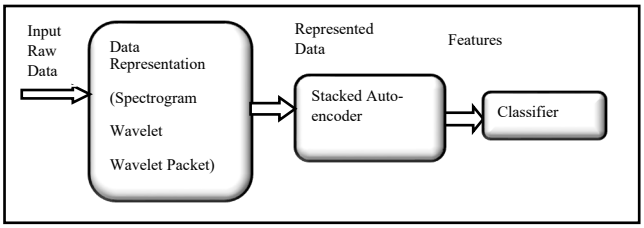


Fig.2. Procedures of Sparse Auto-Encoder signal representation

In the same context of feature learning, autoencoder will generate useful features at the output, rather than producing classes, by decreasing the dimension of the input data into a lower dimension. However, the new lower dimension data will be dealt with our features that contain essential and discriminative information on the data, which will help in better classification results. Sparse autoencoder enhances us to leverage the availability of data.

4. Bio Signal Representation

We suggested three signal representations be applied on raw biological data to ensure fidelity and precision of our bio signal. Moreover, introducing raw data directly to first auto-encoder stage resulted in accuracy less than 50%. The first data representation was the spectrogram for bio raw signal. The spectrogram is interpreted to be the illustration of the spectrum of frequencies of our surface electromyography signal in a visible method. Numerically, Spectrogram can be estimated by calculating the square of the magnitude of Short-Time Fourier Transform (STFT). It can be called short-term Fourier transforms rather than spectrogram. In short time Fourier transforms, the long-time signal is divided into equal length segments and shorter in time. Short time Fourier transforms is relevant to Fourier transform. Then, the frequency and phase for each segment to be estimated separately. Based on the above, we can deduce that spectrogram can be treated as Fourier transform but for shorter segments rather than estimating it from the full long signal at once[31].

Assume that we have a discrete time signal x with a finite duration (limited signal) and a number of samples N. The Discrete Fourier Transform (DFT) can be expressed as follows:

$$\hat{x}(k) = \sum_{n=0}^{N-1} x(n) e^{-i\frac{2\pi k}{N}n} \quad , k = 0, \dots\dots, N-1 \qquad (7)$$

Knowing that the Fourier transform is estimated at frequency $f = \frac{k}{N}$

The original signal x can be restored back from $\hat{x}$ by applying the inverse Discrete Fourier Transform as follows:

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} \hat{x}(k) e^{-i\frac{2\pi k}{N}n}, \qquad n = 0, \dots\dots, N-1 \qquad (8)$$

The above-mentioned two equations can be rephrased as follows:

$$x = \frac{1}{N} F\hat{x} \qquad (9)$$

$$\hat{x} = \bar{F}x \qquad (10)$$

Where F is Fourier matrix of $n * n$ dimensions and $\bar{F}$ is its complex conjugate

$$F = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{i\frac{2\pi}{N}} & e^{i\frac{4\pi}{N}} & \cdots & e^{i\frac{2\pi(N-1)}{N}} \\ 1 & e^{i\frac{4\pi}{N}} & e^{i\frac{8\pi}{N}} & \cdots & e^{i\frac{2\pi 2(N-1)}{N}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{i\frac{2\pi(N-1)}{N}} & e^{i\frac{2\pi 2(N-1)}{N}} & \cdots & e^{i\frac{2\pi(N-1)^2}{N}} \end{bmatrix} \qquad (11)$$

Where the entries of $\hat{x}$ is expressed in terms frequencies coefficients

$$f = 0, {}^{1}/_{N}, {}^{2}/_{N}, \dots\dots {}^{(N-1)}/_{N} \qquad (12)$$

We need to calculate the spectrogram of the signal. Assume that we have a signal x of length N, which is divided into successive equal segments m. Where $m < N$. The matrix of successive equal segments can be expressed as X where $X \in R^{mx(N-m+1)}$. The first column of X matrix equals $[x[0], x[1], \dots\dots x[m-1]]^T$ and its second column equals $[x[1], x[2], \dots\dots x[m]]^T$. The spectrogram for a signal x with window size m can be annotated $\widehat{X}$. The columns which are composing matrix $\hat{X}$ is the discrete Fourier transform

$$\hat{X} = \bar{F}X \qquad (13)$$

$$X = \frac{1}{m} F\hat{X} \qquad (14)$$

The rows of the matrix $\hat{X}$ are representing the signal x in the time domain while its columns are representing the signal x in the frequency domain. So simply spectrogram is a time-frequency representation of the signal x.

The spectrogram was used in many applications especially for speech signal analysis wherein [32] the authors represented the speech signal by different representations like Fourier and spectrogram to conclude that the resolution is mainly dependent on used representations. In [33] the researchers estimated the time corrected version of rapid frequency spectrogram of the speech signal which showed a better ability to follow the alterations in the bio-signal than other published techniques.

The second signal representation used was wavelet of the signal. Wavelet is estimated by shifting and scaling small segmentations of the bio-signal. Fourier transform is an illustration of the signal in a sinusoidal wave by using various frequencies while wavelet is the illustration of the abrupt changes that happen to the signal. Fourier transform is considered a good representation of the signal in case of having a smooth signal. While wavelet is believed to be a better representation, than Fourier, for the sudden changing signal. Wavelet gives the opportunity to represent rapid variations of the signal and help the system extract more discriminative features. We implemented Haar wavelet for our proposed model.

So in brief, a wavelet is an analysis for time series signal that has non-stationary power at many frequencies [34]. Assume that we have a time series signal x_n with equal time spacing δt and $n = 0, \dots\dots N-1$ where the wavelet function is $\Psi_o(\eta)$ that depends on time η. The wavelet signal has zero mean and is represented in both time and frequency domain [35]. Morlet wavelet can be estimated by modulating our time domain signal by Gaussian as follows:

$$\Psi_o(\eta) = \pi^{\frac{-1}{4}} e^{i\eta\omega_o} e^{\frac{-\eta^2}{2}} \qquad (15)$$

Where ω_o is the frequency of the unmodulated signal. The continuous wavelet of a discrete signal x_n is the convolutional of x_n with scaled and shifted version of $\Psi_o(\eta)$

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'}\ \Psi^*\left[\frac{(n'-n)\delta t}{s}\right] \qquad (16)$$

Where * is the complex conjugate and s is the scale.

The wavelet transform was applied in several studies and different fields as in [36] where wavelet implemented in Geophysics field, in [37, 38] for climate, in [39] for weather, in [40] and many other applications. The above equation can be simplified by reducing the number of N. The convolutional theorem permits to estimate N convolutional in Fourier domain by implementing Discrete Fourier Transform (DFT). The Discrete Fourier Transform for x_n.

$$\hat{x}_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{-2\pi i k n/N} \quad (17)$$

Where $k = 0, \ldots, N-1$ which is representing the frequencies. For a continuous signal $\Psi(t/s)$ is defined as $\hat{\psi}(s(\omega)$. Based on the convolutional theorem, the inverse Fourier transform is equal to wavelet transform as follows:

$$\psi_n(s) = \sum_{k=0}^{N-1} \hat{x}_k \hat{\psi} * (s\omega_k) e^{i\omega_k n \delta t} \quad (18)$$

Where the angular frequency ω_k can be expressed as follows:

$$\omega_k = \begin{cases} \frac{2\pi k}{N\delta t} & : k \le \frac{N}{2} \text{ or} \\ \frac{-2\pi k}{N\delta t} & : k > \frac{N}{2} \end{cases} \quad (19)$$

An improved copy of wavelet algorithm was recalled in [41] where the authors presented two techniques. The first one used expansion factors for filtering while the other one is factoring wavelet transform. The researchers in [42] introduced the Morlet wavelet to vibration signal of a machine. The vibration signal of the low signal to noise ratio was represented by wavelet to grant fidelity to the signal and allow extraction better powerful features. This model was implemented in [43] where researchers used wavelet transform to predict early malfunction symptoms that may happen in the gearbox.

As a refinement act, we scaled the wavelet signal by five in wavelet, which in turn promoted the results as shown in Table I. As a comparative study, we utilised wavelet packet for the signal representation. The signal can be represented in both time and frequency domain simultaneously. This representation gains a fidelity to the signal due to its robust representation. Wavelet packet is one of the very widely used signal representation that produces the signal in both time and frequency domain [44]. The wavelet packet shows a very well acted for both nonstationary and transient signals [45-47]. Wavelet packet is estimated by a linear combination of wavelets. The coefficients of linear combination are calculated by recursive algorithm [48]. The wavelet packet estimation can be done as follows:

Assume that we have two wavelets type signal $h(n)$, $g(n)$ and two filters of length $2N$. Let us assume that the following sequence of functions is representing wavelet functions.

$$(W_n(x), n = 0,1,2, \ldots\ldots)$$

$$W_{2n}(x) = \sqrt{2}\sum_{k=0}^{2N-1} h(k) W_n(2x-k)$$

$$W_{2n+1}(x) = \sqrt{2}\sum_{k=0}^{2N-1} g(k) W_n(2x-k) \quad (20)$$

Where $W_0(x) = \varphi(x)$ is the scaling function, $W_1(x) = \omega(x)$ is the wavelet function.

$N = 1, h(0) = h(1) = \frac{1}{\sqrt{2}}$ and $g(0) = -g(1) = \frac{1}{\sqrt{2}}$

The above equations became

$$W_{2n}(x) = W_n(2x) + W_n(2x-1)$$

And

$$W_{2n+1}(x) = W_n(2x) + W_n(2x-1) \quad (21)$$

Where $W_0(x) = \varphi(x)$ is the scaling function, $W_1(x) = \omega(x)$ is the Haar wavelet function.

Many researchers implemented wavelet packet as in [49]. The authors used wavelet packet to create an index called rate index to detect the damage that may happen to the structure of any beam. In the same context authors of [50] employed wavelet packet and neural networks to detect a fault in a combustion engine. The implemented wavelet packet was six levels for sym10 at sampling frequency 2 kHz.

5. Classifiers

In the implementation of Auto-Encoder as feature learning algorithm, we applied three different classifiers, where the first was Softmax layer, the second was Extreme learning machine, and the third was Support Vector Machine (SVM). We measured the accuracy of Linear support vector machine, Quad support vector machine, Cubic support vector machine, Fine Gauss support vector machine, Medium Gauss support vector machine and Coarse Gauss support vector machine and elected the support vector machine classifier that resulted in the highest accurate result. Furthermore, the appending of classifier fusion layer to nominate best local classifier which in return endorsed the accuracy values. Fig.3 shows the block diagram for our implemented autoencoder feature learning proposed model

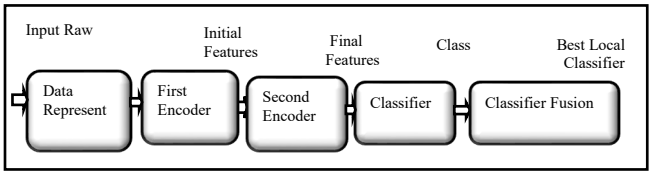


Fig.3. Scheme of proposed Model

Moreover, ANOVA for autoencoder different classifiers was implemented. Where, we assembled average testing accuracies for four signal representation techniques (Wavelet, Wavelet Scale5, Wavelet Packet and Spectrogram) that resulted in P value 0.7487. So as wavelet results should not be counted, due to its low accuracy values, so, we suggested a second trial which was to group average testing accuracies for three signal representing techniques (Wavelet Scale5, Wavelet Packet and Spectrogram) that resulted in P value 0.3405. Both P values showed that there was no sensible variation between any of the implemented three

classifiers as P value was higher than 0.05 in both cases. Fig.7 shows different P values for different classifiers.

In addition to the above, we estimated the confidence interval for each classifier. Our confidence interval was designed for confidence score 60%. Our assessed interval was bounded by higher and lower limit. In other words, we were confident or assured of any new accuracy by percentage 60% as long as it is located in the previously estimated interval.

6. Implementation

In this part, the data acquisition methods we followed will be expressed more extensively, and simulation outcomes will be exhibited and discussed.

6.1. Data Acquisition

The surface Electromyography signal was read by using FlexComp Infiniti™ device. Two sensors were placed on the forearm of the participant of type T9503M. The placement of two electrodes on participant's forearm is as shown in fig.4

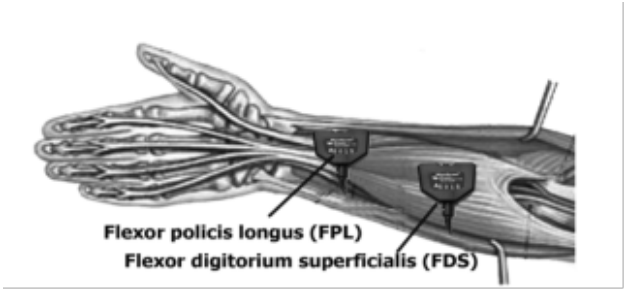


Fig.4. Placement of the electrodes

The Electromyography signal was collected from nine participants. Each participant performed one finger movement for five seconds then had a rest for another five seconds. Every finger motion was reiterated six times. The same sequence was repeated for the second finger activity. Amplification of the signal by 1000 was applied and a sampling rate of 2000 samples for each second was implemented.

The collected Electromyography signal was used to categorise between predefined ten finger motions, as shown in Fig.5 via using our suggested model. Three folded cross validation was applied on our collected Electromyography signal. Accordingly, 2/3 of the collected data was assigned to training set while remaining 1/3 to be used by testing set.

Our surface electromyography signal was filtered to ensure fidelity and removal of any noise that may be inserted into the collected bio-signal. The average training or testing accuracy was estimated by simulating our proposed model for every subject apart then summed the accuracies for all subjects and divided the result by the total number of subjects.

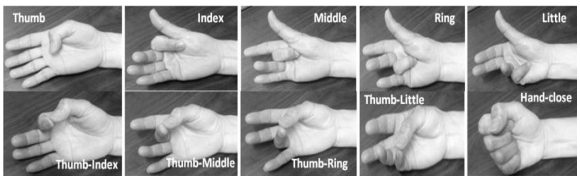


Fig.5. Targeted Ten different finger motions

6.2. Results

We implemented 400 nodes for the first layer of autoencoder and 300 nodes for the second one. As for the transfer function of the encoder, it was the pure linear type.

Table I shows autoencoder feature learning testing and training accuracy where the bio-signal was presented by spectrogram, wavelet, Wavelet scale 5 and wavelet packet. Three different classifiers were executed. The first was a SoftMax layer. While the second classifier was an extreme learning machine, we examined the performance of various activation functions for extreme learning machine classifier like sigmoid, the rectified linear unit and radial basis function. As that, the third classifier was support vector machine. We examined the performance of linear support vector machine, quadratic support vector machine, cubic support vector machine, fine gauss support vector machine, medium gauss support vector machine and coarse gauss support vector machine. Then, we selected the highest support vector machine that showed better classification ability to be our implemented support vector machine.

Table I. Auto Encoder Testing and Training Accuracy

Signal Representing	Average Training Accuracy	Average Testing Accuracy	Classification Algorithm	Simulation Time
Spectrogram	95%	73%	SoftMax Layer	830.93 Seconds
Spectrogram	95.5%	79.14%	ELM (Sigmoid)	379.10 Seconds
Spectrogram	95.5%	82.78%	ELM (ReLU)	379.54 Seconds
Spectrogram	97.237%	83.56%	ELM (RBF)	353.756 Seconds
Spectrogram	89.48%	77.27%	Cubic SVM	1000.409 Seconds
Wavelet	91.42%	45.71%	SoftMax Layer	210.04 Seconds
Wavelet	79.95%	59.88%	ELM (Sigmoid)	265.23 Seconds
Wavelet	83.62%	64.73%	ELM (ReLU)	276.149 Seconds
Wavelet	81.98%	62.70%	ELM (RBF)	257.83 Seconds
Wavelet	70.29%	52.29%	Linear SVM	341.57 Seconds
Wavelet (Scale 5)	98.85%	82.13%	SoftMax Layer	668.355 Seconds
Wavelet (Scale 5)	95.65%	85.416%	ELM (Sigmoid)	402.647 Seconds
Wavelet (Scale 5)	96.98%	86.827%	ELM (ReLU)	276.697 Seconds
Wavelet (Scale 5)	95.59%	85.58%	ELM (RBF)	444.067 Seconds
Wavelet (Scale 5)	96.55%	83.85%	Cubic SVM	495.124 Seconds
Wavelet Packet	98.69%	84.176%	SoftMax Layer	429.52 Seconds
Wavelet Packet	93.3%	86.79%	ELM (Sigmoid)	419.27 Seconds
Wavelet Packet	96.42%	89.41%	ELM (ReLU)	419.27 Seconds
Wavelet Packet	95.59%	87.78%	ELM (RBF)	540.22 Seconds
Wavelet Packet	98.16%	89.707%	Quad SVM	579.713 Seconds

outstanding in our application for all signal representation except for wavelet packet. Both quadratic support vector machine and extreme learning machine, with the rectified linear unit as an activation function, showed a very close performance for wavelet packet signal representation. Extreme learning machine was improved when we replaced sigmoid activation function by Radial basis function and the rectified linear unit. The rectified linear unit activation function for extreme learning machine presented a superior behaviour than radial basis function and sigmoid activation functions for wavelet, Wavelet scale 5 and wavelet packet. However, the rectified linear unit offered better performance than sigmoid and lower accuracy than radial basis function for spectrogram.

Cubic and Quad support vector machine started to result in a good testing accuracy for Wavelet scale 5 and wavelet packet only. Simulation time for support vector machine is relatively longer than other compared classification algorithms. SoftMax layer resulted in a very poor classification for wavelet signal representation, as the testing accuracy was less than 50%. Softmax started to prove its classification ability for Wavelet scale 5 and wavelet packet. Fig.6 shows different P values for different classifiers and Fig.7 shows confidence intervals for each classifier where it was calculated twice. The grey bars were calculated for different classifiers with three signal representation methods (Wavelet Scale5, Wavelet Packet and Spectrogram). While, yellow bars calculated for different classifiers with four signal representation methods (Wavelet, Wavelet Scale5, Wavelet Packet and Spectrogram). The narrowest interval was 2.53% for the extreme learning machine. While the widest one was 6.10% for support vector machine and softmax layer interval reached 5.77%.

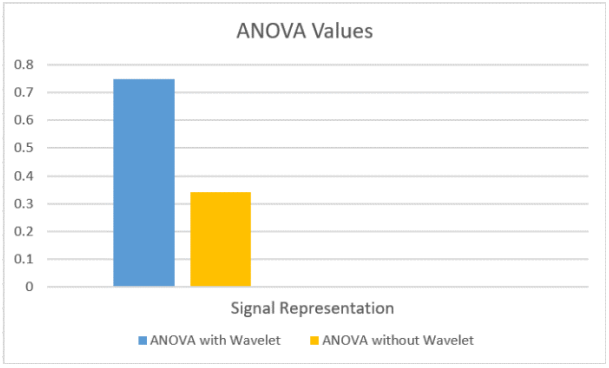


Fig.6. ANOVA values for different Classifiers

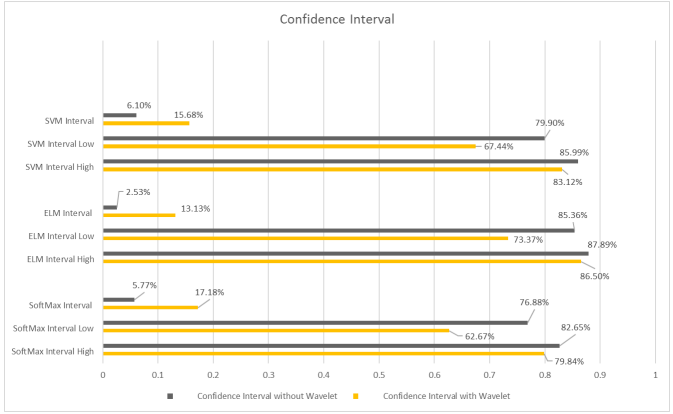


Fig.7. Confidence Intervals for used classifiers

We concatenated a layer of classifier fusion after classification layer. The function of this added layer is to nominate the best-implemented classifier based on the outcomes of accuracies values. This added classifier fusion layer in return enriched our accuracies as displayed in Table II. On the other side, adding classifier fusion layer relatively increased the simulation time than without fusion layer.

Table II: Auto Encoder Classifier Fusion Testing and Training Accuracy

Signal Representing	Average Training Accuracy	Average Testing Accuracy	Classification Algorithm	Simulation Time
Spectrogram	99.53%	91.05%	Classifier Fusion	2118.846 Seconds
Wavelet	96.99%	86.80%	Classifier Fusion	745.67 Seconds
Wavelet (Scale 5)	99.24%	89.02%	Classifier Fusion	1378.41 Seconds
Wavelet Packet	98.70%	92.25%	Classifier Fusion	1656.724 Seconds

7. Conclusion

Sparse autoencoder is just one hidden layer algorithm. Therefore, to establish the concept of deep learning, and take advantage of stacking more than a layer, as the testing set accuracy was less than 50% for one stage only of the autoencoder, we implemented stacked autoencoder that led to verifying deep learning concept and enriching the results accuracy. In addition, applying some signal representation like calculating spectrogram, wavelet and wavelet packet, instead of using raw signal, and introducing the output of these signal representation to the first stage of autoencoder enhanced the performance of the system. Extreme learning machine showed a satisfactory performance on the level of testing accuracy and simulation time. Softmax layer classification resulted in the most mediocre testing accuracy although it consumed longer simulation time than extreme learning machine. Support vector machine produced an excellent testing accuracy but consumed long simulation time. Applying signal representation like Spectrogram, wavelet or wavelet packet improved both training and testing accuracy a lot as both accuracies were much less than 50% when we fed first stage autoencoder by raw data. Multiplying wavelet scale by 5 enhanced the results a lot. As a conclusion, applying any signal representation either in the time domain, frequency domain, or both had a good impact on our training and testing accuracies. We introduced the rectified linear unit as an activation function for extreme learning machine besides already existing functions such as radial basis function and sigmoid. The rectified linear unit was superior in its testing accuracy than both radial basis function and sigmoid one for wavelet, Wavelet scale 5 and wavelet packet signal representation. Moreover, it resulted in a lower testing accuracy than radial basis function but better than sigmoid for spectrogram signal representation.

learnt, by itself, the best features suitable for the application under examination. In addition, since feature extraction and reduction methods varied according to the application so, feature extraction and reduction algorithms were not fixed and needed more experience. In other words, deep learning system should be adaptable to any set of data if the data was accurate and well represented. This brought a new challenge on the scene in regarding representing the data. The data should be represented in a high precision way to expect a good result from implementing deep learning technique.

References

[1] R. Dechter, "Learning While Searching in Constraint-Satisfaction-Problems," presented at the Proceedings of the 5th National Conference on Artificial Intelligence, Philadelphia, 1986.

[2] L. Y. Deng, D, "Deep Learning: Methods and Applications," *Foundations and Trends in Signal Processing,* vol. 7, pp. 3-4, 2014.

[3] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning,* vol. 2, no. 1, pp. 1-127, 2009.

[4] Y. C. Bengio, A.; Vincent, P., "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 8, pp. 1798–1828, 2013.

[5] J. Schmidhuber., "Learning complex, extended sequences using the principle of history compression," *Neural Computation,* vol. 4, pp. 234–242, 1992.

[6] G. E. Hinton, "Deep belief networks," *Scholarpedia,* vol. 4, no. 5, p. 5947, 2009.

[7] A. H. Noriyoshi Uchida, Noboru Sonehara, and Katsunori Shimohara, "EMG pattern recognition by neural networks for multi fingers control," presented at the Engineering in Medicine and Biology Society, 1992 14th Annual International Conference of the IEEE, 1992.

[8] E. M. a. L. M. Alex Andrews, "Optimal Electrode Configurations for Finger Movement Classification using EMG," presented at the 31st Annual International Conference of the IEEE EMBS, 2009.

[9] S. K. Rami N. Khushaba, Dikai Liu, Gamini Dissanayake, "Electromyogram (EMG) Based Fingers Movement Recognition Using Neighborhood Preserving Analysis with QR-Decomposition," presented at the 2011 Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2011.

[10] C. A. a. C. C. Gunter R. Kanitz, "Decoding of Individuated Finger Movements Using Surface EMG and Input Optimization Applying a Genetic Algorithm," presented at the 33rd Annual International Conference of the IEEE EMBS, Boston, Massachusetts USA, 2011.

[11] G. B. Ali H. Al-Timemy, Javier Escudero and Nicholas Outram, "Classification of Finger Movements for the Dexterous Hand Prosthesis Control With Surface Electromyography," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 17, no. 03, pp. 608-618, 2013.

[12] W. C. Mochammad Ariyanto, Khusnul A. Mustaqim, Mohamad Irfan, Jonny A. Pakpahan, Joga D. Setiawan and Andri R. Winoto, "Finger Movement Pattern Recognition Method Using Artificial Neural Network Based on Electromyography (EMG) Sensor," presented at the 2015 International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (ICACOMIT), Bandung, Indonesia, 2015.

[13] C. W. Daniele Rav`ı, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo and Guang-Zhong Yang, "Deep Learning for Health Informatics," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 4-21, 2016.

[14] T. K. Julius Hannink, Cristian F. Pasluosta, Karl-G ¨unter Gaßmann, Jochen Klucken, and Bjoern M. Eskofier, "Sensor-Based Gait Parameter Extraction With Deep Convolutional Neural Networks," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 85-93, 2016.

[15] T. T. Phuoc Nguyen, Nilmini Wickramasinghe, and Svetha Venkatesh, "Deepr: A Convolutional Net for Medical Records," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 22-30, 2016.

[16] C. W. Daniele Rav`ı, Benny Lo, and Guang-Zhong Yang, "A Deep Learning Approach to on-Node Sensor Data Analytics for Mobile or Wearable Devices," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 56-64, 2016.

[17] Y. Bengio, "Learning Deep Architectures for AI," in "Foundations and Trends in Machine Learning," Universit´e de Montr´eal, Canada2009.

[18] D. J. F. BRUNO A OLSHAUSEN, "Sparse Coding with an Overcomplete Basis Set: Strategy Employed by V1?," *Pergamon,* vol. 37, no. 23, pp. 3311-3325, 1997.

[19] V. a. H. Nair, Geoffrey E, "In Advances in Neural Information Processing Systems," presented at the Advances in Neural Information Processing Systems, 2009.

[20] A. a. N. Coates, Andrew, "The importance of encoding versus training with sparse coding and vector quantization.," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 921–928.

[21] Cheng-Yuan Liou, Jau-Chi Huang, Wen-Chie Yang, "Modeling word perception using the Elman network," *Neurocomputing,* vol. 71, no. 16-18, pp. 3150-3157, 2008.

[22] C.-Y. Liou, Cheng, C.-W., Liou, J.-W., and Liou, D.-R., "Autoencoder for Words," *Neurocomputing,* vol. 139, pp. 84–96, 2014.

[23] G. E. H. a. R. R. Salakhutdinov. (2006) Reducing the Dimensionality of Data with Neural Networks. *SCIENCE*.

[24] F. V. Maryam M Najafabadi, Taghi M Khoshgoftaar, Naeem Seliya, Randall WaldEmail author and Edin Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data,* vol. 2, no. 1, pp. 1-21, 2015.

[25] S. K. a. L. H. Zhou G, "Online incremental feature learning with denoising autoencoders," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 1453–1461.

[26] L. H. Vincent P, Bengio Y, and Manzagol Pierre Antoine, "Extracting and composing robust features with denoising autoencoders," in *25th international conference on Machine learning*, Helsinki, Finland, 2008.

[27] H. L. Pascal Vincent, Isabelle Lajoie, Yoshua Bengio and Pierre-Antoine Manzagol "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *The Journal of Machine Learning Research,* vol. 11, no. 3, pp. 3371-3408 2010.

[28] X. Z. Chen M, Weinberger KQ and Sha F, "Marginalized denoising autoencoders for domain adaptation," in *29th International Conference in Machine Learning*, Edingburgh, Scotland, 2012.

[29] B. A. a. B. Y. Glorot X, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *28th International Conference on Machine Learning* 2011, pp. 513–520.

[30] L. Y. a. F. Weisen, "Application of stacked denoising autoencoder in spamming filtering," *Journal of Computer Applications,* vol. 35, no. 11, pp. 3256-3260, 2015.

[31] I. D. Ervin Sejdic, Jin Jiang "Time-frequency feature representation using energy concentration: An overview of recent advances," *Digital Signal Processing,* vol. 19, no. 1, pp. 153-183, 2009.

[32] R. R. Mergu and S. K. Dixit, "Multi-resolution speech spectrogram," *International Journal of Computer Applications,* vol. 15, no. 4, pp. 28-32, 2011.

[33] S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *The Journal of the Acoustical Society of America,* vol. 119, no. 1, pp. 360-371, 2006.

[34] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *Information Theory,* vol. 36, no. 5, pp. 961 - 1005, 1990.

[35] M. Farye, "WAVELET TRANSFORMS AND THEIR APPLICATIONS TO TURBULENCE," *Fluid Mechanics,* vol. 24, pp. 395–457, 1992.

[36] H. W. a. K.-M. Lau, "Wavelets, Period Doubling, and Time–Frequency Localization with Application to Organization of Convection over the Tropical Western Pacific," *Journal of the Atmospheric Sciences,* vol. 51, no. 17, pp. 2523–2541., 1994.

[37] D. G. a. S. G. H. Philander, "Secular Changes of Annual and Interannual Variability in the Tropics during the Past Century," *Journal of Climate,* vol. 8, pp. 864-876, 1995.

[38] B. W. a. Y. Wang, "Temporal Structure of the Southern Oscillation as Revealed by Waveform and Wavelet Analysis," *Journal of Climate,* vol. 9, pp. 1586-1598, 1996.

[39] N. Gamage, and W. Blumen, "Comparative analysis of lowlevel cold fronts: Wavelet, Fourier, and empirical orthogonal function decompositions," *Monthly Weather Review,* vol. 121, pp. 2867–2878, 1993.

[40] P. F. Sallie Baliunas, Dmitry Sokoloff and Willie Soon, "Time scales and trends in the central England temperature data (1659–1990)," *BrAVELET ANALYSIS OF CENTRAL ENGLAND TEMPERATURE* vol. 24, pp. 1351-1354, 1997.

[41] A. Calderbank, I. Daubechies, W. Sweldens, and B.-L. Yeo, "Wavelet transforms that map integers to integers," *Applied and computational harmonic analysis,* vol. 5, no. 3, pp. 332-369, 1998.
[42] J. Lin and L. Qu, "Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis," *Journal of sound and vibration,* vol. 234, no. 1, pp. 135-148, 2000.
[43] J. Lin and M. Zuo, "Gearbox fault diagnosis using adaptive wavelet filter," *Mechanical systems and signal processing,* vol. 17, no. 6, pp. 1259-1269, 2003.
[44] M. V. Wickerhauser, "INRIA lectures on wavelet packet algorithms," 1991.
[45] R. P. a. G. Wilson, "Application of 'matched' wavelets to identification of metallic transients," in *Proceedings of the IEEE-SP International Symposium Time-Frequency and Time-Scale Analysis*, Victoria, BC, Canada, Canada, 1992: IEEE.
[46] E. S. a. M. Fabio, "The use of the discrete wavelet transform for acoustic emission signal processing," in *Proceedings of the IEEE–SP International Symposium*, Victoria, British Columbia, Canada, 1992: IEEE.
[47] T. P. T. Brotherton, R. Barton, A. Krieger, and L. Marple, "Applications of time frequency and time scale analysis to underwater acoustic transients," in *Proceedings of the IEEE–SP International Symposium*, Victoria, British Columbia, Canada, 1992: IEEE.
[48] M. Y. G. a. D. K. Khanduj, "Time Domain Signal Analysis Using Wavelet Packet Decomposition Approach " *Int. J. Communications, Network and System Sciences,* vol. 3, pp. 321-329, 2010.
[49] J.-G. Han, W.-X. Ren, and Z.-S. Sun, "Wavelet packet based damage identification of beam structures," *International Journal of Solids and Structures,* vol. 42, no. 26, pp. 6610-6627, 2005.
[50] J.-D. Wu and C.-H. Liu, "An expert system for fault diagnosis in internal combustion engines using wavelet packet transform and neural network," *Expert systems with applications,* vol. 36, no. 3, pp. 4278-4286, 2009.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 94-99 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Technology

ASTESJ

ISSN: 2415-6698

Modelling and Analysis of Radial Flux Surface Mounted Direct-Driven PMSG in Small Scale Wind Turbine

Theint Zar Htet *, Zhengming Zhao, Qing Gu, Jing Li

State Key Lab of Power Systems, Dept. of Electrical Engineering, Tsinghua University, Beijing,100084, China.

ARTICLE INFO

Article history:
Received: 13 October, 2017
Accepted: 02 November, 2017
Online: 15 November,2017

Keywords :
Airgap flux density
Cogging Torque
Maxwell 2D
PMSG
Wind Turbine

ABSTRACT

This paper presents the modelling and analysis of permanent magnet synchronous generator (PMSG) which are used in direct driven small scale wind turbines. The 3 kW PM generator which is driven directly without gear system is analyzed by Ansoft Maxwell 2D RMxprt. The performance analysis of generator includes the cogging torque in two teeth, induced coil voltages under load, winding current under load, airgap flux density distribution and so on. The modelling analysis is based on the 2D finite element techniques. In an electrical machine, an accurate determination of the geometry parameters is a vital role. The proper performance results of 3kW PMSG in small scale wind turbine can be seen in this paper.

1. Introduction

In this century, wind power generation has become the most powerful influences in all over the world[1]. The variable speed wind turbines have been more popular in nowadays because it has some advantages compare to fixed speed wind turbines such as less mechanical stress, higher efficiency, higher power quality and so on[2].

The several advantages of variable speed wind turbines are reduced mechanical stress and optimized power capture.In variable speed wind turbines, two types of generators are usually employed in the small scale wind turbine such as induction generators (IGs) and permanent magnet synchronous generators (PMSGs)[3].

Recently, in the WTs, the PMSGs have been more attractive the IGs because of direct drive and high efficiency. Moreover, the PMSGs have the high torque density and the power factor characteristics. The direct-drive train wind power on system generation can be seen in Figure 1.Four main components are contained in this system and the PMSG connects to the grid through a back to back converter, converting the mechanical energy produced by the wind turbine to the electrical energy[4].

Two type of rotor position can be used in PMSG family such as outer rotor and inner rotor. In this study, the small outer rotor surface mounted generator is adopted for modelling and analysis because of the easier construction. In conventional machines, the magnetic field is in the radial direction. These types of machines are referred to as radial flux machines (RFM) whereas the magnetic field of axial flux machine (AFM) is in the axial direction [5].

The small scale generators can be used for remote areas. The American Wind Energy Association (AWEA) defines small scale wind generator as those whose power reaches up to 100 kW.

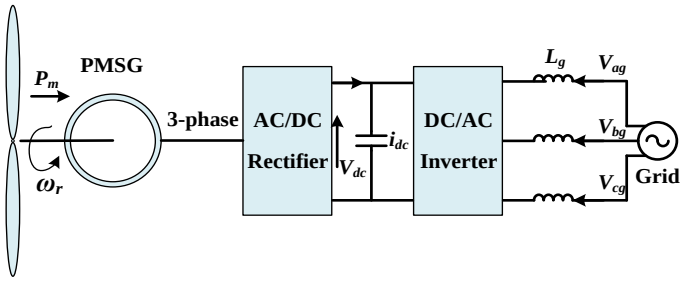


Figure 1. Configuration of a direct-drive PMSG system

The 2D analysis of outer rotor surface mounted PM generator is studied in this paper. The output power is 3 kW, rated

*Corresponding Author: Theint Zar Htet, Tsinghua University, +8618811515613
Email: theintzarhtet@gmail.com

mechanical speed is 250 rpm and number of poles are 34, respectively.

2. Modelling of PMSG by Maxwell 2D RMxprt

2.1. The geometry model of PMSG

In this study, outer rotor radial flux surface mounted direct-driven PMSG are employed. The radial flux outer-rotor PMSG, 17 pole pairs, are considered. The advantages of outer rotor PMSG compared to inner-rotor PMSG is that the magnets are easily attached to the rotor surface.Permanent magnet synchronous generators (PMSGs) have two parts; stator and rotor. Rotor has permanent magnets to create constant magnetic field in electrical machine[6]. Other time and space varying magnetic field is created with the stator windings. The geometry model of outer rotor radial flux surface mounted PMSG is shown in Figure 2.

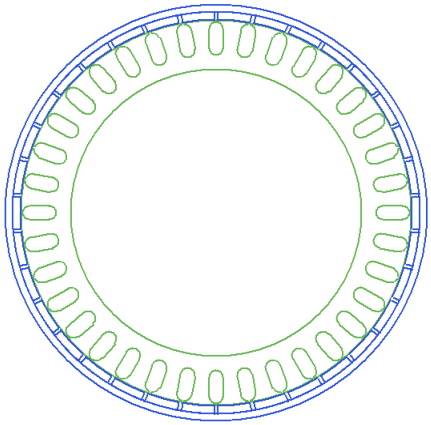


Figure 2. The geometry model of outer rotor PMSG

2.2. Maxwell 2D Model of PMSG

In this model, 3 kW small scale DDPMSG for wind turbine are designed. The main parameters of the DDPMSG are shown in Table 1.

Table 1: Major Parameters of DDPMSG

Parameters	Unit	Quantities
Rated output power	kW	3
Rated phase voltage	V	220
Control Type	-	AC
Rated power factor	-	0.9
Rated speed	rpm	250
Number of poles	-	34
Number of phases	-	3
Permanent magnet	-	NdFeB
Rotor Position	-	Outer Rotor

Parametrical analysis procedure are considered. Firstly, the stator steel type, rotor steel type and the magnet type should be determined from RMXprt library and the diameters of stator and rotor, the thickness of yoke and the airgap are roughly defined. And then, design calculations and modelling analysis are started. Table 2 represents the model design results of 3kW PMSG which are based on electrical machine theory.

Table 2: Design Results of DDPMSG

Parameters	Unit	Quantities
Phase		3
Pole		34
Stator Slot		36
Conductors per slot		24
Length of Rotor	mm	60
Outer Diameter of Rotor	mm	512
Inner Diameter of Stator	mm	358
Thickness of Magnet	mm	10
Rotor Pole Embrace	-	0.9

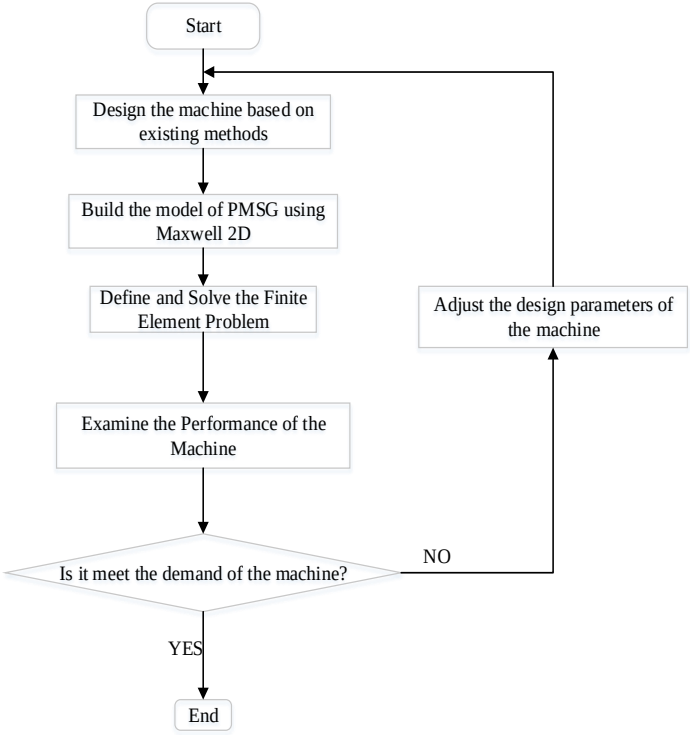


Figure 3. Design flow chart of the model for the PMSG

According to the proposed flow chart of the PMSG shown in Figure 3, a PMSG of 3kW is modelled. In this model, RMXprt (XG196/96) magnets are radially magnetized and are mounted on the surface of the rotor and the steel type of D21-50 is chosen for stator and rotor laminations. To reduce the risk of partial demagnetization, an accurate prediction of magnet loss at the design stage is an important role. It can not only give a better efficiency but also may prevent excessive temperature rise in magnets [6]. Considering the almost fixed magnetic field, the rotor core is made of solid steel and the stator core is made of electrical steel with 36 slots for three phase windings. The

prototype model of 3 kW DDPMSG in wind energy conversion system (WECS) can be seen in Figure 4.

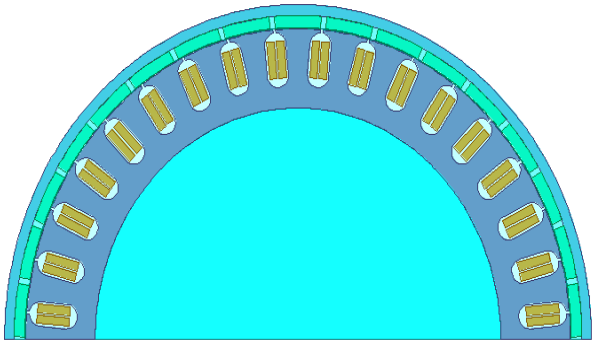


Figure 4. Maxwell 2D Model of PMSG Prototype

3. General Characteristics of PMSG

Nowadays, the DDPMSG is widely used in wind power application which has some advantages compare to other type of generators. The Maxwell 2D stator winding construction of this model is presented in Figure 5.

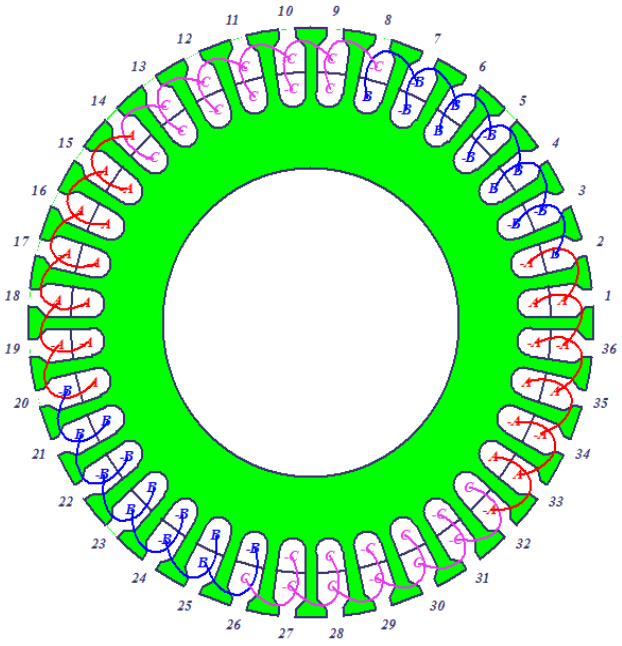


Figure 5. The stator winding connection of 3 kW PMSG

Moreover, the hysteresis (B-H) curves which express the relation between the flux density B and the magnetic field intensity H of the stator and rotor steel is an important characteristics of PMSG[7]. The stator and rotor steel type B-H curve of the FEM model is shown in Figure 6. Figure 7 represents the relationship between the flux density B and the magnetic field intensity H of magnet XG 196/96 from FEM analysis.

3.1. Magnetic Flux and Flux Density Distribution

With the application of high energy-product rare-earth PM to electric machines, many 2-D analytical models have been proposed for the direct driven PM machines and most of them focus on the field distribution and airgap flux density distribution. ANSYS Maxwell 2D RMxprt software is universal finite element analysis software of magnetic field calculation [8].

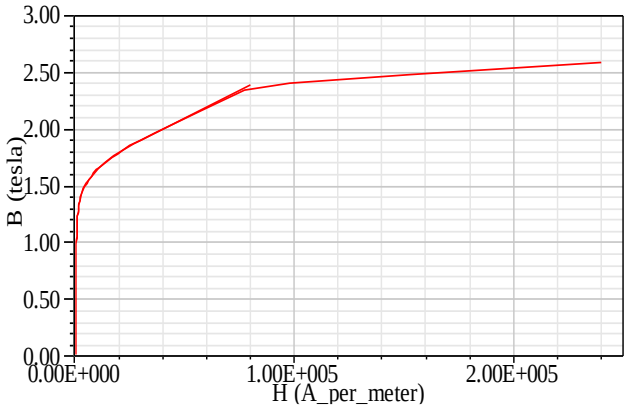


Figure 6. B-H Curve of Stator and Rotor Steel D21-50 RMxprt

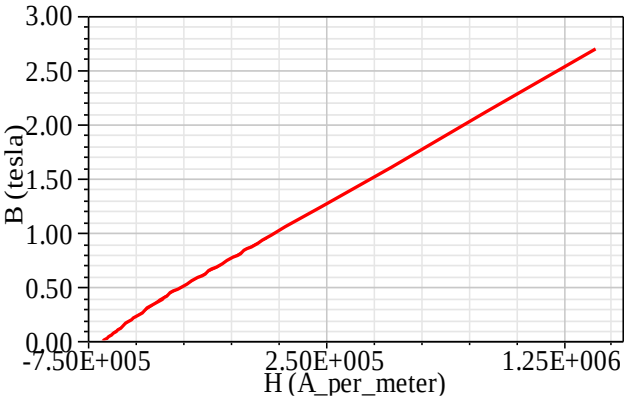


Figure 7. B-H Curve of Magnet (XG196/96) RMxprt

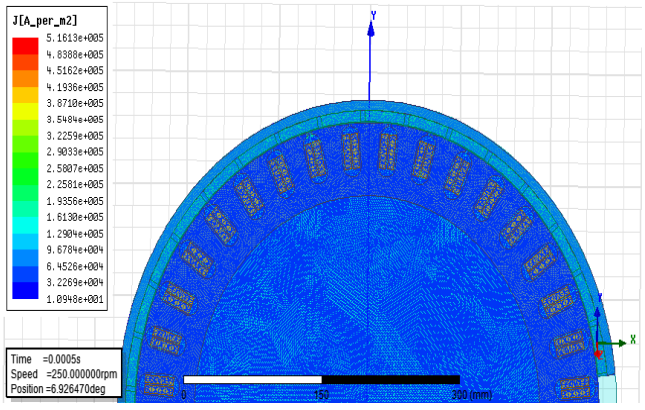


Figure 8. Meshing Plot of J Vector by Maxwell 2-D PMSG Model

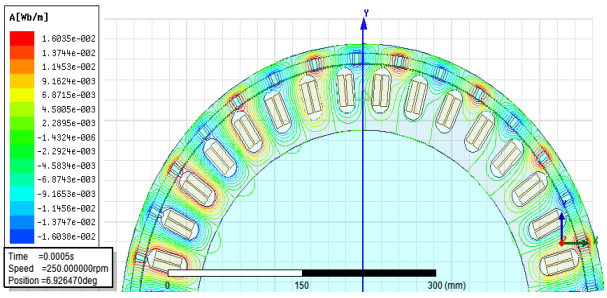


Figure 9. Flux Distribution in Maxwell Design

The model meshing plot of J vector can be seen in Figure *8. And* then, The Maxwell 2D analysis of the important parameters of PMSG such as the flux distribution, field distribution and magnetic flux density distribution are shown in Figure 9, Figure 10 and Figure 11, respectively.

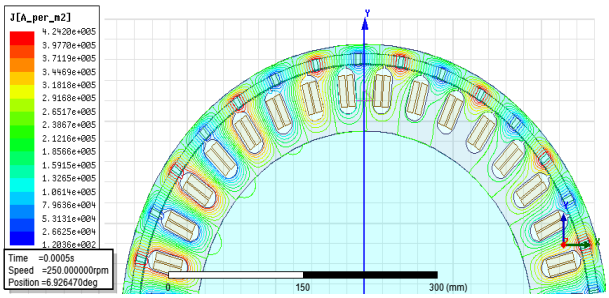


Figure 10. Field Distrubtion of J vector

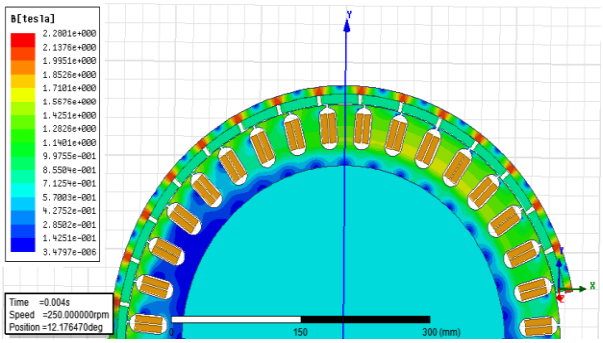


Figure 11. Magnetic flux density distribution of Maxwell 2-D

3.2. Mathematical Models of PMSG

Mathematical model of PMSG is usually given in rotating reference frame of machine. The quadrature (q) axis and the direct (d) axis PMSG equations are given below. Figure 12 and Figure 13 show equivalent circuit of PMSG in d-q frame [9].

$$v_{qs} = R_s i_{qs} + \frac{d\psi_{qs}}{dt} + \omega\psi_{ds} \quad (1)$$

$$v_{ds} = R_s i_{ds} + \frac{d\psi_{ds}}{dt} - \omega\psi_{qs} \quad (2)$$

$$\psi_{ds} = (L_{ld} + L_{md})i_{ds} + \psi_{mag} = L_{ds} i_{ds} + \psi_{mag} \quad (3)$$

$$\psi_{qs} = (L_{ld} + L_{mq})i_{qs} = L_{qs} i_{qs} \quad (4)$$

$$T_e = \frac{3}{2}\frac{P}{2}\psi_{mag} i_{qs} \quad (5)$$

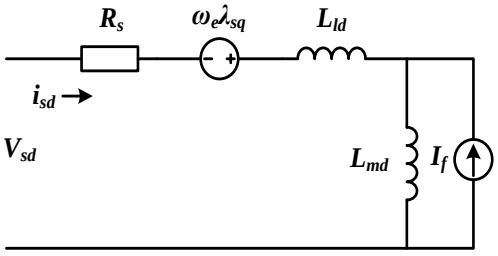


Figure 12. The d- axis equivalent circuit of PMSG

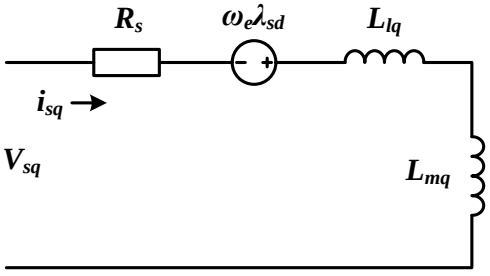


Figure 13. The q- axis equivalent circuit of PMSG

3.3. Finite Element Model Analysis

The two dimensional finite element model of the PMSG of 3kW with surface mounted PM is set up by using Maxwell 2D RMxprt. There are four main modelling processes such as inputting the data of the geometrical size of the PMSG, defining the stator, rotor, airgap, windings and PM properties, defining the boundary condition and generating finite element mesh.

After finishing these four steps, the proposed PMSG could be the modeled and solved though the FEA at different operation modes.

4. Performance Analysis of DDPMSG

The small scale wind turbine is directly connected to PMSG which receives wind turbine torque as a mechanical input and converts it into the electrical input; however, this electrical energy has a variable amplitude and frequency [10]. The FEA model of electromagnetic field is built by Ansoft Maxwell 2D and the simulation time is taken some hours [11]. To visualize magnetic fields and predict magnetic forces, Maxwell 2D program can be used.

The modelling analysis of 3 kW PM generator for wind energy conversion system is presented. The key characteristics of DDPMSG such as the cogging torque, air-gap flux density, induced coil voltage at rated speed and winding current under load are analyzed by Maxwell 2D RMxprt. Figure 14 shows the curve of cogging torque in two teeth. In PMSG, the cogging torque is generated by the sum of a series of cogging torque harmonics [12]. The cogging torque in PM machine can be expressed as:

$$T_c = \frac{1}{4\mu_0} L_{eff}(r_m^2 - r_s^2)\sum_{i=1}^{\infty} nNB_{nN}^2 \lambda_{nN}^2 \sin(nN\theta_r) \quad (6)$$

Where N is the least common multiple of the number of poles and stator slots of the PMSG. The cogging torque causes noise and vibration in PMSG at low speed and the direct drives applications. The frequency of cogging torque is the proportional to the least common multiple of the number of poles and the stator slots of the PMSG. The results obtained by using 2D Maxwell model provide sufficient accuracy with measured ones in a PMSG prototype.

In this FEM model, the proper simulation results of DDPMSG as shown in Figure 14 to Figure 19.The variation of cogging torque with respect to electrical degree can be seen in Figure 14 that generated the 3 kW FE model of Maxwell 2D.

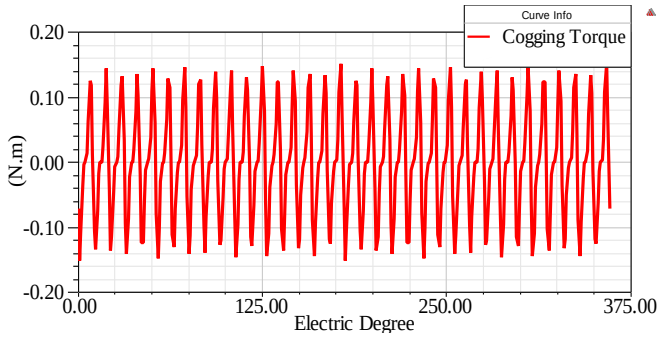


Figure 14. Cogging torque in two teeth of 3kW PMSG

The relationship between the winding voltage under load condition and electrical degree can be determined by the FEM analysis represented in Figure 15. The induced winding voltage of rated speed is shown in Figure 16.

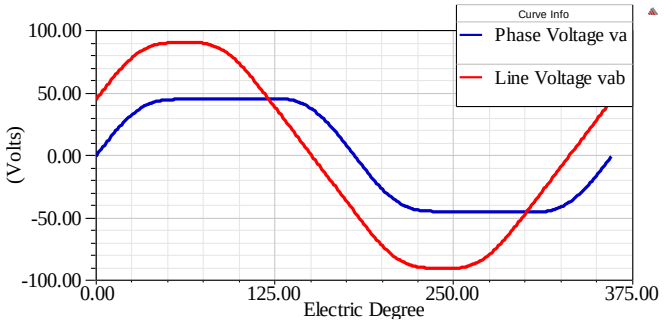


Figure15. Winding Voltage under Load

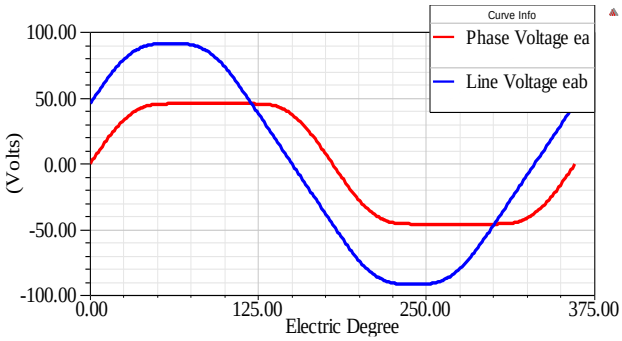


Figure 16. Induced Winding Voltage of Rated Speed

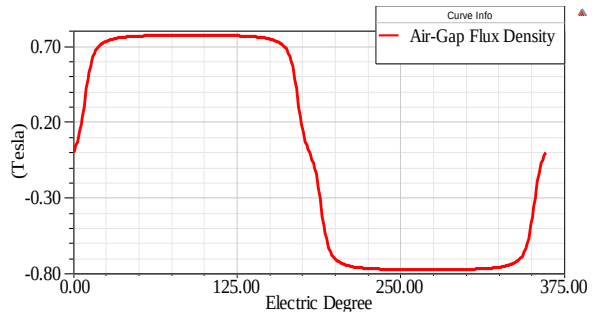


Figure 17. Air –Gap Flux Density

In this model, the maximum air-gap flux density distribution can be generated about 0.7 T by Maxwell 2D.The important parameter of air-gap flux density distribution of PMSG as shown in Figure 17. The performance curves of induced coil voltage at rated speed and the winding current under load are represented in Figure 18 and Figure 19,respectively.

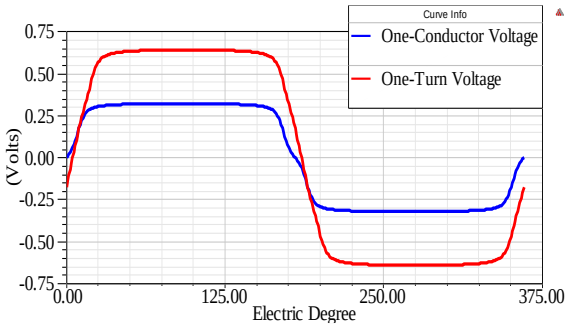


Figure 18. Induced Coil Voltage at Rated Speed

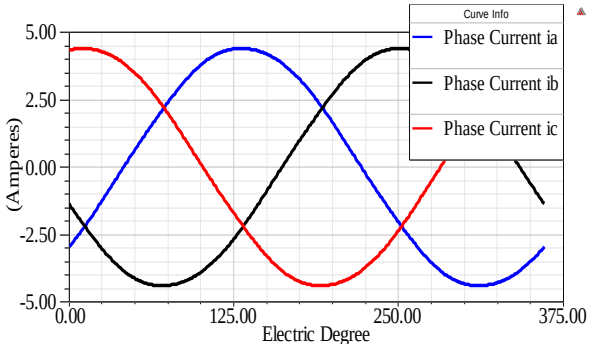


Figure 19. Wndinding Current under Load

Moreover, the transient analysis of 3 kW DDPMSG are discussed in this paper. The transient plots of main parameters in wind generator like the induced winding voltage of phase A,B and C,the comparison of solid losses and core loss,the moving torque of PMSG and the flux distributions in machine are generated by Ansoft Maxwell 2D RMxprt. All of the transient analysis curves are shown below.

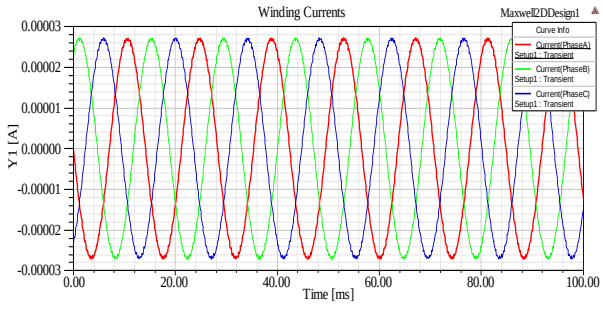


Figure 20. Induced Windind Voltages of Phase A, B, and C

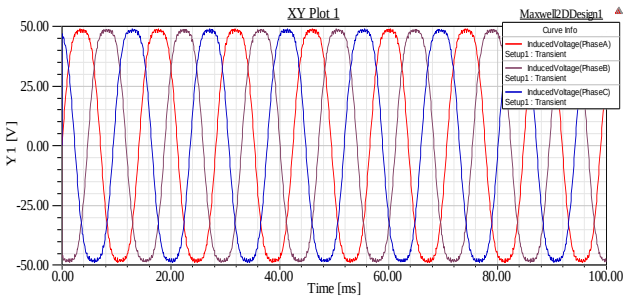


Figure 21 .Comparison of Core loss and Solid loss

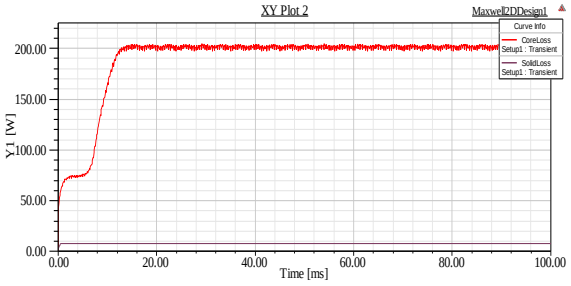


Figure 22. Transient Moving Torque of PMSG

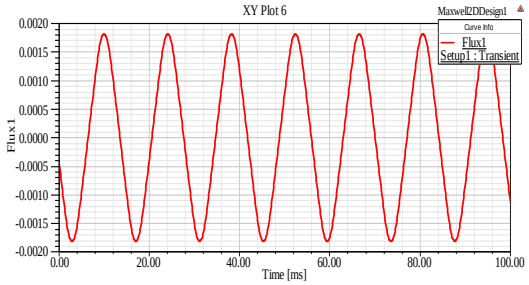


Figure 23. Flux analysis of Generator

5. Conclusions

The 2D modelling analysis of surface mounted PM generator in small scale wind turbine is studied in this paper. There are two main parts of the performance analysis in electrical machine such as the FEM performance curves and the transient curves. In this paper, the key machine parameters analysis are discussed by Ansoft Maxwell 2D RMxprt. To construct the machine model, Maxwell 2D RMxprt is to be a competent and valuable tool. Moreover, the analysis results obtained by using 2D finite element model provided sufficient accuracy with a PMSG prototype. According to the FEM modelling analysis and transient analysis, the design of 3 kW PMSG in this paper is suitable for small wind power applications.

Nomenclature

T_c	Cogging torque
L_{eff}	Effective length of the PMSG
r_m	Inner diameter of the magnet
r_s	Outer diameter of the stator
L_{mq}	Q-axis magnetizing inductance
L_{md}	D-axis magnetizing inductance
ω	Electrical angular speed
T_e	Electromagnetic Torque of PMSG
L_{ds}	D-axis stator inductance
L_{qs}	Q-axis stator inductance
R_s	Stator resistance
v_{qs}	Q-axis stator voltage
v_{ds}	D-axis stator voltage
ψ_{ds}	D-axis flux linkage
ψ_{qs}	Q-axis flux linkage
ψ_{mag}	Permanent magnet flux linkage
i_{ds}	D-axis stator current
i_{qs}	Q-axis stator current

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The lead author is grateful to her supervisor for supporting to develop this study by providing the suitable suggestions.

Reference

[1] Umashankar, D.P.K.a., *Wind Energy Systems and Applications*. 2013, Narosa.
[2] Yasa, Y. and E. Mese. *Design and analysis of generator and converters for outer rotor direct drive gearless small-scale wind turbines*. in *Renewable Energy Research and Application (ICRERA), 2014 International Conference on*. 2014.
[3] Alnasir, Z. and M. Kazerani, *An analytical literature review of stand-alone wind energy conversion systems from generator viewpoint.* Renewable and Sustainable Energy Reviews, 2013. **28**: p. 597-615.
[4] Yan, J., et al., *Control of a grid-connected direct-drive wind energy conversion system.* Renewable Energy, 2014. **66**(Supplement C): p. 371-380.
[5] Faiz, J., et al. *Design of a radial flux permanent magnet wind generator with low coercive force magnets*. in *2016 2nd International Conference on Intelligent Energy and Power Systems (IEPS)*. 2016.
[6] Nair, S.S., et al., *Computationally efficient 3D analytical magnet loss prediction in surface mounted permanent magnet machines.* IET Electric Power Applications, 2017. **11**(1): p. 9-18.
[7] Ducar, I.M. and C.P. Ion. *Design of a PMSG for micro hydro power plants*. in *Optimization of Electrical and Electronic Equipment (OPTIM), 2012 13th International Conference on*. 2012.
[8] Zhang, Z., et al. *Study on circle-annulus meshing methods using ANSYS software and its application in motor electromagnetic field analysis*. in *Electrical and Control Engineering (ICECE), 2011 International Conference on*. 2011.
[9] Ugalde-Loo, C.E., et al. *Analysis and efficient control design for generator-side converters of PMSG-based wind and tidal stream turbines*. in *Power Systems Computation Conference (PSCC), 2014*. 2014.
[10] Fazli, N. and J. Siahbalaee. *Direct torque control of a wind energy conversion system with permanent magnet synchronous generator and Matrix Converter*. in *2017 8th Power Electronics, Drive Systems & Technologies Conference (PEDSTC)*. 2017.
[11] Alam, F.R. and K. Abbaszadeh, *Magnetic Field Analysis in Eccentric Surface-Mounted Permanent-Magnet Motors Using an Improved Conformal Mapping Method.* IEEE Transactions on Energy Conversion, 2016. **31**(1): p. 333-344.
[12] Liu, T., S. Huang, and J. Gao. *Optimization of a direct-drive wind power generation system by using PMSG*. in *Power Engineering, Energy and Electrical Drives (POWERENG), 2011 International Conference on*. 2011.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 100-106 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Switched-Capacitor Low-Pass Filter with Dynamic Switching Bias OP Amplifiers

Hiroo Wakaumi*

Tokyo Metropolitan College of Industrial Technology, 140-0011, Japan

ARTICLE INFO

Article history:
Received : 28 September, 2017
Accepted : 8 November, 2017
Online : 18 November, 2017

Keywords :
Switched Capacitor Circuit
Filter
CMOS
Operational Amplifier
Dynamic Switching

ABSTRACT

A switched capacitor low-pass filter employing folded-cascode CMOS OP Amps with a dynamic switching bias circuit capable of processing video signals, which enables low power consumption, and operation in wide bandwidths and low power supply voltages, is proposed. In this filter, charge transfer operations through two-phase clock pulses during the on-state period of the OP Amps and a non-charge transfer operation during their remaining off-state period are separated. Through simulations, it was shown that the low-pass filter with an OP Amp switching duty ratio of 50 % is able to operate at a 14.3 MHz high-speed dynamic switching rate, allowing processing video signals, and a dissipated power of 68 % of that observed in the static operation of the OP Amps and a full charge transfer operation without separation of a cycle period. The gain below -31 dB in the frequency response, which is suitable, was obtained at over 6 MHz within a stop-band. Especially high attenuation in 5 MHz was achieved under the optimized condition of load capacitances (4 pF) of OP Amps.

1. Introduction

The switched capacitor (SC) techniques are appropriate for realizing various filters that can be integrated in monolithic ICs (Integrated Circuits) from using the CMOS (Complementary Metal-Oxide-Semiconductor) technology. In conventional active RC filters, monolithic ICs cannot be realized from using resistors. On the contrary, the CMOS SC techniques suitable for realizing analog signal processing ICs, have promising use in video signal bandwidth circuits in particular, because these can replace resistors to switched capacitor pairs with small capacitances. It has been demonstrated that SC techniques using CMOS operational amplifiers (OP Amps) are useful for implementing analog functions such as filtering [1-5]. Although CMOS OP Amps are suitable for such filter ICs, the use of several OP Amps results in large power consumption. Especially, the power consumption of OP amplifiers in high speed operation becomes large because they have wideband properties. Therefore, the use of them is currently limited to the use in low-frequency passband of at most a few hundred kHz (that is, applications of low speed signal processing, such as analog voice signals).

*Corresponding Author: Hiroo Wakaumi, Tokyo Metropolitan College of Industrial Technology, Tokyo, 140-0011 Japan
E-mail: wakaumi@metro-cit.ac.jp ; waka781420j@ab.auone-net.jp

Until now, several approaches have been considered to decrease the power consumptions of OP Amps, including the development of ICs that work at low power supply voltages [6]. A clocked current bias scheme for folded-cascade OP Amps suitable for achieving a wide dynamic range has typically been proposed to decrease the power consumption of the OP Amp itself [7, 8]. Because the circuit requires complicated four-phase bias-current control pulses and biasing circuits, it results in a large layout area and is not suitable for the high-speed operation. A control method using power supply switching has been proposed for audio signal processing as another approach in decreasing the power consumption of OP Amps [9]. Because large capacitors for the power supply terminals are intrinsically loaded, the switching speed is limited to a low speed of 1 MHz at most. Therefore, this type of control circuit is not suitable for application to video signal processing ICs, which are required to operate at over 10 MHz switching frequency.

Recently, the author proposed a folded-cascade CMOS OP Amp with a dynamic switching bias circuit (DSBFC OP Amp), of simple configuration, to provide low power consumption while maintaining high speed switching operation suitable for processing video signals [10]. However, low-power signal processing circuits such as wideband filters with such DSBFC OP

Amp, suitable for processing video signals, have not been developed yet.

In this paper, a configuration of SC Butterworth low-pass filter (LPF) with DSBFC OP Amps [11] as an example of application of the DSBFC OP Amp is proposed, which enables low power consumption and is suitable for achieving wide bandwidths and operation in low power supply voltages, and its availability is also evaluated for the performance of frequency response and power dissipation. Further, the effect of OP Amp load capacitances on the frequency response is evaluated. This paper is an extension of work originally presented in LASCAS 2017 [11].

2. SC Filter Theory

The discrete-time transfer function of second-order SC infinite impulse response (IIR) LPF is shown as follows using the z-transform.

$$H(z) = \frac{K\ (1+z^{-1})^2}{1+\sum_{k=1}^{2} b_k z^{-k}} \quad (1)$$

Here, K, b^k, and z^{-1} represent the gain constant, the filter coefficient in a recursive loop, and the one-step delaying operation, respectively. All of operation circuits are composed of active sampled data processing circuits with a sampling circuit, switching circuits and capacitors.

3. SC Filter Circuit Design

A second-order IIR LPF with the Butterworth frequency characteristic was designed because it is easy to design due to its flat gain characteristic in the passband. The Butterworth LPF is also superior to a Chebyshev filter for processing video signals owing to its ripple less characteristic within the passband. The filter order of second was selected to achieve a gain of -30 dB at a stop-band over 6 MHz. The other design condition was set as follows. That is, a sampling frequency fs=14.3 MHz, which is equal to four times as much as the NTSC color sub-carrier frequency 3.58 MHz, and a cutoff frequency fc=2 MHz, respectively, were chosen, that enable the LPF to process video signals. Under this condition, the inverted discrete-time transfer function is given by

$$H(z) = -\frac{0.1174(1+z^{-1})^2}{1-0.8252z^{-1}+0.2946z^{-2}} \quad (2)$$

The circuit configuration realizing this transfer function is shown in Figure 1. In order to enable easily to determine the capacitance value of each capacitor, the coefficient of A is set to be equal to that of B. Capacitors can be basically divided into two groups. In one group (C, D, E, and G), charges are supplied to OP Amp 1. In another group (A, B, and I), charges are supplied to OP Amp 2. Even if a capacitance of each capacitor group is multiplied by constant times, the transfer function does not change. Therefore, capacitances of integral capacitors B and D are here selected as a reference capacitance in each group and each coefficient of B and D is normalized to 1. At this time, 1 as every normalized coefficient of A, B, and D is obtained because A and B are the same coefficient. In Figure 1, when the coefficients of A, B, and D are normalized to 1, other coefficients are determined as follows.

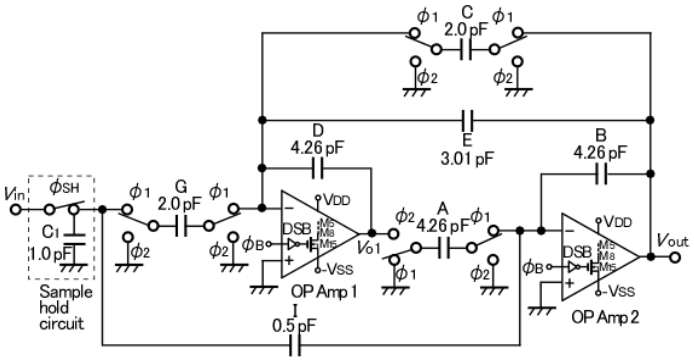


Figure 1: Configuration of the 2nd-order SC LPF with DSBFC OP Amps

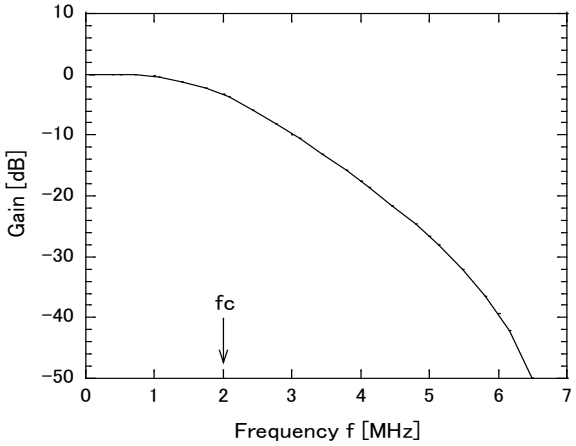


Figure 2: Theoretical frequency response of the 2nd-order SC Butterworth LPF

I=K=0.1174

G=2K+2I=0.4694

C=1+b_1+b_2=0.4694

E=1-b_2=0.7054

Here, b_1=-0.8252 and b_2=0.2946. When the smallest coefficient of I=0.1174 is replaced as a reference capacitance of 0.5 pF, each capacitance in the SC IIR LPF IC is set in proportion to the above coefficients as shown in Figure 1. Because its input signal is desirable to be maintained by a sample/hold circuit for stabilizing, this sampling circuit is also applied in the SC LPF. At this time, the transfer function is multiplied by the following zero-order hold function due to a sample-hold effect.

$$H_s(j\omega)=\frac{\sin(\omega T_s/2)}{(\omega T_s/2)} \quad (3)$$

Here, Ts represents a cycle period of sampling and switching pulses. Therefore, when the function of (2) is replaced using $z=e^{j\omega T_s}$, the magnitude of the transfer function of the second-order SC LPF considering the sample-hold effect is given by (4).

$$|H(j\omega)|=\frac{\sin(\omega T_s/2)}{(\omega T_s/2)} \cdot \frac{0.2347(1+\cos\omega T_s)}{\sqrt{1.7678-2.1368\cos(\omega T_s)+1.6513\cos(2\omega T_s)}} \quad (4)$$

The theoretical frequency response including the sample-hold effect is shown in Figure 2. The SC LPF configuration was designed referencing a SC biquad circuit with integrators in the

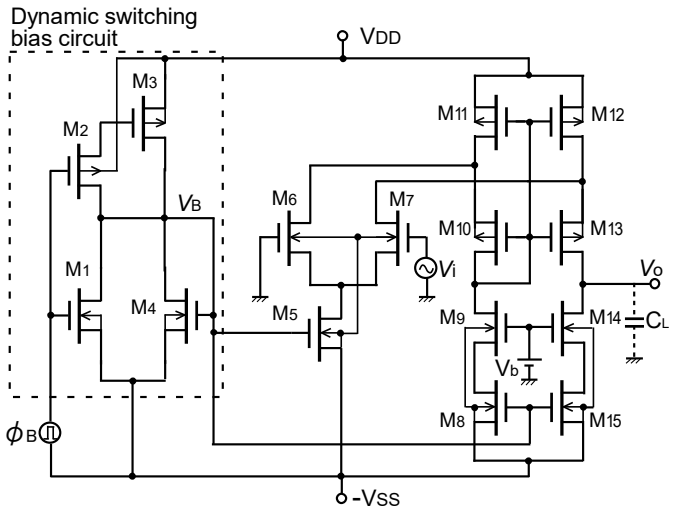


Figure 3: Configuration of the DSBFC OP Amp

reference [12]. It consists of a sample-hold circuit with a holding capacitor C_1 and a CMOS sampling switch controlled by ϕ_{SH}, CMOS switches ϕ_1, ϕ_2, capacitors A, B, C, D, E, G, and I, and two FC CMOS OP Amps with the DSB circuit (DSBFC OP Amps). A configuration of the DSBFC OP Amp enabling low power consumption, which is different from conventional ordinary OP Amps [10], is shown in Figure 3. With respect to the DSBFC OP Amps, the same CMOS channel width / length as that in the DSBFC OP Amp shown in the reference [10] was employed. The sampling switch was designed to a channel width / channel length W/L=35/2.5 (μm/μm) for each of p-MOSFET and n-MOSFET. The holding capacitor C_1 has a small capacitance of 1 pF. CMOS switches with a W/L=25/2.5 (μm/μm) are turned on and off by non-overlapping two-phase clock pulses ϕ_1 and ϕ_2, swinging from -2.5 V to 2.5 V. These sampling and CMOS switches are designed to have a balanced structure with each equal length and width of p-channel and n-channel MOSFETs (component of these CMOS switches) to delete a feed-through phenomenon, which is caused by gate clock pulses due to a capacitive coupling between gate and CMOS-switch output terminals. Major CMOS process parameters are given as a gate insulating film thickness t_{ox}=50 nm, an n-MOSFET threshold voltage V_{Tn}=0.6V, and a p-MOSFET threshold voltage V_{TP}=-0.6 V.

The operation principle of this LPF is simply described in the following. The output signal V_{o1} of OP Amp 1 is obtained as an additional output of an integrated signal of V_{in} using a negative integrator (D, G SC circuit, and OP Amp 1), an integrated signal of V_{out} using a negative integrator (D, C SC circuit, and OP Amp 1), and a signal multiplied V_{out} by E/D. The output signal V_{out} is an additional output of an integrated signal of V_{o1} using a positive integrator (A SC circuit, B, and OP Amp 2), and a signal multiplied V_{in} by I/B. V_{out} is basically fed back to an input of OP Amp 1 like this. V_{in} is also integrated twice and added after being decreased by an appropriate capacitance ratio. Due to these integration using positive / negative integrators, addition and feedback operations, the function of LPF is achieved.

The operation waveforms of the SC LPF are shown in Figure 4. In this SC LPF, charge transfer operations through the clock

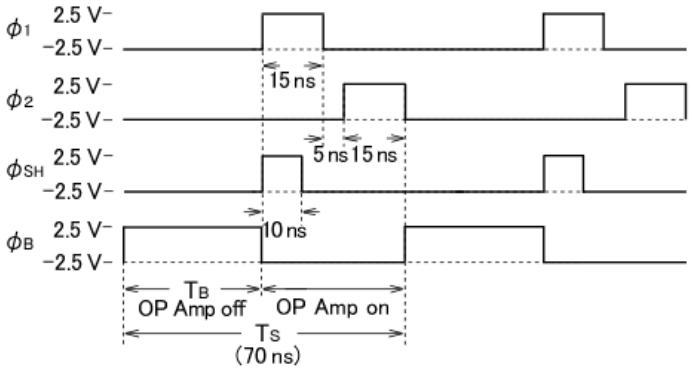


Figure 4: Operation waveforms of the 2nd-order SC LPF

pulses ϕ_1 and ϕ_2 are achieved during the on-state period of the DSBFC OP Amps. The off-state period T_B (the remaining period of the one cycle period Ts) is separately provided to realize low power dissipation for the SC LPF. An input signal is sampled during the sampling phase of ϕ_{SH} (10 ns) and the first period of clock phase ϕ_1, while its corresponding charge is stored on the holding capacitor C_1, and is transferred to an output terminal V_{out}, charging all capacitors. The voltage at the off transition of ϕ_{SH} is kept on C_1 during the remaining period of clock phase ϕ_1. During subsequent clock phase ϕ_2, each charge of two capacitors C and G is discharged and each charge of remaining capacitors is redistributed. These charge transfer operations are achieved during the on-state period of the OP Amps. During this period, the OP Amps turn on by setting a bias voltage of V_B at an appropriate level enabling M3 and M4 to operate in the saturation region, and operate normally as operational amplifiers. ϕ_B is set to low just before ϕ_1 changes to high.

Subsequently, ϕ_B becomes 2.5 V at the off-state transition of the OP Amps, at the same time ϕ_2 is switched to off. During this off period T_B, the OP Amps turn off by setting V_B at nearly -2.5 V enabling M1 to operate in a low impedance and M3 in a high impedance. Therefore, during this off period, the OP Amps do not dissipate at all. When T_B is relatively long as compared to the one cycle period Ts, the power dissipation is expected to become lower than that observed in ordinary static operation for the SC LPF using conventional OP Amps. If half GB (Gain Bandwidth product) OP Amps for the SC LPF are used in the static operation (without DSB operation), rise and fall times to stable states of filter output signals will increase to much larger than twofold, because slowly changing transition occurs at the end of transition. Therefore, an expected proper filter performance will not be able to be obtained when such OP Amps are used.

4. Simulation Results

The performance of the SC LPF was investigated by simulation using the SPICE (Simulation Program with Integrated Circuit Emphasis) program package. Operation waveforms for an input signal of 1 MHz with an amplitude of 0.3 V and an output load capacitance of 1 pF are shown in Figure 5. For the passband frequency signal, the same level output signal as the input one was obtained. The frequency response of the SC LPF is shown in Figure 6 (a) in the dynamic switching operation of the DSBFC OP Amps and (b) in the on-state (static operation) of the DSBFC

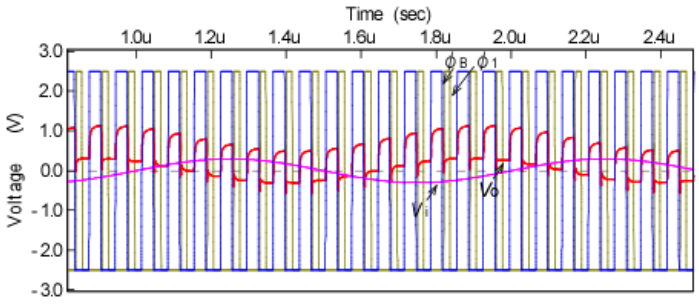
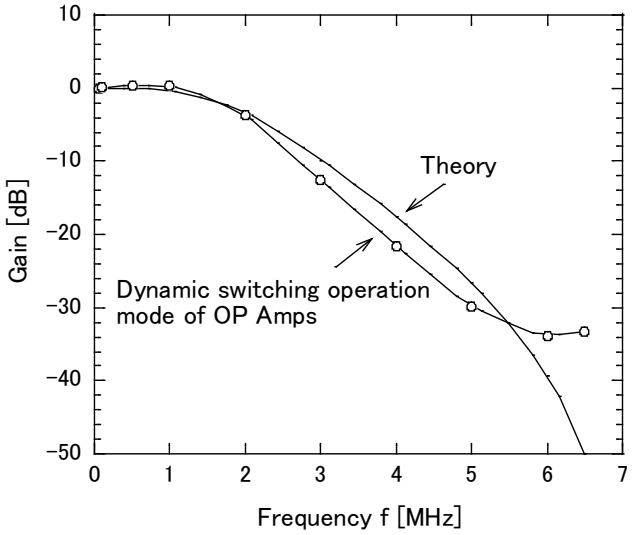
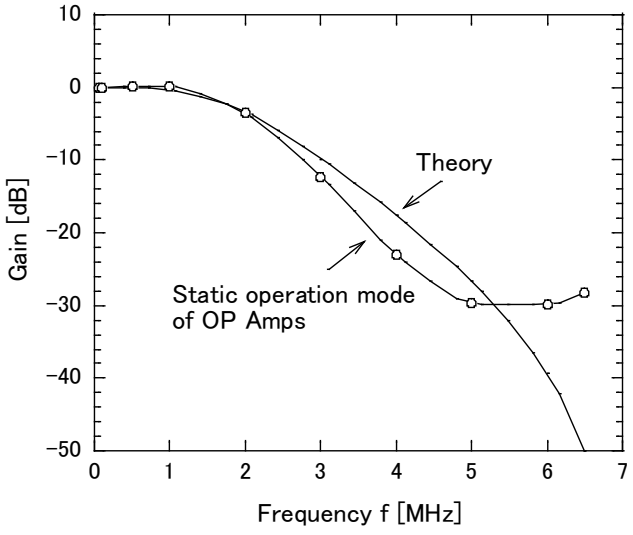


Figure 5: Simulation waveforms for the 2nd-order SC Butterworth LPF. Input signal=0.3 $V_{0\text{-}P}$, Input signal frequency $f$in=1 MHz



(a)



(b)

Figure 6: Frequency response of the 2nd-order SC Butterworth LPF. (a) Dynamic switching operation mode of the DSBFC OP Amps, T_B=35 ns, (b) Static operation mode of the DSBFC OP Amps, ϕ_B=-2.5 V

OP Amps at ϕ_B= -2.5 V. The uncertainty of gain characteristics in simulation is nearly 1 dB at the frequency of 6.5 MHz. The frequency responses in case of both modes are almost the same for the in-phase state, in which the maximum output signal is

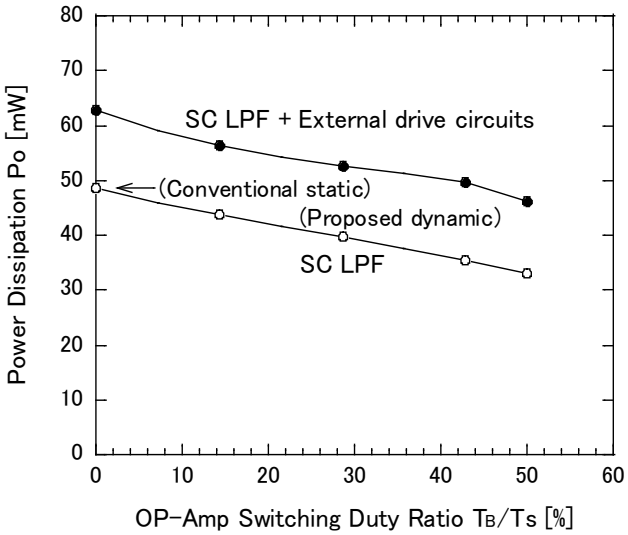


Figure 7: Power dissipation vs. OP Amp switching duty ratio in the 2nd-order SC LPF. $f$in=1 MHz. A constant pulse width of 15 ns for ϕ_1 and ϕ_2 was used

obtained depending on a phase between the input signal and the sampling pulse. The response was near the theoretical one from 100 kHz up to near 5 MHz, in the high frequency range over 6 MHz, it deteriorated due to a sampling phase effect. The gain below -33 dB was obtained at over 6 MHz within a stop-band. This was superior to that in the static operation. Therefore, it is thought that there is almost no gain deterioration caused by employing the DSB operation of the OP Amps. Though the stop-band gain in the second order LPF is not low enough as shown above, it is expected that the roll-off in the frequency response will become steeper by increasing the filter order, and so the stop-band gain will greatly decrease. Therefore, the achievement of a wide stop-band with a high attenuation will become possible.

Power dissipation vs. OP-Amp switching duty ratio with ϕ_1= ϕ_2=15 ns is shown in Figure 7. The power dissipation of the SC LPF itself excluding that of external drive circuits for ϕ_{SH}, ϕ_1, ϕ_2, and ϕ_B decreased in proportion to the off period T_B of the Op Amps as expected. In the operation mode of T_B=35 ns (=50 % switching duty ratio) and ϕ_1= ϕ_2=15 ns, the power dissipation of the SC LPF (32.9 mW) was decreased to 68 % as compared to that in the static operation of the OP Amps (48.5 mW). In the full charge transfer operation without separation of the one cycle period Ts (ϕ_1= ϕ_2=30 ns), the power dissipation of the LPF was 48.5 mW, the same as that in the static operation of the OP Amps with the above separated charge transfer mode. Most of the power dissipation of the SC LPF corresponds to total power consumption dissipated in the OP Amps themselves. The power consumption of the external drive circuits was nearly 13~14 mW. Thus, even when two DSBFC OP Amps are applied to the SC LPF, the dynamic operation of the DSBFC OP Amps enabling low power dissipation as compared to their static operation is also useful for reducing the power dissipation of SC LPF. When the SC LPF is operated at a lower dynamic switching rate, because it enables T_B/Ts to become larger than 50 %, the power dissipation of the SC LPF is expected to decrease in proportion to T_B still more. This means that the SC LPF with high-speed DSB operation OP Amps is advantageous compared to the SC LPF using static operation OP Amps with a lower GB.

Table 1: Performance comparison of SC LPFs with CMOS technology

Performance parameters	Fifth-Order SC LPF with CMOS OP Amps [13]	2nd-Order SC LPF with DSB CMOS OP Amps
Process technology	0.35 µm CMOS	2.5 µm CMOS
Power supply voltage	5 V	5 V
Sampling frequency fs	103.68 MHz (Double sampling)	14.3 MHz
Cutoff frequency fc	20 MHz	2 MHz
Power consumption	125 mW	32.9 mW (Dynamic : T_B/T_S=50 %)

When the filter order is increased, a SC LPF with DSB OP Amps uses OP Amps of the number of filter order. Therefore, the power dissipation of this SC LPF is expected to increase in proportion to its filter order. The power dissipation for the reported fifth-order SC LPF employing conventional CMOS OP Amps with a 5-V power supply and a 0.35-µm CMOS technology was 125 mW as shown in the performance comparison of Table 1 [13]. If a fifth-order SC LPF using the DSB OP Amps with the sampling frequency fs=14.3 MHz is achieved, its power dissipation will be estimated to be 82.3 mW. Considering that the power consumption of OP Amps parts in each LPF is dominant, a comparison between these values is possible. Obviously, the estimated power dissipation of a revised version of the proposed SC LPF with DSB OP Amps is much less than that of the above conventional fifth-order SC LPF.

5. Effect of Load Capacitances

Because the DSBFC OP Amp switches dynamically, its output becomes a quasi-floating state during the off-state period. In the off-state period of the OP Amp, though MOSFETs M5, M8, and M15 turn to the off-state completely, MOSFETs M11 and M12 become the on-state strongly because over the threshold voltage between each gate and source is applied. At this time MOSFETs M6, M7, M9, M10, M13, and M14 become the on-state weakly. The output terminal *Vo* of the OP Amp is set to a voltage depending on the load capacitance through the capacitive coupling between the drain and the gate of the MOSFET M13. Therefore, when a large output swing in *Vo* occurs at the off-state transition, there is a fear that the output voltage during the subsequent on-state period of the OP Amp suffers the influence of this transition. So, a dynamic offset voltage Voff (the difference of the on-state and the off-state output voltages of the OP Amp) at the off-state transition of the OP Amp vs. load capacitance C_L was tested and is shown in Figure 8. Obviously, the dynamic offset voltage depends on the load capacitance and decreases as its load capacitance becomes large because C_L compared to the drain-gate capacitance in M13 becomes large. The change of Voff against C_L of the SC LPF resembles to that of the OP Amp (Figure 9). This means that Voff of the SC LPF is mainly determined by the OP Amp's dynamic off-state transition. Gain vs. OP Amp load capacitance (of two OP Amps) for the SC LPF is shown in Figure 10. In this case, the phase between the input signal V_{in} and the sampling signal ϕ_{SH} was fixed to a constant value in each input signal to avoid the sampling phase effect. The gain became the minimum at a load capacitance of nearly 4~5 pF for the input signal frequency of 5 MHz. The reason is explained in the following. In small load capacitances, V_{off} is not only large, but also its variation is not negligible depending on the on-state output voltage. This phenomenon causes the on-state output voltage's difficulty to

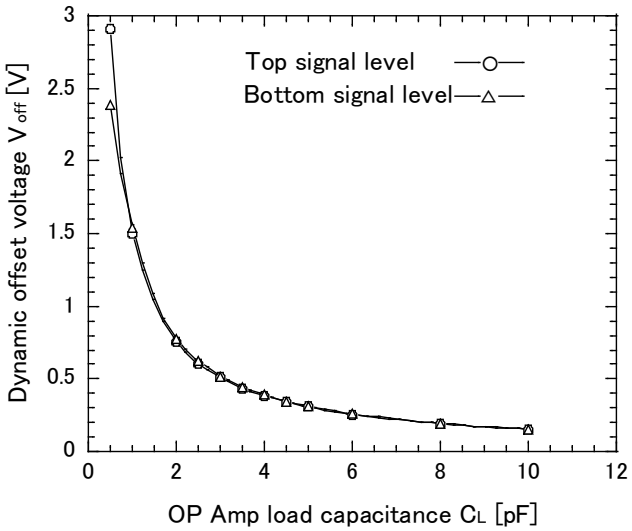


Figure 8: Dynamic offset voltage vs. load capacitance in the DSBFC OP Amp. Input signal=0.00085 $V_{0\text{-}P}$, Input signal frequency *f*in=0.5 MHz, Cutoff frequency *f*c at C_L of 1 pF =4.13 MHz

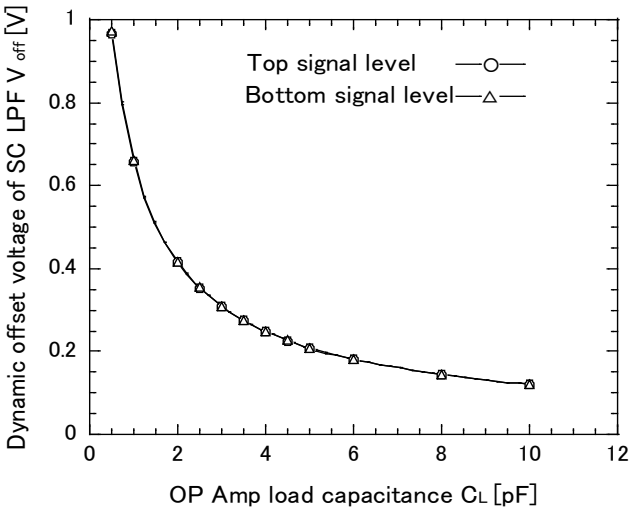


Figure 9: Dynamic offset voltage vs. OP Amp load capacitance in the SC LPF. Input signal *V*in=0.3 $V_{0\text{-}P}$, Input signal frequency *f*in=5 MHz

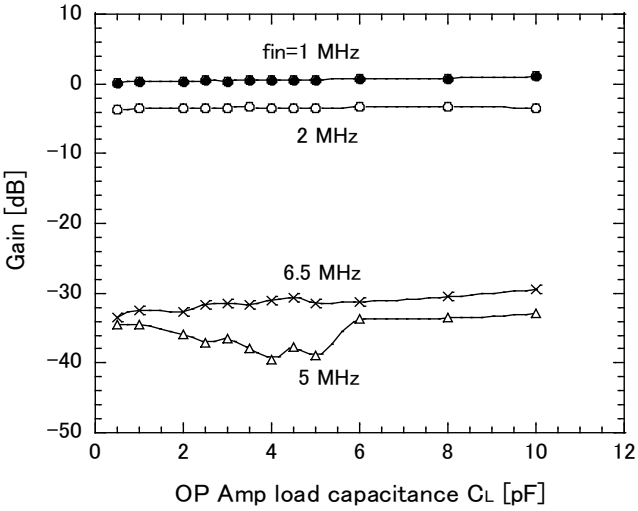


Figure 10: Gain vs. OP Amp load capacitance in the SC LPF

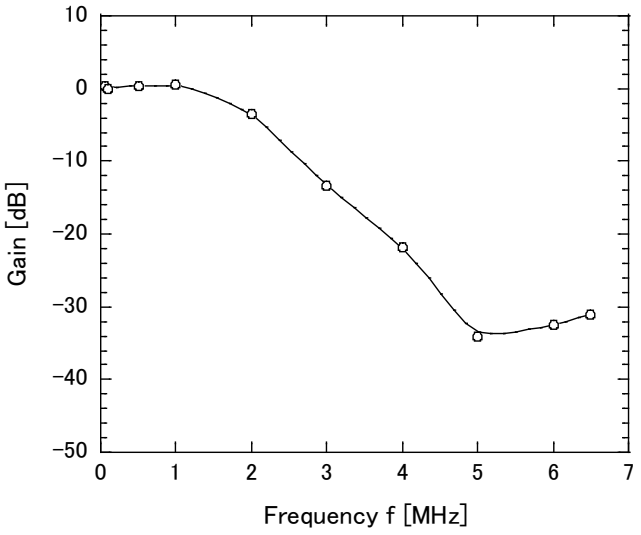


Figure 11: Frequency response of the SC LPF. OP Amp load capacitance C_L=4 pF

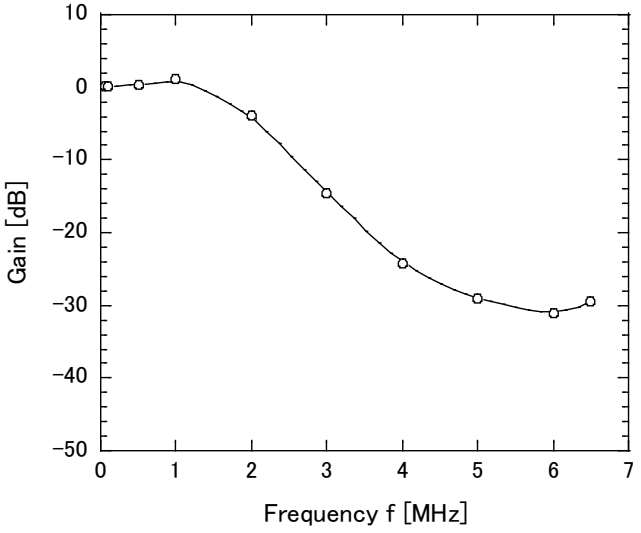


Figure 12: Frequency response of the SC LPF. OP Amp load capacitance C_L=10 pF

reach its steady state completely within the short time period of 15 ns (in ϕ_1 or ϕ_2), while permitting its output voltage to change slightly (bringing about the SC LPF gain's deterioration). In large load capacitances, it becomes also hard to reach its steady state within 15 ns (causing the gain's deterioration) due to OP Amp's bandwidth deterioration although V_{off} and its variation are very insignificant. However, at an optimum capacitance of nearly 4 pF, its gain deterioration becomes slight due to small offset transition with slight offset voltage variation and fast transition to the steady state. For other input signals of low frequencies (1, 2 MHz), the variation effect of V_{off} is negligible due to its large output signal (the change of gains is hardly seen) although its variation depending on the output signal voltage is relatively large in small load capacitances. Of course, since V_{off} and its variation for these low frequencies are negligibly small in large load capacitances, the change of gains with C_L is dominated by only OP Amp's bandwidth deterioration. For the 6.5 MHz input signal, though the gain is basically determined due to the sampling phase effect, it is slightly changing with C_L due to OP Amp's bandwidth deterioration.

Under the optimized load capacitance of 4 pF, the frequency characteristic of gain for the SC LPF was tested. As shown in Figure 11, its high frequency gain of near 5 MHz improved drastically (over 4.3 dB compared to the gain in the load capacitance C_L of 1 pF). Though its gain increased slightly (up to -31 dB) in 6.5 MHz, the amount of its gain increase is slight. On the contrary, when C_L is increased to 10 pF larger than the optimized value, the frequency response of the SC LPF deteriorated in high frequencies over 5 MHz (Figure 12). Thus we can see that the SC LPF gain characteristics can be improved by the optimization of load capacitances of DSB OP Amps. Typical characteristics are listed in Table 2.

Table 2: Typical characteristics of SC LPF with DSB OP Amps (V_{DD}=V_{SS}=2.5 V)

Performance parameters	Simulation results
Sampling and switching frequency fs	14.3 MHz
Input signal amplitude	0.3 $V_{0\text{-}P}$
Cutoff frequency fc	2 MHz
Gain at over 6 MHz within a stopband	<-31 dB (DSB mode)
	<-27 dB (Static mode)
Power consumption	48.5 mW (Conventional static)
	32.9 mW (Proposed dynamic: T_B/Ts=50 %)

Conclusions

A switched capacitor low-pass filter employing folded-cascode CMOS OP Amps with a dynamic switching bias circuit capable of processing video signals, which enables low power consumption, operation in wide bandwidths and low power supply voltages, was proposed and its performance was evaluated. In this SC LPF, charge transfer operations through two-phase clock pulses during the on-state period of the OP Amps and non-charge transfer operation during the remaining off-state period of the OP Amps were separated. Through SPICE simulations, it was shown that the SC LPF is able to operate at a 14.3 MHz high-speed dynamic switching rate, allowing processing video signals, and a dissipated power of 68% of that observed in the static operation of the OP Amps and the full charge transfer mode without separation of the one cycle period. The power consumption in the SC LPF body except for the external drive circuits was that of OP Amps. When rearranging these results, it became clear that a lower-dissipated-power SC LPF employing DSB folded-cascode CMOS OP Amps compared to conventional SC LPFs with static operation OP Amps can be realized. The gain below -31 dB in the frequency response, which is suitable, was also obtained at over 6 MHz within a stop-band. Especially high attenuation in 5 MHz was achieved under the optimized condition of load capacitances (4 pF) of OP Amps.

Thus, the dynamic charge transfer operation during the on-state period of the OP Amps and non-charge transfer operation during their off-state period is useful for high speed operation, and reducing the power dissipation of the SC LPF. This circuit should be useful for the realization of low-power wide-band signal processing ICs including over one of multi-order low-pass, high-pass and band-pass filters. The DSB circuit achieving such operation can be applied to not only folded-cascode but telescopic, two-stage, and rail-to-rail OP Amps.

References

[1] R. Gregorian and W. E. Nicholson, "CMOS Switched-Capacitor Filters for a PCM Voice CODEC," IEEE J. Solid-State Circuits, **SC-14**(6), 970-980, Dec. 1979.

[2] R. Dessoulavy, A. Knob, F. Krummenacher, and E. A. Vittoz, "A Synchronous Switched Capacitor Filter," IEEE J. Solid-State Circuits, **SC-15**(3), 301-305, June 1980.

[3] A. Iwata, H. Kikuchi, K. Uchimura, A. Morino, and M. Nakajima, "A Single-Chip Codec with Switched-Capacitor Filters," IEEE J. Solid-State Circuits, **SC-16**(4), 315-321, Aug. 1981.

[4] J.-T. Wu, Y.-H. Chang, and K.-L. Chang, "1.2 V CMOS Switched-Capacitor Circuits," in 1996 IEEE International Solid-State Circuits Conference Digest of Technical Papers-42nd ISSCC, 388-389, 479.

[5] S. Chatterjee, Y. Tsividis, and P. Kinget, "Ultra-Low Voltage Analog Integrated Circuits," IEICE Trans. Electronics, **E89-C**(6), 673-680, June 2006.

[6] Z. Kun, W. Di, and L. Zhangfa, "A High-Performance Folded Cascode Amplifier," in 2011 International Conference on Computer and Automation Engineering-ICCAE 2011, IPCSIT, **44**, 41-44, 2012.

[7] D. B. Kasha, W. L. Lee, and A. Thomsen, "A 16-mW, 120-dB Linear Switched-Capacitor Delta-Sigma Modulator with Dynamic Biasing," IEEE J. Solid-State Circuits, **34**(7), 921-926, July 1999.

[8] H. L. Chen, Y. S. Lee, and J. S. Chiang, "Low Power Sigma Delta Modulator with Dynamic Biasing for Audio Applications," in 50th Midwest Symposium on Circuits and Systems 2007-MWSCAS 2007, 159-162.

[9] H. Wang, J. Xu, and X. Wu, "A Low Power Audio Delta-Sigma Modulator with Opamp-Shared and Opamp-Switched Techniques," in Proceedings of the International MultiConference of Engineers and Computer Scientists 2010-IMECS 2010, Ⅱ, 1296-1300.

[10] H. Wakaumi, "A Folded-Cascode OP Amplifier with a Dynamic Switching Bias Circuit," Engineering Letters, **23**(2), 92-97, June 2015.

[11] H. Wakaumi, "A Switched-Capacitor Filter with Dynamic Switching Bias OP Amplifiers," in 8th Latin American Symposium on Circuits and Systems-LASCAS 2017, 33-36, Feb. 2017.

[12] M. Takebe, J. Iwata, N. Takahashi, and H. Kunieda, Switched Capacitor Circuit, Tokyo: Gendai Kohgaku-Sha, Apr. 2005.

[13] U.-K. Moon, "CMOS High-Frequency Switched-Capacitor Filters for Telecommunication Applications," IEEE J. Solid-State Circuits, **35**(2), 212-220, Feb. 2000.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 99-105(2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Detection of ExoMars launcher during its passage over Europe with Space Surveillance radar breadboard

Stéphane Saillant*, Marc Flécheux, Yann Mourot

Department of Electromagnetism and Radar, ONERA, Palaiseau 91123, France

A R T I C L E I N F O

Article history:
Received: 15 November, 2017
Accepted: 01 February, 2018
Online: 12 March, 2018

Keywords:
Radar
Bistatic
Kinematic Analysis
Space surveillance
ExoMars
Space launcher

A B S T R A C T

A bistatic radar breadboard for space surveillance has been developed by ONERA for the European Space Agency. This item was operated during the launch of the ExoMars mission on March 14th 2016. The spacecraft, attached to the Proton launcher, was well detected in real-time during its passage over South Europe. This paper presents the setting up of an experiment to detect this particular type of targets with the radar breadboard. The results of its operation as space surveillance system as well as a specific kinematic analysis of the ExoMars spacecraft as viewed from the radar.

1. Introduction

This paper is an extension of works originally presented in 2017 IEEE Radar Conference (RadarConf17) [1].

ABISS project (Antenna BIstatic for Space Surveillance) is relative to the breadboarding of a ground-based surveillance radar demonstrator for European Space Agency (ESA) [2]. This paper presents the setting up of an experiment to detect certain types of targets with this Space Surveillance radar demonstrator

The first section presents a description of the radar demonstrator and its principle of operation for Space Surveillance. Results obtained during the different campaigns of measurements are presented.

The second section is relative to the experiments which have been done during ExoMars launch in March 2016. A kinematic analysis of the radar parameters has been realized from the measurements of detection of this particular target.

2. Radar demonstrator

The European Space Surveillance radar breadboard is a bistatic system that is located on two separate sites in the northern part of France around Paris as shown in Figure 1.

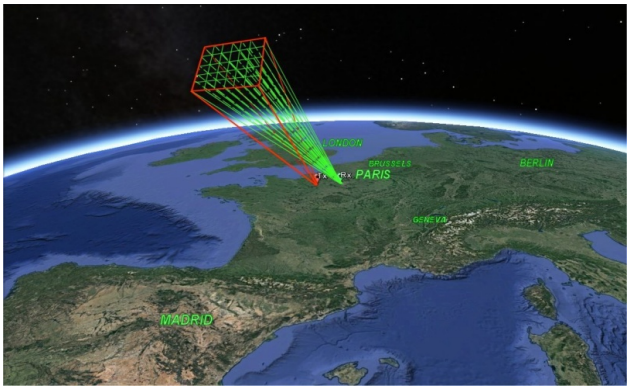


Fig.1. Bistatic radar configuration (*Tx*: transmitting site in red, *Rx*: receiving site in green)

This radar operates in L-band for the detection of space objects over a steerable Field of Regard (FoR) defined as 30° in azimuth by 25° in elevation oriented preferably to the South.

One of the advantages of such a system is its operation with a continuous wave (CW) signal, which is easier for spectral management as it is strictly limited to the carrier frequency.

*Stéphane Saillant, Email: stephane.saillant@onera.fr

The breadboard has been designed on the basis of simulations of performances, as a radar which should have a reference detection range at 500 km for a reference target of 0 dBm² as Radar Cross Section (RCS) for a revisit time of 10 seconds, meaning that the beam of the radar should revisit the same azimuth and elevation position within 10 seconds.

These are the reference values that have been used in the power budget computation for the design of the system. The principle of the detection for such radar is only based on Doppler measurement; there is no evaluation of the distance of the target.

The transmitting system (*Tx*, presented in figure 2 and figure 3) is constituted of one antenna array of surface area 1.5 m² with 49 radiating elements, and 49 CW solid state power amplifiers, controlled in both phase and gain, associated to each radiating element.

All electronics equipments for transmitting and control are in a shelter. The antenna array is set up under a radome on the roof oriented to the South. The positioner which hosts the antennas array can be steered in +/-90° in azimuth around a central position oriented in South direction, and 0° to 90° in elevation.



Fig.2. The transmitting system the shelters with the antenna on the roof under a radome)



Fig.3. Transmitting system (the shelters with the antenna on the roof without the radome)

The receiving site (*Rx*) is within ONERA centre in Palaiseau. The antenna panel, with a surface area of 10.5 m² and 64 receiving antennas, is installed on the roof of an office building as shown on Figure 4.

The antennas array can be steered in +/-90° in azimuth around a central position oriented in South direction, and it can be inclined from 0° to 90° in elevation. The receiving system is constituted of 64 analog front-edge stages in one cabinet, 4 digital receivers of 16 channels and a cluster of 10 PC servers to process the data.

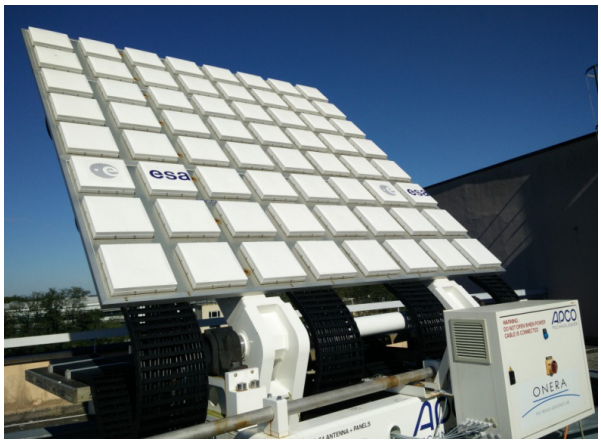


Fig.4. The receiving antenna panel in Palaiseau (ONERA centre)

In order to achieve the simulated detection performances, the whole Field of Regard of 30° in azimuth by 25° in elevation as shown in Fig.5, should be revisited every 10 seconds.

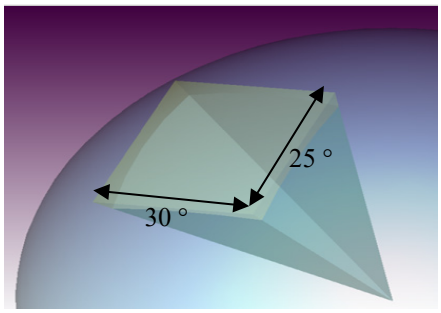


Fig.5. The Field of Regard (FoR) of the radar

According to simulations taking into account the geographical constraints and the maximization of objects detection, the difference of pointing between the two beams oriented South is few degrees between the two boresight directions.

The aperture of the transmitting beam plotted on Figure 6 is 11° azimuth by 11° elevation. Thus, the entire Field of Regard is covered by 10 scanning positions of this beam as presented on Fig.7.

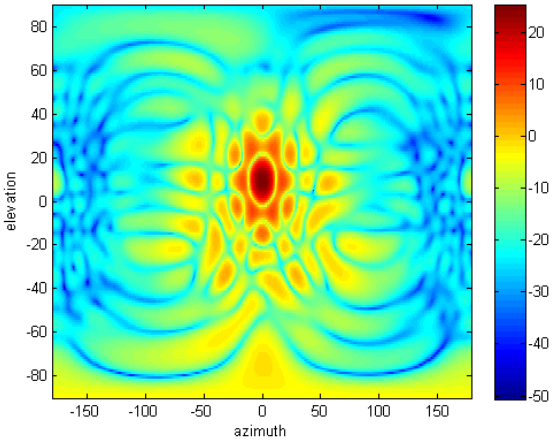


Fig.6. Pattern of the transmitting beam formed with transmitting array (11° aperture in azimuth and elevation).

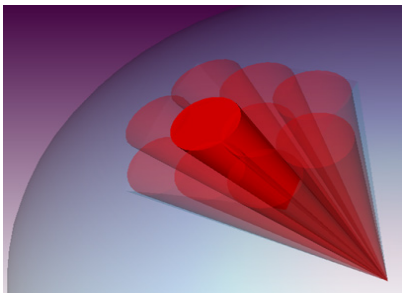


Fig.7. The 10 beams positions necessary to cover the Field of Regard

The Field of Regard is completely covered with an average gain greater than 22 dB as presented on Fig.8.

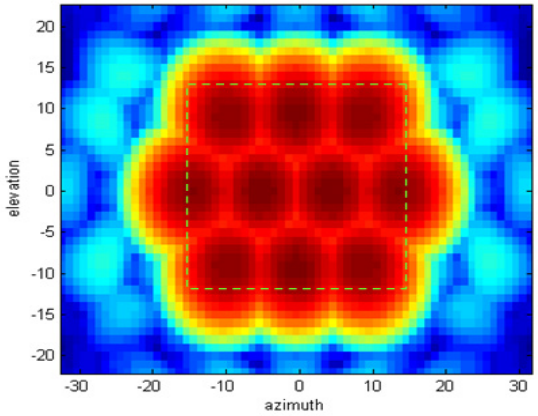


Fig.8. Optimization of the coverage of the Field of Regard for detection requirement

An evaluation of the number of receiving beams to form to cover each transmitting beam is done with an overlapping factor of 0.8 with a beam width of 4° (the aperture of receiving beam). The Figure 9 shows the results of the evaluation of receiving beams to form for extreme positions of the transmitting beam in the four corners and in the middle of the Field of Regard.

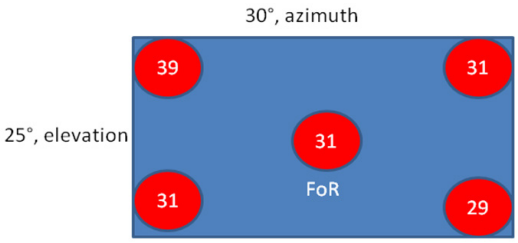


Fig.9. Number of *Rx* beams required in different locations of transmitting beam (Red areas) in the Field of Regard (Blue area: 30° azimuth x 25° elevation).

The most critical position which requires most of receiving beams to compute is when transmitting beam is focalized in the position corresponding to the upper left corner of the Field of Regard. Maximum number of beams to form for this steering direction of the transmitting beam is around 39, according to the following result represented on Figure 10 with a drawing which takes into account the distortion of the illuminated area.

The full receiving antenna array (including coupling and losses of materials) has been simulated. The global gain in the central boresight of the panel is 31.8 dB as shown in the 3 dimension display of the Figure 11.

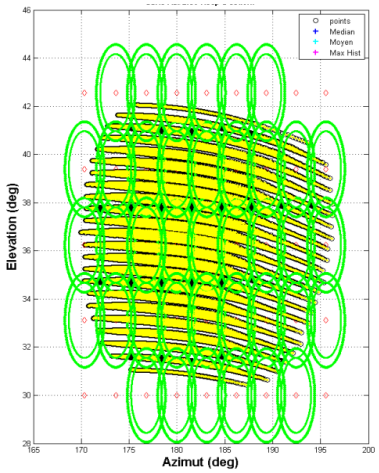


Fig.10. Transmitting beam deformed (yellow) and viewed from receiving site with the coverage of associated receiving beams (green)digitally formed

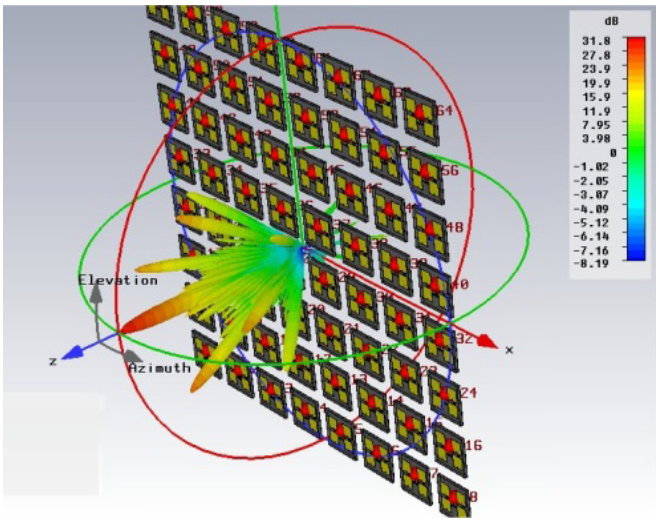


Fig.11. 3D simulation of the full receiving antenna panel (normal direction pointing)

The average gain has been evaluated over the whole Field of Regard of the radar. It is greater than 29.9 dB by computing all the beams formed for each of the 10 positions of the transmitting beam.

The signal processing implemented in the radar consists in Doppler/acceleration measurements (*i.e.*, speed estimation), detection and monopulse compression for precise direction finding. As the Field of Regard is covered by 10 positions of the transmitting beam in 10 seconds, one second of signal is analyzed at each position. Consequently, the coherent integration time is 1 second.

Thus, detections of objects that pass through the Field of Regard are characterized by Doppler traces versus time. The passage of one space object is signaled by a series of closely spaced radar plots as it can be observed on Fig.12.

Each extracted plot contains all measurements parameters, namely azimuth, elevation, Doppler (speed), radial acceleration and time of detection.

Tracking of a previously detected target is realized with consecutive plots using the Track-While-Scan (TWS) method, which combines both search and track functions. The orbital

trajectory can be built from the track from latitude and longitude positions of the transmitting site and receiving site of the radar.

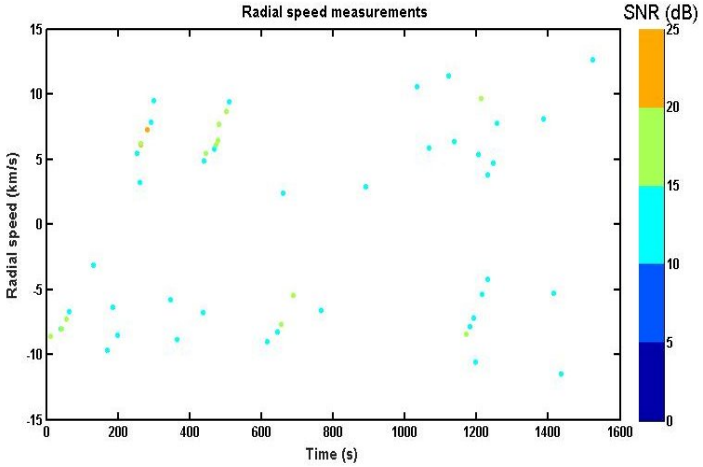


Fig.12. Doppler traces (speed) versus time (many traces of detected space objects are characterized by alignments of closed detected plots at 50 s, 250 s, 430 s, 620 s and 1200 s)

The breadboard radar has been regularly operating for more than one and half year.

Around 700 space objects have been detected and their associated tracks have been generated. The tracked objects are plotted as red dots over the Space track catalogue display on the Figure 13. More than 60% of detections are related to targets at least 3 meters in diameter at ranges less than 700 km. Objects tracked over 800 km to 1000 km are generally bigger (more or less 5 meters in diameter).

More than 40% of detections are identified as rocket debris.

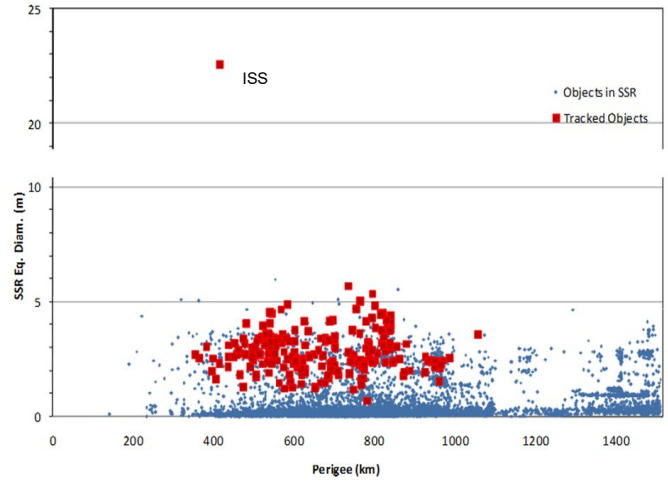


Fig.13. Estimated size of the detected objects with ABISS radar (Red) - Equivalent diameter (m) versus perigee (km) given by Space Track catalogue (Blue dots)

3. Operation of ABISS radar during ExoMars mission

The first ExoMars mission started with the launch of the spacecraft in collaboration with Russian agency Roscosmos. The launch is operated with a Proton rocket from Baikonur cosmodrome, Russian launch center.

Before separating from ExoMars, the Proton launcher vehicle needs three steps to place the spacecraft on its required Earth escape trajectory as shown on Figure 14:

- One first complete revolution around the Earth on a low parking orbit after the lift-off concluded by a boost to change its altitude,
- A second complete revolution on an intermediate eccentric orbit with a new burn to place the launcher on its transfer orbit just above North-West of Spain (in term of ground track),
- A third quasi-complete revolution with a final maneuver to eject the spacecraft on the way to Mars, above Central Africa.

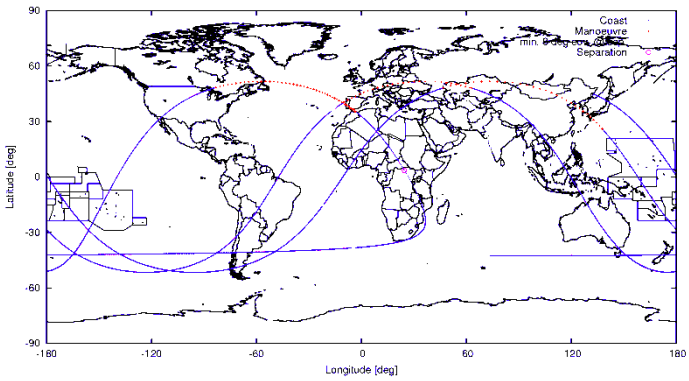


Fig.14. Estimate trajectory of the launcher (Maneuver sections in red, separation from ExoMars above Africa at magenta dot) (Sources: blogs.esa.int, http://www.russianspaceweb.com)

The maneuver to move from intermediate eccentric orbit to its transfer orbit during the second phase of the flight is exactly realized in the area where the Field of Regard of ABISS radar can be steered. The radar was set up to steer its Field of Regard so as to cover the area where this maneuver would take place as it can be seen on Figure 15.

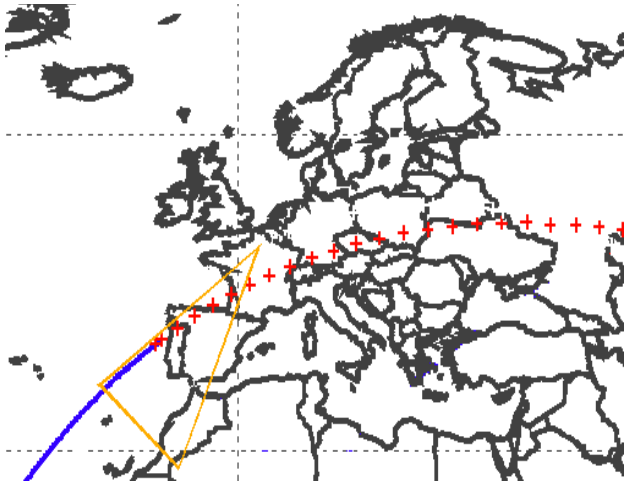


Fig.15. Trajectory of the launcher in the Field of Regard of ABISS radar (orange) during the second passage

The spacecraft was launched on March14th 2016 at 10:31 CET. The burn for this change of orbit was operated around 5 hours after lift-off.

The last stage of the launcher, with the ExoMars spacecraft, was detected by the radar approximately at the expected time when it passed through the Field of Regard [3]. At the beginning of the detection the object was just coming off the perigee of its second revolution and a new burn was operated to go on its transfer orbit.

Few plots were directly extracted from the signal processing but a corrected trajectory was calculated after a correlation with a refined trajectory. As the ExoMars launcher is a specific object in terms of kinematic behavior, we decided to have a detailed analysis of radar parameters concerning this detection [1].

Other space objects were detected during the measurements campaign for ExoMars observation. One of them is particularly useful because it appeared almost at the same time as the passage of the launcher as it can be observed on Fig.16. This space object has been identified by data processing as the rocket-debris named SL-14 R/B with a well-known actualized TLE (Two lines Elements).

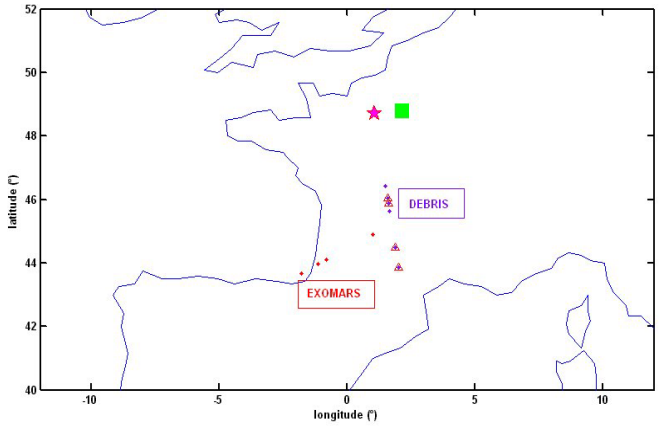


Fig.16. ExoMars launcher and rocket-debris SL-14 R/B detected simultaneously – 14th march 2016 – 10:31 CET – Triangles represent the confirmed positions of detected plots (dots)

Rocket-debris SL-14 R/B, which was detected just before ExoMars passage, is used as a calibration source for speed and acceleration measurements. The visualization of its trajectory in radar axis on Figure 17 shows the good correlation between the plots positions and the TLE of the object.

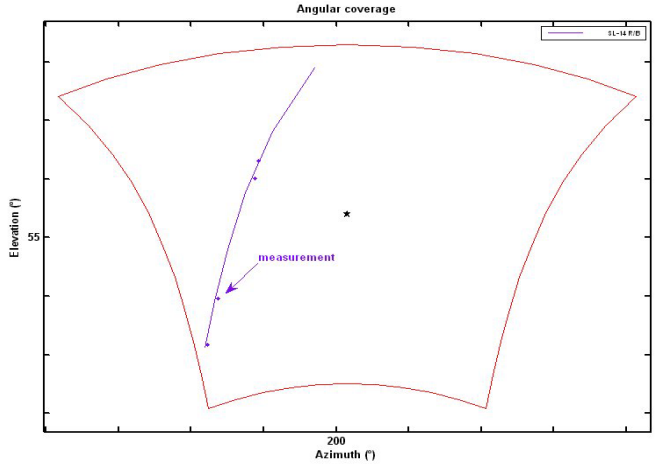


Fig.17. Rocket-debris SL-14 R/B passage in radar axis in UTC reference

As the passage of this object through the Field of Regard was confirmed by the successful tracking and approved by the identification process, the calibration of the radar was realized by evaluating the shift observed on radial speed ΔV (as shown in Fig.18) and radial acceleration parameters$\Delta\gamma$ (as shown in Fig.19) for each detected plot compared to TLE.

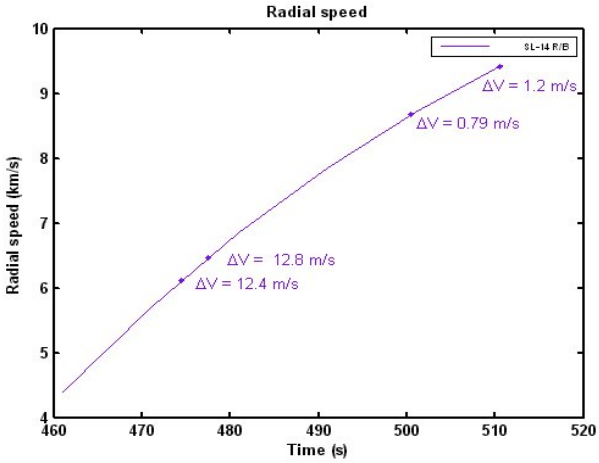


Fig.18. Rocket-debris SL-14 R/B radial speed and error margins

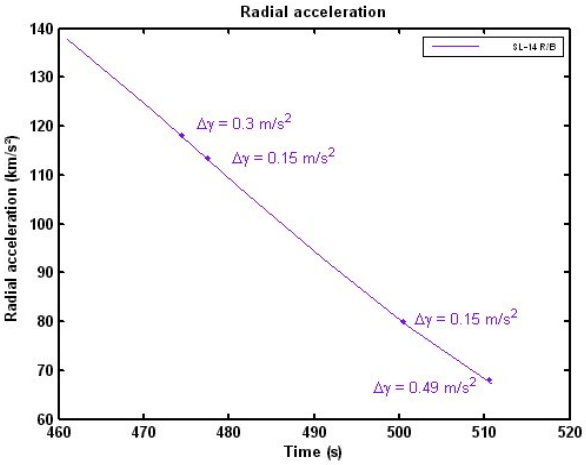


Fig.19. Rocket-debris SL-14 R/B radial acceleration and error margins

Measurements are in accordance with the parameters of movement of the target, no major bias can be observed in comparison with the values deduced from the TLE. The good correlation of acceleration versus speed confirms the synchronization of detections as presented in Fig.20. This could be used to estimate the time shift to apply for recalibration of ExoMars detections. Indeed, biases are expected to be observed along the trajectory of ExoMars because its orbit is not keplerian especially during the thrust phase for the change of altitude of the spacecraft.

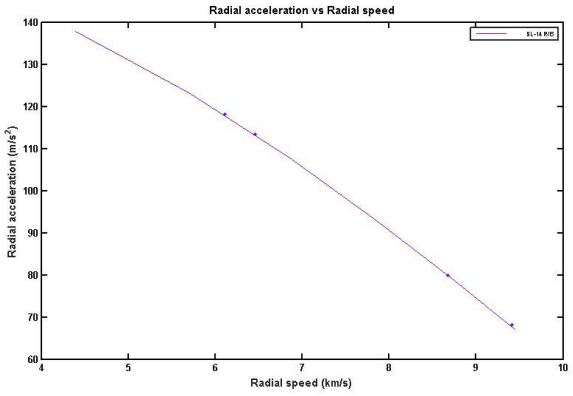


Fig.20. Radial acceleration versus radial speed of SL-14 R/B

Four detected plots of the ExoMars launcher appeared during the time window of the passage through the Field of Regard as spotted on Figure 21.

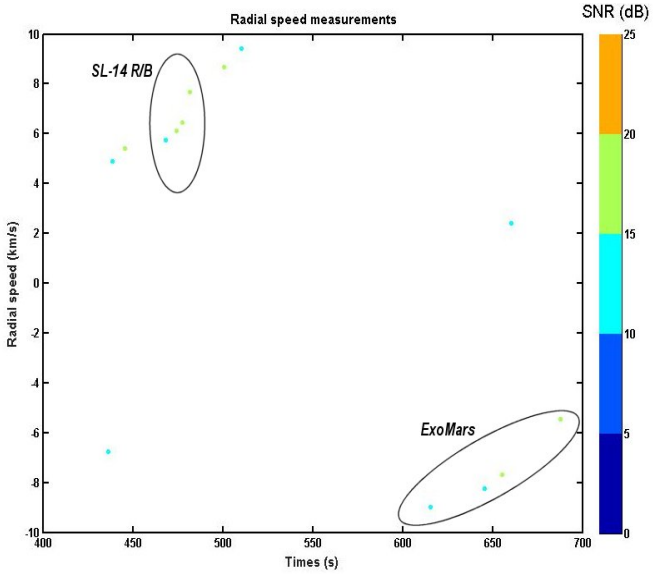


Fig.21. Detected plots around the time of ExoMars passage in the Field of Regard - Radial speed versus time (The two targets of interest have been circled - 4 plots characterize the passage of ExoMars)

The synchronization of these detections was obtained by shifting the reference time by +6 seconds (UTC + 6 s). Only two of them were matched to the estimated trajectory, especially for the beginning of the passage when the object is still in an eccentric orbit at its perigee around 600 km as shown with an estimated trajectory of classical orbit on Figure 22.

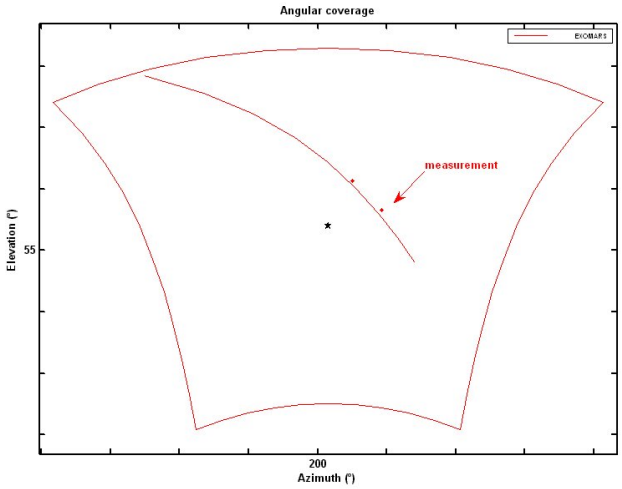


Fig.22. ExoMars launcher passage in the radar axis in UTC + 6 sec

The other two plots are not visible due to the thrust phase of the launcher (change from the second orbit to the third). The figure of radial acceleration ($\Delta\gamma$) versus radial speed (ΔV) for the two matched plots shows an important shift between measurements and predicted parameters from the TLE as observed on Figure 23.

These two biases measured on the parameters of radial speed and radial acceleration, are presented respectively on Figure 24 and Fig.25.

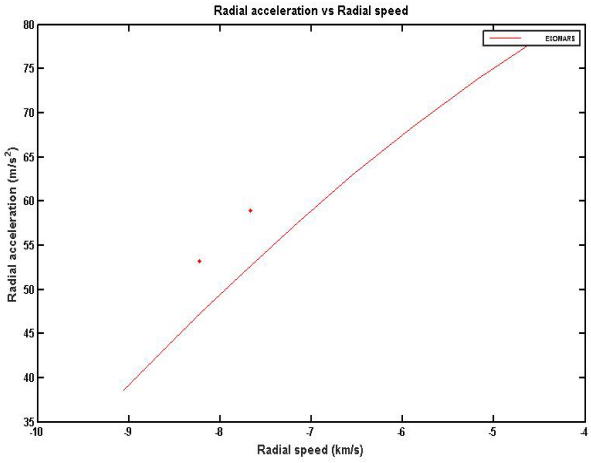


Fig.23. Radial acceleration ($\Delta\gamma$) versus radial speed (ΔV) of ExoMars

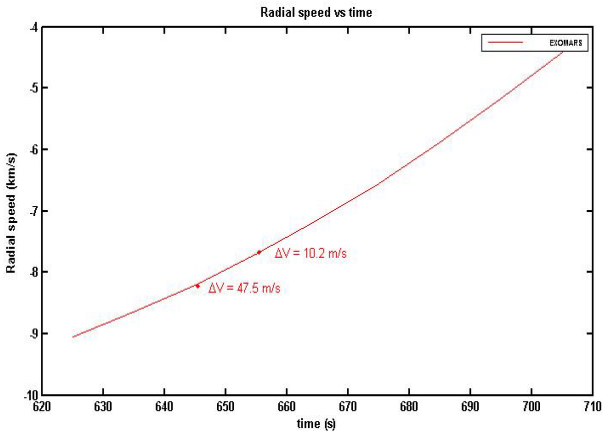


Fig.24. Radial speed of ExoMars for the first 2 detected plots

It can be noticed that the observation of the launcher was excellent during the thrust phase for orbit change, in accordance to the predicted trajectory. As shown on Fig.25, the two last plots of detection of the launcher are outside the limits of the coherence domain with a classical orbit consideration. This is due to the increase of the acceleration for the thrust.).

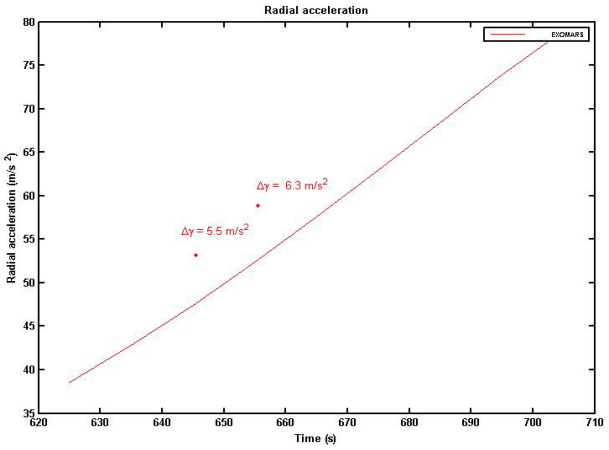


Fig.25. Radial acceleration of ExoMars for the last 2 detected plots

4. Conclusion

The challenge was to detect the ExoMars spacecraft still attached to the launcher during its passage of over South Europe with the demonstrator of Space surveillance radar ABISS. The results obtained during this campaign of experiments have shown that such a system has the capacity to detect as well as to match a specific target in the radar area of surveillance during its thrust phase as long as it stays within the domain of coherence of a classical orbit consideration in terms of kinematic.

Conflict of Interest

The authors declare no conflict of interest.

References

[1.] Saillant, S., " Kinematic analysis of the ExoMars launcher change of orbit as detected during its passage over Europe ", Proceedings of IEEE Radar Conference (RadarConf17), Seattle (2017).

[2.] Saillant, S., " Bistatic L-Band Radar To Monitor Space", Proceedings of International Radar Conference, Lille (2014).

[3.] Saillant, S., " ExoMars spacecraft detection with European Space Surveillance bistatic radar", Proceedings of CIE International Conference On Radar, Guangzhou (2016).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 103-108 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Statistical Approach for Gain Bandwidth Prediction of Phoenix-Cell Based Reflect arrays

Hassan Salti*, Raphael Gillard

Institute of Electronics and Telecommunication of Rennes, INSA de Rennes CS 70839 – 35708 Rennes Cedex 7, France

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
Reflectarray
Bandwidth
Prediction

ABSTRACT

A new statistical approach to predict the gain bandwidth of Phoenix-cell based reflectarrays is proposed. It combines the effects of both main factors that limit the bandwidth of reflectarrays: spatial phase delays and intrinsic bandwidth of radiating cells. As an illustration, the proposed approach is successfully applied to two reflectarrays based on new Phoenix cells.

1. Introduction

This paper is an extension of work originally presented in the 11th European Conference on Antennas and Propagation (EuCAP) where a novel single layer stub-patch Phoenix cell is suggested as a broadband and easy solution to fabricate reflectarray (RA) elements [1].

The Phoenix cell concept was firstly introduced in [2] and many other Phoenix cells were derived later [3-8]. As their name suggests, Phoenix cells are characterized by rebirth capabilities, which means that their geometry comes back to its initial state after a complete 360° phase cycle. This guarantees a smooth evolution of cell geometries over the RA panel and prevents perturbations in the radiation pattern. Phoenix cells are also characterized by a quasi-linear phase response which classifies them as broadband RA cells. Nevertheless, as for other broadband cells, their phase response is still not totally perfect and the proper assessment of the residual phase error versus frequency is still missing in the literature.

In this paper, this last issue is addressed and a new statistical approach for estimating the bandwidth of RA based on Phoenix cells is proposed. The suggested approach relies on the standard deviation of phase errors over the RA panel and combines the effects of both bandwidth limiting factors: the dispersion of spatial phase delays with frequency [9] and the intrinsic limited bandwidth of cells themselves [10]. This standard deviation is shown to provide a promising figure of merit, better than those in [9, 10] where only the maximum phase error due to spatial delays is taken into account.

The paper is organized as follows. In section 2, the new statistical approach is defined. In sections 3 and 4, a bandwidth estimator for RA gain is derived. Finally, in section 5, it is validated by comparison with the simulated bandwidth of a test-case RA based on two different Phoenix cell topologies.

2. Statistical Bandwidth Estimator

Consider an *N*-cell circular RA of diameter *D* with a feeding antenna positioned at distance *F* normally above the array center. The phase of the wave radiated by cell *i* at central frequency f_0 is defined as:

$$\phi_i^{Rad}(f_0) = \phi_i^{Inc}(f_0) + \phi_i^{Sh}(f_0) \quad (1)$$

where $\emptyset_i^{Sh}$ is the phase-shift produced by the cell and $\emptyset_i^{Inc}$ is the phase of the incident wave defined as:

$$\phi_i^{Inc}(f_0) = -2\pi f_0 F\sqrt{1+(\rho_i/F)^2}/c \quad (2)$$

In these equations, ρ_i is the radial distance between the center of the array and cell *i* and *c* is the velocity of light in vacuum.

In order to produce a desired radiation pattern, the required radiated phases $\emptyset_i^{Rad}$ are usually specified at f_0 and the subsequent

*Hassan Salti, Department of Electrical Engineering, Australian College of Kuwait, P.O. Box 1411, Safat 13015, Kuwait, Email: h.salti@ack.edu.kw

phase-shifts $\emptyset_i^{Sh}(f_0)$ are directly deduced from (1) and (2). The cells are selected accordingly and appropriately distributed over the RA.

When the frequency is shifted to $f=f_0+\Delta f$, the phase radiated by cell i is changed by $\Delta\emptyset_i^{Tot}$ as:

$$\Delta\phi_i^{Tot}(f) = \Delta\phi_i^{Inc}(f) + \Delta\phi_i^{Sh}(f) \quad (3)$$

where $\Delta\emptyset_i^{Inc}$ is the predictable deviation of the phase of the incident wave defined as:

$$\Delta\phi_i^{Inc}(f) = -2\pi\Delta f F\sqrt{1+(\rho_i/F)^2}/c \quad (4)$$

while $\Delta\emptyset_i^{Sh}$ is the cell-dependent phase deviation due to the cell's dispersive phase response. At the RA level, these phase errors are responsible for a decrease of the gain at f, and thus for the limited bandwidth.

The bandwidth estimator we propose is derived from the standard deviation of the total phase error. Let $\Delta\emptyset^{Tot}(f)$, $\Delta\emptyset^{Inc}(f)$, and $\Delta\emptyset^{Sh}(f)$ be the statistical variables related to the total, incident and phase-shift errors at f respectively. $\sigma^{Tot}(f)$, $\sigma^{Inc}(f)$ and $\sigma^{Sh}(f)$ are the respective standard deviations. Using (3), the standard deviation of the total error $\sigma^{Tot}(f)$ can be expressed as:

$$\sigma^{Tot}(f) = \sqrt{[\sigma^{Inc}(f)]^2 + [\sigma^{Sh}(f)]^2 + 2\,\mathrm{cov}[\Delta\phi^{Inc}(f), \Delta\phi^{Sh}(f)]} \quad (5)$$

Assuming $\Delta\emptyset^{Inc}(f)$ and $\Delta\emptyset^{Sh}(f)$ are uncorrelated, which is the case when the synthesis process is done at f_0 only, as usually applied in the literature [11-15], the covariance term reduces to zero and $\sigma^{Tot}(f)$ reduces consequently to:

$$\sigma^{Tot}(f) = \sqrt{[\sigma^{Inc}(f)]^2 + [\sigma^{Sh}(f)]^2} \quad (6)$$

3. Bandwidth Estimator: Incident Phase Errors

3.1. Standard Deviation of Incident Phase Errors

Defining:

$$S_i = \sqrt{1+(\rho_i/F)^2} \quad (7)$$

(4) can be reformulated as:

$$\Delta\phi_i^{Inc}(f) = -2\pi\Delta f F S_i/c \quad (8)$$

Due to the mathematical properties of standard deviation, $\sigma^{Inc}(f)$ can be derived as:

$$\sigma^{Inc}(f) = 2\pi|\Delta f|F\sigma_S/c \quad (9)$$

where: $$\sigma_S = \sqrt{E[S^2] - (E[S])^2} \quad (10)$$

E[S] and *E[S²]* are the first statistical moments and can be calculated as:

$$\begin{aligned} E[S] &= (2/D)\int_0^{D/2}\sqrt{1+(\rho/F)^2}\,d\rho \\ &= 0.5\sqrt{1+(D/2F)^2} \\ &\quad + (F/D)\ln[(D/2F)+\sqrt{1+(D/2F)^2}] \end{aligned} \quad (11)$$

$$\begin{aligned} E[S^2] &= (2/D)\int_0^{D/2}[1+(\rho/F)^2]\,d\rho \\ &= 1+0.33(D/2F)^2 \end{aligned} \quad (12)$$

Note that σ_S does not depend on frequency but only on the dimensions of the RA, which is consistent with other criteria in the literature [10]. Note also that, in (11) and (12), a rectangular lattice is considered and ρ_i is consequently supposed to vary uniformly in the [0, D/2] range. In addition, though ρ_i is normally a discrete variable, it is assumed here to vary continuously. This assumption is reasonable since the inter-element spacing is usually a small fraction of λ_0 (which is much lower than D/2). Furthermore, as in [9] and [10] and for simplicity reasons, S_i and the resulting σ_S are calculated for a centered fed RA. Different expressions could easily be established for offset configurations.

3.2. Bandwidth Estimator

We now investigate how the gain deteriorates with respect to σ^{Inc}. To do so, we consider the gain at broadside for a test-case RA with 12mm spacing (i.e. $0.5\lambda_0$ at the center of the [11-14] GHz frequency band). Edge tapering is supposed to be -12dB and different antenna configurations are considered: F/D=0.6 and 0.8 with D varying from 0.28m≈ $12\lambda_0$ to 1m ≈ $42\lambda_0$. For each couple (F/D; D), the reflected field (phase and magnitude) in the aperture is calculated and the associated gain is derived using simple array theory (as in [16]). Simultaneously, $\sigma^{Inc}(f)$ is also computed as the standard deviation of all phase errors. Finally, Figure 1 gives the representation of the normalized simulated gain $G(f)/G(f_0)$ versus the corresponding standard deviation σ^{Inc}.

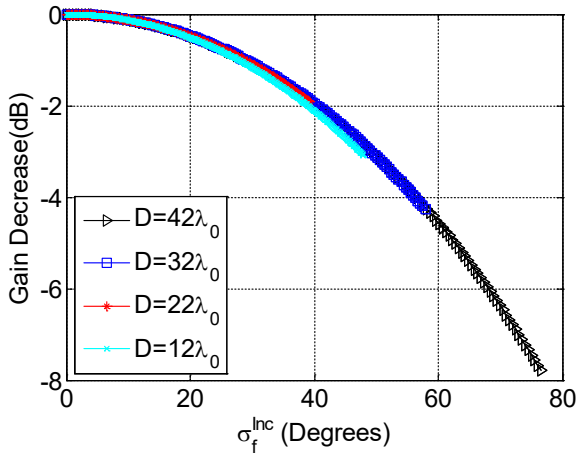


Figure 1. Simulated gain decrease with respect to the standard deviation of incident phase errors (F/D=0.8 or F/D=0.6).

An important conclusion from Figure 1 is that all curves superimpose whatever the particular values of *D, F* or *f*. This demonstrates that σ^{Inc} is a reliable estimator for the bandwidth since it directly reflects the gain decrease. Figure 1 also shows that a 1dB gain-drop approximately corresponds to σ^{Inc}=π/6. As a consequence, the upper frequency f_{max} of the -1dB bandwidth may be derived simply by replacing $\sigma^{Inc}(f)$ by π/6 in (9), leading to:

$$B_{Inc} = 2(f_{\max} - f_0)/f_0 = c/(6Ff_0\sigma_S) \qquad (13)$$

where

$$f_{\max} = f_0 + c/(12F\sigma_S) \qquad (14)$$

Equation (13) will thus be used as a bandwidth estimator. At this stage, it does not depend on any particular cell topology but only on the spatial phase delay error. As will be seen now, this initial estimator can advantageously be replaced by a more sophisticated one that also accounts for the phase dispersion of the used RA cells. In what follows, the case of cells with an ideal phase response is considered.

4. Bandwidth Estimator: Total Phase Errors

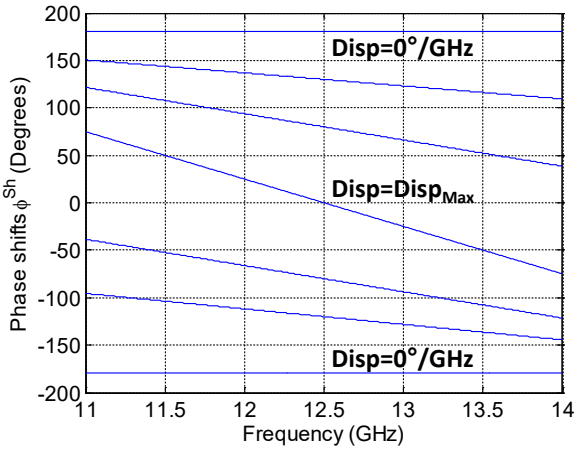


Figure 2. Linear phase response of an ideal Phoenix cell.

The ideal cell we suggest here is quite representative of Phoenix cells as will be shown in section 5. It is supposed to provide a phase range of at least 360° at f_0 and perfect linear variations with respect to frequency. Figure 2 shows the phase response versus frequency of the ideal cell. As such, the phase-shift error for cell *i* at frequency *f* can be expressed as:

$$\Delta\phi_i^{Sh}(f) = \phi_i^{Sh}(f) - \phi_i^{Sh}(f_0) = Disp_i.\Delta f \qquad (15)$$

where $Disp_i$ is the dispersion that is supposed to vary uniformly in the interval [0-$Disp_{max}$]. More specifically, $Disp_i$ is considered to be equal to 0°/GHz for phase-shifts $\emptyset_i^{Sh}(f_0)$= ±180° and to reach its maximum (i.e. $Disp_{max}$) when $\emptyset_i^{Sh}(f_0)$= 0°. Note that the perfect Phoenix cell is obtained when $Disp_{max}$= 0°/GHz as all its phase states would be perfectly parallel.

Assuming all phase-shifts are equally probable on the radiating aperture, $\sigma^{Sh}(f)$ can be expressed as:

$$\sigma_f^{Sh} = |\Delta f| Disp_{\max}/\sqrt{12} \qquad (16)$$

Using (9) and (16) in (6), a generalized bandwidth estimator $\sigma^{Tot}(f)$ accounting for both types of errors is derived:

$$\sigma_f^{Tot} = |\Delta f|\sqrt{(2\pi F\sigma_S/c)^2 + (Disp_{\max}/\sqrt{12})^2} \qquad (17)$$

Finally, $\sigma^{Tot}(f)$ is set to π/6 as already done in Part 3.2 to obtain the generalized bandwidth estimator:

$$B_{Tot} = B_{Inc}/\sqrt{1 + 0.75(Disp_{\max}.f_0B_{Inc}/180°)^2} \qquad (18)$$

To validate (18), the gain decrease simulation described in Section 3.2 is repeated (as in [16]), now accounting for both types of phase errors. The associated simulated -1dB gain bandwidth is then extracted and compared to the theoretical value predicted from (18). In this study, *F/D* is set to 0.8, *D* varies from 0.16m to 1m (i.e. ~7λ_0 to 42λ_0) and $Disp_{max}$ varies from 0°/GHz to 100°/GHz. Figure 3 shows that the difference between simulation and theory is less than 3%, even for the highest dispersion and the smallest diameter values.

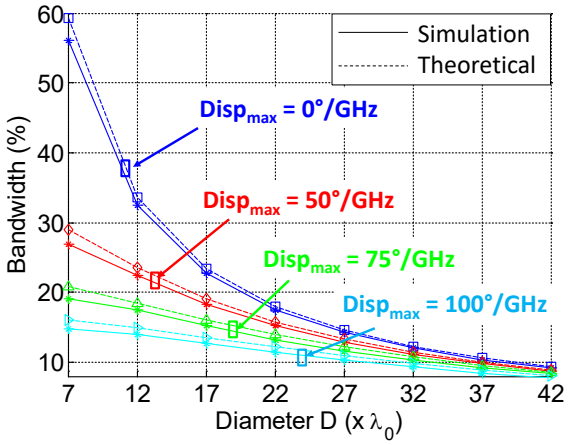


Figure 3. Total phase errors effects: simulated and theoretical bandwidths (F/D=0.8).

As a conclusion, (18) appears to be a reliable bandwidth estimator. In practice, it can be used to define the maximum acceptable cell dispersion for a given application. As an example, for a RA with *D*=22λ_0 and *F/D* = 0.8, the cell dispersion should be less than 50°/GHz to ensure a 15% bandwidth.

5. Practical Validation

In practice, the phase response of Phoenix cells is not purely linear as in the previously-used ideal cell model. Therefore, to assess the validity of our approach, the actual performance of two recently-proposed Phoenix cells is assessed and compared to those obtained using (18).

The two cells are designed to operate around a central frequency f_0 = 12.5GHz, with $\lambda_0/2$ spacing at f_0. Both cells are printed on a Duroïd substrate with 2.17 dielectric constant and backed by a ground plane. The substrate height is fixed to 4 mm.

This corresponds to approximately $\lambda_0/(4\sqrt{\varepsilon_r})$, which means the reflected phase is close to 0° if the cell is transparent.

To extract the phase responses, both cells are simulated using ANSYS-HFSS software assuming normal incidence and local periodicity. As Phoenix cells allow for smooth evolution of cell geometries over the RA panel, it is assumed that the phase responses obtained by simulation are valid for finite reflectarray configurations [2].

5.1. Cell 1: Slot – Patch Phoenix Cell

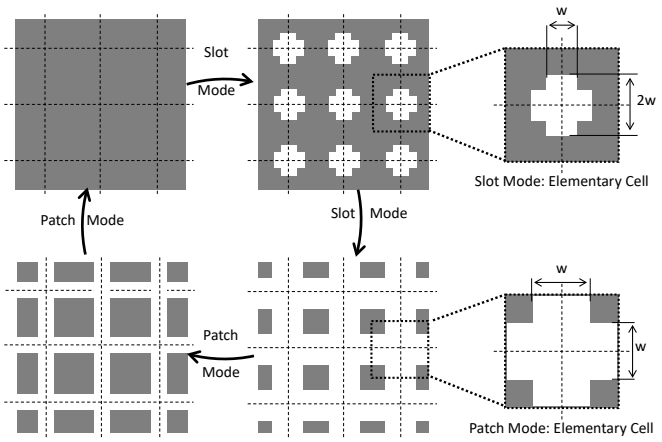


Figure 4. Slot-Patch Phoenix Cell: Rebirth cycle

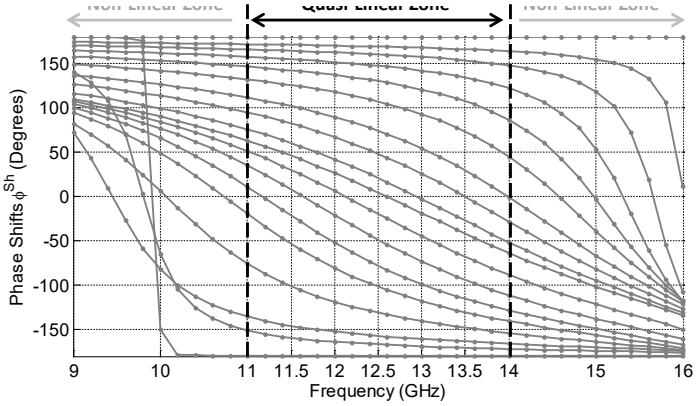


Figure 5. Slot-Patch Phoenix Cell: Phase response.

The first cell, the Slot – Patch Phoenix cell was initially introduced in [8] and its performance was improved in [7]. The cell cycle is illustrated in Figure 4. The initial cell consists of a ground plane providing a 180° phase shift, whatever the frequency. The first mode of operation, or slot mode, is obtained by opening a crossed-shaped slot with variable length and width in the ground plane. For simplicity reasons, the length of the cross is fixed as twice its width. The slot mode ends when the slot arms reach the borders of the cell, thus defining square patches. The operating mode then switches to a patch mode. In this second mode, the length of the pre-opened slot is frozen and only the width of the slot w is decreased. The patch mode ends when the slot vanishes, taking back the cell to its initial geometry and opening the door for a new cycle.

The obtained phase response is presented in Figure 5. At the central frequency, the slot mode provides phase shifts between 0° and 180° while the patch mode completes the remaining phase range between 0° and -180°. The use of complementary modes provides a phase response that is quasi-linear within a 24% bandwidth around 12.5GHz. The maximum dispersion is 53°/GHz. This phase response fits well with the ideal model used to derive (18), although the linearity is not perfect.

5.2. Cell 2: Stub – Patch Phoenix Cell

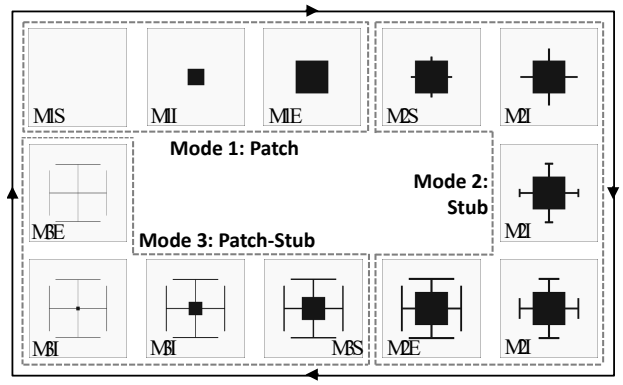


Figure 6. Patch-Stub Phoenix cell: Rebirth cycle (MαS: Mode α's Start; MαI: Mode α's Intermediate states; MαE: Mode α's End).

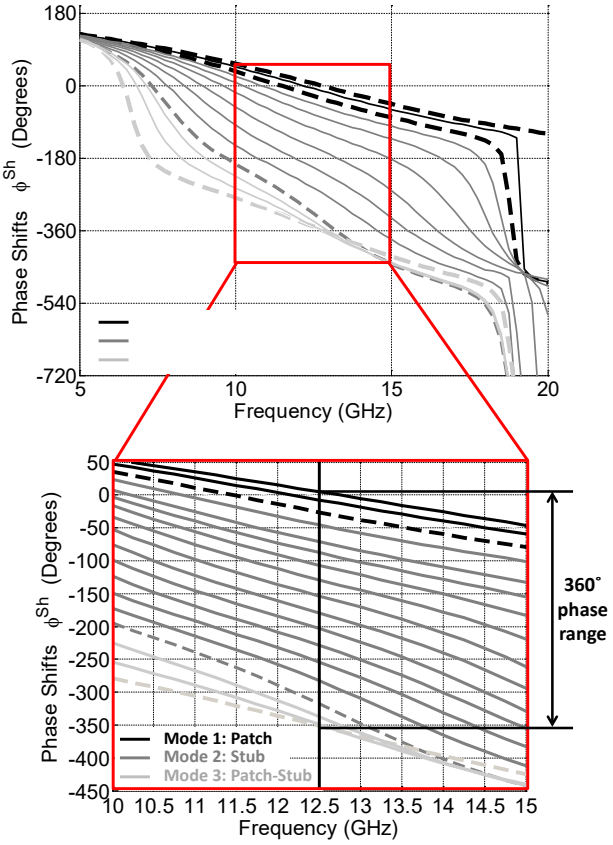


Figure 7. Patch-Stub Phoenix cell: Phase response.

The second cell, the Stub-Patch Phoenix cell was introduced recently in [1]. It improves the bandwidth further due to the three possible operating modes it offers: a patch mode, a stub mode, and a combined patch-stub mode.

As illustrated in Figure 6, the cycle starts with a pure Duroïd substrate backed by a ground plane providing an initial phase shift close to 0° at f_0. The first mode of operation, namely the patch mode, is obtained by inserting a square patch at the center of the cell (cf. Figure 6 – Mode 1). The phase shift produced by the cell is controlled by increasing the patch size until it reaches a maximum predefined value.

In mode 2, namely the stub mode, the patch size is frozen to this maximum value and four open-circuited stubs are grown perpendicularly to the patch from the center of its edges (cf. Figure 6 – Mode 2). A T-shape is used for the stubs when the total metal length approaches the limit allowable by the inter-element spacing. In this mode, the phase shift is thus controlled by the length of the stubs.

The mode then switches to mode 3, namely the patch-stub mode, during which the stub-loaded patch shrinks gradually until both patch and stubs disappear completely allowing the cell to rebirth and to start a new cycle (cf. Figure 6 – Mode 3). In this mode, the phase shift is controlled by the shrinking ratio.

The phase response of the suggested cell in all modes is summarized in Figure 7. Dashed curves represent the start/end of a mode (i.e. MαS/MαE) and continuous lines represent intermediate states (i.e. MαI). As can be noticed, a phase range of 360˚ is achieved at f_0. Within a frequency band ranging from 10 to 15GHz (40%), the phase response is almost linear. The maximum phase dispersion is obtained at the transition from mode 2 to mode 3 and is equal to 55˚/GHz. The minimum dispersion is obtained at the beginning of mode 1 and is equal to 21˚/GHz. Compared to cell 1, cell 2 exhibits a better linearity of phase response. On the other hand, its minimum dispersion is not zero as required by the previously-used ideal cell model. However, this model still applies if we replace the maximum dispersion by the relative maximum dispersion, defined as the difference between the maximum and minimum dispersions. For cell 2, it is then equal to 34˚/GHz.

5.3. Performance Evaluation and Comparison

The previously-described Phoenix cells are now consecutively used as the radiating element in our test-case RA (F/D=0.8, variable D). The bandwidth is calculated by simulations as in [16] and compared to that given by (18). For this theoretical study, the maximum dispersion $Disp_{max}$ in (18) is set to 53°/GHz for cell 1 and 34°/GHz for cell 2.

Figure 8 summarizes all theoretical and simulation results. The results show a remarkable agreement between simulation and theoretical curves for a given cell. The slight discrepancy is mainly due to the linearity assumption in the ideal model which is not fully respected by realistic Phoenix cells. It is less than 5% for cell 1 and 3% for cell 2. As expected, a smaller error is obtained for cell 2 as it offers a better linearity of phase response.

Formula (18) is hence a reliable bandwidth estimator, even for realistic phase-shifting cells, provided that they are characterized by a quasi-linear response. As a consequence, it can be advantageously used to define the maximum dispersion allowed for a Phoenix cell to comply with given bandwidth specifications or to predict a Phoenix cell's performance in a RA configuration.

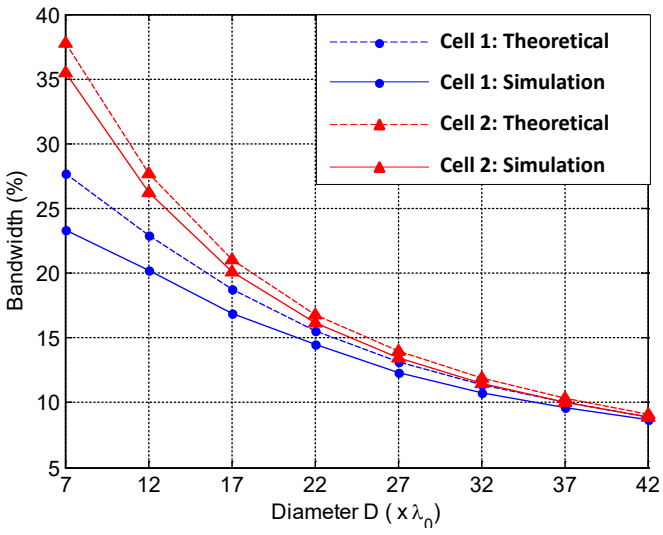


Figure 8. A Realistic study: simulated versus theoretical bandwidth of realistic Phoenix cells (F/D=0.8)

6. Conclusion

Bandwidth limitation of RA results from both the effect of various path delays between cells and source on one side, and intrinsic narrow bandwidth of cells themselves on the other side. The first phenomenon had been significantly investigated in the literature. For the second one, the usual solution relies on the use of broadband cells providing linear and parallel phase states. As this ideal characteristic is never met perfectly, this paper has defined a new approach to assess the effect of imperfection in the phase response of broadband RA cells.

We firstly proposed the standard deviation of phase errors over the array as an efficient criterion to predict RA's bandwidth when accounting for both its limiting phenomena. This criterion has then been formulated and validated for an ideal cell model with linear but non-parallel phase states. Finally, it has been successfully applied to realistic and novel Phoenix-cells. The suggested approach has thus been demonstrated as a powerful tool to help the designer in the selection of appropriate cells before entering the complex RA optimization process.

References

[1] H. Salti, R. Gillard, "A Single Layer Stub-Patch Phoenix Cell for Large Band Reflectarrays" in 11th European Conference on Antennas and Propagation, Paris, France, 2017. https://doi.org/10.1109/LAWP.2011.2108633

[2] L. Moustafa, R. Gillard, F. Peris, R. Loison, H. Legay, E. Girard, "The Phoenix Cell: A New Reflectarray Cell With Large Bandwidth and Rebirth Capabilities" IEEE Antennas and Wireless Propagation Letters, **10**, 71-74, 2011. https://doi.org/10.1109/LAWP.2011.2108633

[3] R. Deng, F. Xu, F. Yang, M. Li, "Single-Layer Dual-Band Reflectarray Antennas With Wide Frequency Ratios and High Aperture Efficiencies Using Phoenix Elements" IEEE Transactions on Antennas and Propagation, **65**(2), 612-622, 2017. https://doi.org/10.1109/TAP.2016.2639023

[4] R. Deng, S. Xu, F. Yang, M. Li, "Design of a Low-Cost Single-Layer X/Ku Dual-Band Metal-Only Reflectarray Antenna" IEEE Antennas and Wireless Propagation Letters, **16**, 2106 - 2109, 2017. https://doi.org/10.1109/LAWP.2017.2698099

[5] Q. Wang, Z. Hai Shao, Y. J. Cheng, P. K. Li, "Ka/W Dual-Band Reflectarray Antenna for Dual Linear Polarization" IEEE Antennas and Wireless Propagation Letters, **16**, 1301 - 1304, 2017. https://doi.org/10.1109/LAWP.2016.2633289

[6] R. Deng, F. Yang, S. Xu, and M. Li, "A low-cost metal-only reflectarray using modified slot-type Phoenix element with 360˚ phase coverage" IEEE Transactions on Antennas and Propagation, **64**(4), 1556-1560, 2016. https://doi.org/10.1109/TAP.2016.2526258

[7] H. Salti, R. Gillard, "Slot-Patch Cell with Low Phase Distortion for Large Band Reflectarrays" in Proc. of the Fifth International Conference on Digital Information and Communication Technology and its Applications, Beirut, Lebanon, 2015. https://doi.org/10.1109/DICTAP.2015.7113196

[8] T. Makdissy, R. Gillard, E. Fourn, E. Girard, H. Legay, "A patch-slot combination approach for large band reflectarrays" in IEEE European Microwave Conference, Amsterdam, Netherlands, 2012. https://doi.org/10.23919/EuMC.2012.6459392

[9] M. E. Bialkowski, K. H. Sayidmarie, "Bandwidth considerations for a microstrip reflectarray" Progress In Electromagnetics Research B, **3**, 173–187, 2008. http://dx.doi.org/10.2528/PIERB07120405

[10] D. M. Pozar, "Bandwidth of reflectarrays" Electronics Letters, **39**(21), 1490-1491, 2003. https://doi.org/10.1049/el:20030990

[11] D. Cadoret, A. Laisné, R. Gillard, and H. Legay, "A new reflectarray cell using microstrip patches loaded with slots" Microwave and Optical Technology Letters, **44**(3), 270-272, 2005. http://onlinelibrary.wiley.com/enhanced/exportCitation/doi/10.1002/mop.20608

[12] S. D. Targonski and D. M. Pozar, "Analysis and design of a microstrip Reflectarray using patches of variable size" in IEEE Symposium on Antennas and Propagation. Seattle, WA, USA, 1994. https://doi.org/10.1109/APS.1994.408184

[13] H. Salti, R. Gillard, R. Loison, and L. Le Coq, "A Reflectarray Antenna Based on Multiscale Phase-Shifting Cell Concept," IEEE Antennas and Wireless Propagation Letters, **8**, 363-366, 2009. https://doi.org/10.1109/LAWP.2008.2006073

[14] M. Bozzi, S. Germani, and L. Perregrini, "Performance comparison of different element shapes used in printed reflectarrays," IEEE Antennas and Wireless Propagation Letters, **2**, 219–222, 2003. https://doi.org/10.1109/LAWP.2003.819687

[15] Kai Zhang, Yangyu Fan, Jiadong Xu, Chen Qu, "Design of broadband, low cost single layer reflectarray using phoenix cell," in TENCON conference, Xi'an, China, 2013. https://doi.org/10.1109/TENCON.2013.6718981

[16] H. Salti, E. Fourn, R. Gillard, "Minimization of MEMS breakdowns effects on the radiation of a MEMS based reconfigurable reflectarray," IEEE Transactions on Antennas and Propagation, **58**(7), 2010. https://doi.org/10.1109/TAP.2010.2048861

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 106-118 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Performance of Location and Positioning Systems: a 3D-Ultrasonic System Case

Khaoula Mannay[*,1,2], Jesus Urena[1], Álvaro Hernández[1], Mohsen Machhout[3]

[1] *Department of Electronics, University of Alcala, 28805, Alcalá de Henares, Madrid, Spain*

[2] *EμE Lab Faculty of Sciences of Monastir, National Engineer School of Tunis, University of Tunis EI Manar, 1002, Tunis, Tunisia*

[3] *EμE Lab, Faculty of Sciences of Monastir, University of Monastir, 5019, Monastir, Tunisia*

A R T I C L E I N F O

Article history:
Received: 01 November, 2017
Accepted: 21 February, 2018
Online: 12 March, 2018

Keywords :
Indoor Ultrasonic LPS
LPS distribution
3D positioning
PDOP
Positioning error
Accuracy
LPS fusion

A B S T R A C T

The necessity of navigation in people and mobile robots (MR) through specific environments (indoors or outdoors) has become more and more relevant nowadays. For indoors, generally speaking, the positioning systems can be divided into 2D (two dimensions) or 3D (three dimensions) approaches, where Ultrasonic Local Positioning Systems (ULPS) are often a common solution for MRs in 2D. This work proposes the extension of an already developed 2D ULPS to a 3D ULPS, where the compact design and the suitable performance of the initial 2D ULPS have been maintained. The ultrasonic beacons have been re-arranged to avoid co-planarity, then improving the third coordinate estimation. Furthermore, this work proposes the use of up to four ULPSs together to cover the 3D region of interest. Two configurations have actually been considered, one involving three ULPSs and another based on four. A heuristic Position Dilution of Precision (PDOP) estimation has been carried out, by taking into account two ways of obtaining the 3D-position: a) all beacons from the three different ULPSs are processed simultaneously, so all measurements are considered in the same set of positioning equations; and b) every ULPS is detected and considered separately and, later, the different estimated positions are merged. The second option is more likely to happen in a real scenario and, furthermore, the fusion of the independent positions obtained from each one of the ULPS improves the final position accuracy.

1. Introduction

The necessity for positioning and navigating objects and people indoors and outdoors is expanding more and more every day. For outdoor environments, the Global Positioning System (GPS) is widely used. Nevertheless, whether there is a lack of RF signals, indoors or in constrained outdoor environments, other approaches have already emerged as a supporting and/or alternative technology, thus providing the so-called Local Positioning Systems (LPS). The demand for these LPSs, and in general terms for indoor positioning, has also become more relevant due to the worldwide spreading of smart devices, and their corresponding location applications.

Different sensory technologies have already been used in the development of LPS, such as Wi-Fi, infrareds (IR), ultrasounds (US) or radio-frequency (RF). The final decision on one technology or another is mainly related to the type of application, the environment, the required accuracy, as well as other secondary parameters [1] [2] [3]. Among them, ultrasounds-based approach is often considered whether the required accuracy is in the range of centimetres, not only for 2D but also for 3D deployments.

This work has been initially based on the previous LOCATE-US LPS (ULPS), firstly developed by the GEINTRA-US/RF Research Group from the University of Alcalá for 2D positioning [4]. A performance analysis of an ULPS is presented here when it is adapted for 3D positioning in a certain region. The proposal consists of using several ULPSs, which cover the area under scanning from different points of view in order to achieve an accurate estimation of the receiver's position in 3D. The main goal of the paper is the study of different configurations of these ULPSs (number and location in the environment) and compare them in terms of the accuracy obtained in a grid of test points that cover all the 3D region of interest. As the positioning accuracy depends on

[*]Khaoula Mannay, Email: khaoula.mannay@gmail.com

the position itself to be determined, for every particular configuration it is necessary to study the positioning errors in all the points for all the cases. The rest of the manuscript is organized as follows: Section 2 presents a general background about LPSs, their main characteristics and technologies; Section 3 describes the previous LOCATE-US LPS, whereas the proposed 3-D ultrasonic LPS (ULPS) is presented in Section 4; Section 5 explains the 3D positioning algorithm implemented for the ULPS; Sections 6, 7 and 8 shows some simulation results; Section 9 presents a short discussion; and, finally, conclusions are considered in Section 10.

2. LPS Background

Local Positioning Systems (LPS) try to position one or multiple mobile objects accurately in an indoor environment. They are often used in applications, such as environment monitoring, people tracking, robot localization, resource management, or location-based services. The moving object is usually equipped with a small receiver (or transmitter depending on each LPS design), which acquires the emissions coming from the beacons forming the LPS, in order to estimate its own position. The nature of the transmitted signals can be different, depending on the technology involved: IR, RF, US, etc. [5]. According to this, Fig. 1 shows a general scheme for a LPS, where the beacons are fixed in the environment and the receiver is moving and computing its own position.

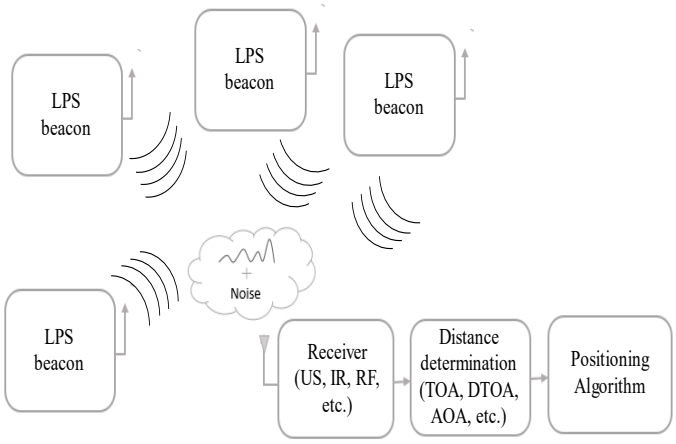


Fig. 1. General scheme of a LPS based on transmitting beacons and a moving receiver.

Furthermore, LPSs can also be classified into two groups: absolute for those where the mobile object is able to estimate its own position with respect to a reference point at any time; relative, when the mobile object can only know its position in a relative way to other receivers (or nodes) existing in the same coverage area [5].

Common LPS applications are those dedicated to 2D positioning and navigation of mobile robots and/or vehicles, as well as those focused on 3D positioning of smart devices and vehicles, such as smart phones, drones, etc. [7] [8]. Furthermore, Location-Based Services (LBS) are also potentially commercial applications for the market, not only to deliver context-dependent information accessible with a mobile device, but also to obtain information or navigate in the corresponding indoor environment. Finally, LPSs become particularly interesting in some emergency scenarios, such as positioning medical staff or equipment in hospitals, assisting rescue services in critical situations, or those proposals applied to intelligent transportation systems and/or industry manufacturing [9].

The final accuracy of a LPS in the position determination often depends on the system configuration, the sensory technology and the type and difficulty of the coverage area. Related to this, proposed LPSs are typically parametrized by the geometric dilution of precision (GDOP), where the distribution of the distance error between the estimated position and the true position is computed. Another key parameter is the range of coverage, where common values are between 5 meters and 50 meters. In this sense, for large coverage areas, scalability is key to guarantee an average positioning performance, as the positioning estimation degrades with the distance between the transmitters and receiver [10] [11] [12].

As far as the positioning algorithms are concerned, all of them are based on the determination of a variable from the beacon-receiver transmission. Some previous works based on ultra-wide band (UWB) have been recently proposed for indoor positioning by using the time differences of arrival (TDOA) from the RF transmissions between a reference beacon and the others. Typical accuracy values for these approaches are in the range of metres or even centimetres, although they present significant drawbacks, such as complexity or multi-path effects on the achieved accuracy [13] [14].

On the other hand, wireless local area networks (WLAN) have also been applied to indoor positioning [15]. Although they can also be based on TDOA, most cases deal with the received signal strength (RSS), reaching accuracies from 3m to 30m approximately [16] [17]. A similar case is the use of radio-frequency identification (RFID) [18]. In the case of infrareds, they often require a direct line-of-sight (LOS) communication between devices along a very short distance [19], where proximity, differential phase shift, and angle of arrival (AoA) are predominant.

Finally, ultrasounds-based systems also deal with TDOA [20], when there is no synchronization between the beacons and the receiver, and, afterwards, a hyperbolic positioning algorithm allows the position of the receiver to be estimated. The global accuracy of ultrasonic LPS (ULPS) is high and the structure is simple, but they often present drawbacks, such as multipath effects, Doppler, etc. [21] [22] [23].

3. General view of LOCATE_US system

The 2D LOCATE-US ULPS, designed by the GEINTRA-US/RF Research Group from the University of Alcala [4] [24], is a compact, light and portable ultrasonic beacon architecture [5]. The ULPS is formed of five coplanar ultrasonic beacons (B_i, where i=1,2,…5), placed at the four corners of a 0.707m x 0.707m square and at the centre, as is shown in Fig. 2. The five beacons have the same orientation (usually the ULPS is placed in the ceiling) and they cover roughly the same area on the floor. As the emission pattern of transducers is 120º [5], the ULPS can cover an area of 40m² roughly for a height of 3.5m. Any receiver inside the coverage area (attached to smartphones, a mobile robot (MR), a drone, ...) can compute its position in an independent and autonomous way [4] [24].

In order to cover wide indoor areas, several ULPSs can be easily deployed. Particular calibration techniques have been proposed in [27] to facilitate this deployment.

The emitters in the ULPS use a code division multiple access (CDMA) and a time division multiple access (TDMA) protocols. Every emitter is encoded with a different code, with good auto-correlation properties and low mutual interference properties with the others. The ultrasonic transmissions of the different emitters can be configured in terms of sampling frequency, modulation schemes and code patterns to be emitted. The obtained accuracy for the measurement of distances is in the centimetre range (we assume a typical deviation of 1cm). The distances measured can be up to about 20m. All that is enough for MR applications in 2D spaces, and if extensive areas must be covered we can use a set of single ULPSs [4].

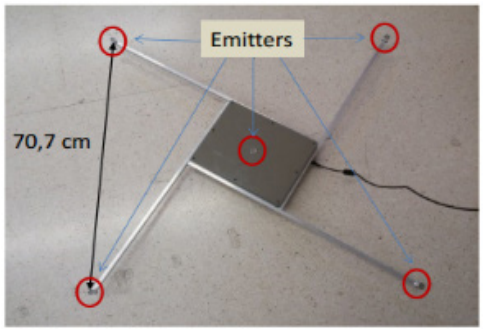


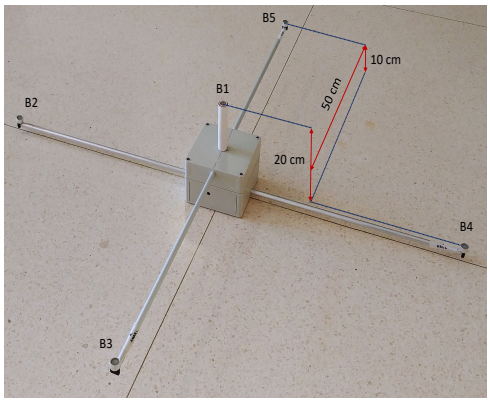
Fig. 2. General view of the 2D LOCATE-US LPS

4. Proposed 3D ULPS

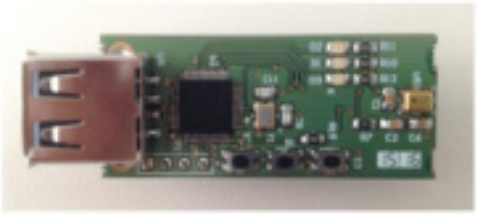
The 3D ULPS described hereinafter is an extension of the 2D LOCATE-US, so it is also formed by five ultrasonic beacons B_i placed at the four vertices of a 0.707m x 0.70 m square and at the centre, as is shown in Fig. 3.a). All of them present slight variations in the z coordinate to improve this coordinate estimation and to avoid co-planarity; but still keeping the same properties of the previous 2D ULPS, such as the common orientation and coverage area. The new beacon distribution is: B2 and B4 are in the base plane, B3 and B5 are 10cm high from the base plane, and finally B1 is placed at 20cm high from the base plane [24].

The ultrasonic beacons are wired-synchronized to enable simultaneous and periodic emission [4] [24]. The ultrasonic transmission are encoded with orthogonal 1023-bit Kasami sequences, in order to mitigate any effect coming from multiple access interference (MAI) as much as possible [25]. These codes have been selected due to their suitable auto-correlation and cross-correlation properties. For their transmission, a binary phase shift-keying (BPSK) modulation has been carried out, with a carrier placed at the central frequency of the transducer bandwidth, f_c=41.67kHz. Two carrier cycles per modulation symbol have been applied, and ultrasonic transmissions are carried out periodically every 50ms to reduce multipath effects.

The involved ultrasonic transducer is the Prowave 328ST160 [26], together with an ad-hoc front-end designed for this application. The five beacons B_i are managed by a FPGA-based system, in charge of controlling the global operation of the ULPS through an Ethernet link [24].The type of binary sequence and its length, the number of carrier periods, the type of carrier, and the time interval between different emissions can be configured by the PC. A digital pass-band filter with a 40 kHz central frequency and a 10 kHz bandwidth has been included to constrain information to the ultrasonic transducer bandwidth [26].



a)



b)

Fig. 3. a) General view of the 3D LOCATE-US LPS;
b) Designed ultrasonic receiver.

With regard to the reception, the moving device estimates its position asynchronously by hyperbolic trilateration from the TDOA measurements between a reference beacon and the others. To detect the TDOAs, the mobile device demodulates the received signal and correlates it with the corresponding emitted Kasami sequences, based on a generalized cross-correlation (GCC) of the received signal. A main lobe appears at the arrival instant for every Kasami sequence, so it is possible to calculate the associated TDOAs. Finally, a Gauss-Newton minimization method computes the position. Fig. 3.b) shows the general aspect of the used ultrasonic receiver [24].

5. 3D Positioning

In positioning systems, it is important to know the accuracy of the estimated position, which depends on the quality of measurements (ranging distances, signal strength, etc.), and on the geometry of the positioning system (beacons' geometry) with respect to the mobile node. The Position Dilution of Precision (PDOP) includes such dependencies and can be obtained empirically using (1).

$$PDOP \approx \frac{\sqrt{\sigma_x^2+\sigma_y^2+\sigma_z^2}}{\sigma_m} \qquad (1)$$

Where σ_x^2, σ_y^2 and σ_z^2 are the position variances in the three axes X, Y and Z, respectively; and σ_m is the standard deviation in the distance measurements (assumed to be 1cm hereinafter). As has been already mentioned, the technique used to position objects in 3D is based on the TDOAs, thus requiring synchronization between beacons (emitters), but not with receivers.

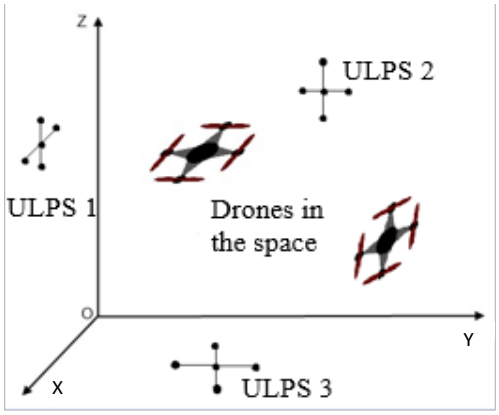


Fig. 4. Example of 3D configuration using three ULPSs and two receivers (drones).

In order to position a mobile target in a 3D space (see Fig. 4), a single ULPS is not enough to cover all the space with enough accuracy, assuming an 8x8x8m^3 volume. The setup with a single ULPS implies that the performance significantly degrades with the height variation, due to the poor Vertical Dilution of Precision (VDOP), which determines the performance by only taking into account the typical deviation in the *z* coordinate.

Thereby, the use of only one ULPS is not a suitable solution. As an example of that, Fig. 6 shows the case in which one ULPS is placed at the centre of the floor at coordinates (4m, 4m, 0m), whereas Fig. 8 depicts the case of one ULPS placed at the lower corner with coordinates (0m, 0m, 0m). Both Figs. represent the cloud of obtained position points, assuming the receiver in the X-Y plane (with steps of 1m in both axes), for three different heights: *z*=6m in the half upper volume of the room; *z*=4m in the middle of the room; and *z*=2m in the half lower volume of the room.

Additionally, Figs. 7 and 9 show the different PDOP values for the same X-Y planes and heights (*z*=2m, 4m, 6m). Note that the PDOP values are high in general terms (above 100). The contour map is a representation of the PDOP in planes at different height. The values of the PDOPs have been calculated for every point in the grid (according to the cloud of points with the estimated positions after the simulations).

Each color represent a particular value of PDOP; and the greater the PDOP the greater the positioning error in this point (even in the case that all the distances has been measured always with a typical deviation of 1 cm). With this representation one can have an idea about the error we can wait in each region of the environment in a real situation with a particular ULPS arrangement.

On the other hand, the receiver has been placed at every point in a 9x9x9 grid (8x8x8m^3 volume using 1m interval). At each point, a hundred simulations have been run by using an hyperbolic trilateration with the Gauss-Newton Positioning Method (GNPM) [10]; the standard deviation in the distance measurements is σ_m =1cm, which is consistent with ultrasonic measurements of distances.

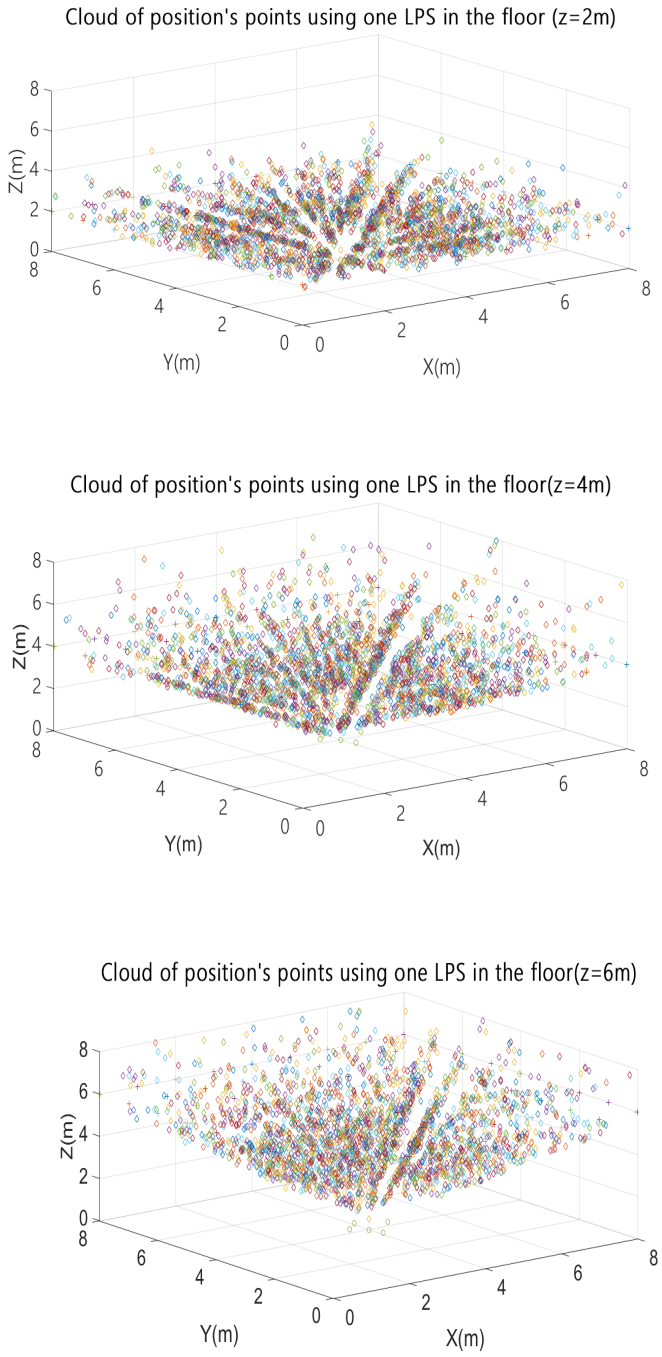


Fig. 6. Cloud of position points for one ULPS placed at the centre of the floor (4m,4m,0m) for X-Y planes at different heights (z=2m, z=4m, z=6m).

As can be observed in Fig. 7, the PDOP values differ from one plane to another, but, in general, it is smaller in the centre of every plan, and then it increases towards the sides of the room. For the first plane *z*=2m (lower half of the room), the PDOP varies from 100 to 900.

Whereas these values increase in the middle of the room (*z*=4m), where the PDOP is between 250 and 800; and, finally, they range from 450 to 850 in the third plane *z*=6m (upper half of the room).

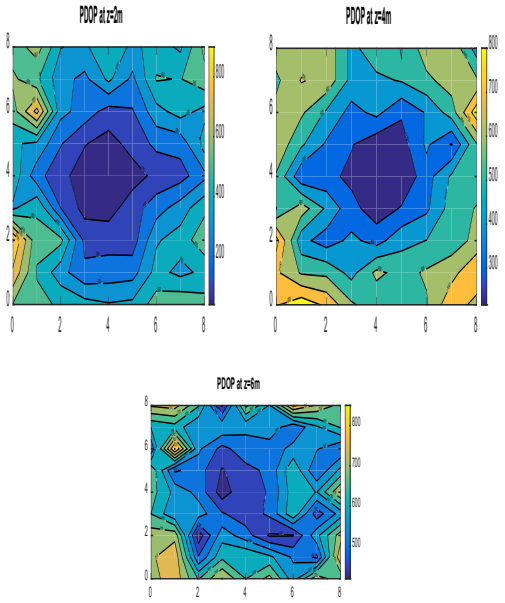


Fig. 7. Colour map of PDOPs for an ULPS placed at the centre of the floor (4m, 4m, 0m) for different X-Y planes (z=2m, z=4m, z=6m).

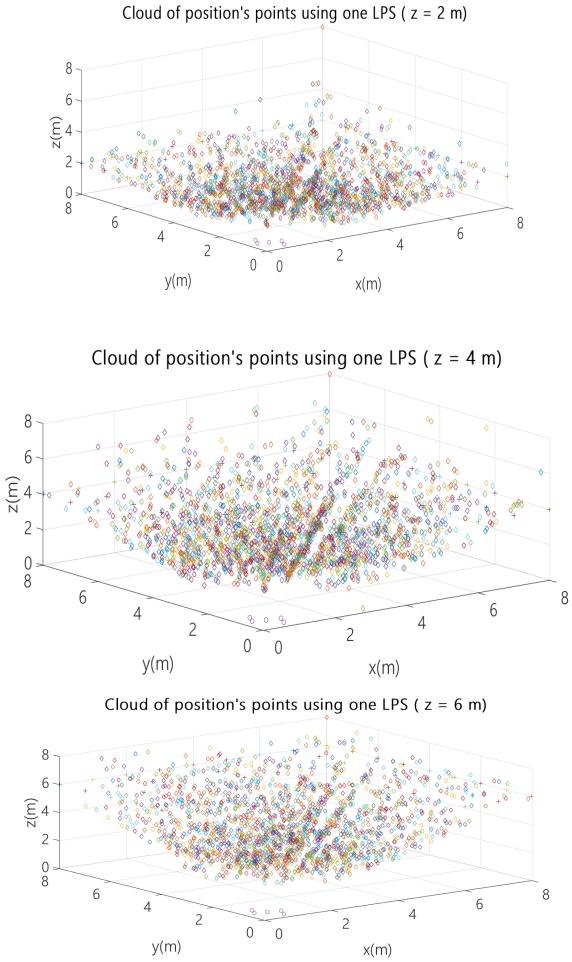


Fig. 8. Cloud of position points using an ULPS placed at (0m, 0m, 0m) corner for X-Y planes z=2m, z=4m and z=6m.

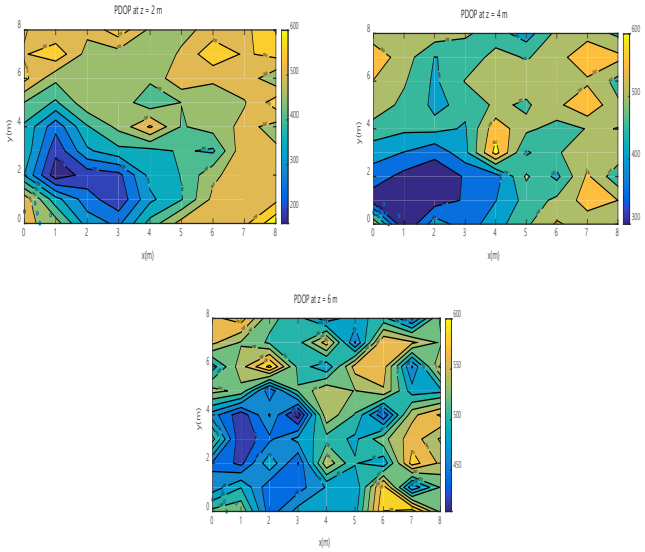


Fig. 9. Colour map of PDOPs for an ULPS placed at (0m, 0m, 0m) corner for X-Y planes (z=2m, z=4m, z=6m).

Similar conclusions can be derived from Fig. 9 with the LPS placed at corner (0m, 0m, 0m) and pointing at the centre of the room. For the first plane z=2m, the PDOP values are between 200 and 600; for z=4m, they are between 300 and 600; and between 420 and 600 for z=6m.

In order to improve these results, as well as to enhance the coverage area, the use of several ULPSs placed at different points of the room has been considered. These tests have been carried out by using two different configurations:

- Configuration A: three ULPSs placed at the centres of X-Y plane z=0m, Y-Z plane x=0m and X-Z plane y=8m, respectively. All these ULPS are emitting perpendicularly to the plane in which they are placed.
- Configuration B: four ULPSs placed at corners (0m, 0m, 0m), (8m, 8m, 8m), (0m, 8m, 0m) and (8m, 0m, 8m), respectively. Each LPS is emitting in the direction of the cube diagonal corresponding to the corner at which it is placed.

In practical situations, the 3D position can be computed with one or an array of microphones to cover all the incoming signals. In this way, two options have been considered here:

- Simultaneous measurements: all the distances (derived from TDOAs), from all the ULPSs that must be synchronized, are obtained at the same time, so the positioning algorithm involves as many equations as measured distances. For the three ULPSs configuration, the positioning algorithm requires fifteen equations, whereas, in the four ULPSs configuration, twenty equations.
- Independent measurements for each ULPS: whether five distances from one single ULPS are obtained at the receiver, it obtains a 3D position. In parallel, several 3D positions can be computed (one per every detected ULPS). To combine all these 3D positions, the Maximum Likelihood Estimation (MLE) is applied. In the case of having available three independent measurements q_1, q_2 and q_3 for a certain position q, provided that the positioning error may be modelled as $p(q_i|q)=N(q, \sigma_i)$, the

merged estimate q_{MLE} for position q can be obtained by following the procedure (2):

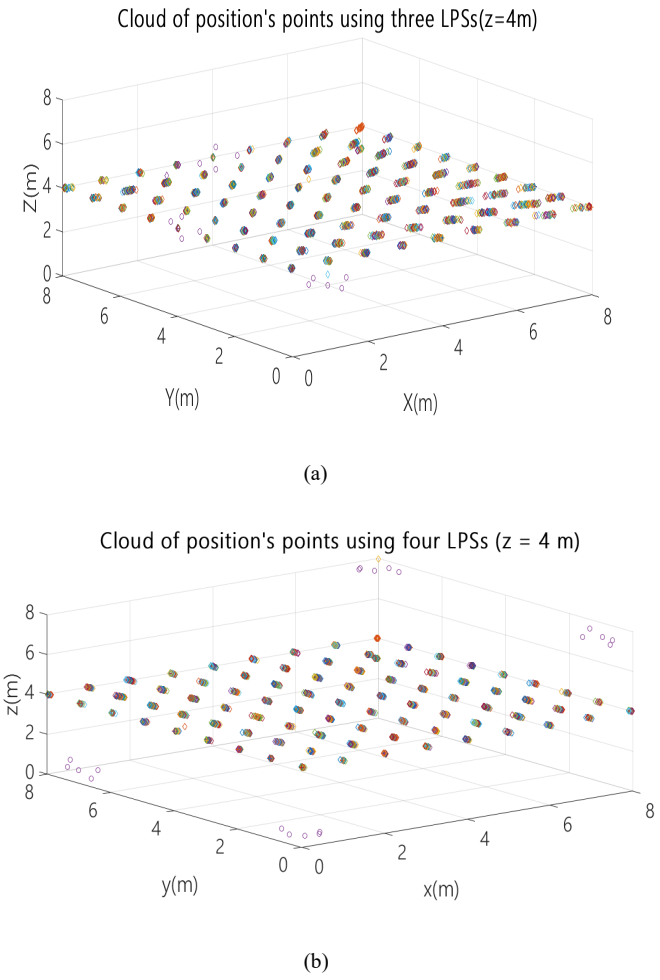


Fig. 10. Cloud of position points for the X-Y plane z=4m using: (a) configuration A with three ULPSs; (b) configuration B with four ULPSs.

$$q_{MLE} = \frac{\sigma_1^{-2} \cdot q_1 + \sigma_2^{-2} \cdot q_2 + q_3^{-2} \cdot z_3}{\sigma_1^{-2} + \sigma_2^{-2} + \sigma_3^{-2}} \qquad (2)$$

Where q_1, q_2 and q_3 are three independent measurements for position q in the case of detecting three different ULPSs. Since the statistical information is additive, the new standard deviation σ will be (3):

$$\sigma^{-2} = \sigma_1^{-2} + \sigma_2^{-2} + \sigma_3^{-2} \qquad (3)$$

Where σ_1, σ_2 and σ_3 are the standard deviations for the corresponding position measurements q_1, q_2 and q_3.

6. Results for configuration A (three ULPSs on the walls)

In order to compare the performance of each ULPS arrangement and the total number of distance measurements available, for both configurations A and B, one hundred simulations have been conducted at every point in the 3D grid (9x9x9m^3) with a step in x, y and z of 1m. The standard deviation considered in the distance measurements is σ_m=1cm.

6.1. Simultaneous measurements from all ULPSs

In this case, the 3D receiver position is estimated by the GNPM at each point, assuming there are available simultaneous measurements of TDOAs from all the ULPSs at the receiver. For configuration A, the space is covered by three ULPSs, as shown in Fig. 10. Two cases have been studied: only two of them are available; all the three are available.

6.1.1. Two ULPSs available

Firstly, in addition to the ULPS placed at the position (4m, 4m, 0m), a second ULPS has been placed at the centre of the wall y=8m, at position (4m, 8m, 4m). Figs. 11 and 12 show the cloud of positions obtained, as well as the PDOP for planes z=2m, z=4m and z=6m.

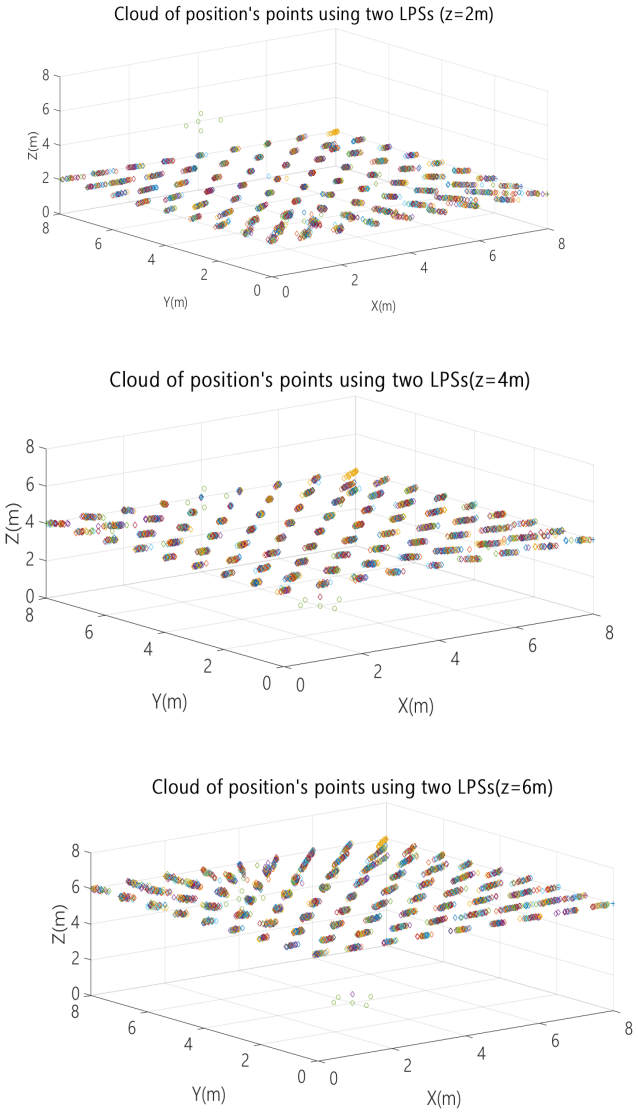


Fig. 11. Cloud of position points for two ULPSs: one placed at (4m, 4m, 0m) and the other at (4m, 8m, 4m), for different X-Y planes (z=2m, z=4m, z=6m).

Comparing with the previous results achieved with only one ULPS (Figs. 6 to 9), the improvement is clear: now the PDOP varies from 5 to 100, providing larger volumes with lower values. The lowest PDOP values are around the centre of the room.

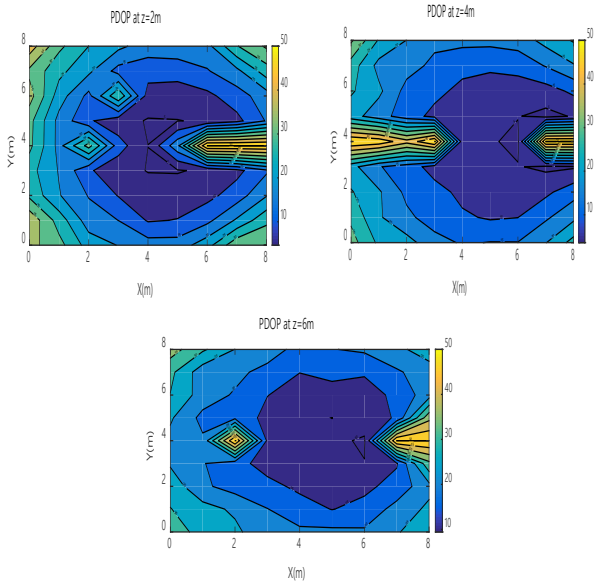


Fig. 12. Colour map of PDOP values for two ULPSs placed at (4m, 4m, 0m) and the other at (4m, 8m, 4m), for different X-Y planes (z=2m, z=4m, z=6m).

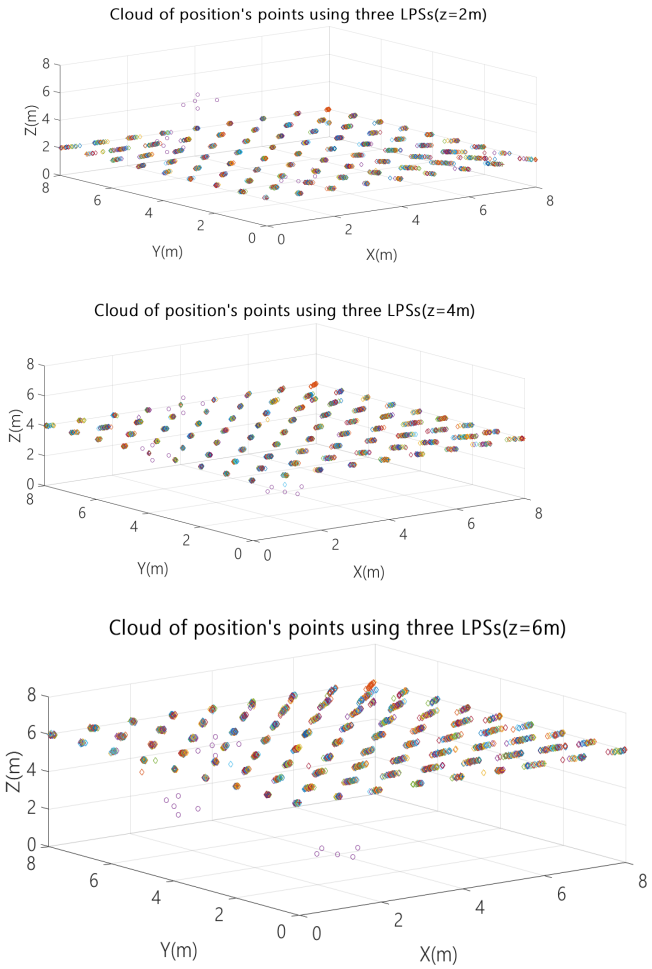


Fig. 13. Cloud of position points for three ULPSs placed at (4m, 4m, 0m), (4m, 8m, 4m) and (0m, 4m, 4m) for different X-Y planes (z=2m, z=4m and z=6m).

6.1.2. Three ULPSs available

A third ULPS has been added at coordinates (0m, 4m, 4m). Figs. 13 and 14 show the same clouds of points and PDOPs as before, for X-Y planes *z*=2m, *z*=4m and *z*=6m.

By adding the third ULPSs, the cloud of the 100 position points simulated is more accurate around the real positions considered in the grid. The PDOP values decrease below 30 in general terms, and below 15 in almost all the space, independently of the height.

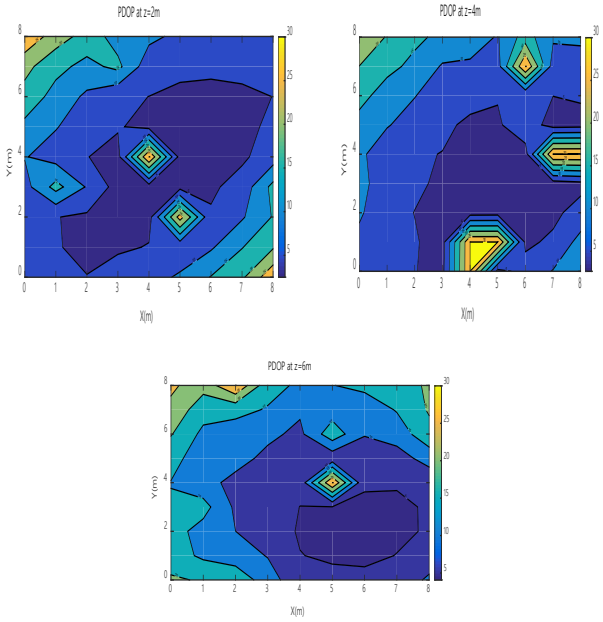


Fig. 14. Colour map of PDOP values for three ULPSs placed at (4m, 4m, 0m), (4m, 8m, 4m) and (0m, 4m, 4m) for different X-Y planes (z=2m, z=4m, z=6m).

6.2. Independent measurements for each ULPs

6.2.1. Fusion of two ULPSs

The configuration analysed here is the same as that one in Section 5.1.1. Now, the 3D positions from the two different ULPSs are obtained separately, and, afterwards, these positions are combined by using a MLE to estimate a final position. The process has been repeated a hundred times at each point in the aforementioned grid. The final PDOP values are shown in Fig. 15, varying from 5 and 100, also providing large areas with low values. It is worth noting that these results are similar to those in Fig. 12 for the case of simultaneous measurements.

6.2.1. Fusion of three ULPSs

When having a third ULPS placed at (4m, 0m, 4m), the PDOP values have been calculated for a hundred simulations at each point in grid, and they are shown in Fig. 16. Again, the distribution of the PDOP values with the MLE merging method are very close to the case when using simultaneous measurements (see Fig. 14).

In order to summarize the results for the case analysed in this section, for the whole volume under analysis the Cumulative Distribution Function (CDF) of the position error has been obtained, by taking into account all the points in the grid (100 simulations per each position). Fig. 17 shows the CDF for simultaneous (a) and independent (b) measurement approaches.

According to Fig. 17, for simultaneous measurements, in the 90% of the cases the error is below 0.7m when using only one ULPS, below 0.2m with two ULPSs, and below 0.1m for three ULPSs. On the other hand, for independent measurements, in the 90% of the cases the error is below 0.9m for one ULPS, 0.8m for the fusion of two LPSs, and 0.7m for the fusion of three ULPSs.

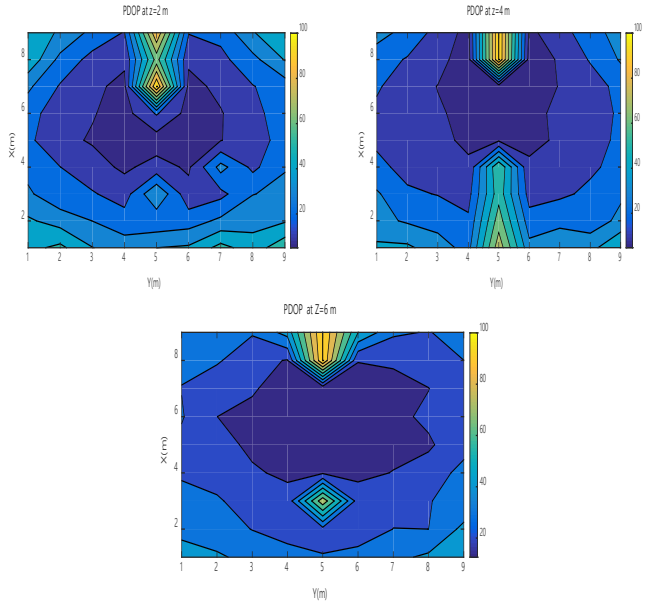


Fig. 15. Colour map of the PDOP values obtained by the MLE when using two ULPSs at (4m, 4m, 0m) and at (4m, 8m, 4m) for X-Y different planes (z=2m, z=4m, z=6m).

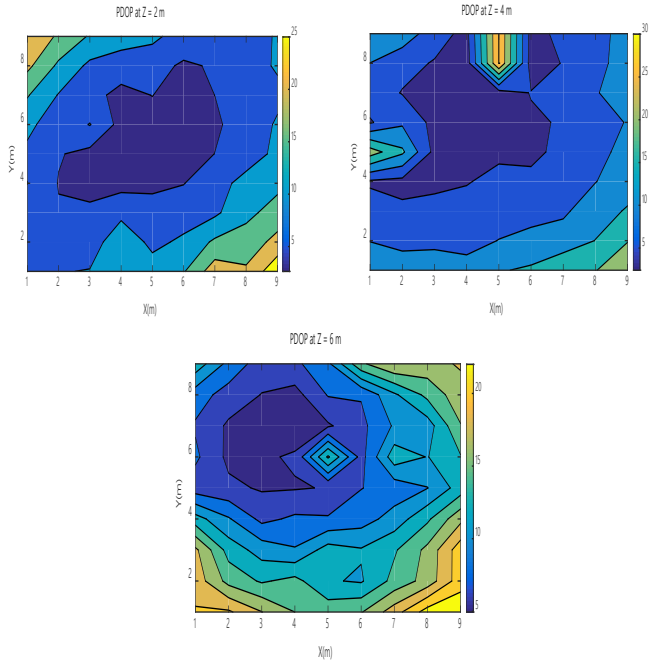
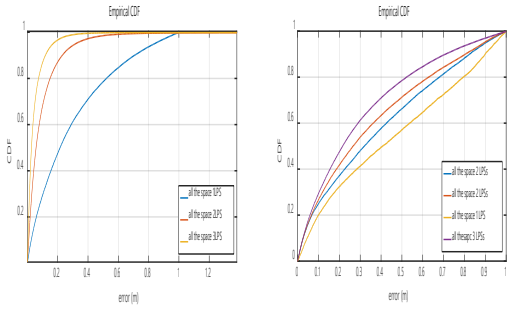


Fig. 16. Colour map of the PDOP values obtained by the MLE when using three ULPSs at (4m, 4m, 0m), (4m, 8m, 4m) and (0m, 4m, 4m) for X-Y different planes (z=2m, z=4m, z=6m).



(a) (b)

Fig. 17. CDF for the position error in the whole volume: a) using simultaneous measurements; and b) using independent measurements (and fusion).

7. Results for configuration B (with four ULPSs)

As has been already explained in Section 4, the whole volume is covered by four ULPSs in the configuration B. The goal is to improve the results presented in Figs. 8 and 9, where only one ULPS located at position (0m, 0m, 0m) was used.

7.1. Simultaneous measurements

7.1.1. Two ULPSs

Together with the ULPS located at (0m, 0m, 0m), a second one is added at the opposite corner, that is, located at (8m, 8m, 8m). The resulting clouds of points and PDOP after 100 simulations at each grid point are shown in Figs. 18 and 19, respectively, for planes z=2m, z=4m and z=6m.

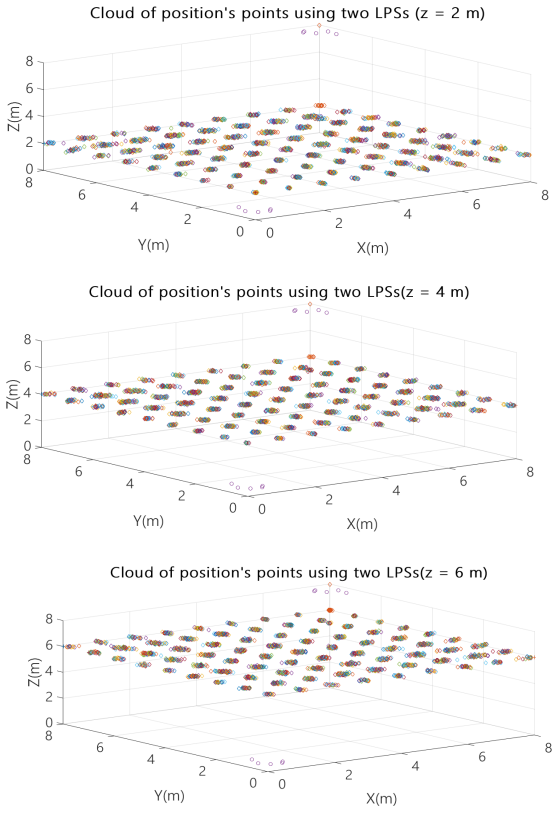


Fig. 18. Cloud of position points for two ULPSs at corners (0m, 0m, 0m) and (8m, 8m, 8m) for planes (z=2m, z=4m, z=6m).

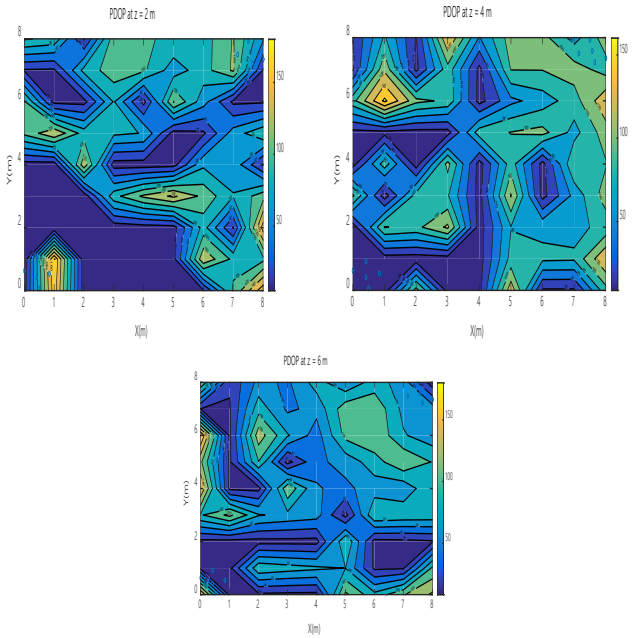


Fig. 19. Colour map of PDOP values for two ULPSs at (0m, 0m, 0m) and at (8m, 8m, 8m) for X-Y planes (z=2m, z=4m, z=6m).

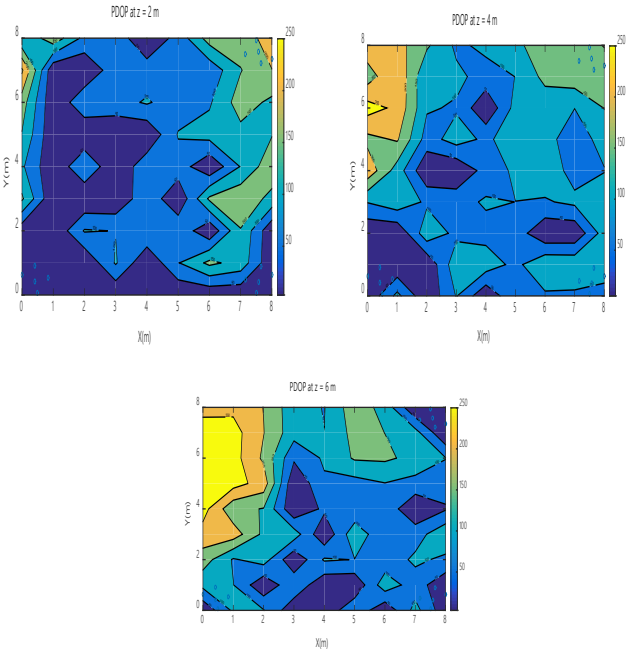


Fig. 21. Colour map of PDOPs for three ULPSs at (0m, 0m, 0m), (8m, 8m, 8m) and (8m, 0m, 8m) for X-Y planes (z=2m, z=4m,z=6m).

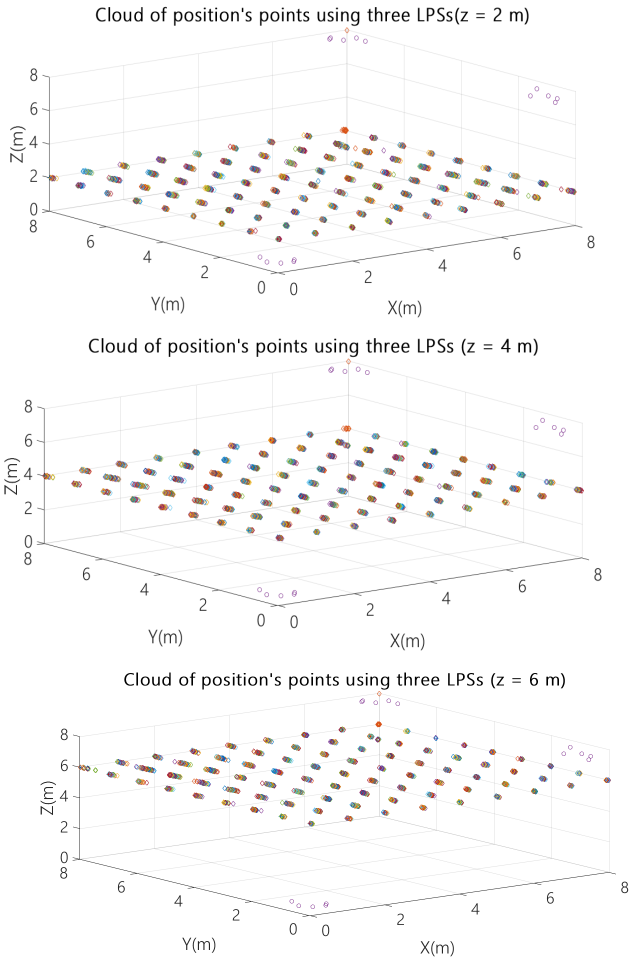


Fig. 20. Cloud of position points for three ULPSs, placed at (0m, 0m, 0m), (8m, 8m, 8m) and (8m, 0m, 8m) for X-Y planes z=2m, z=4m and z=6m.

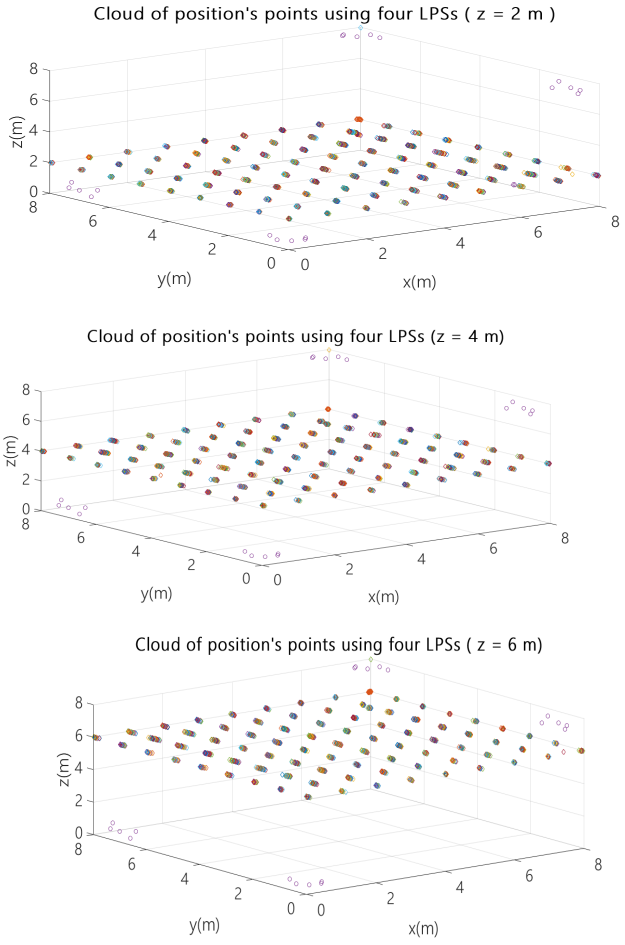


Fig. 22. Cloud of position points for four ULPSs at (0m, 0m, 0m), (0m, 8m, 0m), (8m, 0m, 8m) and (8m, 8m, 8m) for X-Y planes (z=2m, z=4m, z=6m).

It can be observed that there is an improvement in the error for all the grid points, compared to the case of only one ULPS, but still the PDOP values are in the interval from 10 to 180.

7.1.1. Three ULPSs

With a third ULPS placed at the corner (0m, 8m, 0m), the obtained results are shown in Figs. 20 and 21, also for the three planes *z*=2m, *z*=4m and *z*=6m.

The 3D positions are now more accurate, since the errors decrease especially in the neighbourhood of the ULPSs. The PDOP values are below 50 in large areas of the analysed planes.

7.1.2. Four ULPSs

Finally, a fourth ULPS is inserted at (0m, 8m, 0m), in addition to the previous three ones. The corresponding results are plotted in Figs. 20 and 21.

The errors have been considerably reduced with PDOPs between 5 and100, including large areas below 20.

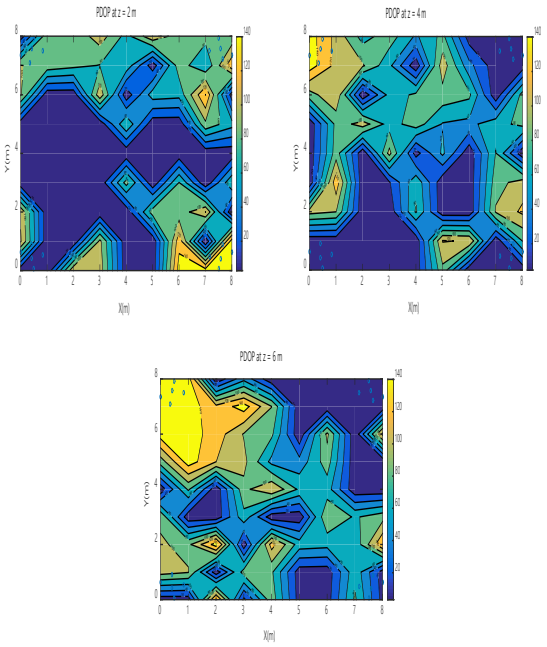


Fig. 23. Colour map of PDOP values for four ULPSs at (0m, 0m, 0m), (0m, 8m, 0m), (8m, 0m, 8m) and (8m, 8m, 8m) for X-Y planes (z=2m, z=4m, z=6m).

7.2. Independent measurements

In the second case analysed with the configuration B, each ULPS is considered independently. Figs. 24, 25 and 26 represent the PDOP values for two, three or four ULPSs,

respectively, after the fusion of the independent calculations for the positions.

Note that for two ULPSs the PDOP values are still as high as 420 in relatively large areas around the centre of the space. These areas are greatly reduced with three ULPSs (see Fig. 25) and even more with four ULPSs (see Fig. 26).

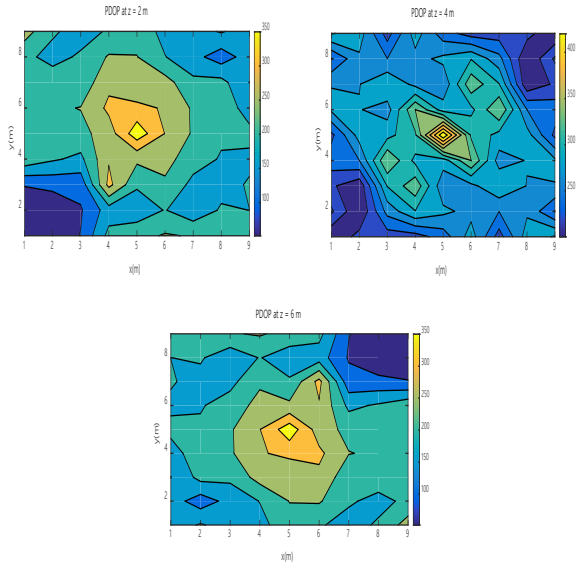


Figure 24. Colour map of the PDOP values obtained by the MLE when using two ULPSs placed at (0m,0m,0m) and (8m,8m,8m) for different X-Y planes (z=2m, z=4m, z=6m).

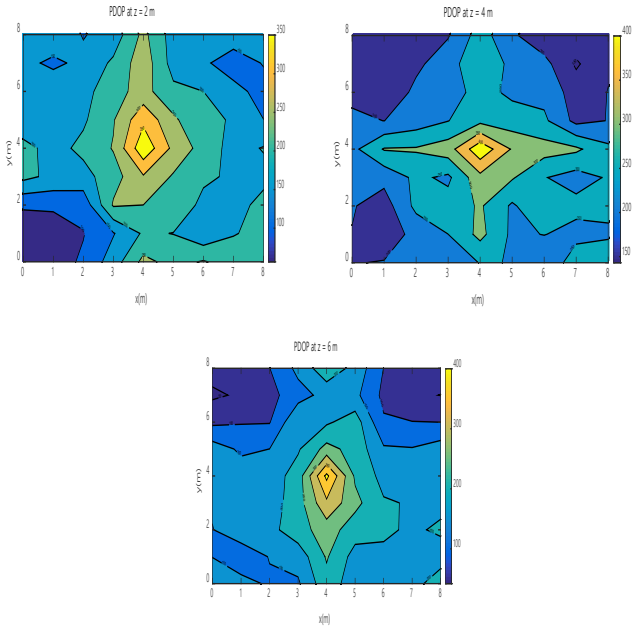


Figure 25. Colour map of the PDOP values obtained by the MLE when using three ULPSs placed at (0m,0m,0m), (8m,8m,8m) and (0m,8m,0m) for different planes (z=2m, z=4m and z=6m).

7.2.1. Error CDF

The error CDF, again obtained in all the points (and for a hundred simulations at each) for these distributions of ULPSs previously described, can be seen in Fig. 27. The upper plot considers simultaneous measurements, whereas the second one involves the fusion of independent measurements from each ULPS. In both, different CDF plots are shown for one, two, three and four ULPSs.

For the simultaneous emission, the errors for the 90% of the cases are below 0.75m using one ULPS, 0.15m for two ULPSs, 0.1m for three ULPSs and, finally, 0.07m for four LPSs (see Fig.

27.a). In case of fusion of data from independent ULPSs, these errors are below 0.85m for one ULPS and below 0.7m for four ULPSs (see Fig. 27.b).

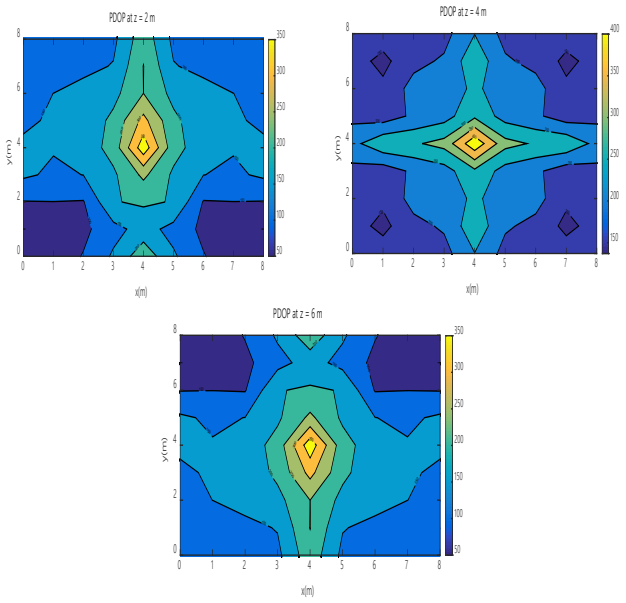
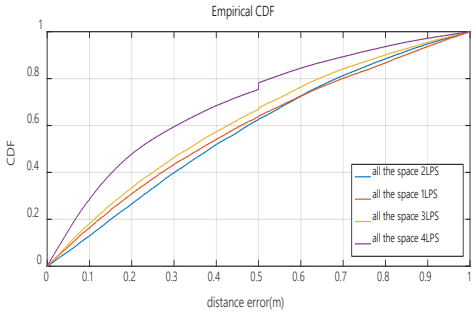
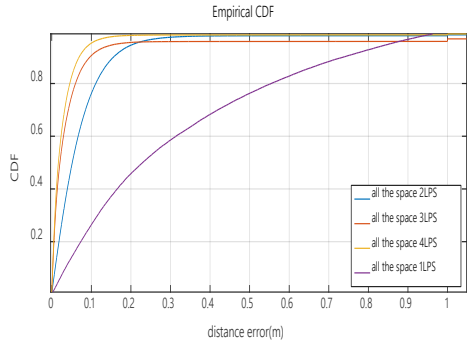


Figure 26. Colour map of the PDOP values obtained by the MLE when using four ULPSs placed at (0m,0m,0m), (8m,8m,8m), (8m,0m,8m) and (0m,8m,0m) for different X-Y planes (z=2m, z=4m, z=6m).



(a)



(b)

Figure 27. Error CDF in all the space, for 1, 2, 3 or 4 ULPSs and considering: a) simultaneous measurements with all the ULPSs; and b) independent measurements from each ULPS.

Finally, for the best case (simultaneous measurements with four ULPSs), the CDF has been split and represented in Fig. 28 for the points in every analysed plane *z*=0m, 1m, 2m …8m. Note that the error is quite similar in all the planes and, for the 90% of the cases, it is below 6-8 cm. That is, the performance of the system is similar in the whole volume.

8. Experiment results

In order to validate the proposal with real data, firstly, configuration A presented in Section 5 has been considered by using one ULPS. The tests have been carried out by positioning a receiver at each one of the 16 specific positions in a grid on the floor, as can be observed in Fig. 29. The ULPS is placed in the ceiling at a height of 3.5m. This workspace is not very complex and only the points of the corners of the grid have some problems due to signal attenuation and multipath.

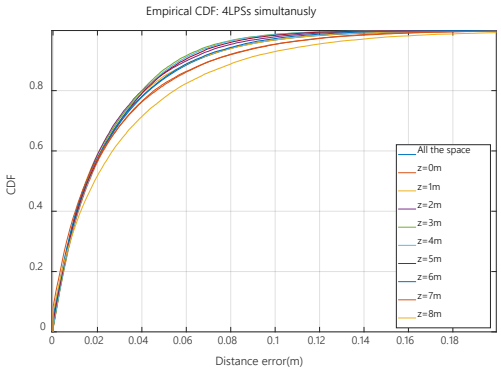


Fig. 28. Error CDF in every X-Y plane (z=1,2,…,8) for 4 ULPSs with simultaneous measurements.

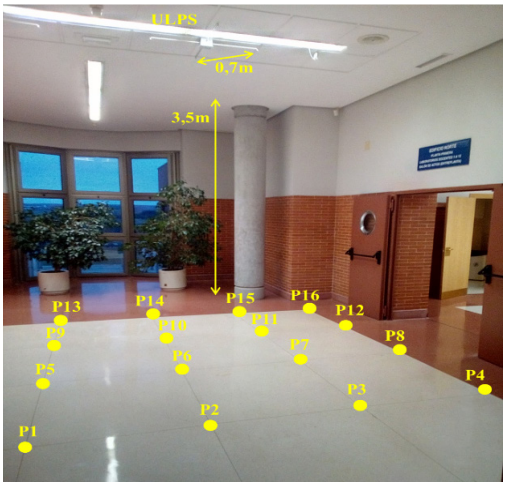


Fig. 29. Experimental setup formed by one ULPS placed in the ceiling at a height of 3.5m and the receiver placed at 16 points in a 4x4m² grid on the floor.

A set of 100 measurements have been carried out for every position, where the final estimated positions are represented by blue crosses in Fig. 30. On the other hand, the red circles are the real positions. An error ellipse containing the 95% of estimates has also been represented for every cloud of points. The green diamonds are the beacons' projections on the floor. It can be observed that the estimated positions at the test points below the beacons have a lower error than those further away. Nevertheless, all the positions are reasonably estimated.

Fig. 31 shows the error CDF obtained by simulating only one ULPS for the grid of points at z=3.5m. It can be observed that, for the 90% of cases, the positioning error is less than 1 cm. On the

other hand, Fig. 32 represents a similar CDF for experimental data, considering all the measurements or grouping them into two subsets: points P6, P7, P10 and P11 (placed at the center of the grid, just below the ULPS) and the rest of points. The differences between simulated and experimental data can be due to several factors: inaccuracies in the position of the receiver when tests were performed, and propagation effects not considered in simulations (attenuation, multipath, non-Gaussian noise, etc.).

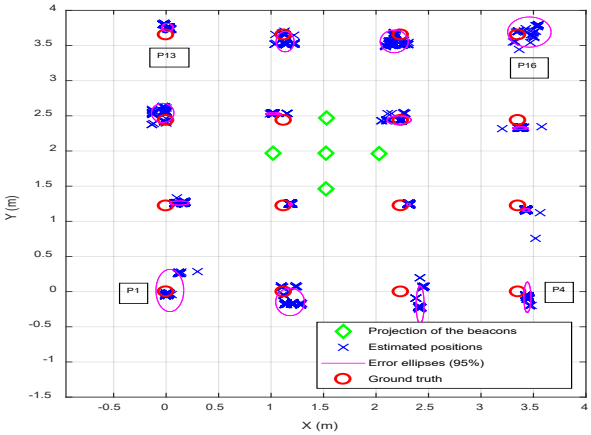


Fig. 30. Experimental test positions (red circles), estimated positions (blue crosses), and beacons' projections (green diamonds)

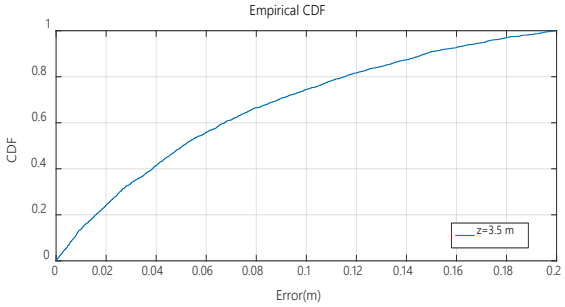


Fig. 31. Error CDF obtained by simulation for the X-Y plane (z=3.5m) using only one ULPS.

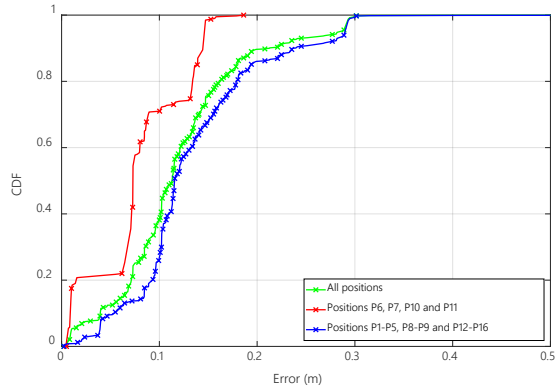


Fig. 32. Error CDF for the experimentally estimated positions in the X-Y plan (z=3.5m) using one ULPS.

Note that, according to the different CDFs in Fig. 32, the error is less than 20cm in the 90% of cases when considering all the test positions. This error is smaller (about 15cm) for those points just below the ULPS, whereas larger (in the range of 25cm) for the test positions that are far away from the ULPS' projection.

9. Discussion

When an ULPS is used in 2D, its final position can be fixed at the ceiling with all the beacons emitting downwards and the targets with receivers pointing upwards. In that case, the coverage area on the floor is a circle about 40m², assuming the ULPS is placed 3.5m high. For an extended 2D workspace, it is needed to add more ULPSs at the ceiling at different positions to cover a larger area on the floor. Anyway, these different ULPSs are always placed in the ceiling, they work independently and the obtained accuracy is similar for all the 2D floor positions considered.

Nevertheless, for a 3D positioning it is necessary to cover all the space and not only the floor surface. This is why a set of ULPSs must be placed at different positions and with different orientations (see Fig. 4) to cover all the space, thus having enough accuracy in all the volume of interest.

Two different configurations for 3D positioning in a cubical region of 8x8x8m^3 have been studied in this work. The first is the configuration A with three ULPSs pointing towards the centre of the cube and placed at the centres of the X-Y plane (z=0m), the Y-Z plane (x=0m), and the X-Z plane (y=8m). The second is the configuration B with four ULPSs pointing towards the centre of the cube and placed at corners (0m, 0m, 0m), (8m, 8m, 8m), (0m, 8m, 0m) and (8m, 0m, 8m).

Furthermore, two alternatives have been considered for each configuration:

- To have measurements from all the ULPSs simultaneously (with all the emitters synchronized). In this case, all the measured distances can be provided to the positioning algorithms.
- To have measurements independently for each ULPS (avoiding the need of synchronization among the different ULPSs). In this case, a position is estimated for every ULPS and afterwards all the results are merged.

The use of simultaneous measurements from different ULPSs is better in terms of the obtained accuracy, but it requires synchronization among ULPSs, as well as a wide receiver coverage that can be difficult in practical systems. On the other hand, with independent measurements for each ULPS, these issues can be avoided at the cost of reducing accuracy (that can still be enough for many applications).

10. Conclusions

The performance of a 3D ultrasonic positioning system (ULPS), firstly developed with a structure and characteristics for 2D positioning, has been studied and adapted to be deployed for 3D positioning. The study has covered the use of different ULPSs configurations, considering simultaneous or independent measurements. The comparison has been carried out and based on the estimation of the PDOP, by using a grid that covers all the space of interest (a cube of 8x8x8 m^3).

The achieved accuracy varies in the two cases analysed (case A and case B), depending on the number of ULPSs involved (1 to 4 ULPSs) and their positions/orientations (at the corners, in the ceiling, in walls....). Some preliminary experimental tests have also

shown the performance for a single ULPS. The positioning error is less than 20cm for the 90% of considered measurements. This centimetre accuracy is in accordance with other positioning systems based on ultrasounds as can be observed in [5], which also compares this technology with others also applied to this type of systems.

Acknowledgment

This work has been possible thanks to the Spanish Ministerio de Economía, Industria y Competitividad (TARSIUS, ref. TIN2015-71564-c4-1-R, and SOC-PLC, ref. TEC2015-64835-C3-2-R) and the University of Alcala (LOCATE-US, ref. CCG2016/EXP-078). Authors thank Dr. David Gualda for his support with experimental measurements.

References

[1] Z. Li, L. Feng, A. Yang, "Fusion Based on Visible Light Positioning and Inertial Navigation Using Extended Kalman Filters", Sensors (Basel). 2017 May; 17(5): 1093. doi: 10.3390/s17051093

[2] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[3] K. Curran, E. Furey, T. Lunney, J. Santos, "An Evaluation of Indoor Location Determination Technologies", Journal of Location Based Services, vol. 5, no. 2, pp. 61-78, 2011.

[4] D. Gualda, J. Ureña, J. C. García and A. Lindo, "Locally-Referenced Ultrasonic-LPS for Localization and Navigation", Sensors, vol. 14(11), pp. 21750-21769, 2014.

[5] R. Mautz, Indoor Positioning Technologies, Habilitation Thesis, ETH Zurich, 2012.

[6] A. Alarifi, A Al-Salman, M. Alsaleh, A. Alnafessah, S. Al-Hadhrami, M. A. Al-Ammar, H. S. Al-Khalifa, "Ultra-Wideband Indoor Positioning Technologies: Analysis and Recent advances", Journal of Sensors 2016

[7] N. Parnian, *Integration of Local Positioning System & Strapdown Inertial Navigation System for Hand-Held Tool Tracking*, Habilitation Thesis, University of Waterloo, 2008.

[8] S. Sameshima, E. P. Katz, "Experiences with Cricket/Ultrasound Technology for 3-Dimensional Locationing within an Indoor Smart Environment Harry", Technical Report, Carnegie Mellon University, 2009.

[9] K. Pahlavan, X. Li, and J. P. Makela, "Indoor geolocation science and technology", IEEE Communications Magazine, vol. 40, no. 2, pp. 112-118, 2002.

[10] T. Roos, P. Myllymaki, H. Tirri, P. Misikangas, J. Sievanen, "A probabilistic ap-proach to WLAN user location estimation" International Journal of Wireless Information Networks, July 2002, Volume 9, Issue 3, pp 155–164

[11] N. Bodhi, M, Goraczko, *The Cricket Indoor Location System*, PhD Thesis, Massachusetts Institute of Technology, 2005.

[12] K. Mannay, J. Ureña, Á. Hernández, D. Gualda, J. M. Villadangos," Analysis of performance of Ultrasonic Local Positioning Systems for 3D Spaces", 2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN). pp. 1-4. (ISBN 978-4-86049-074-4).

[13] Y. Zhou, C. L. Law, Y. L. Guan and F. Chin, "Indoor Elliptical Localization Based on Asynchronous UWB Range Measurement", IEEE Trans. on Instrumentation and Measurement, vol. 60, no. 1, pp. 248-257, 2011.

[14] E. García, P. Poudereux, A. Hernández, J. J. García and J. Ureña. "DS-UWB indoor positioning system implementation based on FPGAs", Sensors and Actuators A: Physical, vol. 201, pp. 172-181, 2013.

[15] J. Prieto, S. Mazuelas, A. Bahillo, P. Fernandez, R. M. Lorenzo and E. J. Abril, "Adaptive data fusion for wireless localization in harsh environments", IEEE Trans. on Signal Processing, vol. 60(4), pp. 1585-1596, 2012.

[16] M. Ocaña, M. A. Sotelo, L. M. Bergasa, R. Flores, "Low level navigation system for a POMDP based on WiFi and ultrasound observations", Proc. of the 2005 IEEE International Symposium on Computational Intelligence in Robotics and Automation, (CIRA'05), pp. 335–340, 2005.

[17] Y. Álvarez, M. E. De Cos Gómez, J. Lorenzo, F. Las-Heras, "Evaluation of an RSS-based indoor location system", Sensors and Actuators, A: Physical, vol. 167 (1), pp. 110–116, 2011.

[18] A. D. Koutsou, F. Seco, A. R. Jimenez, J. O. Roa, J. L. Ealo, C. Prieto, J. Guevara, "Preliminary localization results with an RFID based indoor guiding system", Proc. of the 2007 IEEE International Symposium on Intelligent Signal Processing (WISP'07), pp. 1–6, 2007.

[19] E. M. Gorostiza, J. L. Lázaro Galilea, F. J. Meca Meca, D. Salido Monzú, F. Espinosa Zapata and L. Pallarés Puerto, "Infrared sensor system for mobile-robot positioning in intelligent spaces", Sensors, vol. 11(5), pp. 5416-5438, 2011.

[20] D. Ruíz, J. Ureña, J. C. García, C. Pérez, J. M. Villadangos, E. García, "Efficient trilateration algorithm using time differences of arrival", Sensors and Actuators A: Physical, vol. 193, pp. 220-232, 2013.

[21] J. M. Villadangos, J. Ureña, J. J. García, M. Mazo, A. Hernández, A. Jiménez, D. Ruíz, C. De Marziani, "Measuring Time-of-Flight in an Ultrasonic LPS System Using General Generalized Cross-Correlation", Sensors 2011, vol. 11, pp. 10326-10342, 2011.

[22] D. Gualda, Mª C. Pérez, J. Ureña, J. C. García, D. Ruiz, E. García and A. Lindo, "Ultrasonic LPS Adaptation for Smartphones", International Conference on Indoor Positioning and Indoor Navigation (IPIN'13), pp. 1-6, 2013.

[23] A. Lindo, MC. Perez, J. Urena, D. Gualda, E. Garcıa and J. Manuel Villadangos, "Ultrasonic signal acquisition module for smartphone indoor positioning", Proc. of 2014 IEEE Conference on Emerging Technology and Factory Automation (ETFA), pp. 1-4, 2014.

[24] A. Hernández, E. García, D. Gualda, J. M. Villadangos, F. Nombela, J. Ureña, "FPGA-based Architecture for Managing Ultrasonic Beacons in a Local Positioning System", IEEE Trans. on Instrumentation and Measurement, vol. 66, no. 8, pp. 1954-1964, 2017.

[25] T. Kasami, "Weight distribution formula for some class of cyclic codes," Technical reportR-285, Coordinated Science Lab. University of Illinois, April 1968.

[26] Pro-Wave Electronics Corporation, *Air Ultrasonic Ceramic Transducers 328ST/R160*, Product Specification, 2014.

[27] D. Gualda, J. Ureña, J.C. García, J. Alcalá, A.N. Miyadaira. "Calibration of Beacons for Indoor Environments based on a Map-Matching Technique". *2016 International Conference on Indoor Positioning Systems.* pp. 1-7, Alcalá de Henares, Spain, October 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 107-118 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Technology

ASTESJ

ISSN: 2415-6698

Cognitive Cybernetics vs. Captology

Zdenko Balaž*, Davor Predavec

Zagreb University of Applied Sciences, Electrical Engineering Department, Zagreb 10000, Croatia

ARTICLE INFO

Article history:

Received: 29 September, 2017
Accepted: 06 November, 2017
Online: 20 November, 2017

Keywords:

Captology
Cognitive Cybernetics
Intelligent Systems
Expert Systems
Learning Mechanisms
Motivation

ABSTRACT

In acronym Captology - Computers as Persuasive Technology, a persuasive component (lat. persuasibilibus - enticing) refers to the persuasive stimulation by intelligent technologies. Latter being transitive and interactive as intelligent systems, they have imposed, by their persuasivity, a 'cult of information', after which information has become a type of goods that as a utilitarian resource must be exploited quickly and efficiently. Such a widely accepted fact resulted as hype, presenting a perspective that the approach to a large amount of information and faster 'digestion' of their content will enable users to quickly get desired knowledge.

Recent investigations about persuasion processes have shown its dependence on intelligent technology factors (design, interactive computer products, web, desktop and others). Such technologies are also used to influence people's attitudes, beliefs, learning, and behaviour. Development strategies for global computer production and sales head in that direction and confirm latter statement with the promoted 3-P model: persuasive, permissive and pervasive components. Cognitive level of human integrated development is increasingly overshadowed by the contribution of artificial intelligence through its products, i.e. 'smart' creations, and by the array of shortcomings and problems that the same interactive technology brings. This paper presents a parallel between captological component of intelligent and interactive technologies on one side and illustrates examples of captological influences proved by confirmed trials within cognitive science through computer simulations of human thinking on the other side. Many studies have shown that the success of persuasion depends on the factors which have been exposed by cognitive cybernetics. Next to it, people's behavior system is transforming through the very development of society. Therefore, the influence of latter can be either positive or negative, while its extremes are already escalating in direction of a new trans-singularity and post humanism theories based on the principles of extropy.

1. Introduction

This article extends the work originally presented at 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, May 2017, [1-4]. The presented paper confirmed how human mind development is being contributed by the 'smart' creations of intelligent technologies as artificial intelligence (AI) products. Researches have been devoted to methodological directions and developing concrete tools for self-analysis and synthesis, simultaneously providing exhaustive theories about human nature being experienced as a 'new philosophy', but also as an 'anti-philosophy', defined on a new scientific course called cognitive cybernetics.

Cognitive cybernetics is the expansion of the 'classical' cybernetics, which includes scientific comprehensions of cognitive control and regulation of artificially-technical and socio-technical systems, biological organisms, organizations, processes, structures etc. The cognitive aspect of all evaluated systems elaborates and applies cognitive models, cognitive-functional and procedural system implementation, problem-solving based on natural methods by analogue cognitive computer techniques and lastly, the realization of system solutions and their organization in the context of management approaches and methodologies. This work is based

*Corresponding Author: Zdenko Balaž, Našička street 59, 10110 Zagreb, Croatia, +385 99 3111 460 Email: zbalaz@tvz.hr

on the cognitive development and implementation of ITK's (Ingenieurgesellschaft für Technische Kybernetik, Rülzheim, Germany) education modules in cooperation with National University of Vietnam, Hochschule Offenburg and Electrical Engineering Department at Zagreb University of Applied Sciences.

Cognitive Cybernetics recognize motivation as an internal state that excites, directs, and maintains the behavior of an individual. In education and training, motivation is especially important both for learning and for seeking the purpose and meaning. In academic activities or preferences, latter is described as meaningful and effort-worthy exertions, while tend to be realized and applied within importance of higher levels competences and skills. Those who are more motivated have a greater positive attitude on everything, and so on education; they persist longer in more difficult tasks, process information at a deeper level and excel in learning experiences. The classic distribution of motivation on intrinsic and extrinsic motivation in education is an incentive for activities which:

- represent the challenges of optimally assessed goals,
- can be controlled for goal achievement, which is emotionally fulfilling,
- establish the autonomy of action through the existence of feelings that can be controlled and have the effect on own development,
- encourage curiosity for the new, surprising experiences that increase the intrinsic motivation,
- include creativity and imagination where the motivation awareness is confirmed by imagination.

Approaches for defining motivation in education are:

- Behavioral approach in which understanding of motivation begins with a careful analysis of incentives and awards,
- Humanistic approach that emphasizes the understanding of the need for self-reliance through the stimulation of internal resources (sense of competence, self-esteem, and autonomy),
- The cognitivistic approach that accepts the fact peoples' activity and curiosity are inherent, and they seek information to solve relevant tasks.

An integrated approach for creation of education based on cognitive cybernetics foundation, considers:

- socio-cognitivistic focus on individual's behavior and potential interest for the impact on beliefs and expectations,
- the socio-cultural atmosphere created by participation in the community, in activities that maintain the identity of the community's interrelationships.

Therefore, motivation in the community is strengthened by the education to which it belongs and appreciates it. Specialist tutoring programs are the foundation for acquiring an expert status, where the very understanding of problem-solving is needed at that level, for which is necessary to gain the knowledge by understanding. Elementary forms of understanding are prescribed by the concept of objective mind or spirit, which guarantees the possibility of intersubjective communication at the basic level. 'Higher forms of understanding' occur when one internal difficulty or contradiction confronts with what is already known. It is then necessary to summon up and include the entire life structure, (establishing a relationship between life expression and its cognitive content), which can be found in its hermeneutical understanding, [5-8]. Research results show that understanding develops through information processing and modifies over a longer period. At a somewhat higher flexible cognitive level, it is possible to cognitively gain skills through which information is processed both intuitively and analytically. This knowledge, along with the persuasive context and information technology, closes the 'hermeneutic circle of artificial intelligence', Figure 1.

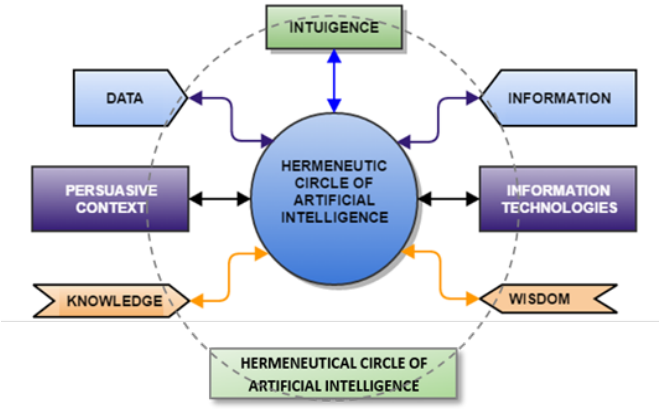


Figure 1 - Hermeneutical circle of artificial intelligence

2. Cognitive cybernetics – the future of education

Researches that deal with the captological components of intelligent and interactive technologies are confirmed by cognitive cybernetics through simulation examples of human thinking. Approaches to the study include analytical levels by evaluating the system almost in all technological branches and domains with the assumption of significant and dramatic advances in neuroscience, genomics, robotics, nanotechnology, computers and artificial intelligence with the help of other interdisciplinary studies, shown in Figure 2.

Studies and researches have further led to transhumanism extremes that advocates human body modification, including mental and psychological changes, all to control desires, mood and mental states. The everlasting fight against aging and illnesses is therefore not the only task of transhumanists, but also a try to improve the man at all levels, emphasizing simultaneous mental and physical advancements would be desirable.

If transhumanism associated with humanism is recognized as writing it as 'trans + humanism', then from the enlightened humanism comes the emphasis of a progress (its ability and desirability, not its inevitability), personal responsibility to create a better future and progress to a new society, but at the same time the awareness of accidents that can be caused by technology misuses in such society. For these reasons, it is necessary to differentiate the forms of meaning (i.e. primary and secondary), on which science assignments are based, and values that are recognized in the world, [9], addressed below in Figure 3.

Treatment by cognitive science has selected, among other things, certain factors of one's individual ability expectations to achieve desired goals and their value, respectively.

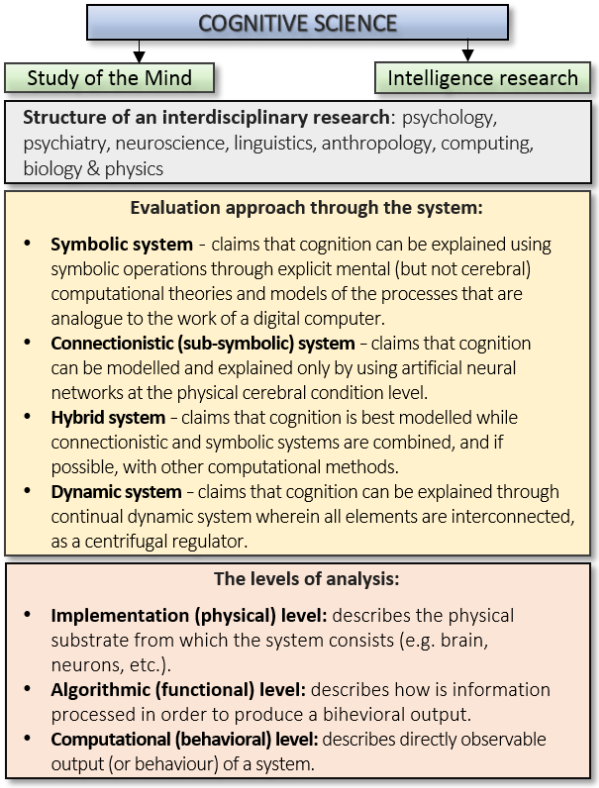


Figure 2 - Interdisciplinary study of the mind and intelligence through level analysis over the system

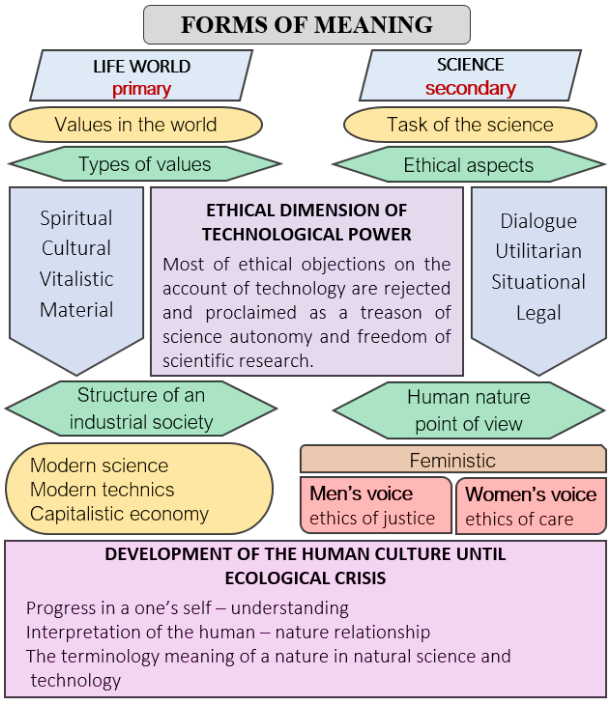


Figure 3 - Primary and secondary forms of meaning and ethical dimensions of technological power

Apart from this, it is accepted the fact that a man is a cognitive, reasonably-minded being, who accomplishes its own understanding at two levels. The first one is subjective level or the level of psychological participation, and the second is objective level or the level of intellectual participation. Hence, it is possible to conclude that hermeneutics is a universal theory of understanding in which no confusion is present because hermeneutic understanding is neither interpretation nor explanation. In this way, a differentiation between the objective understanding of the meaning on one side and subjective understanding of motives and intentions on the other side can be made. Theoretically, it is possible to attain adequate learning processes from the book, without producing any external signs of acceptance and application of such learned matter. The effects of learning efficiency, whose outcome must have been understanding, depend on many factors such as physical and biological states, prior knowledge, daily oscillations in the brain chemistry, peptide hormone levels, and ultimately – the captology. Therefore, in understanding the recognition of its either elementary or higher forms for the purpose of education system, it is important to point out the most important motivational variables, namely:

- needs,
- beliefs,
- goals as outcomes to be achieved,
- internal interest,
- emotions.

These variables are linked and processed through natural analogue systems, confirming the new concept in cognitive cybernetics – 'Brainware', [10].

2.1. Needs

Although there are many theories, all rely on the most famous one – the Maslow's hierarchy of needs that differentiates the needs of a lower order or the needs of lack - survival, security, affiliation, and self-esteem, and the needs of higher levels or the growth needs - intellectual success, aesthetic assessment, and self-realization. The latter needs can never be fully met, since when one desire is fulfilled, the other one immediately takes its place. Newer approaches to the motivation influenced by intelligent technologies focus on highlighting not only individual needs (a generated need to feel capability and competence), but also on interaction with the world, having the choice and feeling of controlling life, while at the same time on the challenge of connectedness through sustainability in the world. Such a theory, though seemingly presents the absurd paradigm because it reduces the basic human needs to autonomy, competence and connectivity, relies on the persuasion which is the second name for the captology present in cognitive cybernetics, Figure 4, [11].

2.2. Beliefs

Beliefs as a variable of cognitive cybernetic processing come down to the achievements of:

- knowledge,
- capabilities,
- causes and control,
- self - value.

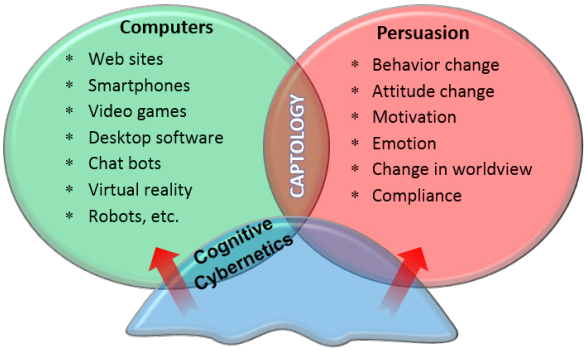


Figure 4 - The captology of cognitive cybernetics in computer persuasion, [11]

Knowledge beliefs are defined as beliefs about the structure, stability, and security of knowledge and as best ways of acquiring knowledge that influence learning strategies use.

Capability beliefs are one of the most powerful beliefs that affect motivation. There are two types of capability views: entity and incremental. The entity view refers to the belief that capability is a fixed characteristic and can not be changed. Although capabilities differ between individuals, their representation is potentially stable when acquired, but capability as a set of skills may change. By means of continuous work, learning and exercising, it is possible to increase one's knowledge and thus improve the capability.

Causes and control beliefs are based on attempt to understand the meaning of one's own and/or another peoples' behavior. This theory explains how individual explanations, justifications, and excuses about oneself and others affect motivation.

Self-value beliefs can be oriented to control or avoid failure. Orientation toward control seeks out the tendency of addressing value to achievement and the ability to set the attitude that achievement can be enhanced by targeting the goals. By mere control, one's own skills and abilities are being improved.

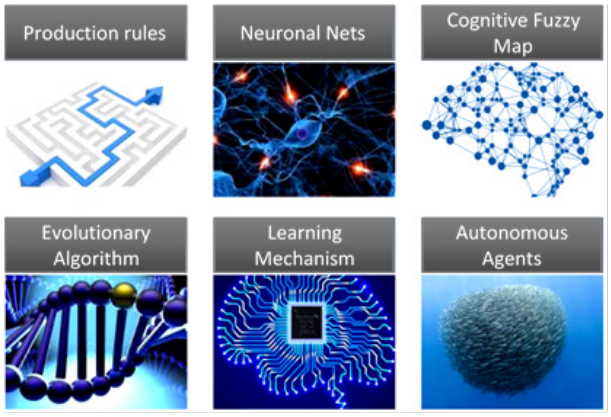


Figure 5 - Concept of cognitive solutions in environment, [11]

Orientation to avoid failure supports the entity's view of ability and sets further its goals related to performance. One can have a lack of strong ability sense and value separated from the performance, thus cognitive cybernetics serve as a simulation for technological and technical solutions (production rules, evolutionary algorithms, autonomous agents etc.), for the purpose of understanding and functioning of biological states (neural networks, learning mechanisms, cognitive fuzzy maps, etc.). In this way, mapped into education systems, strategies for avoiding failure are being worked out so they don't lead to a failure or to prevent the continuation of failure. Such strategies (shown in Figure 5) do not accept excuses or incapability, therefore the possibility to give up or accept a failure as a belief that problems arise due to low capabilities is not considered.

2.3. Goals

Goals are the outcomes or achievements to be attained and they enhance performance because they direct one's attention on the current tasks and divert from the distractors. Goals encourage devotion, for as the goal is as greater, such is the commitment. They increase perseverance, for when there is a clear goal, it is more difficult to give up. Furthermore, goals also encourage the development of new knowledge and strategies until they prove as successful ones.

Types of goals being set affect the extent of motivation that is needed to achieve them. Aims that are specific, moderately tough and likely to be achieved soon, often increase motivation and persistence. Just as goals, there are target orientations related to belief patterns about goals, also associated with achievements. Goals differ from the type of control, performing, avoiding the effort and social goals that need to be recognized for motivation.

The concept of cognitive solutions that today's civilizational environment treats as an agent systems by cognitive coverage of processes and programs, integrates human factors and machines (technologies) as inspirational cognitive goals, Figure 6.

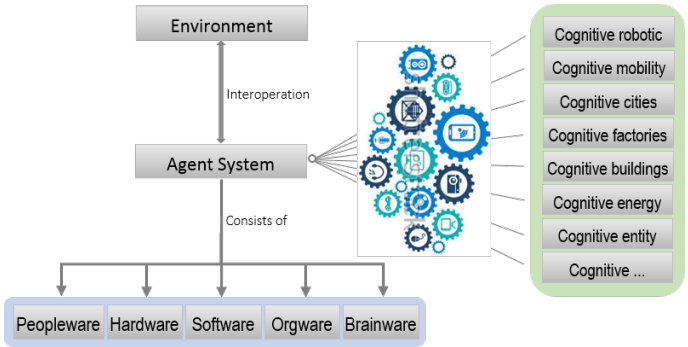


Figure 6 - Concept of environment cognitive solution set as cognitive goals, [11]

2.4. Internal interest

There are two kinds of interest: personal and situational. Personal interests relate to more persistent personal aspects, such as a long-term preference for the particular subject, while situational interests relate to short-term aspects of activity, such as a text or material that attract and retains attention. Interest in the education grows with a sense of competence, even if initially there was no interest in a certain activity. Cognitive industry, by

its evolution, is an inexhaustible source of interests, and as such through modern civilization imposing itself as a leading variable. Within cognitive cybernetics processing and understanding, interest is related to entities (synonym for the subject, article, individual, being, form of life, body, creature, organism etc.), which in ontological terms is connected with different and independent existence plus applicability, and is profiled as a cognitive entity platforms, entity ecosystems, browsers, entities, services, service portals, interoperability servers, maintenance portals, integrated server, entity application as a tool, etc., Figure 7.

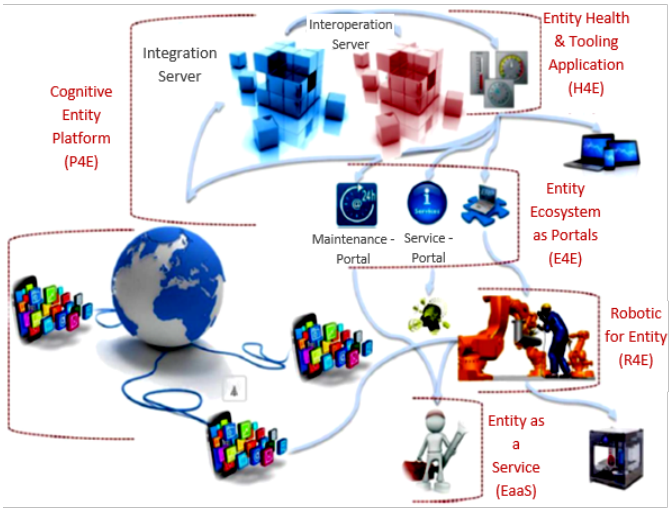


Figure 7 - Platforms of creating cognitive solutions as goal variables, [11]

2.5. Emotions

Emotions, especially pleasant ones, have very positive effects on ability beliefs (accepting an incremental view of ability), target orientation (associated with learning objectives), self-regulation, information processing (emotions are encoded with information), etc. Positive emotions encourage positive thoughts and memories, along with evaluation of education, hence such a positive effect is adequately treated for evaluation and performance of education based on the concepts of a system cognitive modelling.

In the beginning of education, emotions are very important, especially when it comes to the feelings of excitement, pleasure, fun, and curiosity, because these emotions can trigger a situational interest. It is only necessary to control them in order to prevent exaggerating effects which can lead to negative emotions, having a somewhat more complex relationship with motivation.

Additional stimulus to emotional motivation is the over-occupancy that refers to an intense form of intrinsic motivation in which an individual is in a state of profound insight, concentration and focus on the challenging activity. The experience of occupancy depends on autonomy. Intrinsic emotional motivation was often experienced in activities that were chosen willingly rather than imposed by others. Therefore, competence (skills and techniques) is so trained that it becomes automatic and is responsible for the optimal challenge (the optimal ratio of challenges and skills), creating a special emotional concentration that is not obscured with time and space.

3. Cognitive cybernetics in mystical artificial intelligence implosion

When artificial intelligence emerged in such terminology in the middle of the last century, it was the fastest growing scientific discipline. At the beginning of this century it started to be mystically missing in implosion, which can be attributed to a certain transformation. Development of artificial intelligence encompasses four periods of time, [10]:

- The period of enthusiasm and great expectations (1952 - 1969),
- The Age of Cognitive Reality (1966 - 1973),
- Knowledge-based Systems (1969-1979),
- The period of applicability (it begins in 1980s and continues as a process to date).

The fourth period is a process that goes through its five stages of transformation:

- The first stage is application of artificial intelligence in industry, from 1980s,
- The second stage is a transformation through neural networks, introduced in 1986,
- The third stage of transformation has placed artificial intelligence in the science, and is accepted since 1987,
- The fourth stage of transformation introduced super-intelligent agents, in 1990s,
- The fifth stage of transformation is the mystical implosion of artificial intelligence that is still ongoing.

Mysticism in this matter is trying to point out that implosion is a way in which artificial intelligence disappeared within itself, and at the same time in some certain connectedness, cognitive cybernetics emerges. In that mysticism lies the impossibility of terminological discernment and a clear access to technological level called 'Industry 4.0', as well as transition and emergence of the new era of 'Evolution Industry - Industry X.X.', in which cognitive cybernetics has been already implemented. Therefore, regarding such a rapid pace, it can be concluded this already 'is' the future so widely present as a subject of discussions.

Today, this future is also called the 'Society 5.0', (germ. 'Gesellschaft 5.0'). The Idea of Society 5.0. was launched from the 'Fifth Scientific and Technological Fundamental Development Plans' in Japan by the end of 2015. The idea is to bring digitalization closer to all segments of society through the creation of a so-called 'super-smart society', whose core is the social impact of technology and which is focused on cognitive cyber-products. Production processes and production itself of these cyber-products are created by developed and self-regulated autonomous cycles, partly realized through the very products.

Toward adoption of the Society 5.0 strategy, on CeBIT meeting in Hannover, 25th of January, 2017, the Japanese Prime Minister (presented Japan as the partner country) and the German chancellor agreed about opinions that although the market

prerequisites for further development brought by technology such as open borders, democratic values and a free trade, they stated: '*Millions of people do not know what to expect from the upcoming changes, and those who should change the policy and the digital sector in Europe which relies on cognitive cybernetics are often slow in key decisions*'.

3.1. Transformation approaches of cognitive cybernetics

The mystical disappearance of artificial intelligence presented above and characterized as an implosion, further shows that the emergence of cognitive cybernetics can be studied through three approaches to that transformation, as addressed below.

A. Anthropological approach

The anthropological approach of artificial intelligence into cognitive cybernetics transformation relies on the study of a human species development (life and culture), which looks at the presence and future of the human race, and links:

- nature and culture,
- past and present,
- the uniqueness of the human race through diversity of its forms.

The relationship between nature (biology) and culture defines the development of human species through cortex specialization, which enables understanding, communication, and learning. Thusly created (learned) content is called culture. As a fundamental concept of anthropology, culture is a set of behaviors that people learn and share in a given period and in a chosen natural and social environment.

The relationship between past and present although in the evolutionary sense compared to other beings follows the same natural rules, enriches people with extraordinary ability to learn and pass on the knowledge from generation to generation. This approach has accelerated the cultural evolution which has become faster than the biological evolution. In this sense, profound differences are being outlined, not only in terms of anatomy and physiology, but also in terms of the very habits so different that the essential unity of the human race is not seen anymore, and therefore the potential for innovation is harder to be recognized.

The uniqueness of human race through diversity of its forms is a consequence of the lifestyle difference not only in different times and areas, but also within the same society due to regional specialties based on the economy and different living conditions as entities. Within the same area, differences exist in the way of certain communities' lifestyle, not only immigrants' ones, but also among the sexes, ethnic, age and class groups. Responses about purposefulness inquiries given from anthropology are used to understand, create and inspire cognitive cybernetics.

B. Ontological approach

Ontological approach to the concept of cognitive cybernetics relies on strictly traditional philosophical discipline as the central branch of metaphysics that studies the being, if it is a being, as such. Modern analytical ontology is, above all, the theory of general categories such as object, attribute, event, etc. Ontology is a study of nature of the things we find in our environment, around the world, and what relationships are among them. In addition, it is related to existence, objective property of things, space, time, causality, and opportunities that result from all of this. Since cybernetics within science explains the general principles of the process management, regulation, obtaining, storing, transforming and transmitting information, both in living beings and in technical systems, therefore as a scientific discipline explores a dynamic self-regulating and self-organizing systems, while through ontologic patterns also defines system properties. In this way, accessibility is recognized in environment and by means, changes as responses within. Activity as operation is directed towards achieving the goal through takeover of the initiative. Socialization that works together with other systems and people to achieve goals also helps others in achieving theirs. Furthermore, mobility opens the possibility of acting in any environment and lastly, a learning ability based on understanding human intelligence and communicating with machines ensures exceptional availability.

Ontological approach enables expansion and connection with boundary theory areas of a system, management, information, coding, logic, formal languages and grammar, games, mathematical algorithms and programming, and robotics, which include all four existing types of cybernetics:

- Theoretical or general cybernetics,
- Applied or applicable cybernetics,
- Technical / technological cybernetics,
- Economic cybernetics.

Their most important features are the features of the system:

- a great complexity,
- stochasticity and
- auto-regulation.

Recognizing these cybernetic characteristics and features enables ontological observation through its theoretical, technical and economic views.

Theoretical point of view represents the logical structure of hypotheses, knowledge, discoveries, scientific facts, theories and laws. It involves observing, collecting, measuring and rendering data, and lastly, data experimentation, for ensuring optimal conditions under which data, information and knowledge can be obtained. The economic or organizational standpoint provides rational technology intended for further research in all essential elements.

C. Captological approach

The captological approach to cognitive cybernetics includes the cognitive character of persuasive technologies and as a superior scientific field introduces expansion limes on ethical principles. This approach to cognitive cybernetics through its purposefulness in industrial ecology, an emerging scientific interdisciplinary field, unites its development by combining natural, technical and social sciences into a unique system from

global to local level. Foundation of this concept is a cognitive captological integration based on analogy between intelligent technologies, processes in nature (biosphere) and processes in society (technosphere), Figure 8.

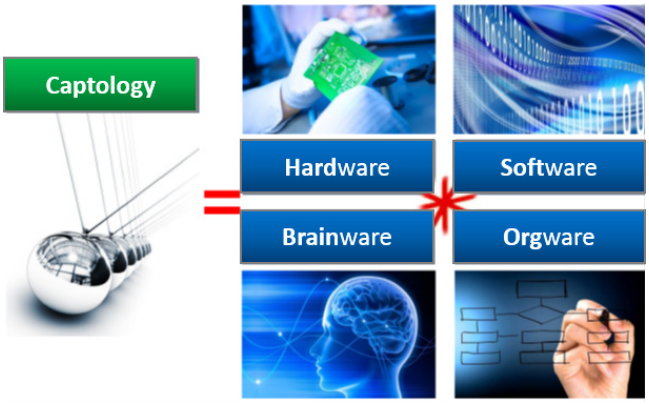


Figure 8 - Captology as a threat to innovation's access and openness

The biosphere, with its laws and principles, refers to coniferous life around the planet, summed up in earthly life as a whole in which all transformations take place. Technosphere is an artificial sphere of all kinds of technology around the planet (all technical and technological aids, creations and systems), with its own artificial time functions in which everything is immersed. Today, the ultimate development and application of intelligent technologies is experienced as a planetary intelligence system made by the Internet, mobile communications, applications and virtual structures of technological simulations. Cognitive cybernetics might enable the creation of awareness of innovations that will drive evolution of life on Earth only in a mental and spiritual direction, [12-15].

3.2. The role of human factors in cognitive cybernetics

Cognitive cybernetics has advanced to such a degree that it can do all the imaginable management and monitoring in a hardware sense. For its efficient application, education is needed to acquire special skills, knowledge and competences required for unerring behavior of decision-makers in approaches to accept innovations and consider scrupulosity. Herein come invaluable characteristics of the ability to use a 'common sense':

- Expressive abstraction and conceptual analysis for quick and successful problem solving in unexpected situations
- Unlimited flexibility.

Certain problems can be identified through persuasion perfidy in the context of treatment and understanding data and information. Research has confirmed that 'third-party tools for experiments on users from database software algorithms evaluate and rank all information accessed through the network - currently 95% for which an average educated user reaches!? Thus, it is not surprising fact that over 300 trials and about 200 signals are continually being run on a daily basis, watched and measured to help maintain 'freshness' and the speed of collection, securing, and transferring data. Although scarcely mentioned, but shows persuasion as useful, the captological arsenal is revealed as:

- Cognitive prejudices, (between more than 170 cognitive prejudices emphasize the 'unwillingness of taking risk in time deficit', and which has been led by persuasive message),
- The weapon of persuasion (this is about persuasive techniques that rely on the 'principle of scarcity' and the 'theory of authority acceptance'),
- Productivity laziness (confirmed by the fact of human mind laziness and the tendency to skip or careless reading which, due to the persuasive guidance, results in unwanted consequences),
- Decision-making fatigue, (associated with the user exhaustion by persuasive over-offerings that encourage the use of lower resistance paths).

The brain capacity for learning is enormous and created for more demanding intellectual activities than the ones commonly included into. The human brain possesses two mutually independent systems of receiving and thinking. The leading one is in the left half of the brain and it analyses, writes, reads, talks and receives logically. The main center for feelings in the right half of the brain is more complete, emotional and associative, or better said - intuitive. Selected activities of the left and right hemispheres of the brain help to its learning by the pairs model (challenges and feedbacks).

The right side of the brain is used for intuitive handling of problems based on two components:

- Simple provisional rules that exploit the proviso to hit the most important information but ignore everything else, hence allowing a quick reaction,
- Predispositions of evolved brain abilities - skills acquired by genetics and learning to transform predisposition into the ability.

As the human mind develops regarding to inexhaustible desire to try and change, not just by wanting to always follow only 'reliable' answers, the question is whether the brain can be stimulated to develop when captologies is not new but uniform, deadening situations.

3.3. Conducted trials

In addition to aforementioned persuasive addictions, the results of 'tests/exercises' can be joined from originally elaborated materials of the course Intelligent systems, conducted at the Zagreb University of Applied Sciences, Electrical Department. It is basically about mechanisms and elements of conclusion through knowledge bases and databases of an expert system. Conclusion is gained by applying acquired knowledge about distributed artificial intelligence in correlation with human intelligence and social predisposition. The task is precisely described by instructions each student receives in the envelope, and before starting, teacher reads the instructions aloud before all

students at the initial location of the exercise. The task is shared to two groups of students who, after receiving the instructions, go at two different locations. Students have drawn attention for having the opportunity to cooperate with each other, which is the goal of exercise, at last. There are five simple questions needed to be answered in manner that all students participate and create common, identical answers. The guideline text is divided into three parts (task description, approach to solving and solution guidelines). Conducting this exercise within four educational periods in three academic years, 148 students participated, and the task was successfully solved not even once.

The reason why such a simple task has not yet been solved lies in the hyper-offer of data (information). Students are mostly taught that use of information technology encourages rapid data inspection. By doing such, reading of the task is reduced to mere browsing and is very superficial, which makes easier overlook of the solving approach. Too much data drives away from deeper and more permanent entry into any single argument, fact or idea. An overwhelming number of information and characters nullify themselves. Students are apt to access the search by going to links (connections) irrelevant to problem solving, but seemingly interesting information is the very beginning of current potential termination of available concentration. This outwardly produces impatience, a lack of attention and the absence of co-operation, respectively. As in reviews and surveys of a vast data, users are prompted to quickly come and go; the same effect was achieved through the lumped text of this task. In this journey of a rapid collection and data transfer, the cognitive efficiency is lost, unfortunately. A lot of confusing information and attention-shifting parameters fragment, hinder and distort the flow of thoughts that end up as distraction. The use of mind is under influence of intelligent technology and therefore under the pressure. Part of the human mind in charge for critical thinking, analysis of reality and which should be used daily as a way of thinking in solving problems, but also in social interactions when it serves to connect with the outside world, doesn't come to the expression in this task, and there was no created understanding – mind is lost in captology for it does not follow the evaluation sequence.

A. An example of understanding when meeting other people

As an example of understanding, a study was conducted related to learning of understanding when meeting other people, already discovered by William Outhwaite, who had been looking at such situations through four categories:

1. The category of physical facts that can be noticed (how people look and what is the consequence of that). These are understandable phenomena in relevant sense, when being seen as a sign of something not physical from the past or something related to the mental state.
2. The category of mind states in others that can be identified now or can be inferred based on the previous comprehension.
3. The category of others' actions (by noticing what others do and what they consider to be doing) is described with three sub-categories:
   a) Recognizing an action by approaching from distance,
   b) Accepting of what is being said about on the same elemental level,
   c) Complex judgment of the real importance of what is being done or what is thought to be said;
4. The category of questions why certain things are being done, and what motivates people to do them.

Latter categories try to establish a difference (especially between the states of third and fourth category), or the differences between motives and the 'hermeneutic understanding'. Psychological understanding of the human mind mental states can be observed based on three conclusions about peoples' motives and their intentions, which are the following: a) visible signs, b) explicit statements, and c) knowledge of the 'situational facts', confirmed by the psychology of the situation that has been discussed since Aristotle, Figure 9. From the aforementioned examples a complexity of the understanding comprehension can be seen, which is theoretically evaluated; and people have a social process of meeting embedded in a sub-conscious algorithm operable in real-time.

B. An example of learning understanding while reading aloud

To confirm the conducted research, it is useful to consider the acquisition of understanding on research example of literacy development. Therein, it comes to conclusion that the central process of learning by reading is understanding the text and active search of meaning and purpose, as established by G. J. Whitehurst and C. J. Lonigan. The reading process or understanding what was read comes down to understanding of coding and decoding, like digital data transmission. This also presents the concept of an expert system functioning, which was followed herein for explanation and elaboration.

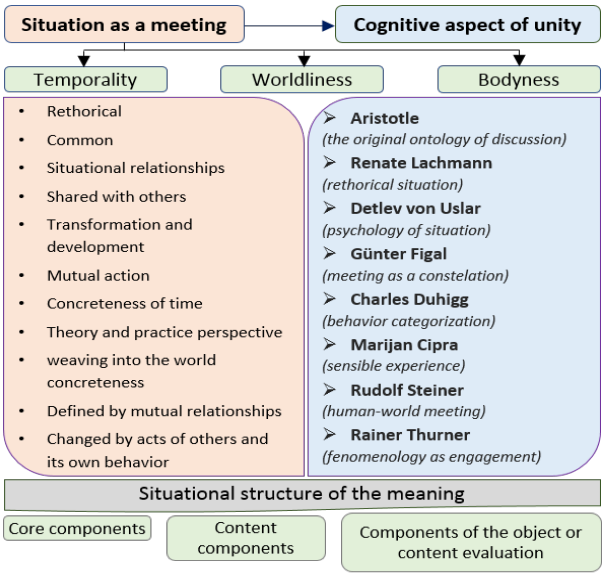


Figure 9 - Analysis of the situation psychology as a situational structure of significance

Researches on reading effects has set the groundwork for reading mechanisms evaluation, which are:

- a comparison,
- a recall from the memory,
- linking what is seen or heard, (or heard before),
- identifying essential data and rejection of redundant ones.

On these grounds, data processing while reading developed by R. J. Marzano and D. E. Paynter happens simultaneously with five concurrent processes, known as a 'Model of 5 babushkas', Figure 10, which comprises:

- Understanding the purpose of reading,
- Estimating the compliance of read data with the purpose,
- The meaning of the smallest thinking units (claims),
- Identifying the meaning of words,
- Recognizing the meaning of a read wholeness.

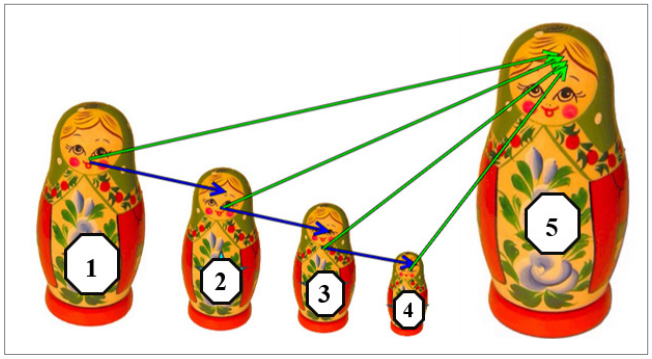


Figure 10 - Data processing while reading, (the model of 5 babushkas)

Reading, listening, watching and pronouncing are the ways of learning and each has a different contribution to learned matter. In accordance with learning mode, the left and right hemisphere along with emotional center of the brain are activated. Understanding the reading comprehension (reading aloud) is an understanding strategy that along with understanding, implies the voice sensitivity and motivation development. The above-mentioned examples, especially data processing while reading, is an important foundation for analyzing different types of reading, such as:

- mature reading,
- reading a linear text - a printed book,
- reading a linear text - e-book,
- reading by text search – a computer,
- reading hypertext – a computer,

and where reading should be distinguished from a mere search, under the influence of captology and turned into an addictive need, because it maintains a worthless dominance over digital information flow. The imposed 'Internet ethics' through different kinds of media results in persuasive addiction manifested by:

- encouraging and treating a 'thirst' for small and short-term information slots,
- curious dependence on a continuous data stream updated in the real-time; an 'updated status' which lost its purpose slightly after a few moments of release, and
- the desire to speed up the flow of information.

Aforementioned manifestations hold not any contribution to storage of learned, because nothing has been learned. Moreover, a disturbance has occurred, and no recognized brain cell nor emotional center has been activated, which today can be safely confirmed by psychometric investigations for more serious (i.e. legislative) use, still deliberately delayed, while minor trials so needed have not yet been clearly recognized and are ongoing through new researches.

C. An example of a thought material storage

In support of the researches conducted, there is also an example of storing thought material in the memory, which requires a careful and balanced information flow. An example can be visualized for the practical acceptance. It is easily possible by photography, Figure 11, to identify a chair (or chairs) that represents a thought material which can be carefully and permanently stored in a visualized mental space.



Figure 11 - The HAZU[1] revival hall, Opatička 18, Zagreb - an example with a careful and balanced storage of thought material

An example of a mess explains a fragmented, disturbed, and a ruined flow of thoughts that end up as a distraction. Hardly anyone will be able from the photos, (Figure 12 and Figure 13), to recognize the thought material at the glance, the 'seat', which should be stored in the permanent memory. Careful and balanced storing of thought material in a chaos is impossible, because from a visualized matter a particular psycho-culture is participated, lately projected in all spheres of life.

The term 'psycho-culture' comes from the culture (lat. *colere* - cultivate, nourish, *cultos* - processing, nourishing, cultivating), which represents an integrated system of attitudes, beliefs and behavior patterns characteristic for the members of society and which are not a result of biological heritage, but a social product created, transmitted and maintained through communication and

learning. The function and meaning of the culture are facilitation, extension and development of the human society.



Figure 12 - The courtyard of the house, Trešnjevka in Zagreb - an example of thought material storing represented by mess and chaos



Figure 13 - The environment of the building, Hrgovići settlement, Trešnjevka in Zagreb - an example of thought material storing represented by mess and chaos

Culture arises from the peoples' need to master the nature and to be free from animalistic life, so that man does not cultivate only external nature but also his own, thus achieving the process of humanization. Under various influences and so under the influence of intelligent technologies, culture is subjective to collective programming of the mind by which members of one community differ from others. Psycho-culture associated with Cicero's cultivation of the soul from ancient times (lat. *cultura anima*), I. Kant evolves and develops it through his 'categorical imperative' as a noble trait, but also recognizes the formation of a collective mentality in society achieved by a dominant ideology and which directs social and political processes. Each person has its own ideology, a system of values and life orientations that direct through life and determine the behavior influenced by intelligent technologies and intuigence (intuition + intelligence), [16-23]. Socio-cultural neuroscience provides better and better insights into complex interactions between socio-cultural and neuronal structures and processes, and thus within biological mechanisms that underpin socio-cultural phenomena and human behavior including all components of Kierkegaard's existentialism.

For these reasons it is useful to point out that the structure of neural network, part of the big brain cortex, represents the biological basis of our mental abilities. In adulthood, new knowledge and skills are easily adopted without changing the structure of the neural network. On the other side, educational, psychological, social and emotional environments during such development change the very structure of the neural network, especially in the brain parts crucial to adopt the most complex brain functions. This includes the affective modulation of emotional expression, the conceptualization of one's own mind, mentalization, cognitive flexibility and a work memory.

4. Continuation of research

The continuation of researches presented in the paper is based on the prepared teaching material, [24-28]. In the new education period part of the Intelligent Systems program is transformed into a new area, [29-34]. This new area will be presented at the 1st International Conference on Intelligent Human Systems Integration: Integrating People and Intelligent Systems (iHSI 2018), which will be held at the JW Marriott Marquis in Dubai, United Arab Emirates, January 7-9th, 2018.

Authors in their work have come up with recognizing:

I. Human Intelligence Systems, (HIS) - at the very beginning of the 1980s, researches in the field of artificial intelligence was focused on systems associated with human intelligence. Their goal was to incorporate the principles of human intelligence into intelligent platforms for data processing through artificial intelligent systems, which in turn resulted and actively contributed to the industry development and advancement of society, such as:
   a) advanced development and use of mechanical systems and devices in industrial production as intelligent autonomous robots,
   b) the introduction of humanoids into the social and palliative spheres of life.

II. Intelligent Human Systems, (IHS) - researches continued in the 1990s, conducted on already developed intelligent systems foundation with the intent of implementation in education, had included among others:
   a) developing and upgrading intelligent interactive technologies for use in systems with algorithms involving the principles of human thinking,
   b) scientific analysis of social activities and human intelligence through mathematical modelling, scientific brain research and cognitive sciences in general.

Such intelligent systems contain the knowledge of certain domains with sophisticated decision-making processes and the ability to explain their behavior. This feature shows the most important aspect of such systems - their ability to effectively interact with people in social environment, teaching, and assisting in availability and complex data manipulation. Furthermore, their use has resulted as a significant impact on human resources with the effect of increased demand for education, [35-40]. Knowledge (when considering Human Intelligence Systems / Intelligent

Human Systems) on the moral plane is designed to achieve what is achieved physiologically by asking the following questions:

1. How does consciousness, starting from the physical plane, finally reach the moral one?
2. How is physical balance a first step towards moral balance?
3. How is every controlled act the path towards improvement?
4. How are the rules of control also the rules of training?
5. How there is no discussion without calmness?
6. How does raising a will as a freedom and reasonable energy give man the freedom, right and task to choose between good and evil?

Although there are some suggestions today, it is more probable that in the near future answers to above questions need to be considered taking into account the parameters listed in Table 1, [41-45].

Table 1 - Parameters of the future considerations and justifications of Human Intelligence Systems / Intelligent Human Systems areas

Genetic neo-construction	Transhumanism	Ultra-Intelligence
Genetic Technology	Singularity	Super-Intelligence
	Extropy	
Biotechnics	Cyberpunk	Historical apokatastasis
Bionics		
Bioethics	Posthumanism	Human-biological planning
Biopolitics	The epoch of hypermodern	Spiritual programming
Bioconservatism	Nanotechnology	

5. Conclusion

Education for acquiring knowledge, skills and competencies will experience drastic changes in the near future because it will have to rely on cognitive cybernetics. In this paper, the results of research are presented as elaborated concepts of internationally recognized ITK institutions that launch new education programs with their partners. The Electrotechnical Department of Zagreb University of Applied Sciences is one of the partners since proven by its approach to educational standards formation and study programs alignment in polytechnic areas with the needs (existing and future) of the labor market, all in the context of globalization and technological development. Regarding to changes of higher education environment and the needs of labor market, it is necessary, as soon as possible, to harmonize them with the principles of cognitive cybernetics. This alignment must necessarily improve latter with introduction of new cognitive education modules because such activity of educational content should be a research for which human factors are at the same level of perfection and reliability as the available cognitive cyber-technology.

When information processing, problem-solving and decision-making are compared with expert system that contains human cognition and mental potential, it is possible to pull a parallel with intuition and intelligence.

The knowledge base is 'cognitively' dependent, associated with symbolic representation, and is identical to intelligence. Applying databases and knowledge bases, it would be useful to subject it to conclusion mechanism. This in biological sense of identifying with the human brain functioning (the way of thinking) suggests that purposeful use of intuition and intelligence require some sort of a skilled strategy – the intuigence, which has been confirmed by outputs of detailed researches carried out from 2008 to date, and which are presented in this paper.

When the education, psychological, social and emotional environment are linked to the presence and intimacy of intelligent systems and interactive technologies, the hypothesis of their multiple influences is confirmed; in a positive sense – through the knowledge of raising the level of integrated human mind development, and the negative one – the influence on people through neuropsychiatric disorders development. Biomedical and sociologically important concept is one of the main subjects of continuous discussion between advocates of selective theory against advocates of constructive theory of mental processing. Nowadays, the question about how to use the mind and intelligent technologies seems to be the same one. As one part of the human mind is in charge for critical thinking and reality analysis, it is used for problem solving but also in social interactions when it serves to connect with outside world. The second part of the mind is used for imagination and can only be triggered when the analytical opinion is excluded. The notion that imagination and analytical thinking are not treated as two opposing forms of thinking, but as two kinds of opinions with different functions, is the result of intelligent technologies impacts, for which human intelligence should set some directions and advices. For these reasons the question is set: does between cognitive cybernetics and captology exist a cooperation or the conflict?

References

[1] Balaž, Z., Predavec, D., "The Captology of Intelligent Systems" in 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, May 2017. http://dblp.org/db/conf/mipro/mipro2017

[2] Balaž Z., Meštrović K., "Learning and teaching from Artificial Intelligence", (article in Croatian with an abstract in English), Zagreb University of Applied Sciences; Polytechnic & Design, Vol. 2, No 1, pp 9-14; ISSN 1849-1995; Zagreb, 2014. http://polytechnicanddesign.tvz.hr/?page_id=363

[3] Balaž Z., Predavec, D., "The Brain lost in the space of persuasive no-placelessness", (article in Croatian), Public lecture and workshop conducted on 16th Brain week in Croatia, at Electrical department of Zagreb University of Applied Sciences, specialist graduate study, Zagreb, Croatia, 15th of May 2017. https://www.bib.irb.hr/867247

[4] Predavec D., "The Captology of Intelligent Systems", Master's Thesis, Zagreb University of Applied Sciences, 2017.

[5] Afrić, V., "Theoretical basics of hermeneutics", (article in Croatian language), Sociological review, Vol XIV, No. 3-4, UDK 167.7:316, pp. 191-200, 1984. https://hrcak.srce.hr/155800

[6] Heidegger, M., "Ontology (Hermeneutics and Factuality)", Bröcker-Oltmanns; Frankfurt a. M., 1988; Introduction to philosophical hermeneutics, Darmstadt 1991. http://www.iupress.indiana.edu/product_info.php?products_id=84761

[7] Wilber K., A Theory of Everything: An Integral Vision for Business, Politics, Science and Spirituality, SHAMBHALA, Boston, 2001.

[8] Lachmann R., Phantasia Memoria Rhetorica, (book in Croatian) Translated by Davor Beganović, Press: Matica Hrvatska, ISBN 953-150-150-560-8; Zagreb, 2002.

[9] Balaž, Z., "Cognitive Cybernetics - a Future Society 5.0", the paper prepared for International Symposium P&P, Zagreb, Croatia, October 2017.

[10] Haun, M., "Cognitive Computing - Steigerung des systemischen Intelligenzprofils", (article in German) ISBN 8-3-662-52951-5, DOI 10.1007/978-3-662-52952-2, Springer - Verlag Berlin, Heidelberg, 2014.

[11] Balaž, Z., Haun, M., "Cognitive Cybernetics – a future that is started" in 34th International Symposium on New Technologies, SOLARIS Hotel Niko, Šibenik, Croatia, 15th -16th May, 2017. https://www.bib.irb.hr/898570

[12] Teilhard de Chardin, J. M. P., L'Energie Humaine, (book in French), Editions du Seuil, Paris, 1962.

[13] Carr, N., The Glass Cage: How Our Computers Are Changing Us, (book in Croatian), Jesenski i Turk, Translated by Ognjen Strpić, Publisher: Znanje d.o.o..; Biblioteka „42"; Zagreb, November 2016.

[14] Detlev v. U., Psychologie und Welt, (book in German) Verlag W. Kohlhammer, Stuttgart, Berlin; Koeln, Meinz, 1972.

[15] Gardner H., Kornhaber M., Wake W., Intelligence: Multiple Perspectives, Hercourt Brace College Publishers, Fort Worth, Philadelphia, San Diego, New York, Orlando, Austin, San Antonio, Toronto, Montreal, London, Sydney, Tokyo, 1996.

[16] Balaž, Z., The Intuigence and Captological Traps of Artificial Intelligence, manuscript of the book at a review process, Zagreb, November 2016.

[17] Sternberg R. J., Beyond IQ: A Triarchic Theory of Human Intelligence, Cambridge University Press, Cambridge, 1985.

[18] Sternberg R. J., The Triarchic Mind: A New Theory of Human Intelligence, Viking, New York, 1992.

[19] Anderson M., Intelligence and Development: A Cognitive Theory, Blackwell, Oxford, 1988.

[20] Figal G., Der Sinn des Verstehens, Beiträge zur hermeneutischen Philosophie, (book in German) Philipp Reclam jun., Studgart, 1996.

[21] Adorno T. W., Ästetische Theorie, (book in German) R. Tiedemann, Gesammelte Schriften, Frankfurt. M., 1970. - 1973.

[22] Gadamer H. G., Wort und Bild – So Wahr, So Sein, (book in German), Gesammelte Werke, Tübingen, 1993.

[23] Gadamer H. G., The Legacy of Europe; MH, Edition: PARNAS, Press: Targa, Zagreb, 1997.

[24] Balaž Z., Meštrović K., Haun M., Cognitive Cybernetics - Propaedeutic of Polytechnic, the book manuscript on review, prepared issue of teaching material for Manualia Polytechnici Studiorum Zagrabiensis, Zagreb University of Applied Sciences, November 2017.

[25] Balaž Z., Predavec D., "Integration of Cognitive Cybernetics into Intelligent Human Systems?", scientific article prepared for the 1st International Conference on Intelligent Human System Integration, Dubai, 7th – 9th of January 2018.

[26] Balaž Z., "Implemented research study", course: Artificial Intelligence - Expert Systems, Electrical engineering department, Zagreb University of Applied Sciences, Specialist graduate study, 3rd semester, September - December 2013.; and course Intelligent Systems, 2nd semester, September 2015. – December 2016.

[27] Balaž, Z., "Intuigence and captology of Intelligent technologies – challenges and fears", (article in Croatian), sent on Polytechnic & Design journal, Zagreb University of Applied Sciences, October 2016.

[28] Parmy, O., "Jawbone guinea pig economy", FORBES, Vol XIV, No 3-4, UDK 167.7:316, pp. 63-68, III. 2015.

[29] Faucault, M., "The Courage of Truth", (article in Croatian) lectures from Collège de France, (1983.-1984.), ISBN 978-953-7715-61-8, translated from French: Zlatko Wurzberg, Standorf & Mizanotrop, Zagreb, 2015.

[30] Hoesle, V., "Filozofija ekološke krize", (article in Croatian) Moscow lectures, Publisher: Matica Hrvatska, Editor: Jelena Hekman, Press: Targa, ISBN 978-953-150-051-7, Zagreb, 1996.

[31] Grondin, J.: Sources of Hermeneutics, (book in Croatian) Publisher: Matica Hrvatska, Editor: Jelena Hekman, Press: Targa, ISBN 953-150-512-8, Zagreb, 1999.

[32] Jsselsteijn, W.I., Kort, Y., Midden, C., Eggen, B., Hoven, E., Persuasive Technology, ISSN 0302-9743, ISBN 3-540-34291-5, Springer - Verlag Berlin Heidelberg, LNCS 3962, 2006.

[33] Matulić, T., Metamorphosis of culture, Tertium mullenium, ISBN 9789-5324-1161-14, CIP 686343, Zagreb, 2009.

[34] Bubner, R., Izvornik, (book in Croatian) Aestetische Erfahrung, Suhrkamp Verlag, Frankfurt am Main, 1989. Translation: Estetsko iskustvo, Tihomir Engler, MH, Editor: Jelena Hekman, Biblioteka PARNAS, Press: Targa, ISBN 978-953-150-076-0, Zagreb, 1997.

[35] Carr, N., The Shallows: What the Internet Is Doing to Our Brains, (book in Croatian) Publisher: Jesenski i Turk, Translation: Ognjen Strpić, Zagreb, 2011.

[36] Kahneman, D., Thinking, Fast and Slow, (book in Croatian), Translation: Zvonko Pavić, ISBN 978-953-14-1483-8, CIP 857332, Publisher: Mozaik knjiga, Zagreb, 2013.

[37] Heisenberg, K., W., Physics and Philosophy, (book in Croatian), Translation: Stipe Kutleša, Kruzak, Press: S. Brusina – Donja Lomnica, ISBN 953-96477-3-8, Zagreb, 1997.

[38] Voltaire, Treatise on Tolerance, (book in Croatian), Translation: Bosiljka Brlečić, MH, Editor: Jelena Hekman, Biblioteka PARNAS, Press: Targa, ISBN 953-150-082-7, Zagreb, 1997.

[39] Rohr, R., The Naked Now: Learning to See as the Mystics See, (book in Croatian), Edition: Synopsis, Zagreb – Sarajevo, 2016.

[40] Petanjek, Z., "Sinaptički izazovi u djetinjstvu i adolescenciji: Izazovi i smjerovi suvremenog istraživanja mozga", (lecture in Croatian), HAZU: Symposium on Research and Brain Diseases, Zagreb, 16th March, 2016. http://www.hiim.unizg.hr/index.php/24-naslovnica/334-simpozij-hazu-2016-baw

[41] Jakovljević, M., "Mozak, duša i kultura u zdravlju i bolesti: Izazovi pred psihijatrijom 21. stoljeća", HAZU: Symposium on Research and Brain Diseases, Zagreb, 16th March, 2016. http://www.hiim.unizg.hr/index.php/24-naslovnica/334-simpozij-hazu-2016-baw

[42] Goleman, D., Emotional Intelligence, (book in Croatian), Translation: Damir Biličić, Editor: Ivanka Borovac,, Publisher: Mozaik knjiga, 2nd edition, available online, Zagreb 1997.

[43] Hood M. B., SuperSense: Why We Believe in the Unbelievable, (book in Croatian), Translation: Marko Maras, ISBN 978-953-303-175-0, Publisher: Ljevak, 2010.

[44] Paić, Ž., Posthuman Condition: The End of Human and Odds of Other History, (book in Croatian), Publisher: Litteris, ISBN 978-953-7250-41-6, Zagreb, October 2011.

[45] Paić, Ž., The Spheres of Existence: Three Studies on Kierkegaard, (book in Croatian), Publisher: Matica Hrvatska, ISBN 978-953-341-081-4, Zagreb, April 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 109-114 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Impact of Crosstalk on Signal Integrity of TSVs in 3D Integrated Circuits

Shadi MS. Harb*[1], William R. Eisenstadt[2]

[1]*Intel Corporation, Hillsboro, OR 97124, USA*

[2]*University of Florida, Department of Electrical & Computer Engineering, Gainesville, FL 32611, USA*

ARTICLE INFO

Article history:
Received: 02 November, 2017
Accepted: 24 December, 2017
Online: 30 January, 2018

Keywords:
3D Interconnect
Crosstalk
Cross-coupled
TSVs
3D Test
Electrical Characterization

ABSTRACT

Through-Silicon-Vias (TSVs) are utilized for high density 3D integration, which induce crosstalk problems and impact signal integrity. This paper focuses on TSV crosstalk characterization in 3D integrated circuits, where several TSV physical and environmental configurations are investigated. In particular, this work shows a detailed study on the influence of signal-ground TSV locations, distances and their structural configurations on crosstalk. Embedded 3D testing circuits are also presented to evaluate the coupling effects between adjacent TSVs such as crosstalk induced delay and glitches for different crosstalk modes. Additionally, A 3D parallel Ring Oscillators testing structure is proposed to provide crosstalk strength coupling indicator between adjacent TSVs. Simulation results are conducted using a 3D electromagnetic field solver (HFSS) from Ansoft Corporation and a Spice-like simulator (ADS) from Keysight Technologies Corporation based on MIT 0.15µm 3DFDSOI process technology.

1. Introduction

3D Interconnect is the promising technology [1]-[4], which includes Through-Silicon-Vias (TSVs) to connect vertically stacked semiconductor chips with shortest paths, which means lowest inductance and conduction loss, to both signals and power supplies. In spite of these benefits, the signal integrity issues in TSVs become the major challenges in 3D designs [5-6]. The goal of TSVs or 3D Vias development is to acquire high chip density. Therefore, the density of the 3D Vias is also high. In this environment, a crosstalk problem appears between two adjacent signal 3D Vias (Aggressor and Victim). Studies show that the coupling problem is not negligible in TSVs because of the relatively large diameter and small pitch, which results in non-negligible TSV-to-TSV coupling that degrades significantly the 3D circuit performance. Hence, it becomes very essential to precisely model and evaluate the electrical characteristics of TSVs [7]-[9] to analyze signal integrity (SI), and crosstalk of adjacent TSVs under the conditions of various structures and configurations.

*Corresponding Author: Shadi MS. Harb, Intel Corporation, Hillsboro, OR 97124, USA | Email: shadiharb@ieee.org

In this paper, the electrical characteristics of 3D interconnect, based on our previous work [10], is presented to characterize signal integrity effects of 3D crosstalk for different TSVs placement and configurations. 3D Vias based on 0.15 µm 3DFDSOI process from MIT Lincoln lab [11] are used as a device under test (DUT) for crosstalk characterization where a 3D full wave simulator such as HFSS from Ansoft Corporation is used to extract and predict the electrical characteristics of TSVs in the frequency domain (S-Parameters) and a Spice-like simulator such as ADS from KEYSIGHT technologies to evaluate the TSVs transient response in the time domain (Eye-diagram). Additionally, embedded 3D testing applications are proposed to characterize the TSV's signal integrity effects and the impact of TSVs on the 3D circuit performance after fabrication. A 3D circuit test is presented to evaluate the coupling effects between adjacent TSVs such as induced-delay and glitches [12-13] for different crosstalk modes. Additionally, a consecutive triggered parallel Ring Oscillators (ROs) testing structure is proposed to provide a crosstalk coupling indicator between adjacent TSVs.

The paper is organized as follows: section 2 discusses the 3D Full wave modeling for TSV and the simulation setup. A detailed study of crosstalk for different physical and environmental TSVs configurations is given in Section 3. 3D Crosstalk embedded

testing applications are given in Section 4. Section 5 concludes the paper.

2. A 3D Full Wave Modeling for TSVs

In order to evaluate the electrical characteristics of a TSV depending on structural parameters such Via pitch, Via height, and Via size, the 3D interconnect based on MIT 0.15µm 3DFDSOI technology is used as DUT to model and characterize crosstalk in different testing configurations. The vertical connection in this technology is slightly different from the standard Through-Silicon-Via (TSV), which is a square shape via made from Tungsten material and fully surrounded by oxide; thus, it is simply called a 3D Via. The 3D Via pitch is around 3.325 µm (distance between the centers of two 3D Vias), 7.34 µm TSV height, 1.25 µm x 1.25 µm TSV size. The physical size of 3D Via after fabrication is estimated to be around 2 µm for the top dimension and 1 µm for the bottom dimension. The 3D Via was simulated using a 3D full wave simulator (HFSS from Ansoft Corporation), which generates the S-Parameters of the structural model of the Via and a Spice type simulator (ADS from Keysight Corporation) to predict the electrical characteristics of 3D Vias in the time domain (Eye-diagram). Figure 1 presents a pair of TSVs structure using the HFSS simulator.

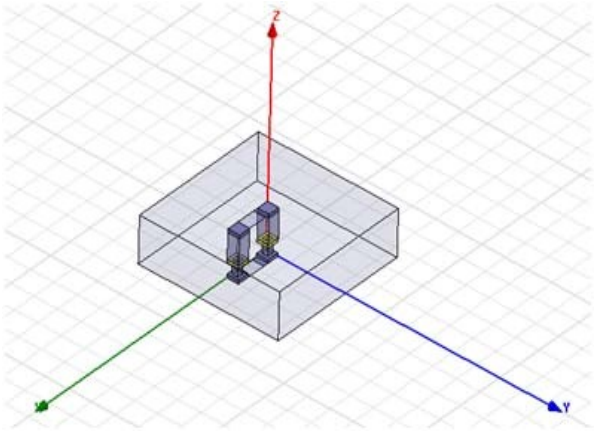


Figure 1. Physical Structure of a pair of TSVs using 3D full wave simulator

Usually, an interconnection line is characterized using S-Parameters. S11 and S21 are the S-parameters reflection and transmission coefficients respectively, which are typical characteristics of an interconnection. The evaluated S11 and S21 magnitudes for the electrical characteristics of a 3D Via based on the default parameters are shown in Figure 2.

The transmitted data stream through the 3D Via was simulated with the evaluated S-parameters from the 3D full wave simulator (HFSS). The Eye-diagram of the transmitted data stream was evaluated for 10^7-1 pseudo random bits sequence (PRBS) using the Spice type circuit simulator (ADS). The source for the simulation has 1.5 Vp-p and 50Ω source termination. The 3D Via is terminated by a shunt connected 50 Ω resistor and 1pF capacitor. The Eye-diagrams of 2 Gbps, and 10 Gbps PRBSs are shown in Figure 3. Also, all PRBSs were assumed that they have 10% rising and falling times

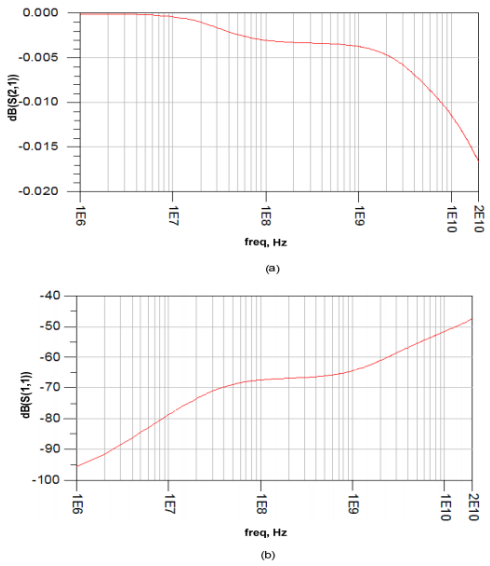


Figure 2. Evaluated S21 magnitude (a) and S11 magnitude (b) of 3D via using HFSS

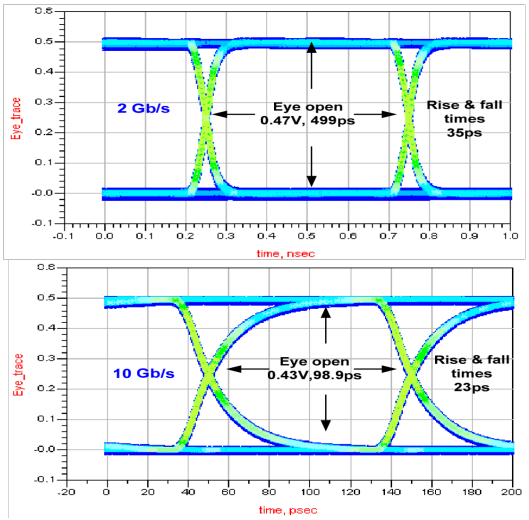


Figure 3. The simulated eye-diagrams of 2Gbps (a), and 10 Gbps (b) PRBSs after passing through the 3D Via

3. A Crosstalk between 3D Vias

3.1. The Influence of 3D Via Locations, and Distances on Crosstalk

Crosstalk is evaluated depending on the distance of two signal Vias and the location of two GND Vias as shown in Figure 4.

Five distances (4 µm, 8 µm, 16 µm, and 32 µm) based on SGSG configuration between the two signal Vias have been simulated. As expected, if the distance between the two signal 3D Vias is larger, the crosstalk level is going down as shown in Figure 5. The effect of the distance of GND Vias with respect to the signal via on the crosstalk is also evaluated as shown in Figure 6 with four distances: 4 µm, 8 µm, 16 µm, and 32 µm. The results show an increase in the magnitude of the crosstalk as the distance of the reference via (GND) increases. Also it is shown in Figure 7 that the crosstalk magnitude of SGGS (i.e. the cross locations of

the signal and ground Vias) configuration is smaller than that of SGSG configuration comparing two same distance cases. The difference of the two cases is almost 10dB. This is a very interesting point for 3D designers to keep in mind, because just changing the 3D Via role can reduce the crosstalk magnitude especially in high frequency applications, where crosstalk problem is very critical to obtain the maximum system performance.

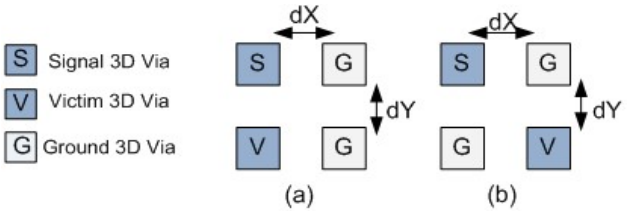


Figure 4. Location of signal Vias and GND Vias for crosstalk evaluation, (a) SGSG, (b) SGGS configurations

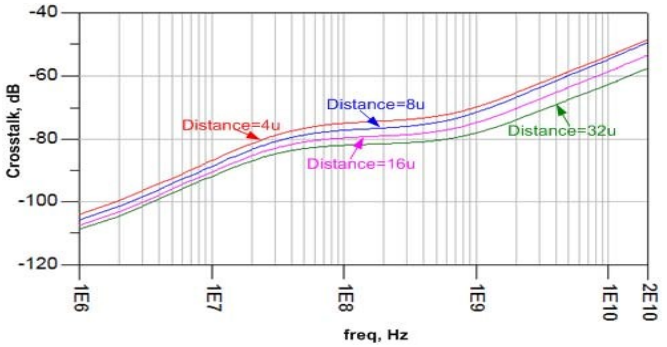


Figure 5. Comparison of five locations of 3D Vias. The distances 4 µm, 8 µm, 16 µm, and 32 µm between two signal Vias are associated to the graphs from top to bottom respectively

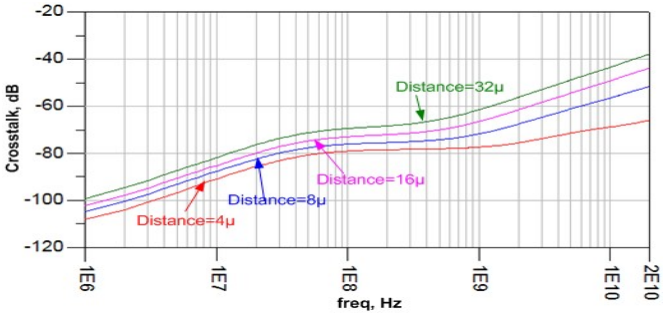


Figure 6. Comparison of three locations of GND Vias. The distances 4 µm, 8 µm, 16 µm, and 32 µm between two GND Vias are associated to the graphs from bottom to top respectively

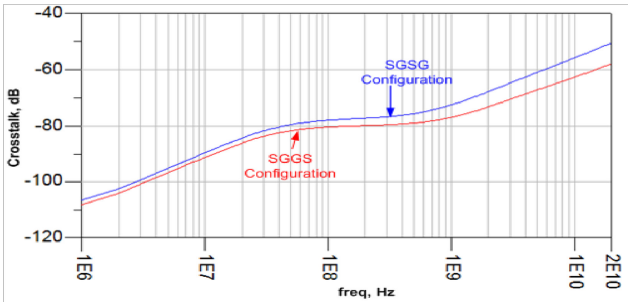


Figure 7. Comparison between SGSG and SGGS configurations for 10 µm distance, SGSG and SGGS are the top and bottom graphs respectively

3.2. 3D Via Crosstalk in Structures with Different Configurations

The geometry of the return current path may be one of the most efficient methods to affect the crosstalk between a signal and victim Via. Four different configurations have been investigated (1) Two Via pairs in a straight line (Figure 8(a)), (2) Two Via pairs placed opposite to each other (Figure 8(b)), (3) A signal Via with two reference via placed opposite to a victim Via with two ground Vias (Figure 8(c)), (4) A signal and victim via, each with three reference Vias as a return current path (Figure 8(d)).

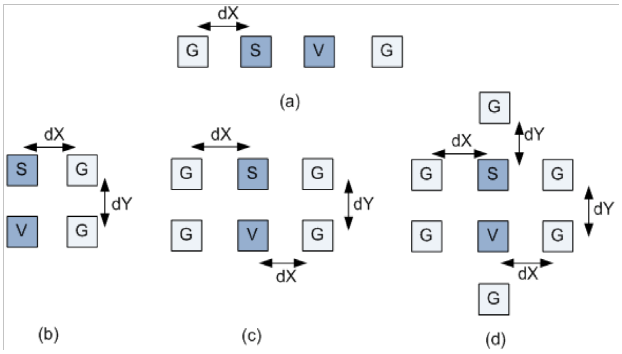


Figure 8. Comparison of three different configurations, (a) Configuration #1, (b) Configuration #2, (c) Configuration #3, and (d) Configuration #4

In configuration 2, the signal Via has a slightly lower inductance than it would in configuration 1 because the second Via is close enough to the signal Via to have a slight impact on its inductance. Configuration 4 has the lowest inductance because it has the most well defined return current path. This lowering of the inductance will also lower the near-end crosstalk as shown in Figure 9. Also it is shown that the highest crosstalk as predicted, comes from configuration 1. Only slightly lower is the crosstalk from configuration 2. Then, there is a significant decrease in crosstalk when the extra two reference Vias are added in configuration 3, and a slight decrease further when the third reference Via is added to the victim and signal Vias in configuration 4. The reduction in crosstalk from adding additional reference Vias is almost 4 dB.

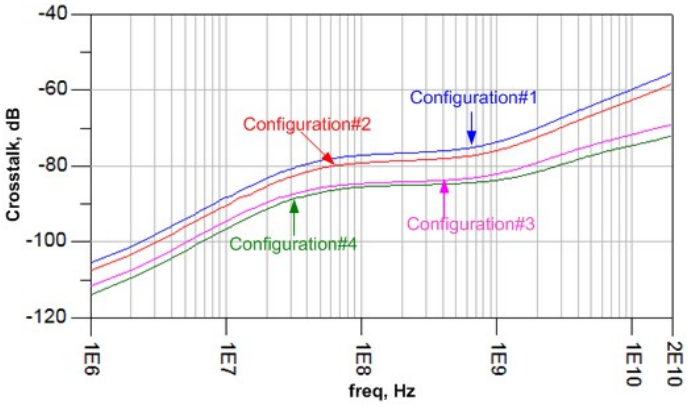


Figure 9. Crosstalk between four different configurations

Until this point, only isolated 3D Vias with respect to ground Vias have been considered. Potential coupling in the 3D Vias in the discontinuity between the 3D via and the transmission line may be an important effect to consider. In Figure 10, a new Ansoft HFSS model is presented that accounts for the discontinuity between the 3D via and the transmission line.

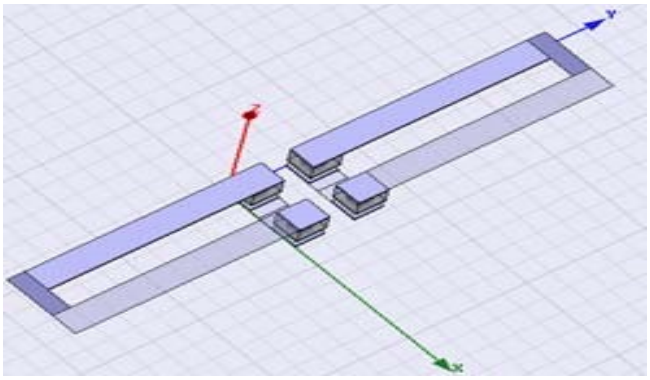


Figure 10. A TSV coupling model that includes the transmission lines

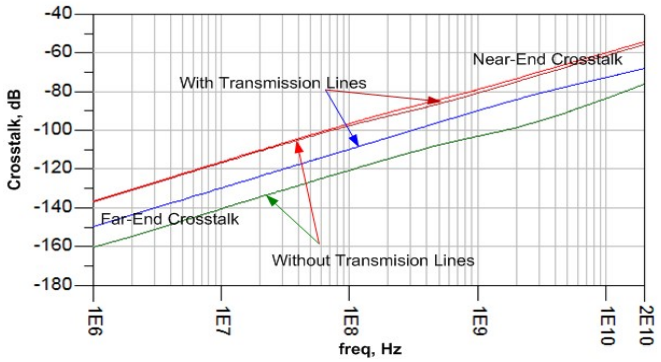


Figure 11. Comparison of crosstalk between TSVs with and without the inclusion of transmission lines

Figure 11 shows that, according to the full wave simulations, there is no measurable difference between the near-end crosstalk as a result of the discontinuity between the transmission lines and the 3D Vias. The difference between the two simulations is very small over the entire frequency range. However, the discontinuity between the transmission lines and 3D Vias will increase the far-end crosstalk. On the other hand, the far-end crosstalk is more minimal than the near-end crosstalk, less significant and never exceeds -50 dB as shown in the simulation.

4. 3D Embedded Crosstalk Test Applications

4.1. 3D Testing Circuit for Crosstalk Induced-Delay and Glitches

In this embedded test application, the coupling effects between adjacent TSVs such as induced-delay and glitches can be investigated for different crosstalk modes. As shown in Figure 12, high speed signals can be fired through three adjacent TSVs at each tier using a multi-edge delay generator circuit. A Mux and Tristate circuits are used to control which signals are active from which tier. The complementary signals are also generated from the delay generator to cover different crosstalk modes. In order to study the effect of phase shifting the aggressor signal on crosstalk induced-delay cancellation, the multi-edge delay generator is used to fine control the delay between adjacent signals.

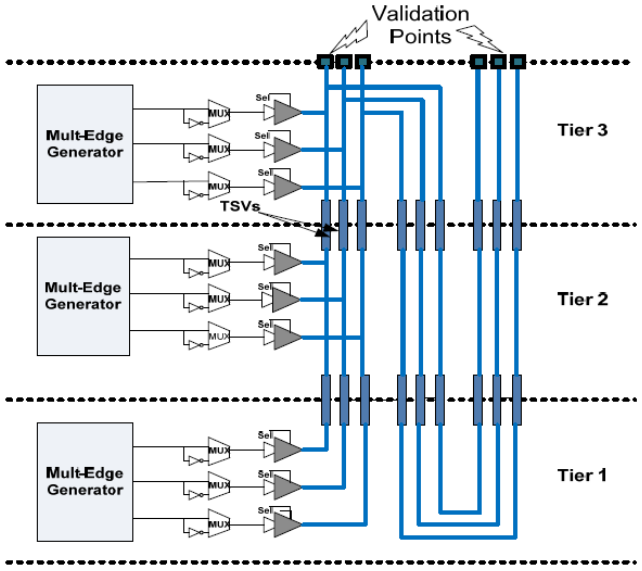


Figure 12. 3D Interconnect crosstalk induced-delay test

Figure 13 shows different crosstalk modes for studying crosstalk coupling on the victim edge. For example; -2X presents the case when both aggressors transition in the same direction as victim. -1X: when one aggressor transitions in same direction and other is quiet. 0X: when both aggressors are quiet or transition in opposite directions or the victim is quiet. +1X: when one aggressor transitions in opposite direction and other is quiet. +2X: when both aggressors transition in opposite direction as victim.

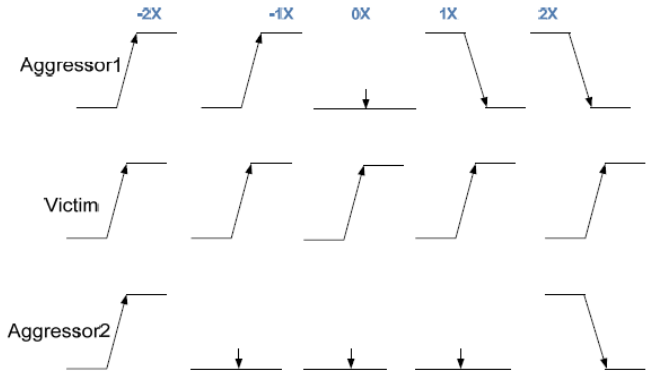


Figure 13. Different Crosstalk Patterns

Figure 14 shows simulated results for the induced-delay crosstalk with different patterns using ADS and 0.25 µm CMOS standard process. The middle graph is the victim line signal with no activity on both aggressor lines. The furthest right and left graphs are -2X and 2X cases respectively, which cause the worst case induced-delay effect. This induced-delay can be mitigated by phase shifting the aggressor signals using the multi-edge delay generator. Figure 15 shows an example of the crosstalk induced-delay cancellation effect after phase shifting the aggressor line 0.8ns for the case of +1X. As graphed, the induced delay due to crosstalk is almost cancelled and the signal aligns again with the 0X case (i.e. without crosstalk).

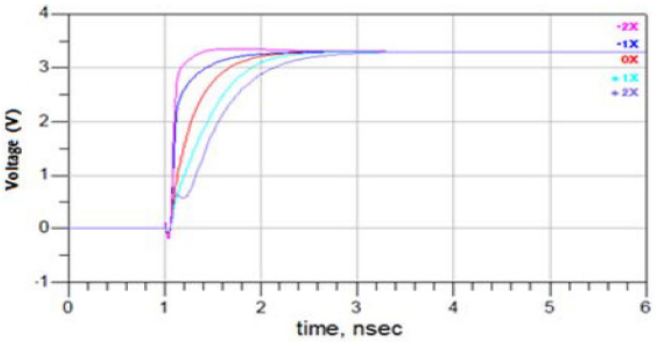


Figure 14. 3D interconnect induced-delay crosstalk for different crosstalk patterns

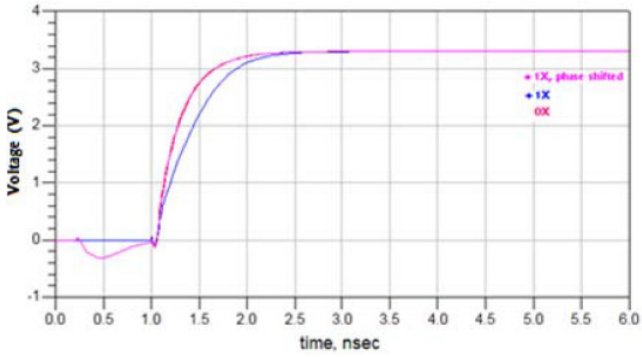


Figure 15. 3D interconnect induced-delay crosstalk cancellation

4.2. Ring Oscillators 3D Crosstalk Test

In this test, a consecutive triggered parallel Ring Oscillators (ROs) structure running same frequency is used to characterize the crosstalk effect between TSVs. Figure 16 shows four triggered oscillators; two oscillators are crosstalk-coupled and the other two are crosstalk-free. The proposed ROs parallel structure creates a delta phase shift difference between each consecutive triggered oscillators, which is equal to the time difference between the delay buffer chain and the oscillation time period of the triggered oscillators. 3D crosstalk detection can be achieved by observing the frequency of crosstalk-coupled oscillators, which is different from the frequency of the crosstalk-free oscillators.

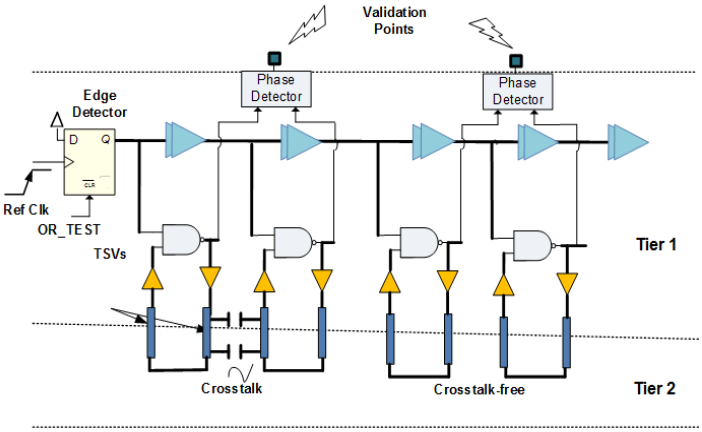


Figure 16: 3D Crosstalk Detection Scheme

Figure 17 shows the output frequency for both crosstalk-coupled (725 MHz) and crosstalk-free oscillators (667MHz).

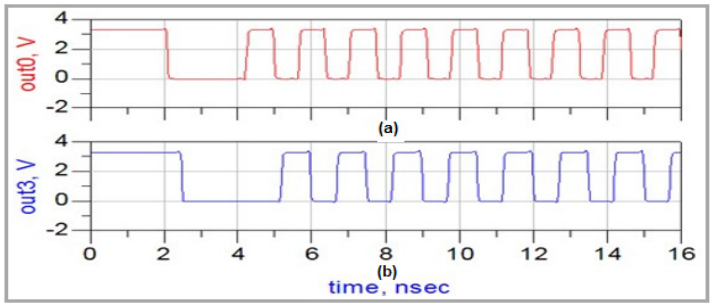


Figure 17: Output frequency for (a) Crosstalk-coupled and (b) Crosstalk-free oscillators

Interestingly, the two crosstalk-coupled triggered oscillators have less oscillation time delay (i.e. faster frequency) than the crosstalk-free oscillators, and the phase difference between the two crosstalk-coupled oscillators diminishes due to the coupling effect as shown in Figure 18. Phase detection at the output of the crosstalk-coupled triggered oscillators can be used as an indicator for strong coupling between TSVs. On the other hand, the edges of crosstalk-free oscillators are still separated by a deterministic phase shift dictated by the time difference between the delay buffer chain and the ring oscillation time period.

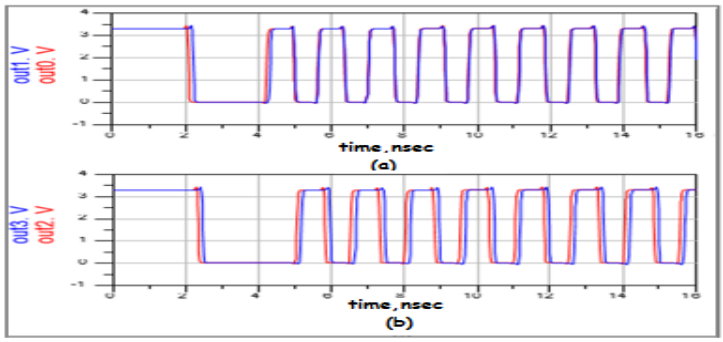


Figure 18: Simulated Results for 3D Interconnect Crosstalk (a) Crosstalk-coupled (b) Crosstalk free

5. Conclusion

In this paper, we presented the signal integrity effects of crosstalk in 3D stacked ICs. A detailed study of TSVs electrical modeling and characterization using HFSS and ADS simulators for frequency and time domains analysis respectively was presented. Simulation results were conducted based on 0.15µm 3DFDSOI process technology from MIT Lincoln lab which present the influence of 3D Vias distances, locations and their structural configurations on crosstalk. The study shows that increasing and decreasing distance of 3D Via signals and grounds respectively can mitigate significantly the effect of 3D crosstalk. In addition, adding more reference Vias and creating a well-defined return current paths have the most impact on mitigating crosstalk. Furthermore, it was shown that the discontinuity between 3D Via and transmission line has negligible impact on near-end crosstalk (NEXT), however; far-end crosstalk (FEXT) might increase but with less significant impact.

Furthermore, a 3D testing circuit application based on a multi-edge signal generator placed at different 3D stacked tiers was studied to evaluate the effect of crosstalk induced-delay and glitches. Additionally, a cross-coupled parallel ROs structure was

also presented to evaluate the crosstalk coupling strength effect compared to ROs structure with crosstalk-free TSVs.

References

[1] A. W. Topol, D. C. L. Tulipe, J. L. Shi, D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini and M. Ieong, "Three-dimensional integrated circuits," IBM Journal of Res. and Dev, vol. 50(4-5), pp. 491–506, 2006.

[2] J.-Q. Lu, "3-D hyperintegration and packaging technologies for micronano systems," Proc. IEEE, vol. 97, no. 1, pp. 18–30, Jan. 2009.

[3] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. D. Franzon, "Demystifying 3D ICs: the pros and cons of going vertical," *IEEE Design & Test of Computers*, vol. 22(6), pp.498–510, 2005

[4] J. Q. Lu, "3-D Hyperintegration and Packaging Technologies for Micro-Nano Systems," in Proc. of IEEE, vol. 97(1), 2009, pp. 18-30

[5] K. Tu, "Reliability challenges in 3D IC packaging technology," Microelectronics Reliability, vol. 51, pp. 517–523, 2011.

[6] T. Song, C. Liu, D. H. Kim, S. K. Lim, J. Cho, J. Kim, J. S. Pak, S. Ahn, J. Kim, and K. Yoon, "Analysis of TSV-to-TSV coupling with high-impedance termination in 3D ICs," in International Symposium on Quality Electronic Design, 2011.

[7] J. Tanskanen, J. Toikka, and E. O. Ristolainen, "Crosstalk of wiring in very small 3D module: IMAPS," 36th International Symposium of Microelectronics, pp. 251-255, 2003.

[8] T. Kang; Z. Yan; W. Zhang; J. Wang, "Research on crosstalk issue of through silicon via for 3D integration", *28th IEEE International Systemon-Chip Conference (SOCC)*, 2015.

[9] T. M. Mak, "Test challenges for 3D circuits," 12th IEEE International OnLine Testing Symposium (IOLTS), pp. 79–79, 2006.

[10] Shadi MS. Harb, William Eisenstadt, "A Study of Characterizing Crosstalk Effects in 3-D Vias," Accepted at the IEEE Latin American Symposium on Circuits and Systems Conference (LASCAS 2017), Bariloche,, Argentina, Feb 20, 2017

[11] MITLL Low-Power FDSOI CMOS Process Design Guide, MIT Lincoln Laboratory, 2008.

[12] Pedram S N M, "Crosstalk-affected delay analysis in nanometer technologies", International Journal of Electronics, Sept 2008.

[13] Gope D and Walker D M H, IEEE 30th International Conference on. IEEE," Maximizing crosstalk-induced slowdown during path delay test", p.159-166 , 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 115-124 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Constructing Learning-by-Doing Pedagogical Model for Delivering 21st Century Engineering Education

Ghassan Frache[1], Hector Nistazakis[2], George Tombras[*,2]

[1]*Abu Dhabi Vocational Education & Training Institute, United Arab Emirates*

[2] *Division of Electronics, Computers, Telecommunications and Control, Dept of Physics, National and Kapodistrian University of Athens, 15784 Athens, Greece*

A R T I C L E I N F O

Article history:
Received: 05 November, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
Higher education
Learning-by-Doing (LBD)
Pedagogical
21st Century skills
Pragmatic

A B S T R A C T

The United Arab Emirates (UAE) is dedicated to establishing the best teaching and learning environment for students and staff across all higher education institutions. To heed this call, the engineering division adopted the Learning-by-Doing (LBD) pedagogical philosophy for 21st Century education at the heart of its strategic directions. This study intends to explore how LBD is understood and practiced in UAE colleges and how 21st Century skills can be explicitly incorporated into its engineering curriculum by using constructive alignment as a pattern for instructional design. This work intends to investigate the general question: "What constructively aligned Learning-by-Doing pedagogical model, with the incorporation of 21st Century skills, needs to be developed to effectively teach and prepare engineering students at the Higher College of Technology, UAE for successful long-term employment in the global working economy?". Using mixed method research design and input from its major stakeholders, student survey questionnaires, engineering instructors and dean structured interviews, overt class observation and syllabi examination have all been utilized for data triangulation. In conclusion, the study developed a collaborative LBD model tailor-fitted to the institution's engineering program and to UAE's culture.

1. Introduction

This study of students' and instructors' perception of the implementation of Learning-by-Doing and 21st Century skills was conducted in Higher Colleges of Technology, HCT, a tertiary education institution in the United Arab Emirates (UAE). This research is an extension of work presented at the 2017 Global Engineering Education Conference [1]. Additionally, the work is part of thesis work for doctoral studies. The 2003 Emiratisation program, developed for the purpose of training Emiratis to become competitive, has been strictly enforced by His Highness Shaikh Mohammad Bin Rashid Al Maktoum, the Vice President and Prime Minister of the UAE (2012) [2]. The Emiratisation program is aligned with the UAE Vision 2021[3], which aims to develop and train Emiratis for taking up jobs in the country. Among other things, Vision 2021 foresees higher education as an environment where students will "enrich their minds with the skills that their nation needs to fuel its knowledge economy" [3, p. 16]. Thus, it can also be observed that part of Vision 2021's directive is for the UAE universities to "listen closely to the needs of Emiratis and of their future employers" and to "balance their teaching to the demands of the workplace" [3, p. 16]. Teaching methods and approaches need to be aligned with this requirement of matching the students' learning with practical needs of jobs in future, and Learning-By-Doing is one method that is gaining a reputation for doing just that [4]. This study is therefore conducted in a tertiary educational institution in the United Arab Emirates (UAE), with the aim of evaluating the pedagogical approach in the context of the framework of 'Learning-By-Doing'.

The concept of a more authentic, relevant learning has been a focus for educators since the time of John Dewey in the early part of the last century. Dewey's concept of "learning by doing" was based on his understanding that people learn best when they are actively involved in tasks that have meaning and importance [4]. In this last century, education has also been shaped by our growing understanding of how people learn. Johnson, Johnson and Holubec [5] argued that the work of Vygotsky in the 1920s, Jean

*Corresponding Author : G.S.Tombras, National and Kapodistrian University of Athens, 15784 Athens, Greece, gtombras@phys.uoa.gr

Piaget's cognitive development stages, Bloom and his now famous taxonomy, Howard Gardner's multiple intelligences and other related research studies have deepened our understanding of human learning, which, as [4] advocated should involve all aspects concerning the mind, hands and heart. Additionally, the principle of Learning-By-Doing is manifested in many of today's learning theories. Experiential learning, under which falls-active learning (whose subsets include cooperative learning and collaborative learning) and service learning – all exemplify the principle of Learning-By-Doing [6]. According to Voogt and Roblin [7, p. 29], in order for students to be successful in the 21st century, not only are certain skills necessary but these specific skills also need to be taught through (and are best supported by) "specific pedagogic techniques, such as problem-based learning, cooperative learning, experiential learning, and formative assessment", which again is encompassed in the approach of Learning-By-Doing [6].

The study aims to answer the overarching research question: What constructively aligned Learning-by-Doing pedagogical model, with the incorporation of 21st Century skills, needs to be developed to effectively teach and prepare engineering students at the Higher College of Technology, UAE for successful long-term employment in the global working economy? The main objectives of the research are as follows:

- To analyze the current understanding of LBD from the viewpoint of the engineering college's dean and instructors.
- To identify what LBD practices are being successfully implemented in the engineering department, from the perspectives of instructors and students in HCT.
- To identify what 21st Century skills are intentionally or incidentally taught and assessed in the practice of LBD, from the perspectives of instructors and students in HCT.
- To identify what pedagogical model might best meet the needs of implementing LBD and 21st Century skills in the engineering faculty at HCT.

The following sub-questions were formulated to guide the two phases of the study and provide conclusions:

RQ1. What are the current understandings of LBD from the viewpoint of the engineering dean and the instructors?

RQ2. From the perspectives of engineering instructors and students, what LBD practices are successfully implemented in the engineering department?

RQ3. From the perspectives of instructors and students, what 21st Century skills are taught and assessed in the practice of LBD?

RQ4. What pedagogical model might best meet the needs of implementing LBD and 21st Century skills in the engineering division at HCT?

2. Research Locale and Participants

The UAE is identified by PISA (Programme for International Student Assessment) as one of the most rapidly improving education systems in the world [9]. The UAE government recognizes that in 21st century economies, knowledge and skills have become so essential that a high value is placed on building a world-class education system that nurtures the minds and hearts of the UAE citizens. The government of UAE plays a dominant role in education, for schools and universities. It typically controls curricula even at private schools [9]. Most instructors are government employees, and most education is publicly financed up to the degree level. There are several types of post-secondary institutions in the UAE. A university offers degrees in several disciplines and usually offers graduate studies, while university colleges focus on one major discipline and are mostly for bachelor degrees. Technical institutes offer two- or three-year diplomas and are common throughout most of the region, accounting for one-third of all post-secondary students.

The UAE has more than 30 foreign academic institutions. The establishment of those institutes is a manifestation of the globalization of higher education and an indication of some of the UAE's attempts to become a knowledge-based society. Public universities in the UAE have separate colleges for males and females, following Islamic traditions. Such is the case with HCT, the locale of the present study.

HCT is one of the public's higher education institutions in the UAE, established in 2006. All the engineering programs are internationally recognized by accrediting bodies specifically the American Accreditation Board for Engineering and Technology (ABET). Because the proposed model is specific to the engineering department of HCT, the participants are its stakeholders: the engineering dean, the instructors teaching major courses in engineering and the engineering students. Only the instructors who have signed an informed consent to participate in the research were interviewed and observed. Table I shows the number of participants who were invited and participated in the study.

Table I: Participants of The Study

Colleges	Year 1	Year 2	Year 3	Year 4	Instructors	Dean
MZC Colleges (women)	18	15	0	0	2	
RUC colleges (women)	31	16			3	
RUC colleges (men)	30	41	32	18	3	
Total	79	72	32	18	8	1

3. Significance of the Study

The researcher believe that while LBD may be translated into various strategies that call for student-centered activities, an explicit definition based on shared understanding, contextualized to how LBD should be applied in the HCT colleges, would make its practice more focused. This has the potential to greatly contribute to achieving the college's mission of providing educational experiences that will infuse its graduates with "the knowledge, skills, and attributes to effectively contribute to the nation-building process and to help them develop a sense of personal and social responsibility". (http://www.hct.ac.ae/about/learning-model).

Additionally, the researcher sees the viewpoints of HCT major stakeholders as vital in developing explicit foundational commitments from the college community in lieu of one solely determined by higher management and handed down to teachers

for mandated implementation. Hence, the researcher takes into account the teachers, the students, and the management's viewpoints. It is anticipated these findings will contribute to the development of an explicit model that will help the teachers focus methodologies, assessments, and other pedagogical activities not only to the principal elements that comprise LBD but also for the promotion of learned 21st Century skills developed amongst students.

4. Implications for the Present Research

Although the literature does not provide any explicit definition of LBD, it links it to a number of important elements that ultimately promote its philosophy. This gives a more definitive idea of what it involves and a basis for investigating how the HCT instructor individually understands and practices the philosophy of LBD under the guidance of the college. The following is an analysis of what LBD involves:

- Learning to do (skills) not just to know, [10] - [11];
- Learning that is experiential [11], [12], active [10], [1],[13], [14], collaborative and cooperative [15];
- Learning that occurs in the context of a goal that is relevant and interesting to the student [2], [16];
- Learning that is planned (not discovered) [16], [2];
- Learning that involves not only quality academic reflection[17], [18] but civic (global) and personal experiences as well [18];
- Learning that considers the students' cultural context [19], [20] and respects every student's experience and builds upon these [11], [16];
- Learning that involves practical experiences in the context of relevant tasks closely related to how students will use it outside the learning environment [1], [5];
- Learning that involves strategies such as presentation, reports, team building, online contact time with students, critical thinking, studio teaching, team projects, and open-ended problem solving [20], [21] - [22];

The literature reviewed leaves no doubt as to the importance of 21st Century skills in today's education, particularly to engineering education. The review has also shown that a number of 21st Century skills take their roots from Dewey's (1899) work, among others.

Also, as mentioned in the research reviewed, over and beyond foundation knowledge taught through content engineering courses, the following are deemed indispensable 21st Century skills applied to an engineering education. The researcher adopts Mishra and Kereluik's re-categorization of 21st Century skills in presenting these.

1. Meta knowledge
   - critical and self-critical abilities/problem solving, [23] - [25];
   - communication and collaboration, [23] - [26];
   - teamwork [23] - [27];
   - creativity/innovation [23];
2. Humanistic knowledge
   - life and job skills [23], [25], [26]
   - ethics and cultural knowledge [26];

Interestingly, the LBD elements as listed above align with a number of the 21st Century skills reflected in the HCT learning model. To fully explore the alignment between LBD elements and 21st Century skills necessary in engineering education as outlined in the learning outcomes of HCT, a brief analysis by the researcher is presented in Table 2.

Table 2: Alignment Between LBD Elements, 21st and HCT Learning Model

LBD elements	21st Century skills	HCT learning outcomes
• Learning to do (skills) not just to know (factual knowledge); experiential, active, cooperative, collaborative learning	creativity/innovation life and job skills/communication and collaboration teamwork	Critical and creative thinking Vocational competencies Communication literacy Teamwork and leadership
• Learning that occurs in the context of a goal that is relevant, meaningful and interesting to the student	life and job skills communication critical and self-critical abilities/problem-solving	Vocational competencies Communication literacy Critical and creative thinking Self-management and independent Learning
• Learning that is planned	all skills should be planned learning outcomes	HCT learning model mandate
• Learning that involves quality reflection	critical and self-critical abilities ethics and cultural knowledge	Self-management and independent learning Global awareness and citizenship
• Learning that considers culture	ethics and cultural knowledge	Global awareness and citizenship
• Learning that involves practical experiences in the context of relevant tasks closely related to how students will use it outside the learning environment	creativity/innovation life and job skills communication and collaboration teamwork	Critical and creative thinking Vocational competencies Communication literacy Teamwork and leadership Mathematical literacy
• Learning that involves strategies such as presentation, reports, team building, critical thinking, steam projects, and problem-solving	communication and collaboration teamwork critical and self-critical abilities	Critical and creative thinking Teamwork and leadership Communication literacy Mathematical literacy

Biggs [28] instructional design, constructive alignment, aside from fully supporting LBD as it espouses experiential learning, sees the usefulness of an LBD structured system of learning in engineering education [29], [30]. The researcher finds this significant in contributing directions on how to go about model construction, giving particular focus to Biggs' alignment of learning outcomes, teaching and learning activities, and assessment.

In developing the model which this study proses to do, the researcher should keep in mind that although LBD is mandated as the teaching principle of Higher Colleges of Technology and is promoted in Biggs' instructional design [28], conventional methods of teaching still have a place in the classroom [23], [31] - [33]. Hence, the researcher notes that some lessons can be taught through interactive, high-level psychological lectures and activities rather than behavioral, experiential ones.

Additionally, following Houghton [34] and Fung [29] propositions, criterion-based assessments will be looked into while keeping in mind Voogt and Roblin's recommendation that while summative and formative assessments are useful in assessing 21st Century skills, new forms of assessment should build on previous assessment practices and should be considered as a starting point.

5. Research Methodology

As mentioned, HCT has mandated the use of LBD and 21st Century skills learning, having adopted these at the heart of its strategic directions. At this point, it is important to delve into how LBD is implemented and how 21st Century skills are infused especially from the collective perspectives of the different stakeholders – students, instructors, and the dean.

Keeping in view the pragmatic stance that reflects the belief that human experiences are multifaceted, and that agreement concerning what is and what should be in any given situation is best negotiated among all concerned parties, the mixed method approach was selected as the best-suited method for this study. The selection of the mixed method approach was in keeping with the stance of Venkatesh, Brown, and Bala [35] who state that "such work can help develop rich insights into various phenomena of interest that cannot be fully understood using only a quantitative or qualitative method" (p. 21). Creswell and Plano Clark [36] claim that using both qualitative and quantitative methods enables the researcher to reach a more comprehensive response to the research questions, as opposed to the possible limitation of using only one method. Using the mixed methods design in this research has enabled reliable and valid data to be elicited, as it helped the researcher in obtaining data from different sources using different approaches – in-depth interviews, observations, quantitative surveys, and document analysis. Migiro and Magangi [37] state that there are three important advantages of this method:

(1) A researcher can use strengths of an additional method to overcome the weaknesses in another method by using both in a research study;

(2) Mixed method research can provide stronger evidence for a conclusion through convergence and corroboration of findings;

(3) Mixed method research can add insight and understanding that might be missed when only a single method is used (p. 3763).

Hussein [38] states that using multiple ways of data collection in the same study increases the credibility of the study (p. 1). He asserts that using mixed methods achieves more accurate and reliable findings with richer information.

The use of mixed methods is therefore grounded in the researcher's belief that findings from each set of participants (students, teachers, and the dean of engineering) will give not only a richer depiction of the findings but also a unique description of what LBD practices and 21st Century skills are currently implemented at HCT.

Quantitative Tools

Quantitative methods are often used in social science to acquire knowledge by manipulating data through sophisticated quantitative approaches, such as multivariate statistical analysis [22]. The engineering students were requested to repond using Likert-scale questions by employing an online survey framework that was used by the researcher. Keeping in mind participant fatigue, the questionnaires were divided into two sets: LBD elements and 21st Century skills. The sets were administered separately, a week after the other.

Qualitative Tools

The qualitative data collection was selected in order to offer the researcher an in-depth perspective regarding the personal experiences. Reference [23] provided a succinct definition that "qualitative research seeks to discover new knowledge by retaining complexities as they exist in natural settings." Reference [24] provides a description of a variety of data collection tools used in qualitative studies: interviews; data analysis; direct observations; and reporting. The choice of tool is influenced by the skill of the researcher, data collection strategy, the type of variable, the accuracy required, and the collection point [21].

Data Analysis

The following data analysis approach was shaped by the type and amount of data collected and made extensive use of description in addition to a number of statistical analyses of the quantitative data. The primary data sources, survey questionnaires, semi-structured interviews, content analysis and over classroom observations, were analysed and reported in terms of patterns in participants' responses in relation to the main and enabling research questions that guided the study.

1) Quantitative Data Analysis

To test whether the actual values are significantly different from the expected values, the chi-square test was applied to the LBD and 21st Century questionnaire. A null hypothesis signifies that there is no statistically significant difference and an alternate hypothesis states the opposite. Based on the results of the above

test, the researcher can either reject or fail to reject the null hypothesis. The techniques and methods used in this study encompass the following: The descriptive analysis by using the mean and standard deviations was conducted for the items determining the 21st Century skills.

2) Qualitative Data Analyses

To analyse the qualitative data, the audio recordings of the interviews and transcripts of the recordings were entered into nVivo software for coding and further analysis.

6. Data Analysis and Results

This section discusses the findings stemming from the quantitative and qualitative data collected. The quantitative methodology tools used are questionnaires given to one particular type of stakeholders, specifically the engineering students. A descriptive analysis was carried by using the frequencies and proportions. The Likert-scaled data of Learning-by-Doing items were tested for statistical significance by using a Chi-squared test. A descriptive analysis using the frequencies and proportions was conducted for the items determining the 21st Century skills. The qualitative methodology tools used are semi-structured, one-on-one interviews among the two sets of target participants: the engineering instructors and the dean.

A. Descriptive statistic Learning-by-Doing survey findings

The first questionnaire was administered to the engineering students seeking to understand how LBD is manifested inside the classroom. Students were given a set of 16 statements and then asked to rate the frequency of these statements. The statements were geared to correspond to LBD principles as opposed to LBD techniques. The survey questionnaire was administered to 184 engineering students. Students' ratings of each of the statements are listed in Table 3. Frequencies, Means, and Standard Deviation data analysis are shown. In addition, a graphical representation for the rating of LBD statement is shown in figure 1.

The first statement provides emphasis on activities done in the classroom which emphasizes collaboration as its main ingredient. Only 26% of the participants stated that this is always the case with 32% stating this is usually done. The second statement aims to find out whether there is a meaningful interaction between students and instructors during discussions. 32% stating this is always done. The third statement emphasizes reflection on activities conducted. This is one of the hallmarks of Learning-by-Doing. More than 50% of the students said it was done at least 4 times in six situations. The fourth and fifth statements are related to each other as they tackle the issue of learning "life skills," another characteristic of LBD. We can safely say more than 50% of students are saying these are being done 4 out of 6 situations. The 6th statement's aim was to ferret out whether "Active Learning" is happening in the classroom. This statement garnered the lowest number 5 rating (12%). The 7th statement refers to the use of simulation in the classroom. The results on this question are encouraging considering it has a high rate of "Usually" (38%). The demonstration of a concept is the emphasis of the 8th statement. Looking at the results, it can be seen that it has the highest points garnered in the data spectrum of "Always" (30%). The 9th statement belongs to the purview of "Active Learning," conducting drills and practice. The results show that this statement has the highest frequency number for the rate of "Usually," garnering a whopping 44%. The 10th statement is part of the reflection phase of LBD where students are supposed to retrospect on what they have learned. The combination of the "Usually" and "Always" frequency score however still ensures that more than 50% of the participants believe that this happens. The 11th statement touches the post-evaluative aspect. LBD encourages different ways of evaluation to encourage the participation of the learner or learners in this phase. We can still conclude that more than 50% of students believe that this happens 4 times out of 6 situations. The 12th statement is part of the real-life learning of LBD were activities that encourage more "learner" immersion is an integral part. Suggested by the statement itself are workshops and field trips. This statement got the highest "Never" and "Almost Never" frequency score. The 13th statement is still within the context of real-life learning. The 14th statement is within the purview of "Active Learning," with at least 50% of the participants believing that this occurs in 4 out of 6 situations. The 15th statement is within the coverage of post-evaluation and reflection. Students are asked to record their impressions on a phase by phase basis. Almost 60% of the participants believe that this happens in at least 4 out of 6 situations. The 16th statement focuses on hand-ons, kinesthetic activities. This statement got the highest "Occasionally" rate, having 40%.

Table 3: Summary of Results for Each LBD Statement

	LBD statements	Frequency %					Sample Size	Mean	SD.P
		Always	Usually	Occasionally	Almost never	Never			
1	There are classroom activities that require students to **collaborate** and learn with and from each other. Examples are **group projects** that emphasize **teamwork**.	26	32	26	12	4	184	20	10.2
2	Discussions in the classroom are **interactive** meaning students as well as the teacher contribute to the topic being discussed.	32	32	25	9	3	184	20	12.0
3	Questions and answers that focus o**n post-evaluation** of learning activities are conducted in the sessions.	25	33	27	10	5	184	20	10.7
4	Exam questions are focused on **scenarios** that require students to apply what they have learned and are not merely limited to ones that call for memorization, definitions, etc.	21	34	30	10	4	184	20	11.4
5	The teacher uses **real life** case studies as a means for teaching the content of the course.	20	32	33	11	5	184	20	11.1
6	Students are presented with **problem-based** questions where students either in group or individually will work out the solutions.	12	37	35	12	4	184	20	13.4
7	Teacher uses **simulation** either digital or manual as a means of teaching a concept.	22	38	30	7	3	184	20	13.3
8	Teacher **demonstrates** a required subject skill first then asks the students to follow suit.	30	35	26	7	2	184	20	13.1
9	Students **do drills and practice** as a means of learning and **mastering a skill** or a concept.	29	44	19	7	2	184	20	15.1
10	Students are encouraged to **reflect** on what they have learned and express this reflection either orally or in written format.	21	34	33	9	3	184	20	12.5
11	In assessing student's work, the teacher uses other means in addition to his/her own assessment. This other means can be **self-assessment or peer review.**	15	36	30	11	7	184	20	11.2
12	Teachers conduct activities that allow students to fully experience the topic. Examples of these type of activities are **field trips and workshops.**	16	29	33	14	8	184	20	9.4
13	The college provides programs that bring students to the **workplace** as part of the student's preparation for professional working life after graduation.	15	34	32	11	8	184	20	10.8
14	Classroom activities that ask the students to **model experiences** or concepts are conducted. Examples of these types of activities are **role-playing,** reenactment or walkthrough (From process to output)	17	33	35	9	5	184	20	12.2
15	Teachers encourage students to **record their impressions** on how they made the project on a **phase-by-phase** basis. This requirement is in addition to the required output of the project.	29	31	26	11	3	184	20	11.1
16	Classroom activities are formulated in such a way that students can be more **active and motivated** in their work. Examples of this type of activity are educational games and other hands-on means.	23	28	36	9	4	184	20	11.9

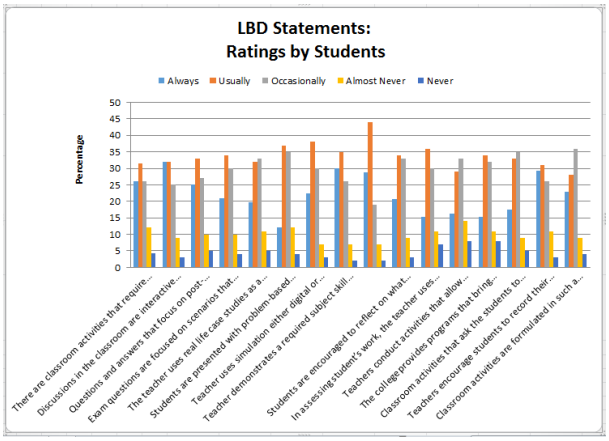


Figure 1. LBD Statements

B. Chi-Square Test on Learning-by-Doing statements

In order to determine whether there is a significant difference between the opinions of the participants regarding the LBD statements, a chi-square test for equal proportions was applied. The null and alternate hypotheses are as follows:

Null hypothesis: There is no significant difference between the opinions of the participants regarding each statement.

Alternate hypothesis: There is a significant difference between the opinions of the participants regarding each statement.

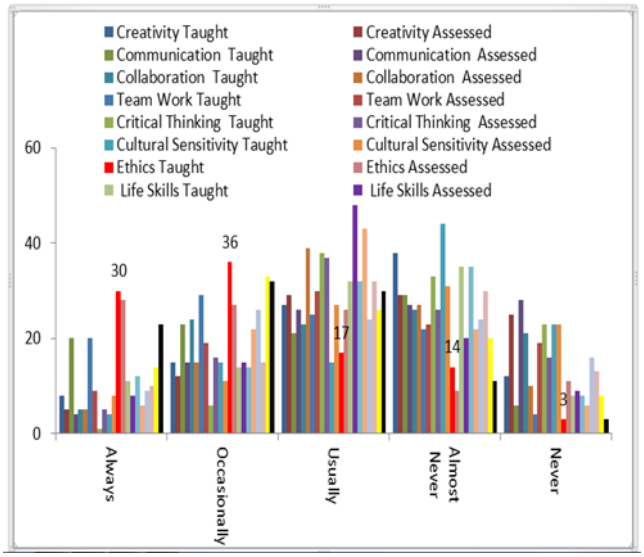
Since the p-value is less than 0.05, we can conclude that there is little significant difference between the participants' opinions.

C. 21stCentury skills survey findings

To answer the 3rd question of the study, the student's point of view with regard to whether 21st Century skills are either intentionally or inadvertently taught or assessed in the classroom was addressed. The manifestation of the skills was established in the questionnaire to crystalize the concept in the mind of the students. Students rated each category based on a 1 to 5 frequency scale. Findings of the survey results are shown in Table 4.

Table 4: Summary of Findings for 21st Century Skills.

	21st Century Skills		Frequency %					Sample Size	Mean	SD.P
			Always	Occasionally	Usually	Almost Never	Never			
1	Creativity	Taught	8	15	27	38	12	132	20	11.0
		Assessed	5	12	29	29	25	132	20	9.8
2	Communication	Taught	20	23	21	29	6	132	19.8	7.6
		Assessed	4	15	26	27	28	132	20	9.3
3	Collaboration	Taught	5	24	23	26	21	132	19.8	7.6
		Assessed	5	15	39	27	10	132	19.2	12.3
4	Team Work	Taught	20	29	25	22	4	132	20	8.6
		Assessed	9	19	30	23	19	132	20	6.8
5	Critical Thinking	Taught	1	6	38	33	23	132	20.2	14.6
		Assessed	5	16	37	26	16	130	20	10.8
6	Cultural Sensitivity	Taught	4	15	15	44	23	128	20.2	13.3
		Assessed	8	11	27	31	23	132	20	9.0
7	Ethics	Taught	30	36	17	14	3	132	20	11.7
		Assessed	28	27	26	9	11	132	20.2	8.4
8	Life Skills	Taught	11	14	32	35	8	132	20	11.2
		Assessed	8	15	48	20	9	131	20	14.7
9	Problem Solving	Taught	12	14	32	35	8	130	20.2	11.1
		Assessed	6	22	43	22	6	129	19.8	13.6
10	Innovation	Taught	9	26	24	24	16	131	19.8	6.4
		Assessed	10	15	32	30	13	131	20	9.1
11	Computer Literacy	Taught	14	33	26	20	8	132	20.2	8.8
		Assessed	23	32	30	11	3	132	19.8	11.2



Creativity: It is observed that 50% of the students think creativity as a skill is being taught and assessed only once in 6 situations. It is interesting to note that about 54% of the participants graded 1 (never) or 2 (almost never) for they now think outside the box by exploring alternative ideas and solutions. It is interesting to note that creativity taught and assessed received 8% and 5% respectively. This shows very little emphasis is devoted to creativity in the classroom by the teachers.

Communication: Communication skills reached the 43% in terms of students believing that it happens 4 out of 6 times in the category of taught and assessed. It would seem that communication is another skill that is not well assessed with 55% of the students thinking that never/almost never done is happening here.

Collaboration: Looking at the taught and assessed categories, we can see that only 5% of the participants think this is done 6 times out of 6 situations and 44% think this is done 4 out of 6 situations. We can conclude that collaboration is not taught nor assessed in the classroom

Team work: Can also be considered as one of the top skills in terms of scoring as it has a more than 40% rating in the "Usually" and "Always" categories. Just like in the other skills it means 50% of the students believe it is happening 4 out of 6 times.

Critical thinking: This skill scored lowe in both categories. These results support the obtained data from the instructor interviews about the teaching and assessment of the critical thinking skill. It is observed that 56% of the students think critical thinking as a skill is being taught and assessed only 1 time in 6 situations.

Cultural sensitivity: This skill received the highest score of 44% for "Almost Never." This highlights that almost half of the participants think that cultural sensitivity is rarely being taught in class.

Ethics: Ethics is one of the top skills that the student consistently rated high. It has 30%, "Always" score based on the taught category, achieving the highest score among all the skills.

Life Skills: Life skills received an average score in both categories. In the taught category Life Skills received 14% in "Usually" and 11% in "Always."

Problem Solving: The problem-solving scores 12% in the taught and 14% that this happens 4 times out of 6 situations. This is surprising result given that most of the exercises in the engineering department require laboratory and math problems. It is observed that about 35% of the participants rated "Occasionally" for classroom activities that require students to examine different processes or paradigms and apply these to different types of problems.

Innovation: Similar to critical thinking and creativity, innovation received a low percentage in "Always." It has 9% in the taught category and 10% percent in assessed category.

Computer Literacy: In the category of "taught" alone, the combination of "Usually" and "Always" is 47% which means almost 50% of the participants agree that it happens at least 4 out of 6 times.

D. Learning-by-Doing as explained by participants

The study presents a discussion on the analysis of the results obtained from the primary research. The study aimed at evaluating the understanding of LBD at Higher Colleges of Technology (HCT). UAE has taken evident measures to enhance the quality of education, in order to prove its educational capabilities on the global platform. Hence, in order to assess if the government is successful in its endeavor, it was important to assess how HCT approaches and implements LBD and 21st Century skills. The research therefore aimed at understanding, from the perspectives of the teachers, students, and the dean of the Higher Colleges of Technology what constitutes of LBD and 21st Century skills, and which activities were in the process of implementation under LBD and 21st Century skills. This is part of primary research that was undertaken by conducting using semi-structured interviews (of teachers and the dean) and surveys (of students). In addition, eight classroom observations were also used to assess first-hand the activities employed in classroom and the approach of teachers while teaching. The classroom observations were then analyzed in conjunction with the interview and the survey findings. In addition, an exhaustive analysis of the HCT curriculum was also undertaken in order to review the LBD and 21st Century related content and activities present in it.

The participants' answers can be grouped into four different classifications; mainly, the practical aspect, real world impact, and definition of LBD. From the data gathered, it can be inferred that the participants have diverse views on Learning-By-Doing practices, and it is stated to have a different meaning depending upon the type of course(s) the faculty is teaching and prior professional experience of the participant educator. Table 5 reflects the answers of the participants when asked about their understanding of LBD.

In the concept of engineering, it was claimed by one of the participants that LBD is more about working on "analytical thinking, thinking and formulating a problem, and understanding the practical application of concepts." Another participant stated that, he looks at Learning-by-Doing as a "fairly permissive term and that it is primarily focused on experiential learning." The same view was shared by the engineering dean where he stated "my understanding on learning-by-doing is it's a fairly lenient term. So the first thing I would say is it hasn't got an exact definition."

Table 5. Summary of findings of LBD understanding

Findings	Occurrence
Different meanings in different context	1
Includes critical thinking	1
Laboratory work	1
Relevant learning experience	1
Embrace all other types of pedagogy	1
Hands on applicability	1
Old concept	1
Practical aspect of learning	4
Integral part of engineering	3
Supports theory	2
Relating to real world	4
Increase productivity of students	1
Involves careful design of activities that allows good foundation for knowledge.	1
Provides an opportunity to explore and develop new ideas.	1
Problem solving	1
Practical aspect	4
Real world	4
Impact of LBD	5
Definition of LBD	2

Almost all educator participants suggested during the interviews the need for a clearer and concise definition of LBD to be produced by the institution. Specifically, as one participant stated - "we need a more expansive definition of Learning-By-Doing that captures what distinguishes as well as what unites all members of engineering divisions in a shared educational enterprise." Table 6 captures the common understandings of LBD highlighted during the interviews.

Table 6. LBD defined

Findings	Frequency	Percentage
Experiential learning	2	25%
Life-long learning	1	12%
Problem solving	1	12%
Applying theoretical knowledge	3	37.5%
Apply learning to real life	3	37.5%
Project-based	1	12%

Almost all participants highlighted the importance of the union of theory and practice with an emphasis on the later. The view of marrying both theory and practice is not something new as it has been part of the LBD research literature tradition. For instance, Dewey [25] suggested the move from formal, abstract education to one that is more experienced-based. At the core of Dewey's LBD philosophy is action. Rather than merely thinking about abstract concepts, LBD involves a direct encounter with the phenomenon being studied. It utilizes actual experience with the phenomenon to validate a theory or concept [25]. It should be emphasized and pointed out that most of the participants believed that LBD is not a "stand-alone" philosophy or practice but instead intrinsically tied in with making the theory more relevant and understandable.

The experiential learning definition received the second highest percentage. This is well in agreement with widely

published literature on LBD and in its roots of formulation as a concept and action. Experiential learning and its subsets, cooperative learning and collaborative learning, exemplify the philosophy of LBD [26]. However, as verified in the interview and follow up questions, most of the participants understood experiential learning as a type of "hands-on" learning that does not really emphasize the promotion of discussion, critical thinking, reflection, introspection, and retrospection. In this aspect, the experiential learning that most of the participants subscribe to has something to do with involving the student in the learning experience in order for the learner to understand the concept, thus this is more akin to active learning [13].

Participants were asked, "What LBD activities do you do or have you done to teach your students?" Diverse practices were encountered in classrooms to achieve LBD. Table 7 summarizes these findings.

Table 7: LBD Practices

LBD Practices	Frequency	%
Laboratory activities/ experiment	5	21.7
Case study	1	4.3
Project-based	3	13
Building	1	4.3
Industry visit	2	8.6
Design process	1	4.3
Demonstration	5	21.7
Peer Demonstration	1	4.3
Exploration	1	4.3
Modeling and non-routine problems	1	4.3
Problem-based	2	8.6

The highest observed practices are laboratory activities and demonstration with 21.7% each. With respect to hands-on courses, the instructors are observed to assess the students and train them in LBD through practical experiments and labs. Certain other answers inferred that experiential learning is not only experiental or lab work. "I think having a reflective analysis and self-criticism or guided criticism of that experiential learning exercise is something that is often missed," the dean of engineering noted. It was also inferred that one of the biggest challenges is the understanding that pedagogy is the philosophy.

E. Participants view on 21st Centure practices

Participants were asked to answer several questions on how 21st Century skills were adopted in their classroom activities. The engineering dean said, "UAE engineering students need to possess them in order to survive in the current corporate world." He used the term "professional skills" for 21st century education to highlight those listed skills. Ongoing research findings reveal that to answer today's challenges, students must have the capacity to apply knowledge in practice by learning to adapt to new situations; critical and self-critical abilities; the ability to use teamwork and communication skills, with these being listed as the top three competencies needed by 21st century engineers [27]. The dean went on to highlight the opportunities and challenges of engineering education in the 21st century, contending that the new professional engineer not only needs to be knowledgeable in his own discipline but also needs a new set of professional skills. Table 8 summaries the 21st Century skills findings.

Table 8: 21st Century Skills Findings

Participant	21st Century skills understanding	21st Century skills used	Assessed
Dean	The participant has a clear understanding of 21st Century skills. He even defined some of the skills to include theirelements in the assessment of students.	50% of courses use the 21st Century skills.	Not really, but should be included with clear rubric during lesson and assessment.
Teacher 1	It seems that the interviewee is unsure of the definition of those skills. He struggles to distinguish between terms such as creativity and innovation.	Innovation, creativity, team work, problem-solving, communicati on, ethics.	It does not seem that the teacher is aware of how to assess his students on any 21st Century skills. None of these skills are clearly assessed.
Teacher 2	It seems the participant is unclear on the definition of 21st Century skills.	Team Work, collaboration , and cultural sensitivity.	Yes, informally. Not with a set of clear rubric.
Teacher 3	The participant was aware of the 21st Century skills and answered some questions about their definitions	Creativity, collaboration , teamwork, critical thinking, problem-solving and innovation.	Yes. In a rubric. Official rubric tool was supposed to be presented to the interviewees.
Teacher 4	"I don't understand what you mean by communication." She seems to understand all the other skills.	Problem-solving, critical thinking, teamwork, collaboration , and life-skills.	Not all are assessed. Problem-solving, critical thinking, teamwork, and collaboration are hard to assess. But others I do assess in LBD activities.
Teacher 5	Don't understand the cultural sensitivity. Some confusion about the definitions of some skills.	All skills are used in LBD activities.	All skills are assessed except ethics and critical thinking. No rubric is available.
Teacher 6	The participant is clear about the definition of the 21st Century skills.	Creativity, communicati on, collaboration , teamwork.	All are assessed. Some might have an indirect assessment. No rubric presented.
Teacher 7	The participant seems to understand 21st Century skills.	Teamwork, collaboration almost all of the 21st Century skills.	All indirectly assessed. No rubric is available.
Teacher 8	The participant was not aware of the distinction between innovation and creativity. Also, he did not understand the use of life skills in classrooms.	Collaboratio n, team work, ICT and life cultural sensitivity is used.	None of the skills are assessed directly.

It is obvious that not all participants understood what 21st Century skills are, let alone what the definition is or what the included activites are. This is despite the fact that a working definition of the skills was sent to them before they started teaching their courses at the beginning of the semester. While the participants understood the general meaning of the terms, they are at a loss when they are asked about them in the context of their inclusion in classroom activities and assessments. This is not surprising since most of the assessments of engineering subjects are technical in nature which means they usually use a quantitative approach. Skills such as collaboration, creativity, innovation, ethics, and cultural sensitivity would seem to require a qualitative evaluation framework in order to be assessed. 90% of the participants admitted that they do not include a majority of these skills as part of their assessment. When pressed with a follow-up question on the reason why, most of them simply said it was not required in the course outline. Some ventured to say it is difficult to assess these skills as they seem to be fraught with subjectivity which to them is a departure to the objective and clear type of assessment engineering students are used to having. However, not all 21st Century skills were unfamiliar to the participants. A clear majority of them stated that technological literacy, problem-solving, teamwork, ethics, and collaboration are mostly included in their assessment tools.

F. View of the participants towards incorporating the Learning-By-Doing and 21st Century skills better in the course outline

The participants also provided their views on how LBD and 21st Century skills were incorporated in the course outline. Mostly, it was stated that the course outline should include clear assessment guidelines for LBD activities and 21st Century skills. Two of the participants emphasized the core concepts and skills students needed to learn and cautioned to avoid unconnected topics which inhibit the development of critical thinking and other 21st Century skills. In addition, the participants also claimed that critical and creative thinking can be incorporated into the course outline, and on giving more importance to improving the communication abilities and promoting the spirit of teamwork and leadership. One of the participants specified four main skill areas that affect creativity and innovation in the current course outline: *fluency, flexibility, elaboration and originality*. Hence it was proposed that making these skills a part of the assessment strategy would better incorporate LBD and 21st Century skills in the course outline. The participants also claimed that the personal and social development of the students may be enhanced by developing their managerial and leadership skills, and by preparing them to implement complex skills such as planning, communicating, problem-solving, and decision-making.

Conclusion

This study reported on data gathered from instructors and the HCT dean's interviews and student questionnaire surveys. It has shown diverse views on the Learning-by-Doing understanding, practices, and definition. While most of the educator and student participants explained their understanding of LBD, the enumeration of the impact verifies the "incomplete" understanding of LBD. On the definition of LBD, participants viewed it as the second half of "theoretical" knowledge which is its application. This understanding is supported by LBD research literature. For instance, [7] suggested the move from formal education to one that is more experienced-based. As engineering is a skill-based field, it was no surprise for the participants to award high percentage for laboratory activities and experiments to be the most used LBD activities. The importance of 21st Century skills was highlighted by all participants during the educator participant interviews. But, most of them failed to use them in the context of their classroom activities. Creativity, critical thinking and innovation skills were highlighted as the most difficult to assess in engineering courses. The underlying theme of the findings from the students' point of view is that nearly 55% of the students believed the LBD is conducted and assessed in the classroom. The rest of the students, 45% seem to believe that LBD is not conducted in the classroom activities. The survey data used in this study was discussed along with other interviews and personal observation data to explore the results further in the construction of the LBD model.

References

[1] G. Frache; H. E. Nistazakis; G. S. Tombras. "Reengineering engineering education: Developing a constructively aligned learning-by-doing pedagogical model for 21st Century education", 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, Greece, 2017, pp. 1119 – 1124

Doi: 10.1109/EDUCON.2017.7942989.

[2] Zaman, S. (2012). 2013 dedicated to Emiratisation. *Gulf News*. Retrieved from http://gulfnews.com/news/gulf/uae/society/2013-dedicated-to-emiratisation-1.1111015.

[3] Vision 2021 United Arab Emirates. Available: http://www.vision2021.ae/en

[4] J. Dewey, *Democracy and education: An introduction to the philosophy of education*. New York: WLC Books, 2009.

[5] D. W. Johnson, R. T. Johnson, and E. Holubec, *Cooperation in the classroom* (8th ed.). Edina, MN: Interaction Book Company, 2008.

[6] T. T. Pham. "Issues to consider when implementing student-centered learning practices at Asian higher education institutions". *Journal Of Higher Education Policy And Management*, 33 (5), pp. 519—528, 2011.

[7] J. Voogt, & N. Roblin, *International Symposium on Education Reform | EDL*. [online]. [Viewed: 2013]: ttp://www.internationalsymposiumoneducationalreform.com

[8] Organization for Economic Cooperation and Development. (2014). Better Skills, Better Jobs, Better Lives: A Strategic Approach to Education and Skills Policies for the United Arab Emirates. Paris, France.

[9] W. Rugh, "Arab education: Tradition, growth, and reform, "*The MiddleEast Journal*, 396-414, 2002.

[10] L. H. Lewis, and C. J. Williams, "Experiential learning: Past and present, " *New Directions For Adult And Continuing Education*, vol. 62, pp. 5-16, 1994

[11] A. Y. Kolb and D. A. Kolb, "Learning styles and learning spaces: Enhancing experiential learning in higher education, " Academy *Of Management Learning & Education*, vol. 4, no. 2, pp. 193-212, 2005.

[12] A. Finkel, "*Innovative Approaches to Engineering Education: The Australian Experience," 2013,* www.caets.org.

[13] R. Pascual, and R. Uribe, "Experiential Learning Strategies in a Mechanical Engineering Senior Course, ", paper presented at *Sixth International Workshop on Active Learning in Engineering Education*. Mexico: Monterrey, 2006.

[14] M. Romi, "Learning by teaching in engineering: a step beyond learning by doing, "Technology*, Education And Development, A. Lazinika and C. Calafate, Eds. Croatia: Intech*, 2009, pp. 337-394.

[15] P. R. Donald and J. L. Faust, "Techniques for active learning", 2010.

[16] L. Kane, "Educators, learners and active learning methodologies", Int. *Journal Of Lifelong Education*, vol. 23 no. 3, 2004.

[17] S. L. Ash and P. H. Clayton, "The articulated learning: An approach to guided reflection and assessment, " *Innovative Higher Education*, vol. 29 no. 2, pp. 137-154, 2004.

[18] R. C. Clark and R. E. Mayer, "Learning by viewing versus learning by doing: Evidence-based guidelines for principled learning environments,

"Performance *Improvement*, vol. 47 no. 9, pp. 5-13, 2008.

[19] M. Ginsburg, *Active-learning pedagogies as a reform initiative: Synthesis of case studies*. Washington, DC: Academy for Educational Development, 2009.

[20] R. Pascual and R. Uribe, "Experiential Learning Strategies in a Mechanical Engineering Senior Course", paper presented at *Sixth International Workshop on Active Learning in Engineering Education*. Monterrey, Mexico, June 2006.

[21] C. M. Vest, "Educating engineers for 2020 and beyond, "*Report of a Convocation "Educating the engineer of* 2020. pp. 160—169, 2005.

[22] P. Morgado, "From Passive to Active Learners: Implementing the Pedagogy of " Learning by Doing" in a Large-sized Design Foundation Class, "Transformative *Dialogues: Teaching & Learning Journal*, v.4 n.2, 2010.

[23] C. M.Vest, C. M, "Educating engineers for 2020 and beyond," *Report of a Convocation "Educating the engineer of* 2020. pp. 160—169, 2005.

[24] L. Morell, "Engineering Education in the 21st Century: Roles, Opportunities, and Challenges," *International Journal of Technology and Engineering Education, 7*(2), 1-10, 2010.

[25] National Academy of Engineering (NAE) (2010) Grand Challenges for Engineering in the Eyes of 21st Century Students. Available from http://www.edstechnologies.com

[26] K. E.Wolfe Understanding the careers of the alumni of the MIT Mechanical Engineering department, 2004.

[27] R. Pascual and R. Uribe, "Experiential Learning Strategies in a Mechanical Engineering Senior Course", paper presented at *Sixth International Workshop on Active Learning in Engineering Education*. Monterrey, Mexico, June 7-9, 2006.

[28] J. B. Biggs, C. S. Tang, and J. B. Biggs *Teaching for quality learning at university*. Maidenhead: Mcgraw-Hill/Society For Research Into Higher Education & Open University Press, 2007

[29] S. Nightingale and A. Carew and J. Fung, "*Application of constructive alignment principles to engineering education: Have we really changed*," In Proceedings of the Australasian Association of Engineering Education Conference, December, pp. 9-13, 2007

[30] N.M. Meyers and D. D. Nulty, "How to use (five) curriculum design principles to align authentic learning environments, assessment, students' approach to thinking and learning outcomes," *Assessment and Evaluation in Higher Education* (34), pp 565-577, 2009

[31] R. Clark and R. E. Mayer, "Learning by Viewing versus Learning by Doing: Evidence-Based Guidelines for Principled Learning Environments," *Performance Improvement*, *47*(9), 5-13, 2008

[32] L. Kane, "Educators, learners and active Learning Methodologies," *International Journal of Lifelong Education, 23*(3), 275-286, 2007.

[33] D. Hung and S. Lee, "Is there an instructional framework for the 21st century?" *Creative Education,* 3(4), 461-470, 2012.

[34] W. Houghton, *Engineering Subject Centre Guide: Learning and Teaching Theory for Engineering Academics*. Loughborough: HEA Engineering Subject Centre, 2004.

[35] V. Venkatesh and S. A. Brown, and H. Bala, "Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems," *MIS Quarterly*, *37*(1), 21-54, 2013.

[36] J. W. Creswell, and V. L. Clark, *Designing and conducting mixed methods research*. Thousand Oaks, Calif.: SAGE Publications, 2011.

[37] S. O. Migiro and B. A. Magangi, "Mixed methods: A review of literature and the future of the new research paradigm." *African Journal of BusinessManagement. 5*(10), 3757-3764. doi: 10.5897/AJBM09.082, 2011.

[38] A. Hussein, "The use of triangulation in social sciences research: Can qualitative and quantitative methods be combined, "*Journal of Comparative Social Work*, *1*, 1-12, 2009.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 119-127 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Automated Text Annotation for Social Media Data during Natural Disasters

Si Si Mar Win[*,1], Than Nwe Aung[2]

[1]*University of Computer Studies, Mandalay, Web Data Mining Lab, 05071, Myanmar*

[2]*University of Computer Studies, Mandalay, Faculty of Computer Science, 05071, Myanmar*

ARTICLE INFO

Article history:
Received: 21 November, 2017
Accepted: 27 February, 2018
Online: 12 March, 2018

Keywords:
LIBLINEAR
Twitter
Annotation
POS
Ngram

ABSTRACT

Nowadays, text annotation plays an important role within real-time social media mining. Social media analysis provides actionable information to its users in times of natural disasters. This paper presents an approach to a real-time two layer text annotation system for social media stream to the domain of natural disasters. The proposed system annotates raw tweets from Twitter into two types such as Informative or Not Informative as first layer. And then it annotates again five information types based on Informative tweets only as second layer. Based on the first and second layer annotation results, this system provides the tweets with user desired informative type in real time. In this system, annotation is done at tweet level by using word and phrase level features with LIBLINEAR classifier. All features are in the form of Ngram nature based on part of speech (POS) tag, Sentiment Lexicon and especially created Disaster Lexicon. The validation of this system is performed based on different disaster related datasets and new Myanmar_Earthquake_2016 dataset derived from Twitter. The annotated datasets generated from this work can also be used by interested research communities to study the social media natural disaster related research.

1. Introduction

Today, online social networking sites like Twitter, YouTube Facebook and Weibo play the important news sources during mass emergencies. Among them, Twitter, the most popular social networking site, provides a wealth of information during a natural disaster. It is often the first medium to break important disaster events such as earthquakes often in a matter of seconds after they occur and more importantly. Recent observation proofs that some events and news emerge and spread first using this media channel rather than other the traditional media like online news sites, blogs or even television and radio breaking news.

People also used social media to share advice, opinions, news, moods, concerns, facts, rumors, and everything else imaginable. Corporations use social media to make announcements of products, services, events, and news media companies use social media to publish near real-time information about breaking news. However, due to questionable source, uncontrollable broadcasting, and small amount of informational tweets among large number of non-informational tweets, Twitter is hardly an actionable source of breaking news.

Tweets from Twitter are highly vary in terms of subject and content and the influx of tweets particularly in the event of a disaster may be overwhelming. It is impractical to generation of efficient features vector based on uniform vocabulary. Therefore, effective feature extraction is first challenge. It is infeasible to automatically classify these varied tweets by using particular annotated corpora for specific messages of every disaster events. Cross event classification is a major challenge.

Another challenge is occurred by the development of supervised learning based systems trained on a single corpus and able to achieve a good performance over a broad range of different events. The annotation of corpus of messages for every disaster by human annotators is obviously time-expensive and practically infeasible on real time manner. Therefore annotation of tweets corpora by human is additional challenge.

In summary, the proposed system is aimed to address these issues by using three main functions: 1. Create annotated disasters corpus of tweets with five labels for Informative tweets on real

[*]S. S. M. Win, U.C.S.M, Email: sisimarwin@gmai.com

time manner. 2. Competitive, easily implementable feature extraction method that act as a benchmark for automated accurate classification approaches for natural disaster related datasets by using natural disaster lexicon. 3. Creation of extended natural disasters lexicon based on publicly available annotated datasets and newly annotated corpus. This paper is an extension of the work originally presented in IEEE/ACIS 16th International Conference on In Computer and Information Science (ICIS) [1]. In the previous work, we identified the tweets into only three labels such as Informative, Not Informative and Other Information as single layer annotation. Therefore, we continue to identify the Informative tweets into more specific information types. Our annotation model in this paper is based on a more relevant and small set of features than our previous work.

The rest of the paper is organized as follow: Section 2 presents the overview of closely related work to this paper. Section 3 explains the methodology that we used in collecting, preprocessing, feature extraction, disaster lexicon creation, and classification scheme used for annotating the tweets. Section 4 describes the architecture of the proposed system. Section 5 expresses the datasets details, experiments and analysis performed. Section 6 summarizes the results from our analysis and highlights the implications of our results. In this section, we also describe the future work of the proposed system.

2. Related Work

This section presents the current state-of-the-art systems, algorithms and methodologies to access the social media data analysis. Social media allows users to exchange small digital content such as short texts, links, images, or videos. Although it is a relatively new communication medium compared with traditional media, microblogging has gained increased attention among users, organizations, and research scholars in different disciplines. There are several researches on social media mining for text based data for classification and prediction of informational posts in different domains.

Among them, the authors in [2] proposed Artificial Intelligence for Disaster Response (AIDR) system to annotate the posts from Twitter into a set of user defined categories such as damage, needs etc. by using hybrid unigram and bigram features. In AIDR system, the tweets were identified into Informative and Non-Informative types during Pakistan earthquake in 2013.

The authors in follow up study automatically and provided human annotated Twitter corpora for 19 different events that occurred between 2013 and 2015. They also experimented their corpora by using the similar features set [3].

The other authors also presented the Tweedr, twitter-mining tool, to retrieve actionable information from Twitter. They applied several different types of features for their CRF clustering. For each token in a tweet, they extracted capitalization, pluralization, whether it is numeric or includes a number, WordNet hypernyms, Ngrams, and part of speech tags to provide specific information about different classes of infrastructure damage, damage types, and casualties [4].

The authors in [5] used six types of features such as Tweet Meta-data Features, Tweet Content Features, User based Features, Network Features, Linguistic Features and External Resource Features for credibility analysis.

They also developed a real time web application, TweetCred, to provide the one of the seven credibility scores of user generated content on Twitter by using 45 features. They tested their application within three weeks period. Their result showed that high credibility tweets were 8% [6].

Moreover, hashtags have been effectively utilized as critical features for various tasks of text or social media analysis, including tweet classification system [7].

In [8], the authors studied the linguistic method to analyze the importance of linguistic and behavioral annotations. They applied the datasets of four crisis events such as Hurricane Gustav in 2008, the 2009 Oklahoma Fires, the 2009 and 2010 Red River Floods, and the 2010 Haiti Earthquake. They observed that the usage of specific vocabulary to convey tactical information on Twitter can achieve higher accuracy than the usage of bag of words (BOW) model for classification of context-aware tweets.

The classification of tweets into Credible or Not Credible was presented in [9]. However, most of the recent research focuses on the information extraction and detection of situational awareness during natural disasters, it is still needed to provide a cohesive pipeline that takes into consideration all of the facets of data extraction.

This system focuses on the content based features set such as Linguistic features, Disaster Lexicon based features, twitter specific features (hashtags and URLs), unigram POS tag features and other salient features from tweet content.

3. Methodology

At the core function of this system is the capability of annotating tweets into predefined information types in real time. We propose, implement and evaluate the approach for determining and assigning a label for each tweet, taking into account terms from the tweet itself and from disaster lexicon. For this study, we first collect the tweets from Twitter. And then we extracted content based features from the collected tweets.

3.1. Data Collection

This function works for tweets collection. It collects messages from Twitter for training and testing using the Twitter streaming API. At first, it collected different annotated datasets published by using annotated tweet_id from Imran et al. [3].

The new data collection process focuses on the exact matching of keywords to acquire tweets and build the query using user defined keywords or hashtags. Using the relevant keywords or hashtags for queries are the best way to extract the most relevant tweets during crisis or disasters. For example, #MyanmarEarthquake hashtag is applied to acquire the news of earthquake that struck in Myanmar.

3.2. Preprocessing

Firstly, this task removes the tweets which already contains the same text in the previous preprocessed tweets to reduce the redundancy and noise by using the cosine similarity.

Secondly, stop-words from tweets are removed to reduce dimensionality of the dataset and thus terms left in the tweets can be identified more easily by the feature extraction process. Stop-words are common and high frequency words such as “a”, “the”, “of”, “and”, “an” “in” etc. [10]. User mention and URLs are also eliminated.

Finally, we used lemmatization instead of stemming to convert all the inflected words present in the text into a base form called a lemma. For the purpose of lemmatization, the proposed system uses Stanford Core NLP.

3.3. Feature Extraction

The most important step in text analysis using supervised learning techniques is generating feature vectors from the text data or documents. This work is intended to build a real time system based on tweets from Twitter, feature extraction is therefore concerned with altering tweet contents into a simple numeric vector representation.

In our previous work, we used hybrid unigram and bigram, unigram Brown cluster, unigram part of speech (POS) tags, number of hashtags, number of URLs and two lexicon based features such as NRC hashtags lexicon and our disaster lexicon as our features set. We found these features outperform the neural word embeddings and only hybrid unigram and bigram features.

According to the constant vocabulary, hybrid unigram and bigram features outperforms the classification of same events (i.e. the training and test datasets are equal). However we need to annotate the unknown disaster events and the contents described for different events may have different vocabulary. Even the same type disaster events may contains the different language style.

The analysis of social media data is heavily rely on the ability to analyze text data. However, there are some unique considerations in the analysis of social media data that make it different than a normal text mining analysis. To overcome the informal social media data to be formal consideration, text in tweets are tokenized using ARK Tweet NLP [11]. This process receives the tweets from preprocessing step, it extracts the features by using ARK POS tagger and different lexicons. The features used in this work are only extracted from tweet contents.

To derive the most relevant feature, this work investigated the three types of feature extraction methods. The first one is BOW model with unigram and bigram based features used in AIDR. The second is the neural word embeddings (WE) model and the last one is the proposed content based features model.

This system proposes the features set based on the following observations:

1. Messages in tweets written by users for same disaster type may have composed of same terms. It is usually the case that the same disasters have the same terms such as shake, strike, magnitude for disaster earthquake.
2. Different natural disaster related tweets may have composed of same terms such as need, pray, pray for, damage, death, destroy, survivor, etc. and may have same syntactical style such as POS tag.
3. Similar words have similar distributions of words to their immediate left and right [11].
4. If a tweet contain more than two hashtags in its content, it may not be information tweet.
5. In crisis related tweets, hashtag may be assumed as topic word or keyword of these tweets.
6. Informative tweets may contain numerical word and URL.

Based on these observations 1 and 2, we decided to create and use Disaster Lexicon and word Ngrams. According to observation 3, we use Brown Word cluster. We also use number of hashtags and hashtag term, URL and numeral features due to the observation 4, 5 and 6. Ngrams POS features are used according to the observation 2. The proposed features used in this system are shown in Table 1.

Table 1. Features used in the proposed system

Feature	Explanation
Brown Cluster Ngrams	Unigram and Bigram of 1000 Brown clusters in Twitter Word Clusters made available by CMU ARK group
Count of disaster related terms	Number of disaster related terms as informative words in a created lexicon for disaster tweets.
Total PMI Score of disaster related terms	Total PMI scores of unigrams and bigrams words that occurred in the tweet and listed as strongly correlated with natural disaster in a disaster lexicon for tweets
Count of non-informational terms	Lexicon creation function of this system also identifies a set of terms which appear only in Not-Informative tweets across all natural disasters datasets.
Total PMI Score of non-informational terms	Total PMI Score for each set of unigrams, bigrams that mostly occur in the Not informative tweet.
Count of numerals	Expected to be higher in situational tweets which contain information such as the number of casualties, emergency contact numbers.
POS tag	Unigram part of speech tags that occur in the Tweet generated by CMU ARK POS-Tagger
Word Ngrams	Unigram and bigram of terms from Disaster Lexicon

To extract neural word embeddings (WE) features for baseline model, this system used Word2vec model in Deep Learning4J [12]. Word2Vec is the representations of words with the help of vectors in such manner that semantic relationship between words preserved as basic linear algebra operations. The following parameters were used while training for Word2Vec: 100 dimensional space, 10 minimum words and 10 words in context. After transforming 100 dimension feature vector of each word in the corpus, this system used t-Distributed Stochastic Neighbor embedding (t-SNE) technique to reduce from 100 dimensions of each feature vector to 10 dimensions feature vector.

3.4. Extended Disaster Lexicon Creation

This system creates the disaster lexicon which contains specific natural disaster related terms with a point wise mutual information (PMI) based score and frequency distribution of these terms based on the set of annotated disaster datasets. This lexicon creation

process follows the method of Olteanu et al. [13]. In this process, we exploit their natural disaster related datasets, the other available natural disaster related datasets [2] and newly annotated dataset such as Myanmar_Earthquake_2016 dataset which are collected by proposed system for lexicon expansion or keywords (disaster related terms) adaptation.

The disaster creation process consists of two main parts. At first, to obtain the most relevant disaster terms, we create various disaster lexicons based on different datasets of same disaster type. For example, this work uses all available earthquake datasets such as 2015_Neapl, 2014_Chile, 2014_Calfornia and 2016_Myanmar earthquakes for creation of Earthquake Lexicon. In this phase, we used equal number of Informative and Not informative tweets for each disaster datasets based on same disaster types.

At second, we combined the different disaster lexicons into one master disaster related lexicon with unique unigram and bigram terms. In this step, we calculate the mean PMI score for the terms which are contained in the two or more lexicons.

The score of a term could be calculated from the PMI value of a term t in an informative context PMI (t, informative) and the same term in a non-informative context PMI (t, non-informative) using the equation:

$$InfoScore = PMI\ (t, informative) - PMI\ (t, non\text{-}informative) \quad (1)$$

Here PMI (t, informative) and PMI (t, non-informative) are calculated using:

$$PMI\ (t, orientation\ ) = log_2 \frac{freq(t, orientation).N}{freq(t).\ freq(orientation)} \quad (2)$$

Where, freq (t) is the number of times term t appears in a tweet, while *N* is total number of terms in the tweet.

This automatically created lexicon is used in feature extraction process of the proposed system.

3.5. Feature Selection

Feature selection is an important problem for text classification. In feature selection, this work attempts to determine the features which are most relevant to the classification process. This is because some of the words are much more likely to be correlated to the class distribution than others. This system applied the information gain based feature selection method which is widely used for text classification.

Information gain (IG) measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. n this method, let P_i be the global probability of class *i*, and $P_i\ (w)$ be the probability of class *i*, given that the document contains the word w. Let *F(w)* be the fraction of the documents containing the word w. The information gain measure *I(w)* for a given word *w* is defined as follows:

$$I(w) = -\sum_{i=1}^{k} P_i\ .\log\ (P_i) + \ F(w).\sum_{i=1}^{k} P_i(\mathrm{w})\ .\log\ (P_i(w)) +$$

$$(1 - F(w)).\sum_{i=1}^{k}(1 - P_i(\mathrm{w}))\ .\log\ (1 - (P_i(w)) \quad (3)$$

The greater the value of the information gain *I(w)*, the greater the discriminatory power of the word *w*.

3.6. Annotation of Social Media Text

This system assess annotation of tweets by using supervised machine learning technique. This technique automatically classifies the information contained in tweets. To perform the annotation task, the proposed system trained a LIBLINEAR classifier operating on extracted features set. LIBLINEAR solves large-scale classification problems in many applications such as text classification. It is very efficient for training large scale. It takes only several seconds to train more than 600,000 examples while a Library for Support Vector Machines (LibSVM) takes several hours for same task [14].

Given a set of features and a learning corpus (i.e. the annotated dataset), the classifier trains a statistical model using the feature statistics extracted from the corpus and then annotates the tweets into Informative or Not Informative. This trained model is then employed in the classification of unknown tweets and, for each tweet, it assigns the probability of belonging to a class: Related and Informative as Informative, Not Related or Not applicable as Not Informative in first layer annotation. And then based on the Informative tweets, this system annotates again these informative into one of five types such as infrastructure damage, caution and advice, dead or injured people, needs and offer and Donations and volunteering as second layer annotation. The annotated datasets required by the system can be obtained from three sources such as AIDR, CrisisNLP which is the collection of tweets from 19 natural and man-made disasters and CrisisLexT26 which is the collection of tweets from 26 Crises [2, 3, 13]. This system uses datasets in English language only.

4. Architecture of the proposed System

The holy grail of text annotation is an automated system that accurately and reliably annotates very large numbers of cases using relatively small amounts of manually annotated training data. This work is intended to develop a two layer annotation system that automatically creates the different disaster datasets with annotated tweets. In this system, annotation is restricted to tweets in English language. Non-English tweets are not considered. Non-English tweets are not considered.

The system, illustrated in Figure, first collects the tweets from Twitter by using user desired query terms or target disaster related terms. After collecting the tweets, it removes the redundant tweets by using tweet_id and then it also eliminates the stop-words. In feature extraction, this system applies Linguistic features such as Brown cluster, Syntactic feature such as POS features, Lexical features using disaster lexicon and the other Twitter centric features such as Hashtags and URLs.

This system also analyzed which features are important in the data to annotation. It applied the annotated corpus to train a classifier that automatically annotates the tweets.

To improve model performance, the best set of 300 features were chosen by using Information gain theory based feature selection method.

In the ground truth annotation process, LIBLINEAR classifier uses these selected features subset for tweets categorization to create annotated corpus with Informative or Non-informative tweets and to provide informative tweets to the users.

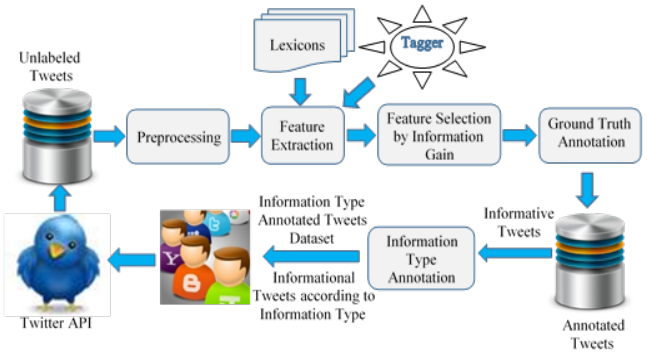


Figure 1: Architecture of the Proposed System

After annotating the collected tweets into one of the five information types, his system provides the informational tweets to users based on their desired type of information.

5. Experiments

This system performs a set of preliminary experiments to evaluate the effectiveness of feature extraction, feature selection model and classifier model on the performance of the proposed approach. For feature extraction, the proposed system applied three models such as neural word embedding, BOW with Unigram and Bigram model and the proposed model.

The final experiment is done under the best development settings in order to evaluate the classifier model with the best feature set. This section presents experiments and results for classification of four annotated datasets. The results along with the experimentation of different datasets are described based on accuracy, precision, recall, and F1 score of classifier model for feature extraction performance.

5.1. Datasets and Setting

In order to evaluate the effectiveness of the proposed social media text annotation strategies for identifying informative tweets during natural disasters events, the experiments of this system used people freely available 10 annotated natural disaster datasets . These datasets are already annotated with different information types.

To reduce the noise in training data, this system discarded all the following tweets.

1. The tweet where an information type clash is observed. An information type clash is a tweet that may happen two or more different type and may ambiguous in the dataset.

2. "Not labeled" tweets.

3. "Animal Management" are also eliminated.

The tweets with similar information types such as "Infrastructure damage" and "Infrastructure and utilities" are combined as "Infrastructure and utilities". "Injured or dead people", "missing or found people", "displaced people and evacuation" and "personal updates" tweets are combined as "Affected individuals" and "donation needs or offer volunteering services" and "Money" are also combined as "Donations and volunteering".

Before training the corpus for second layer annotation, the informative tweets with non-specific information type such as "Other Useful Information" are also discarded.

Detailed information of datasets is described in Table 2 and Table 3. In this table Type 1 refers to the information type "Affected individuals", Type 2 refers to "Infrastructure and utilities", Type 3 means "Donations and volunteering", Type 4 is "Caution and advice", and Type 5 refers to "Sympathy and emotional support".

Table 2. Natural disaster datasets details including disaster type, name, number of informative tweets, number of Not Informative tweets and total tweets.

Type	Disaster Name	Info	Not-Info	Total
Floods	2013_Queensland_floods (QF)	728	281	1009
Bushfire	2013_Australia_bushfire (AB)	691	261	952
Typhoon	2013_Typhoon_Yolanda (TY)	765	175	940
Wildfire	2012_Colorado_wildfires (CW)	685	247	932
Earthquake	2014_Chile_earthquake (ChiE)	1834	179	2013
Floods	2013_Colorado_floods (CF)	589	190	779
Earthquake	2014_Costa_Rica_earthquake (CE)	842	170	912
Floods	2014_Manila_floods (MF)	628	293	921
Floods	2012_Phillipines_Floods (PF)	761	145	906
Floods	2013_Alberta Floods (AF)	684	297	981
Floods	2014_India_floods (IF)	940	396	1336

Table 3. Natural disaster datasets statistics for five information types

Dataset	Type 1	Type 2	Type 3	Type 4	Type 5
QF	207	113	55	114	17
AB	199	65	35	70	33
TY	77	106	383	20	63
CW	44	128	62	69	25
NE	6	165	239	1215	458
IF	30	792	42	51	25
ChiE	55	3	70	14	58

The proposed system performed 10 fold cross validation to test the efficiency of the feature extraction and the model. In the experiments of classification, the proposed system used the set of tweets from five natural disasters such as 2013_Queensland_floods denoted by QF, 2013_Australia Bushfire as AB, the set of tweets for 2013_Typoon_Yolanda as

TY, 2012_Colorado_wildfires as denoted by CW, and the set of tweets from 2012_Costa_Rica_Earthquake as CE respectively.

5.2. Effectiveness of Feature Extraction

To choose the best classification model, we tested the extracted feature set on four different classifiers such as Random Forest, Sequential Minimal Optimization (SMO) which is the fast training algorithm for Support Vector Machine (SVM), Naïve Bayes and our LIBLINEAR classifier that are well known in text classification process. Due to the experiments on the previous work, the performance of Random Forest, Naïve Bayes and SMO was sensitive to the large number of features. Therefore, this system used LIBLINEAR classifier with Information Gain based feature selection method to get better performance and to reduce inconsistent features. This wok uses a well-known WEKA machine learning tools for implementation of Random Forests, Naïve Bayes, SMO, LIBLINEAR and Information Gain based feature selection methods [15]. The results of our previous work are described in [10]. Due to these results, we selected the LIBLINEAR classifier as our classification model.

In this work, the proposed feature extraction method and two baseline methods are evaluated by experiments on ten datasets. To compare the performance of the different feature-models (using LIBLINEAR classifier) under three scenarios such as (i) in-domain classification, (ii) cross event classification and (iii) cross-domain classification, where the classifier is trained with tweets of one event, and tested on another event are considered in this system.

5.2.1. In Domain Classification

In this type of classification, the classifier is trained and tested with the tweets of the same event. To evaluate the in-domain performance of each model, the proposed system followed a 10-fold cross validation process: each dataset was randomly split in 10 different non overlapping training and test sets. The Accuracy, Precision, Recall and F-Measure were calculated as the weighted average of these values over all the 10 test sets.

Table 4. Classification results in terms of Precision, Recall, F-Measure and Accuracy across Informative and Not Informative classes.

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
QF	0.791	**0.813**	0.612	0.812	**0.822**	0.782	0.785	**0.816**	0.687
AB	**0.772**	0.754	0.567	**0.79**	0.765	0.753	0.768	0.758	0.647
TY	0.797	**0.802**	0.685	**0.825**	0.816	0.828	0.788	**0.807**	0.750
CW	0.813	**0.815**	0.676	0.818	**0.820**	0.714	0.711	**0.817**	0.633

Table 4 represents a summary of evaluation for in domain classification by weighted average Precision (P), Recall (R) and F-Measure (F-M) of the classification results on four datasets.

According to the results, BOW with hybrid unigram and bigram model would perform relatively well in in-domain classification, since the training event and test event share a common vocabulary. However, the performances of the proposed features model is as good as BOW method.

5.2.2. Cross Event Classification

In this type of classification, where the classifier is trained with tweets of one event, and tested on another event. The result is significant since it shows that good classification can be achieved even without considering the type of disasters.

Table 5, Table 6, Table 7 and Table 8 also show the cross event classification performance on AB, TY, CW and QF dataset as training and the remaining datasets as testing data using the features sets extracted by the baseline method, proposed method and neural word embeddings method. The results in these tables indicated that the proposed method yields a high accuracy by using the LIBLINEAR algorithm in predicting certain classes in cross event classification.

Table 5. Classification results in terms of Precision, Recall, F-Measure using (i) (BOW), (ii) Proposed (P) (iii) Word Embeddings (WE) for 2013_Australia_Bushfire as training set.

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
QF	0.77	**0.82**	0.76	0.73	**0.83**	0.50	0.74	**0.82**	0.52
TY	0.78	**0.81**	0.77	0.74	**0.83**	0.52	0.76	**0.81**	0.57
CW	0.79	**0.83**	0.75	0.80	**0.83**	0.66	0.80	**0.84**	0.66

Table 6. Classification results in terms of Precision, Recall, F-Measure across two classes using (i) (BOW), (ii) Proposed (P) (iii) Word Embeddings (WE) for 2013_Typhoon_Yolanda as training set.

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
QF	**0.77**	**0.77**	**0.77**	0.64	**0.793**	0.79	0.67	0.755	**0.77**
AB	0.72	**0.77**	0.76	0.554	**0.785**	0.78	0.579	0.748	**0.75**
CW	0.74	**0.796**	0.76	0.589	**0.804**	0.78	0.605	**0.788**	0.76

Table 7. Classification results in terms of Precision, Recall, F-Measure across two classes using (i) (BOW), (ii) Proposed (P) (iii) Word Embeddings (WE) for 2012_Colorado_Wildfire as training set.

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
QF	0.759	**0.795**	0.74	0.73	**0.808**	0.37	0.745	**0.80**	0.36
AB	0.769	**0.797**	0.72	0.756	**0.806**	0.38	0.762	**0.799**	0.35
TY	0.762	**0.810**	0.76	0.673	**0.809**	0.36	0.703	**0.809**	0.37

Table 8. Classification results across two classes using (i) (BOW), (ii) Proposed (P) (iii) Word Embeddings (WE) for 2013_Queensland_floods as training set.

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
AB	0.772	**0.817**	0.76	0.727	**0.831**	0.60	0.743	**0.818**	0.62
TY	0.771	**0.76**	0.76	0.641	**0.793**	0.65	0.672	**0.755**	0.68
CW	0.759	**0.795**	0.75	0.73	**0.808**	0.62	0.741	**0.799**	0.64

According to the experimental results, the performance of the BOW (hybrid unigram and bigram) and WE models is significantly inferior to the proposed model for cross-event classification. This is because the training and testing datasets (related to two different disaster events) have very different vocabularies. On the other hand, the classifier based on the proposed features significantly out-perform these two models in all cases. This implies that the selected features can separate between Informative and not-Informative tweets irrespective of the vocabulary and linguistic style related to specific events. Thus, classifiers can be trained over these features extracted from past disasters, and then deployed to classify tweets posted during future events.

5.2.3. Cross Domain Classification

In cross domain classification, to assign one of the two classes for first layer annotation and one of the five predefined categories (e.g. Affected individuals, Infrastructure and utilities, Donations and volunteering, Caution and advice, Sympathy and emotional support etc.) for second layer annotation to the tweet, the classifier requires sufficient training examples to learn about each pre-defined category. The proposed system used multiple past disasters of various types to train the classifier to robustly identify the different types of tweets for future natural disasters. The experiment for two classes classification (i.e. Informative and Not Informative), the proposed system used the set of tweets from Philippines (PF), Colorado (CF), and Queensland floods (QF) as the training set, denoted by PCQ, the set of tweets for Manila floods as the development set, denoted by MF, and the set of tweets from Alberta and Sardinia floods as two independent test sets, denoted by AF and SF, respectively. The results of this experiment are shown in Table 9.

Table 9. Experiments performed using the combined 2012_Philipinnes_flood, 2013_Colorado_floods and 2013_Queensland_floods as training set

Train /Test	Accuracy	Precision	Recall	F-Measure
PCQ/MF	80.34%	0.8609	0.803	0.808
PCQ/AF	76.47%	0.754	0.76476	0.7414
PCQ/SF	66.58%	0.6463	0.66579	0.6121

In the other experiments over five classes or multi-classes classification, this system combined the three or four datasets of different disaster types as training and the other one for testing data. For example, taking Colorado floods, Costa-Rica-earthquake, Philippine floods, Pablo-typhoon and Australia-Bushfire (CCPPA) as training dataset and the other datasets as individual test data. The classification results are shown in Table 10. Table 11 shows the classification results over the five classes of LIBLINEAR with 10-fold cross validation for six datasets by using three feature models.

Table 10. Classification results in terms of Precision, Recall, F-Measure across all five classes.

Train /Test	Accuracy	Precision	Recall	F-Measure
CCPPA/MF	72.3%	0.7308	0.723	0.719
CCPPA/TY	64.9%	0.72	0.649	0.663
CCPPA/CW	86.4%	86.8%	0.864	0.862
CCPPA/ChiE	79.5%	0.8214	0.795	0.7908

Table 11. Classification results across five classes using (i) (BOW), (ii) Proposed (P) (iii) Word Embeddings (WE)

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
QF	0.56	**0.581**	0.25	0.56	**0.586**	0.415	0.549	**0.576**	0.287
AB	0.537	**0.583**	0.22	0.569	**0.607**	0.465	0.532	**0.582**	0.295
TY	0.694	**0.701**	0.36	**0.724**	0.713	0.596	0.695	**0.705**	0.445
CW	0.653	**0.67**	0.18	0.64	**0.681**	0.427	0.631	**0.666**	0.255
NE	**0.749**	0.712	0.67	**0.761**	0.726	0.681	**0.738**	0.713	0.666
CE	0.757	0.78	0.58	0.783	**0.79**	0.76	0.754	**0.78**	0.658

According to the results in Table 10 and 11, the performance of LIBLINEAR classifier with proposed feature model outperforms the BOW and WE models in most cases and it can identify informational tweets at 67% accuracy on average. The performance of BOW and WE models sometime close to less than 20 %. This indicates that lexical features are critical to solve the ambiguous of information types.

5.3. Effectiveness of Annotation

In this section, the validation of this system on a real disaster study by classifying the data of Myanmar earthquake collected by Twitter API. The 6.8 magnitude earthquake that struck Myanmar on August 24th , 2016 is among the strongest in recent Myanmar history. The shaking was clearly perceived in all Central and Northern Myanmar and caused 4 deaths and several damage to the Pagodas of the area of Bagan. This dataset is crawled for a three days period from August 24th to 26th , 2016 by using the hashtags (#Myanmar, #Bagan, #earthquake, #Myanmarearthquake). And then it was randomly selected 1,800 tweets and was manually annotated based on the available news media in Myanmar such as Myanmar Times, The Global New Light of Myanmar and The Mirror.

Table 12. First Layer Annotation Results of Proposed features by LIBLINEAR

Dataset	Precision	Recall	F1	Accuracy
AB	0.897	0.895	0.895	89.45%
TY	0,912	0,92	0,913	92,02%
IF	0.908	0.908	0.908	90.82%
NE	0.748	0.751	0.749	75.05%

Table 13. Second Layer Annotation Results of Proposed features by LIBLINEAR

Dataset	Precision	Recall	F1	Accuracy
AB	0.682	0.693	0.685	69.31%
TY	0.768	0.786	0.774	78.60%
NE	0.815	0.843	0.828	84.26 %
IF	0.731	0.782	0.755	78.18 %

Base on cross domain classification over all five classes where we train the classifier on one dataset and test on another dataset, the experimental results using 2012-Costa-Rika- and 2014_Chile Earthquake as training data and Myanmar Earthquake as test data confirmed the expected classification of this work.

Myanmar_Earthquake_2016 was successfully annotated with predefined two labels at 75% accuracy on average and five labels also 74 % which is pretty high. The results of each annotation layer for four datasets are shown in Table 12 and Table 13.

5.4. Real Time Annotation

We showed a proof of real time model which takes a direct stream of new tweets as input test set and takes manually annotated tweets from previous disaster as training set, and then uses automated techniques to annotate the new tweets.

In the first layer, this system annotated the tweets into Informative if the tweets contain the information about the target disaster and Not informative for the remaining tweets. In the second layer, each informative tweet from previous layer is annotated into one of the five information types with respect to the information that contained in it.

To do this, we developed and deployed a real time system in the form of a Web application using Java 2 Platform Enterprise Edition (J2EE) and Twitter API binding library (Twitter4j). We analyzed the deployment and usage activity of our application from 24th August, 2016 to 28th August, 2016 which was the day of earthquake and days after an earthquake in Myanmar. For analysis and statistics, we collected the annotated datasets of our system for only three days period since August 24th to August 26th. In real time annotation process, we used the combined 2014_Chile_Earthquake and 2015_Nepal_Earthquake as training set. The number of collection and annotation times per day is 6. Each collection time is 4 minutes for 2000 tweets and annotation time is only 1 minute for each layer. Among the collected tweets, 62% of tweets are retweets and half of them are redundant tweets. Most of them are similar in text. Although the number of tweets collected in each time was 2000, the number of tweets annotated by our system is at most 750 because of the tweet cleaning step in preprocessing. The tweets in annotated datasets from each day are also redundant. The total unique automatically annotated tweets over three days, are nearly 2000.

After manually annotating the tweets, we compared the automatically annotated tweets from our system with manually annotated tweets to obtain the performance of our real time system. The manual annotation process is described in the previous section. According to the analysis results, the annotation accuracy of new tweets by our system is 80% in first layer and 74 % on second layer annotation.

5.5. Findings and Discussion

As mentioned above this system used 10 datasets for training and testing for evaluating classifier models and feature extraction models. Another new dataset for testing again for overall performance of the proposed system. According to the initial experimental results of BOW and word embeddings were very sensitive and depend on the vocabulary. The results of three feature extraction methods, the proposed method always outperforms the other two methods. Therefore, the proposed feature extraction model with LIBLINEAR classifier was chosen for second layer annotation process for categorizing the tweets into five specific frequently found information type.

6. Conclusion

Social media mining for disaster response and coordination has been receiving an increasing level of attention from the research community. It is still necessary to develop automated mechanisms to find critical and actionable information on Social Media in real-time. The proposed system combines effective feature extraction using NLP and machine learning approach to obtain the annotated datasets to improve disaster response efforts. Expanded disaster lexicon is also used to extract the relevant disaster related lexical features for annotation.

The proposed feature extraction method significantly outperforms the standard bag of words model and neural word embeddings model. By using LIBLINEAR classifier based on the proposed method, this system successfully annotated five information types to the Myanmar Earthquake data at 74% accuracy on average. In future, we will investigate the specific variation of terms over different disasters to perform annotation on all disaster types. We hope to formalize disaster lexicon in more detail to improve cross domain classification accuracy.

References

[1] S. S. M. Win, T.N. Aung, “Target Oriented Tweets Monitoring System during Natural Disasters”, In Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference, IEEE, 2017.

[2] M. Imran, C. Castillo, J. Lucas, P. Meier and S. Vieweg, “AIDR: Artificial intelligence for disaster response”, In Proc. of WWW (companion). IW3C2, 2014, pp. 159–162.

[3] M. Imran, P. Mitra, C. Castillo: "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages." In Proceedings of the 10th Language Resources and Evaluation Conference (LREC), pp. 1638-1643. May 2016, Portorož, Slovenia.

[4] Z. Ashktorab, C. Brown, M. Nandi, A. Culotta, “Mining Twitter to Inform Disaster Response”, In Proceedings of the 11th International ISCRAM Conference – University Park, Pennsylvania, USA, May 2014.

[5] A. Gupta, P. Kumaraguru, “Credibility Ranking of Tweets during High Impact Events”, PSOSM’12, 2012.

[6] A. Gupta, P. Kumaraguru, C. Castillo, P. Meier, “TweetCred: Real-time credibility assessment of content on Twitter”, In Proc. Of SocInfo. Springer, 228–243, 2014.

[7] A. Stavrianou, C. Brun, T. Silander, C. Roux, “NLP-based Feature Extraction for Automated Tweet Classification”, Interactions between Data Mining and Natural Language Processing: 145.

[8] W. J. Corvey, S. Verma, S. Vieweg, M. Palmer and J. H. Martin, “Foundations of a Multilayer Annotation Framework for Twitter Communications During Crisis Events”, In Proc. International Conference on Language Resources and Evaluation (LREC'12), May 2012

[9] C. Castillo, M. Mendoza and B. Poblete, “Information Credibility on Twitter”, International World Wide Web Conference Committee (IW3C2), Hyderabad, India, 2011.

[10] C. C. Aggarwal and C.-X. Zhai, “Mining Text Data, Springer”, 2012.

[11] K. Gimpel et al., “Part-of-speech tagging for Twitter: Annotation, features, and experiments”, In Proceedings of the Annual Meeting of the Association for Computational Linguistics, companion volume, Portland, June 2011.

[12] Deeplearning4j Development Team. Deeplearning4j: Open-source distributed deep learning for the JVM, Apache Software Foundation License 2.0. http://deeplearning4j.org.

[13] A. Olteanu, C. Castillo, F. Diaz and S. Vieweg, “CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises”, Association for the Advancement of Artificial Intelligence (www.aaai.org), 2014.

[14] R.E Fan, K.W Chang, C.J Hsieh, X.R Wang and C.J Lin, “LIBLINEAR: A Library for Large Linear Classification”, Journal of Machine Learning Research 9, 2008, pp. 1871-1874.

[15] E. Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data mining: Practical machine learning tools and techniques", Morgan Kaufmann, Fourth Edition, 2016.

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 6, 119-129 (2017)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Control of a three-stage medium voltage solid-state transformer

Claudio Busada[1], Hector Chiacchiarini[1], Sebastian Gomez Jorge[*1], Favio Mengatto[1], Alejandro Oliva[1], Jorge Solsona[1], German Bloch[2] and Angelica Delgadillo[2]

[1]*Instituto de Investigaciones en Ingeniería Eléctrica (IIIE), Universidad Nacional del Sur (UNS)-CONICET and Dpto. Ing. Eléctrica y de Computadoras, UNS, Av. Alem 1253, (8000) Bahía Blanca, Argentina.*

[2]*ICSA S.A. - Argentina*

ARTICLEINFO

Article history:
Received: 26 October, 2017
Accepted: 17 November, 2017
Online: 07 December, 2017

Keywords :
Solid State Transformer
Control
Power Electronics

ABSTRACT

This paper proposes the modeling and control of a Solid-State Transformer using a three-stage conversion topology. First, a rectification stage is used, where a three-phase high-voltage AC signal is converted to a DC level; this stage is then followed by a DC-DC converter, and finally an inverter is used to convert the DC into a three-phase low-voltage AC signal. The adopted topology is modeled using a simplified model for each stage, useful to design their controllers. Based on these models, the controllers are tuned to obtain a good performance to sudden load changes. This performance is tested through simulations.

1 Introduction

It is necessary to mention that this paper is an extension of work originally presented in [1]. This extension includes a detailed description of the proposed models and controllers. These models are used for tuning the controller parameters in order to obtain good performance in presence of sudden load changes.

Solid-State Transformers (SST) are emerging as a new technology capable of replacing power distribution transformers [2]. It is expected that the next generation of the power distribution transformers is based on power electronics semiconductor devices [3]. By using these semiconductors, it is possible to design an apparatus based on power converters with smaller and lighter high-frequency transformers. In this way, it is possible to obtain a smaller and lighter distribution transformer when compared with a traditional transformer of the same power rating. Roughly speaking, SSTs work as follows. In a first stage a sinusoidal signal of, typically, 50 or 60 Hz is converted to a high-frequency signal. Then, the amplitude of this signal is changed, by using a small high-frequency transformer, and finally in a third stage a new signal of the same frequency as the input signal but different amplitude is obtained.

Several topologies for the SST can be found in the literature, with their advantages and disadvantages [4]. In addition, it is possible to find some reviews dealing with the subject [5, 6]. Applications in transportation and smart grid can be found in [7]; whereas a topology based on SiC devices can be found in [8]. Also, a traditional transformer and an SST were compared in [9], and a procedure to obtain a detailed model of an SST topology can be found in [10].

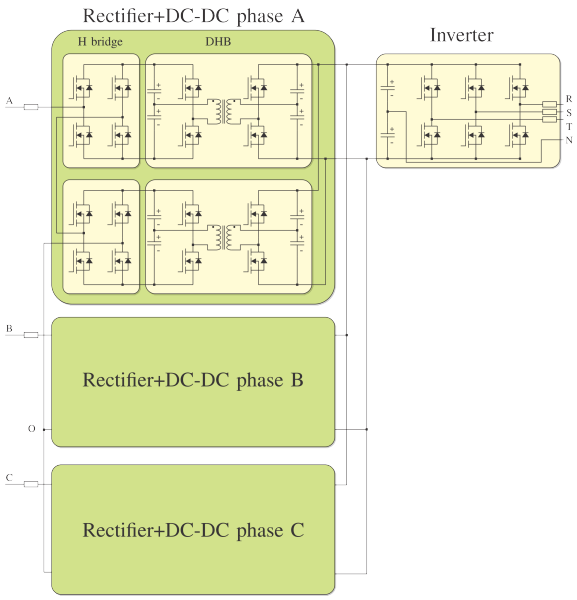


Figure 1: SST topology.

*Corresponding author, sebastian.gomezjorge@uns.edu.ar

A three-stage SST is considered in this paper [7]. Its topology is shown in Figure 1. The three stages are: a rectification stage, where a three-phase high voltage (HV) AC grid voltage is converted to a DC signal, a DC-DC converter stage (multiple stages in parallel) that provides isolation and performs the level adaptation, and finally, an inverter stage that converts the DC signal into a three-phase low voltage (LV) AC sinusoidal wave.

In this paper, simple models for each stage are proposed. Using these models, the required controllers are designed and their parameters are tuned to obtain a good performance in presence of sudden load changes. In order to test the performance of the proposed controllers, simulations results are introduced.

2 Proposed SST topology and model

Figure 2 shows a general diagram of the chosen topology. The SST is composed of a bidirectional multi-level HV three-phase converter, labeled Rectifier, six bidirectional isolated DC-DC converters, labeled DHB1 through DHB6, and a bidirectional LV three-phase converter with neutral, labeled Inverter. This topology has a great degree of modularity, allowing to consider the converters as decoupled systems and simplifying the design of the controllers [7].

Each converter and its simplified model is described below. Since the controllers are implemented using digital control techniques, the models are described in discrete time, assuming a sampling time T_s.

2.1 Rectifier

The rectifier is a three-phase AC-DC converter connected to the HV grid. As shown in Figure 2, the rectifier input port is modeled as a three-phase controlled voltage source $v_{rec} = [v_{reca}\ v_{recb}\ v_{recc}]^T$, coupled to the HV grid voltage $v_{hv} = [v_{hva}\ v_{hvb}\ v_{hvc}]^T$ through an inductor with current $i_{hv} = [i_{hva}\ i_{hvb}\ i_{hvc}]^T$. This coupling filter is modeled using complex vector notation (see Appendix). The zero order hold (zoh) discrete time model of the coupling filter results:

$$\vec{i}_{hv}[k+1] = \frac{T_s}{L_{rec}}\left(\bar{\vec{v}}_{hv}[k] - \vec{v}_{rec}[k]\right) + \vec{i}_{hv}[k], \qquad (1)$$

where $\bar{\vec{v}}_{hv}[k] = 0.5(\vec{v}_{hv}[k+1] + \vec{v}_{hv}[k])$ is the mean value of the HV grid voltage in the time interval $T_s k$, $T_s(k+1)$.

The output ports of the rectifier are modeled as current sources i_{iX}, with $X = 1 \ldots 6$. These currents are computed through power balance. Considering that both H bridges of each phase of the rectifier are given the same reference, it can be found that their currents are equal:

$$i_{i1}[k] = i_{i2}[k] = \frac{v_{reca}[k] i_{hva}[k]}{V_{busH1}[k] + V_{busH2}[k]}, \qquad (2)$$

$$i_{i3}[k] = i_{i4}[k] = \frac{v_{recb}[k] i_{hvb}[k]}{V_{busH3}[k] + V_{busH4}[k]}, \qquad (3)$$

$$i_{i5}[k] = i_{i6}[k] = \frac{v_{recc}[k] i_{hvc}[k]}{V_{busH5}[k] + V_{busH6}[k]}. \qquad (4)$$

Note that in this rectifier model, the independent variables (control inputs) are the three-phase components of $v_{rec}[k]$.

2.2 Dual Half Bridge (DHB)

Since there are six isolated DC buses in the rectifier, there are six DC-DC converters which are implemented using the DHB topology, which are modeled as current sources. The coupling between the output ports of the rectifier and the input port of each DHB is performed through a capacitor. The voltage across each of these capacitors is modeled through the difference equation:

$$V_{busHX}[k+1] = \frac{T_s}{C_H/2}\left(i_{iX}[k] - i_{oX}[k]\right) + V_{busHX}[k], \quad (5)$$

with $X = 1 \ldots 6$. The DHBs themselves are controlled using a phase shift strategy. Therefore, each converter can be modeled through the algebraic equation for their average power transfer [11]:

$$P_{dhbX}[k] = \frac{V_{busHX}[k]\ m V_{busL}[k] \delta_X[k] \left(\pi - |\delta_X[k]|\right)}{8\pi^2 L_d f_{sw}^{dhb}}, \quad (6)$$

where m is the DHB transformer relation, L_d is the total leakage of the transformer referred to the HV side, f_{sw}^{dhb} is the switching frequency of the DHB converter, and δ_X is the phase shift angle between the voltages generated at the HV and LV sides of the DHB under consideration. Dividing this power by $V_{busHX}[k]$, it results

$$i_{oX}[k] = \frac{m V_{busL}[k] \delta_X[k] \left(\pi - |\delta_X[k]|\right)}{8\pi^2 L_d f_{sw}^{dhb}}, \qquad (7)$$

which describes the relation between input port i_{oX} and the control action δ_X. The outputs of all the DHBs are connected in parallel, and feed the LV bus. The relation between the input signals and the output signals is obtained through power balance, as described in the following equation:

$$i_{dhbX}[k] = \frac{i_{oX}[k] V_{busHX}[k]}{V_{busL}[k]}. \qquad (8)$$

Note that in this DHB model, the independent variable (control input) for each DHB is δ_X.

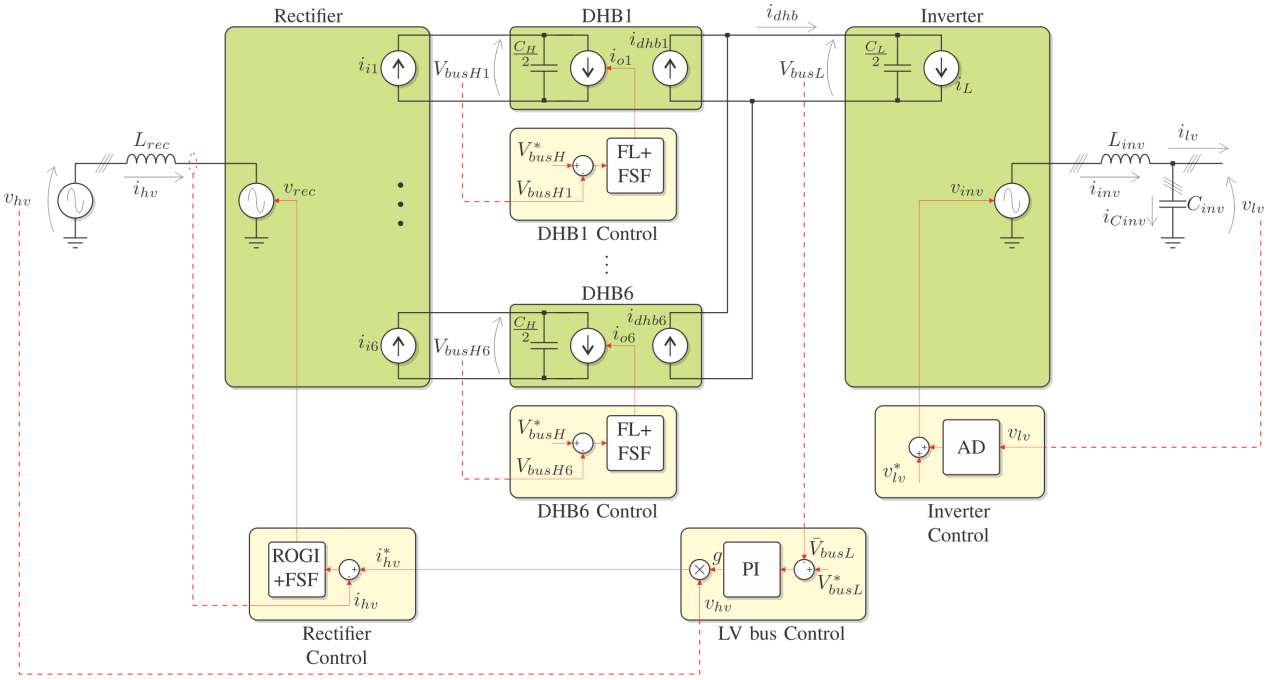


Figure 2: SST topology, model and general control scheme.

2.3 Inverter

The inverter is a three-phase DC-AC converter connected to the LV grid. The input port is modeled as a controlled current source i_L, and the output port is modeled as a three-phase controlled voltage source $v_{inv} = [v_{invr}\ v_{invs}\ v_{invt}]^T$. The load voltage and current are $v_{lv} = [v_{lvr}\ v_{lvs}\ v_{lvt}]^T$ and $i_{lv} = [i_{lvr}\ i_{lvs}\ i_{lvt}]^T$, respectively.

The coupling between the output ports of the DHBs and the input port of the inverter is performed through a capacitor. The voltage across this capacitor is modeled through the difference equation:

$$V_{busL}[k+1] = \frac{T_s}{C_L/2}\Big(i_{dhb}[k] - i_L[k]\Big) + V_{busL}[k], \qquad (9)$$

where

$$i_{dhb}[k] = \sum_{X=1}^{6} i_{dhbX}[k]. \qquad (10)$$

The coupling between the load and the output port of the inverter is performed by an LC filter. Since each leg of the inverter is controlled as a single phase converter, the zoh discretization of this filter is given for each phase of the inverter. This is denoted with subscript Y, which can be equal to r, s or t:

$$x_{invY}[k+1] = A_{inv}x_{invY}[k] + B_{inv}v_{invY}[k] + B_{inv1}i_{lvY}[k], \qquad (11)$$

where $x_{invY}[k] = [i_{invY}[k]\ v_{lvY}[k]]^T$,

$$A_{inv} = \begin{bmatrix} \cos\left(\frac{T_s}{\sqrt{C_{inv}L_{inv}}}\right) & -\sqrt{\frac{C_{inv}}{L_{inv}}}\sin\left(\frac{T_s}{\sqrt{C_{inv}L_{inv}}}\right) \\ \sqrt{\frac{L_{inv}}{C_{inv}}}\sin\left(\frac{T_s}{\sqrt{C_{inv}L_{inv}}}\right) & \cos\left(\frac{T_s}{\sqrt{C_{inv}L_{inv}}}\right) \end{bmatrix}, \qquad (12)$$

$$B_{inv} = \begin{bmatrix} \sqrt{\frac{C_{inv}}{L_{inv}}}\sin\left(\frac{T_s}{\sqrt{C_{inv}L_{inv}}}\right) \\ 1 - \cos\left(\frac{T_s}{\sqrt{C_{inv}L_{inv}}}\right) \end{bmatrix}, \qquad (13)$$

$$B_{inv1} = \begin{bmatrix} 1 - \cos\left(\frac{T_s}{\sqrt{C_{inv}L_{inv}}}\right) \\ -\sqrt{\frac{L_{inv}}{C_{inv}}}\sin\left(\frac{T_s}{\sqrt{C_{inv}L_{inv}}}\right) \end{bmatrix}. \qquad (14)$$

As in the previous converters, the relation between the input signals and the output signals is obtained through power balance, as described in the following equation:

$$i_L[k] = \frac{v_{inv}[k] \bullet i_{lv}[k]}{V_{busL}[k]}, \qquad (15)$$

where the $\bullet$ operator denotes scalar product. Note that in this inverter model, the independent variables (control inputs) are the three-phase components of $v_{inv}[k]$.

3 Converter controller description

Figure 2 shows the simplified block diagrams of the proposed controllers for each converter. In this figure, dashed lines represent measured signals, and solid lines in the controllers represent control signals. In what follows, each individual controller is described.

3.1 Rectifier control

This controller is tasked to make the HV grid current i_{hv} to copy the current reference i^*_{hv}. This is

achieved through a full state feedback (FSF) control with the addition of a reduced order generalized integrator (ROGI) [12]. The ROGI is added to achieve zero steady state tracking error, since the current reference is a grid frequency positive sequence three-phase signal. Considering a one sample time processing delay, typical in digital implementations, complex vector notation, and space vector modulation (SVM), the control action is

$$v_{rec}[k] = v^*_{recSVM}[k-1], \tag{16}$$

$$v^*_{recSVM}[k] = v^*_{rec}[k] - v_{mc}[k], \tag{17}$$

$$v_{mc}[k] = 0.5\,(\max(v^*_{rec}[k]) + \min(v^*_{rec}[k])), \tag{18}$$

where max and min functions are computed between the *abc* components of $v^*_{rec}[k]$. Also,

$$\vec{v}^*_{rec}[k] = -K_{rec}\begin{bmatrix}\vec{i}_{hv}[k] - \vec{i}^*_{hv}[k]\\ \vec{v}^*_{rec}[k-1]\\ \vec{r}_1[k]\end{bmatrix}, \tag{19}$$

$$\vec{r}_1[k+1] = j(1-e^{j\omega T_s})\left(\vec{i}_{hv}[k] - \vec{i}^*_{hv}[k]\right) + e^{j\omega T_s}\vec{r}_1[k], \tag{20}$$

with K_{rec} a 1×3 complex gain vector, and $\vec{r}_1[k]$ represents the state of a ROGI tuned at fundamental grid angular frequency ω. As described in [12], gain vector K_{rec} can be found using the linear quadratic regulator method, or by pole placement using Ackerman's formula, and the matrix description of the system:

$$x_{rec}[k+1] = A_{rec}x_{rec}[k] + B_{rec}\vec{v}^*_{rec}[k], \tag{21}$$

where $x_{rec}[k] = [\vec{i}_{hv}[k]\ \vec{v}^*_{rec}[k-1]\ \vec{r}_1[k]]^T$,

$$A_{rec} = \begin{bmatrix}1 & -\frac{T_s}{L_{rec}} & 0\\ 0 & 0 & 0\\ j(1-e^{j\omega T_s}) & 0 & e^{j\omega T_s}\end{bmatrix}, \tag{22}$$

$$B_{rec} = [0\ 1\ 0]^T. \tag{23}$$

The reference i^*_{hv} is the output of the LV bus controller, and it is related to the instantaneous HV grid voltage through

$$i^*_{hv}[k] = g[k]v_{hv}[k], \tag{24}$$

where g is a scalar variable signal. Therefore, depending on the sign of g and assuming v_{hv} sinusoidal with no harmonic distortion, the rectifier will source or sink a three-phase sinusoidal current to the HV grid, with unity power factor. In this paper the closed loop poles are chosen to achieve a settling time of 4.5[ms].

3.2 DHB control

The objective of the DHBs is to transfer the pulsating power of each of the H bridges of the rectifier to the LV bus, where the resulting power is non-pulsating. To achieve this objective, the controller of each DHB is designed to keep its instantaneous voltage V_{busHX} at its reference level V^*_{busH} (which is the same for all six DHBs). As shown by (7), the relation between current i_{oX} and phase-shift angle δ_X is non-linear. In order to be able to apply linear control techniques, a feedback linearization (FL) is implemented. Then, the resulting linear system is controlled through FSF.

Considering a one sample time processing delay, the control action is

$$\delta_X[k] = \delta^*_X[k-1], \tag{25}$$

where

$$\delta^*_X[k] = 0.5\pi\left(1 - \sqrt{1 - 32 f^{dhb}_{sw}\frac{|i^*_{oX}[k]|}{mV_{busL}[k]}}\right)sign(i^*_{oX}[k]), \tag{26}$$

is obtained from (7) and it is the FL equation for the system. Assuming that $V_{busL}[k] \simeq V_{busL}[k-1]$ (slow varying signal), it can be numerically verified that for a given value of $i^*_{oX}[k-1]$, evaluating (26) at $k-1$, and replacing the result in (25) and (7) yields $i_{oX}[k] \simeq i^*_{oX}[k-1]$. Therefore, the linearized system is modeled by (5) and

$$i_{oX}[k] = i^*_{oX}[k-1]. \tag{27}$$

Adding a discrete time backward Euler integrator to achieve zero steady state error, the control action for each DHB is computed as follows:

$$i^*_{oX}[k] = -K_{DHB}\begin{bmatrix}V_{busHX}[k] - V^*_{busH}[k]\\ r_0[k]\\ i_{oX}[k]\end{bmatrix}, \tag{28}$$

$$r_0[k+1] = T_s\left(V_{busHX}[k] - V^*_{busH}[k]\right) + r_0[k], \tag{29}$$

where K_{dhb} is a 1×3 gain vector, and $r_0[k]$ represents the state of the discrete integrator. Once (28) is computed, the phase shift angle δ^*_X is obtained through (26) and applied to the converter.

Gain vector K_{dhb} can be found using the linear quadratic regulator method, or by pole placement using Ackerman's formula, and the matrix description of the linearized system:

$$x_{dhb}[k+1] = A_{dhb}x_{dhb}[k] + B_{dhb}i^*_{oX}[k], \tag{30}$$

where $x_{dhb}[k] = [V_{busHX}[k]\ r_0[k]\ i_{oX}[k]]^T$,

$$A_{dhb} = \begin{bmatrix}1 & 0 & -\frac{T_s}{C_H/2}\\ T_s & 1 & 0\\ 0 & 0 & 0\end{bmatrix}, \tag{31}$$

$$B_{dhb} = [0\ 0\ 1]^T. \tag{32}$$

This scheme makes each DHB source power to the LV bus if its HV bus voltage is higher than V^*_{busH}, or sink power from the LV bus if its HV bus voltage is lower than V^*_{busH}. In this paper the closed loop poles are chosen to achieve a settling time of 1[ms].

3.3 LV bus control

This controller is tasked to keep the mean value of the LV bus voltage ($\bar{V}_{busL}$ in Figure 2) at the reference level V^*_{busL}. In order to do so, signal V_{busL} is filtered (filter not shown in the figure) and then a proportional integral (PI) controller is used. The output of this controller is the variable gain g. This gain is used to properly scale the measured HV grid voltage, and generate the current reference for the rectifier, defined in (24). For this reason, the dynamics of the rectifier control loop and the LV bus control loop must be decoupled. This is achieved by making the bus control loop significantly slower than the rectifier and DHB control loops.

The LV bus voltage is modeled by (9). Since there is no controlled current source directly connected to the LV bus, the design of the controller assumes that i_{dhb} is the control action. Once the control action i_{dhb} is computed, gain g is obtained through power balance:

$$g[k] = \frac{i_{dhb}[k]\bar{V}_{busL}[k]}{3V^2_{nomhv}}, \tag{33}$$

where V_{nomhv} is the nominal rms value of the HV grid voltage. As stated at the beginning of this section, this controller is slow. Therefore, both the processing delay and the dynamics of the filter used to obtain $\bar{V}_{busL}$ can be ignored without significant error. Considering this, and adding a discrete time backward Euler integrator to achieve zero steady state error, the control action is computed as follows:

$$i_{dhb}[k] = -K_{LV}\begin{bmatrix}\bar{V}_{busL}[k] - V^*_{busL}[k] \\ r_{0LV}[k]\end{bmatrix}, \tag{34}$$

$$r_{0LV}[k+1] = Ts\Big(\bar{V}_{busL}[k] - V^*_{busL}[k]\Big) + r_{0LV}[k], \tag{35}$$

where K_{LV} is a 1×2 gain vector, and $r_{0LV}[k]$ represents the state of the discrete integrator. Once (34) is computed, gain g is computed through (33), and used to generate the HV grid current reference i^*_{hv} through (24).

Gain vector K_{LV} can be found using the linear quadratic regulator method, or by pole placement using Ackerman's formula, and the matrix description of the system:

$$x_{LV}[k+1] = A_{LV}x_{LV}[k] + B_{LV}i_{dhb}[k], \tag{36}$$

where $x_{LV}[k] = [V_{busL}[k]\ r_{0LV}[k]]^T$,

$$A_{LV} = \begin{bmatrix}1 & 0 \\ T_s & 1\end{bmatrix}, \tag{37}$$

$$B_{LV} = [\frac{T_s}{C_L/2}\ 0]^T. \tag{38}$$

In this paper the closed loop poles are chosen to achieve a settling time of 100[ms].

3.4 Inverter control

This controller implements the active damping (AD) of the output LC filter resonance. The implementation of the AD requires to measure the current through C_{inv}. To avoid this measurement, an estimator through high pass filter derivation is used. This estimator only requires the measurement of v_{lv}, and only slightly increases the settling time of the control loop. The LV grid voltage reference v^*_{lv} is added to the control action of the AD. This voltage reference is a three-phase balanced sinusoidal signal, which can be generated internally, or can be obtained through a synchronization algorithm in synchronism with the HV grid voltage.

Since each leg of the inverter is controlled independently, there are three equal controllers. In what follows, they are described with subscript $Y = r$, s or t. Considering a one sample processing time, the control action for the active damping strategy is $v_{invY}[k] = v^*_{invY}[k-1]$, where

$$v^*_{invY}[k] = -K_{lv}\begin{bmatrix}\hat{i}_{CinvY}[k] \\ v_{lvY}[k] \\ v^*_{invY}[k-1]\end{bmatrix} + K^*v^*_{lvY}, \tag{39}$$

where K_{lv} is a 1×3 gain vector, K^* is a gain, and $\hat{i}_{CinvY}[k]$ is the estimated capacitor current, obtained from the zoh discretization of a high pass filter:

$$\eta_Y[k+1] = (e^{-\omega_c T_s} - 1)v_{lvY}[k] + e^{-\omega_c T_s}\eta_Y[k], \tag{40}$$

$$\hat{i}_{CinvY}[k] = C_{inv}\omega_c\Big(v_{lvY}[k] + \eta_Y[k]\Big), \tag{41}$$

with ω_c the cut off frequency of the high pass filter, and η_Y the state of the filter.

Gain vector K_{lv} is obtained by pole placement to damp the LC filter resonance using Ackerman's formula, and the matrix description of the system:

$$x_{lv}[k+1] = A_{lv}x_{lv}[k] + B_{lv}v^*_{invY}[k], \tag{42}$$

where $x_{lv}[k] = [i_{invY}[k]\ v^*_{invY}[k-1]\ v_{invY}[k]]^T$,

$$A_{lv} = \begin{bmatrix}A_{inv} & B_{inv} \\ 0\ 0 & 0\end{bmatrix}, \tag{43}$$

$$B_{lv} = [0\ 0\ 1]^T. \tag{44}$$

Finally, gain K^* is included to compensate the magnitude of the LV grid voltage reference, and is obtained evaluating:

$$TF = C_{lv}(z\,I - A^{CL}_{lv})^{-1}B_{lv}\Big|_{z=e^{j\omega T_s}}, \tag{45}$$

$$K^* = |1/TF|, \tag{46}$$

where $A^{CL}_{lv} = A_{lv} - B_{lv}K_{lv}$, I is the 3×3 identity matrix and $C_{lv} = [0\ 1\ 0]$. Note that the controllers for each phase use the same gain vector K_{lv} and gain K^*. In this paper the closed loop poles are chosen for optimal damping ($\zeta = 0.707$). This results in a settling time of approximately 2[ms].

4 Parameter selection criteria

This section gives criteria for choosing the values of the different parameters of each converter.

4.1 Rectifier coupling inductance L_{rec}

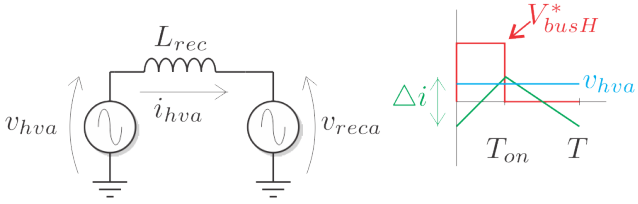


Figure 3: Rectifier phase a and switching interval variables.

The value of L_{rec} is chosen to obtain a desired current ripple $\triangle i$ when injecting zero current to the HV grid. Due to the 5-level structure of the rectifier and the use of unipolar modulation, the effective switching frequency applied to L_{rec} is $4f^{rec}_{sw}$ [1]. Also, this multilevel structure ensures that the voltage difference applied to L_{rec} will never be larger than V^*_{busH} (assuming all the HV buses are kept at that level).

Taking phase a as a reference, Figure 3 shows one switching interval. The current variation in this interval is

$$\triangle i = \frac{T_{on}}{L_{rec}}(v_{hva} - V^*_{busH}). \tag{47}$$

Since zero current injection is assumed, in this interval

$$v_{hva} \simeq V^*_{busH} d, \tag{48}$$

where $d = T_{on}/T$. Therefore, replacing this in (47),

$$\triangle i = \frac{V^*_{busH}}{L_{rec} 4f^{rec}_{sw}}(d^2 - d), \tag{49}$$

where $T = 1/(4f^{rec}_{sw})$ was used. The maximum $\triangle i$ occurs for $d = 0.5$, therefore for this condition, from (49),

$$L_{rec} = \frac{V^*_{busH}}{\triangle i 16 f^{rec}_{sw}}. \tag{50}$$

From Table 1, choosing the peak current ripple $\triangle i/2 = 0.1(\sqrt{2}I_{nomhv})$, the inductance results $L_{rec} = 190$[mH]. From an additional analysis not included in this paper, considering commercially available cores, $L_{rec} = 200$[mH] is chosen.

4.2 DHB transformer leakage inductance L_d

From (6), assuming $V_{busHX}[k] = V^*_{busH}$ and $V_{busL}[k] = V^*_{busL} = V^*_{busH}/m$ the maximum power transfer occurs for $\delta_X[k] = \pi/2$ and results

$$P_{dhbMAX} = \frac{(V^*_{busH})^2}{32 L_d f^{dhb}_{sw}}. \tag{51}$$

Under nominal operation conditions, each DHB will have to transfer a mean power $\bar{P}_{dhb} = S_{nom}/6$. Taking a safety margin of $P_{dhbMAX} = 2\bar{P}_{dhb}$ to catch any transients, and replacing this in (51), the leakage inductance results

$$L_d = \frac{3(V^*_{busH})^2}{32 S_{nom} f^{dhb}_{sw}}. \tag{52}$$

Using the necessary parameters from Table 1 to evaluate (52), it results $L_d = 8.44$[mH]. From an additional analysis not included in this paper, designing L_d considering commercially available cores, it results $L_d = 8.8$[mH] .

4.3 HV bus capacitor C_H

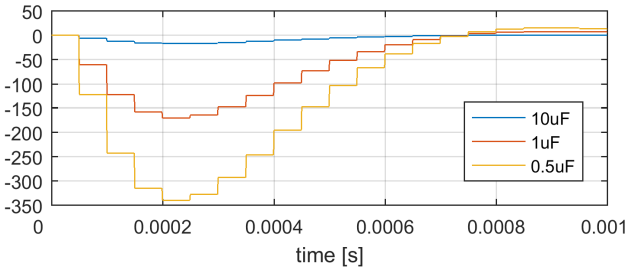


Figure 4: $\triangle V_{busH}$ [V] for a 10% HV grid voltage dip for different values of C_H.

During the event of dips or swells in the HV grid voltage v_{hv}, there will be a transient behavior in each of the HV buses. This can lead to unacceptable under/over voltages in V_{busHX}. The HV buses voltage variations during these transients are determined by the value of C_H, the settling time of each DHB control loop, and the magnitude of the dips or swells. Since the evaluation of these transients involves the DHB control loop, the process is iterative. For a given value of C_H, gain K_{dhb} must be computed. Then, the performance is evaluated through a simple simulation applying the desired dip or swell magnitude. Finally, the process is repeated until the desired performance is attained.

The evaluation is performed considering a step change in the magnitude of v_{hv}. In the following, the worst case scenario for V_{busH1} is analyzed, which is similar for the remaining buses.

The maximum instantaneous power variation in phase a occurs when the rectifier is sinking nominal current, and at the instantaneous peak of $v_{hva}[k]$ its magnitude changes $\triangle v_{hv}$ volts. From (1), assuming that previous to the step $\vec{v}_{hv} \simeq \vec{v}_{rec}$, the magnitude of the change in the HV grid current is

$$\triangle i_{hv} \simeq \frac{T_s}{L_{rec}} \triangle v_{hv}. \tag{53}$$

From (2), assuming that previous to the step $V_{busH1} = V_{busH1} = V^*_{busH}$, the variation in current i_{i1} is

$$\triangle i_i = \frac{\sqrt{2} V_{nomhv} \triangle i_{hv}}{2 V^*_{busH}} = \frac{T_s}{L_{rec}} \frac{\sqrt{2} V_{nomhv} \triangle v_{hv}}{2 V^*_{busH}}, \tag{54}$$

where (53) was used for the last equality. Now the effect of $\triangle v_{hv}$ in V_{busH1}, defined as $\triangle V_{busH}$, is evaluated simulating the closed loop response of the linearized DHB (30)-(32):

$$\triangle x_{dhb}[k+1] = A^{cl}_{dhb}\triangle x_{dhb}[k] + B^{i_i}_{dhb}\triangle i_i, \quad (55)$$
$$\triangle V_{busH}[k] = [1\ 0\ 0]\triangle x_{dhb}[k], \quad (56)$$

where $B^{i_i}_{dhb} = [\frac{T_s}{C_H/2}\ 0\ 0]^T$ and $A^{cl}_{dhb} = A_{dhb} - B_{dhb}K_{dhb}$. Figure 4 shows simulation results for this system for $C_H = 10$ [μF], 1[μF] and 0.5[μF] when a 10% dip is simulated ($\triangle v_{hv} = -0.1V_{nomhv}$) for the DHB control designed with a settling time of 1[ms]. From the results $C_H = 1$ [μF] is chosen, since it results in a 2.5% voltage variation.

4.4 LV bus capacitor C_L

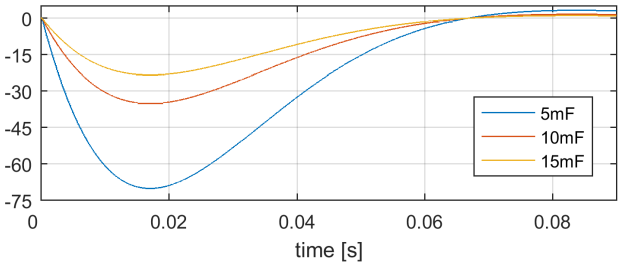


Figure 5: $\triangle V_{busL}$ [V] for a nominal load sudden connection and different values of C_L.

The value of capacitor C_L is chosen so that in the event of a sudden nominal load connection, the LV bus voltage does not go below the minimum value required for normal operation. The minimum LV bus voltage required for operation is

$$V_{busLMIN} = 2\sqrt{2}V_{nomlv} = 622V, \quad (57)$$

where the value is obtained from Table 1. Therefore, a conservative design criterion is to keep $V_{busL} > 700V$ when a nominal load is connected. This is equivalent to obtain a LV bus voltage variation $\triangle V_{busL} < 100V$.

To evaluate the LV bus voltage variation for different values of C_L, gain vector K_{LV} is computed for each given value, and then a simulation is performed. From (36)-(38), the following system is simulated:

$$\triangle x_{LV}[k+1] = A^{cl}_{LV}\triangle x_{LV}[k] + B^{cl}_{LV}\triangle i_L, \quad (58)$$
$$\triangle V_{busL}[k] = [1\ 0]\triangle x_{LV}[k], \quad (59)$$

where $A^{CL}_{LV} = A_{LV} - B_{LV}K_{LV}$, $B^{cl}_{LV} = -B_{LV}$ and $\triangle i_L = S_{nom}/V^*_{busL}$. Figure 5 shows simulation results for this system for $C_L = 5$ [mF], 10[mF] and 15[mF] when a nominal load is suddenly connected and the LV bus voltage control loop is designed with a settling time of 100[ms]. From the results all capacitor values meet the requirements, however $C_L = 10$ [mF] is chosen because of reduced ripple under unbalanced load conditions.

4.5 Inverter filter $L_{inv}C_{inv}$

The typical output impedance of a standard transformer is 2%-5% of its base impedance. Therefore, L_{inv} is chosen so that its impedance at angular frequency ω is 2% of the base impedance:

$$L_{inv} = 0.02\frac{3V^2_{nomlv}}{\omega S_{nom}} = 462\ [\mu H]. \quad (60)$$

On the other hand, capacitor C_{inv} limits the bandwidth of the output. Here, to obtain a fast transient response, the cut-off angular frequency of the output filter is chosen

$$\omega_{cut-off} = 20\omega. \quad (61)$$

Since the resonance frequency of the filter is approximately equal to the cut-off frequency,

$$\omega_{res} \simeq \omega_{cut-off} = \frac{1}{\sqrt{L_{inv}C_{inv}}}, \quad (62)$$

then

$$C_{inv} = \frac{1}{(20\omega)^2 L_{inv}} = 55\ [\mu F]. \quad (63)$$

5 Operation analysis of the control system and simulation results

This section presents simulation results of the proposed SST when a load is suddenly connected and disconnected. Additionally, results for sudden nonlinear load connection are included. The results are obtained using the switching models of all converters. The system parameter summary is shown in Table 1.

Table 1: System parameter summary

RECTIFIER AND LV BUS		
Param.	Value	Description
S_{nom}	20[kVA]	SST nominal power
V_{nomhv}	7621[Vrms]	HV nom. phase voltage
I_{nomhv}	0.875[Arms]	HV nom. phase current
f	50[Hz]	HV grid frequency
L_{rec}	200[mH]	HV filter inductance
V^*_{busL}	800[V]	LV bus reference
C_L	10[mF]	LV bus capacitor
f^{rec}_{sw}	8[kHz]	Switching frequency
DHB		
C_H	1[μF]	HV bus capacitor
C^{dhb}_L	56[μF]	LV DHB capacitor
L_d	8.8[mH]	DHB leakage ind.
V^*_{busH}	6000[V]	HV bus reference
m	7.5	Transformer relation
f^{dhb}_{sw}	20[kHz]	Switching frequency
INVERTER		
L_{inv}	461.2[μH]	LC filter ind.
C_{inv}	55[μF]	LC filter cap.
V_{nomlv}	220[Vrms]	LV phase voltage
f^{inv}_{sw}	20[kHz]	Switching frequency

The controllers were designed so that the rectifier control loop has a settling time of 4.5[ms], the DHB control loop has a settling time of 1[ms], and the LV bus voltage control loop has a settling time of 100[ms]. The settling time of the inverter AD loop is defined by the cutoff frequency of its LC output filter plus the estimation of the current through C_{inv}. This settling time results approximately 2[ms].

5.1 Sudden load connection

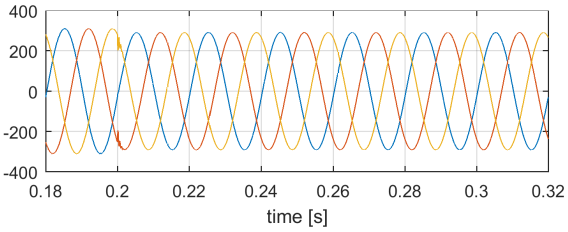


Figure 6: v_{lv} [V] sudden load connection.

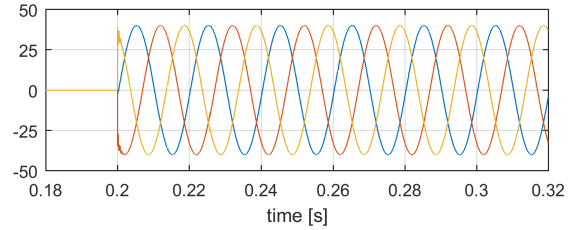


Figure 7: i_{lv} [A] sudden load connection.

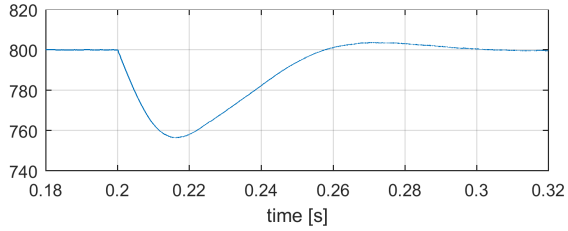


Figure 8: V_{busL} [V] sudden load connection.

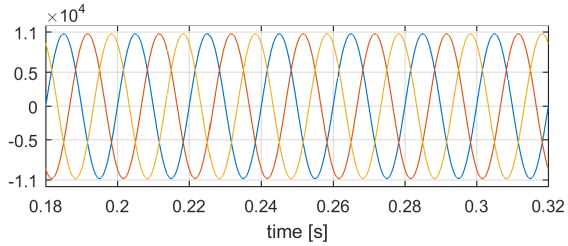


Figure 9: v_{hv} [V] sudden load connection.

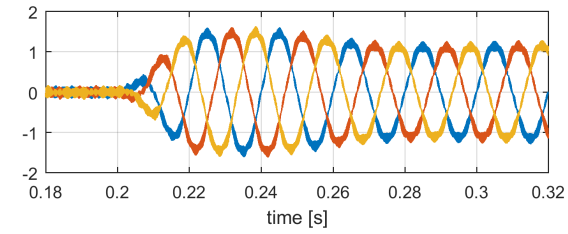


Figure 10: i_{hv} [A] sudden load connection.

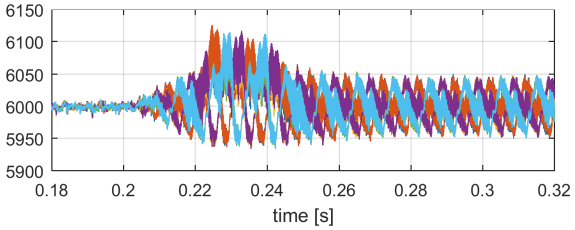


Figure 11: $V_{busH1} - V_{busH6}$ [V] sudden load connection.

The simulation results shown in Figures 6-11 start at $t = 0.18$[s] with the SST in steady state, with no load connected on the LV side. This means that the HV and LV buses start at their reference voltage levels and that a balanced three-phase sinusoidal voltage v_{lv} is generated without taking significant current i_{hv} from the HV grid.

At $t = 0.2$[s] a resistive nominal load is connected at the inverter output. The following sequence of events occurs (refer to Figure 2 for the definition of the variables):

- Figures 6 and 7 show v_{lv} and i_{lv}, respectively. As expected, when the load is connected v_{lv} has a short transient.
- Through power balance, current source i_L takes current from capacitor C_L, reducing voltage V_{busL} as shown in Figure 8. As can be seen, its settling time is approximately 100[ms], as designed.
- The LV bus voltage control loop detects the reduction of $\bar{V}_{busL}$, increasing in turn the value of g, resulting in $g > 0$. As a result, current reference i^*_{hv} magnitude increases.
- Commanded by its current controller, the rectifier sinks active power from the grid, with unity power factor. Figures 9 and 10 show v_{hv} and i_{hv}, respectively. Here the settling time of the current is tied to the settling time of g, defined by the LV bus control loop.
- Through power balance, the current source outputs of the rectifier charges capacitors C_H of the HV buses, increasing their voltage levels. The voltages of the HV buses is shown in Figure 11.
- The control loop of each DHB detects the increase in their respective HV bus voltage V_{busHX}. As a result, it commands each DHB to sink current i_{oX} in order to decrease the instantaneous value V_{busHX} to its reference value V^*_{busH} once again. This results in the short transient increase in the mean value of voltages V_{busHX} seen in Figure 11. Through power balance, each output current i_{dhbX} sources current to the LV bus, charging C_L to V^*_{busL} again, as shown in Figure 8.

By the end of this sequence, in steady state, the system is delivering power to the load and taking current with unity power factor from the HV grid.

5.2 Sudden load disconnection

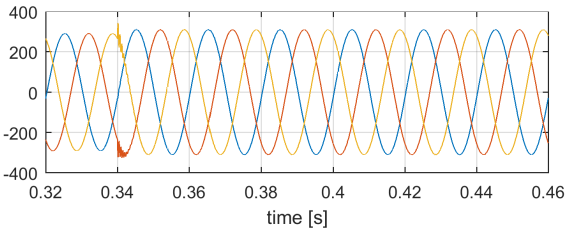


Figure 12: v_{lv} [V] sudden load disconnection.

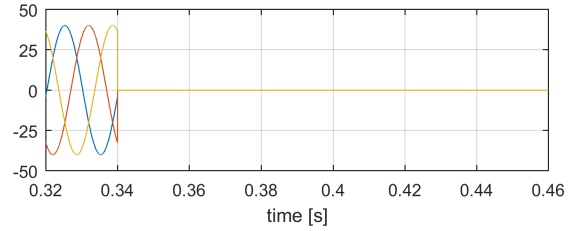


Figure 13: i_{lv} [A] sudden load disconnection.

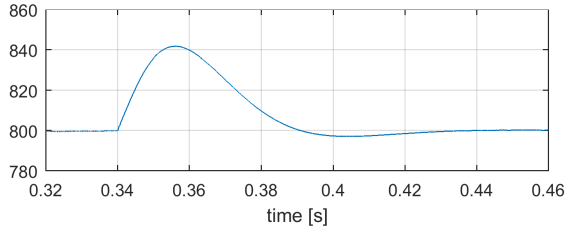


Figure 14: V_{busL} [V] sudden load disconnection.

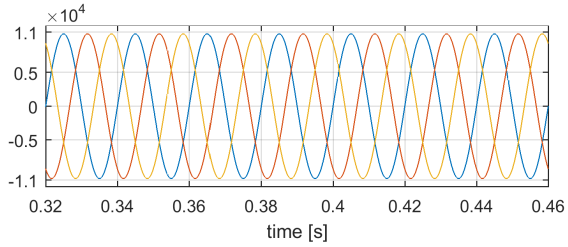


Figure 15: v_{hv} [V] sudden load disconnection.

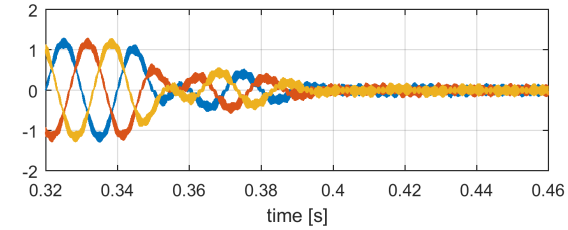


Figure 16: i_{hv} [A] sudden load disconnection.

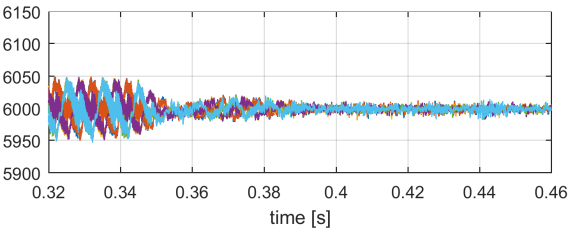


Figure 17: $V_{busH1} - V_{busH6}$ [V] sudden load disconnection.

Starting from the previous condition, if the load is now disconnected, the following sequence of events will occur:

- Figures 12 and 13 show v_{lv} and i_{lv}, respectively. As expected, when the load is disconnected v_{lv} has a short transient.
- Through power balance, current source i_L stops sinking current from capacitor C_L, which charges because the DHBs are still transferring power from the HV grid. This is seen in Figure 14.
- LV bus voltage control loop detects the voltage increase in $\bar{V}_{busL}$, decreasing in turn the value of g, resulting in $g < 0$. As a result, current reference i^*_{hv} magnitude decreases.
- Commanded by its current controller, the rectifier goes from sinking to supplying active power to the grid, with unity power factor. Figures 15 and 16 show v_{hv} and i_{hv}, respectively.
- Through power balance, the current source outputs of the rectifier discharge capacitors C_H of the HV buses, decreasing their voltage levels, as seen in Figure 17.
- The control loop of each DHB detects the decrease in their respective HV bus voltage V_{busHX}. As a result, it commands each DHB to source current i_{oX} in order to increase the instantaneous value V_{busHX} to its reference value V^*_{busH} once again. This results in the short transient decrease in the mean value of voltages V_{busHX} seen in Figure 17. Through power balance, each output current i_{dhbX} sinks current from the LV bus, discharging C_L to V^*_{busL} again, as shown in Figure 14.

By the end of this sequence, in steady state, the system has its buses at their reference values, and the inverter generates v_{lv} without sinking current i_{hv} from the HV grid.

5.3 Sudden non-linear load connection

The simulation of section 5.1 is repeated for the sudden connection of a balanced non-linear load. For each phase of the inverter, the load is composed of a single phase rectifier, feeding a 1[mH] inductor in series an RC load where R and C are in parallel (with R=19.5[Ω] and C=1[μF]). The results of this simulation are shown in Figures 18-23. As can be seen in these results, the proposed topology works well under highly non-linear load conditions.

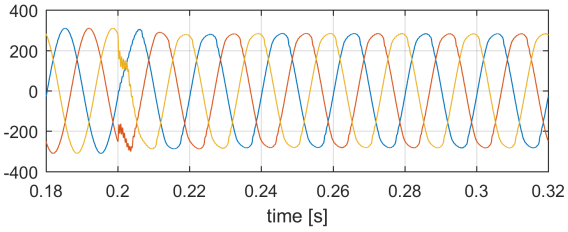


Figure 18: v_{lv} [V] sudden non-linear load connection.

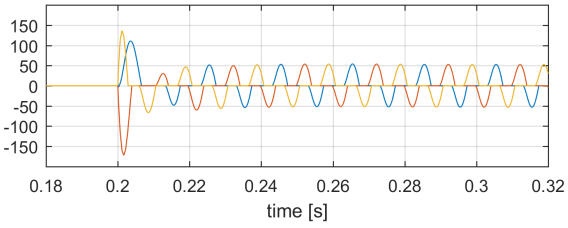


Figure 19: i_{lv} [A] sudden non-linear load connection.

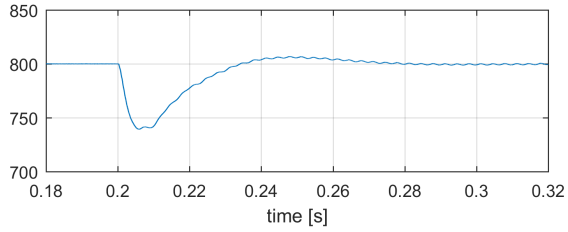


Figure 20: V_{busL} [V] sudden non-linear load connection.

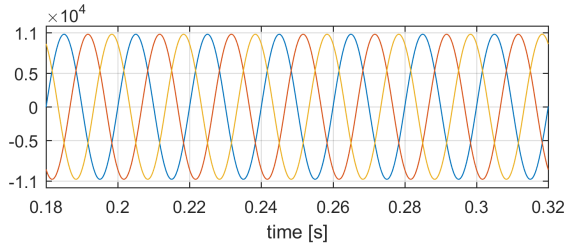


Figure 21: v_{hv} [V] sudden non-linear load connection.

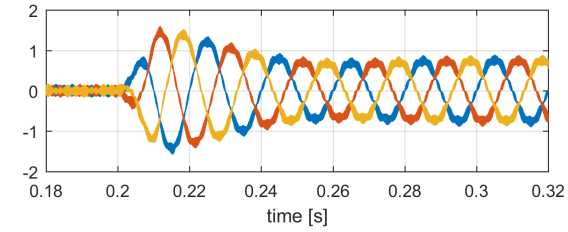


Figure 22: i_{hv} [A] sudden non-linear load connection.

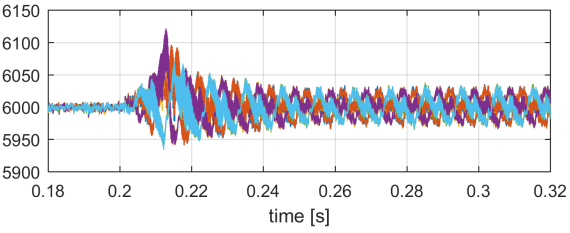


Figure 23: $V_{busH1} - V_{busH6}$ [V] sudden non-linear load connection.

5.4 Sudden non-linear unbalanced load connection

The simulation of the previous section is repeated, using the same non-linear load, but connecting these loads to only two of the three phases. The results of this simulation are shown in 24-29. As can be seen in these results, the proposed topology also works well for unbalanced non-linear loads.

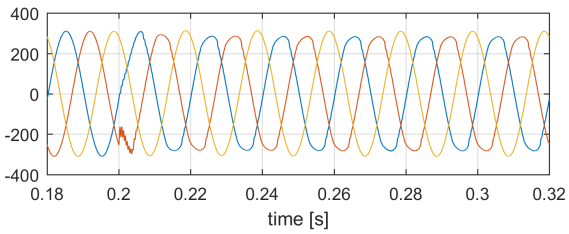


Figure 24: v_{lv} [V] sudden non-linear unbalanced load connection.

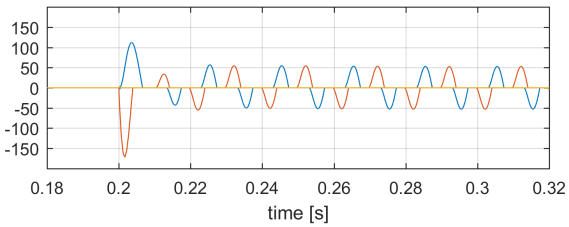


Figure 25: i_{lv} [A] sudden non-linear unbalanced load connection.

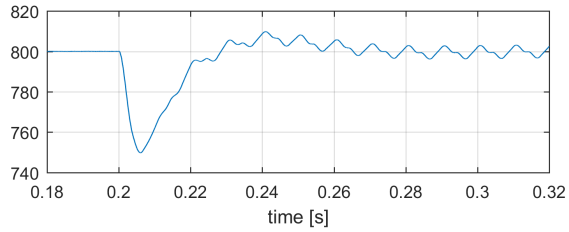


Figure 26: V_{busL} [V] sudden non-linear unbalanced load connection.

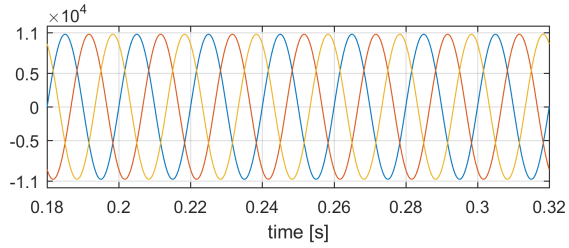


Figure 27: v_{hv} [V] sudden non-linear unbalanced load connection.

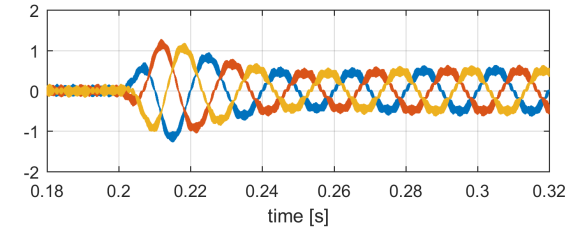


Figure 28: i_{hv} [A] sudden non-linear unbalanced load connection.

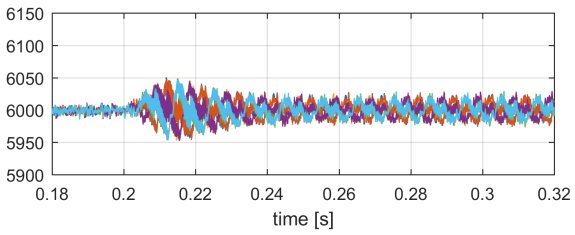


Figure 29: $V_{busH1} - V_{busH6}$ [V] sudden non-linear unbalanced load connection.

6 Conclusions

In this paper a model and a control strategy for an SST topology is proposed. The analyzed SST is designed using three separate stages: a rectifier, a DC-DC converter and an inverter. The proposed design approach is to model each stage using simple models. These simple models help to design and tune the controllers for each stage. The main focus of the tuning procedure is to obtain good performance to sudden load connections. Criteria to choose the main parameters of the SST are also given.

To validate the proposed method, simulation results are presented. The results show that the proposed topology and control strategies perform as expected from design conditions. Moreover, the staged design of the SST allows to decouple the HV side from the LV side in regard to load disturbances, which is an additional feature of the SST when compared to traditional transformers. Simulations were performed for both linear and non-linear loads. Moreover, a non-linear unbalanced load was also tested, with good performance results.

Acknowledgment The authors are grateful to CONICET, UNIVERSIDAD NACIONAL DEL SUR and ANPCyT for their institutional and economic support

Appendix

Given a three-phase signal $f = [f_a\ f_b\ f_c]^T$, its $\alpha\beta$ components are obtained through Clarke's transform:

$$\begin{bmatrix} f_\alpha \\ f_\beta \end{bmatrix} = \frac{2}{3}\begin{bmatrix} 1 & -0.5 & -0.5 \\ 0 & \sqrt{3}/2 & -\sqrt{(3)}/2 \end{bmatrix}\begin{bmatrix} f_a \\ f_b \\ f_c \end{bmatrix}. \quad (64)$$

Then, the complex vector $\vec{f}$ is defined as

$$\vec{f} = f_\alpha + j f_\beta, \quad (65)$$

where $j = \sqrt{-1}$.

References

[1] C. Busada, H. Chiacchiarini, S. G. Jorge, F. Mengatto, A. Oliva, J. Solsona, G. Bloch, and A. Delgadillo, "Modeling and control of a medium voltage three-phase solid-state transformer," in *2017 11th IEEE International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG)*, April 2017, pp. 556–561. DOI: 10.1109/CPE.2017.7915232

[2] J. Van der Merwe and H. d. T. Mouton, "The solid-state transformer concept: A new era in power distribution," in *AFRICON 2009*, 2009. DOI: 10.1109/AFRCON.2009.5308264

[3] G. T. Heydt, "The next generation of power distribution systems," *IEEE Transactions on Smart Grid*, vol. 1, no. 3, pp. 225–235, 2010. DOI: 10.1109/TSG.2010.2080328

[4] S. Falcones, X. Mao, and R. Ayyanar, "Topology comparison for solid state transformer implementation," in *IEEE PES General Meeting*. IEEE, 2010, pp. 1–8. DOI: 10.1109/PES.2010.5590086

[5] X. She, R. Burgos, G. Wang, F. Wang, and A. Q. Huang, "Review of solid state transformer in the distribution system: From components to field application," in *2012 IEEE Energy Conversion Congress and Exposition (ECCE)*. IEEE, 2012, pp. 4077–4084. DOI: 10.1109/ECCE.2012.6342269

[6] X. She, A. Q. Huang, and R. Burgos, "Review of solid-state transformer technologies and their application in power distribution systems," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 1, no. 3, pp. 186–198, 2013. DOI: 10.1109/JESTPE.2013.2277917

[7] J. W. Kolar and G. Ortiz, "Solid-state-transformers: key components of future traction and smart grid systems," in *Proc. of the International Power Electronics Conference (IPEC), Hiroshima, Japan*, 2014.

[8] F. Wang, A. Huang, X. Ni *et al.*, "A 3.6 kv high performance solid state transformer based on 13kv sic mosfet," in *2014 IEEE 5th International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*. IEEE, 2014, pp. 1–8. DOI: 10.1109/PEDG.2014.6878693

[9] J. E. Huber and J. W. Kolar, "Volume/weight/cost comparison of a 1mva 10 kv/400 v solid-state against a conventional low-frequency distribution transformer," in *2014 IEEE Energy Conversion Congress and Exposition (ECCE)*. IEEE, 2014, pp. 4545–4552. DOI: 10.1109/ECCE.2014.6954023

[10] Z. Yu, R. Ayyanar, and I. Husain, "A detailed analytical model of a solid state transformer," in *2015 IEEE Energy Conversion Congress and Exposition (ECCE)*. IEEE, 2015, pp. 723–729. DOI: 10.1109/ECCE.2015.7309761

[11] H. Fan and H. Li, "High-frequency transformer isolated bidirectional dc-dc converter modules with high efficiency over wide load range for 20 kva solid-state transformer," *IEEE Transactions on Power Electronics*, vol. 26, no. 12, pp. 3599–3608, Dec 2011. DOI: 10.1109/TPEL.2011.2160652

[12] C. A. Busada, S. G. Jorge, A. E. Leon, and J. A. Solsona, "Current controller based on reduced order generalized integrators for distributed generation systems," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 7, pp. 2898–2909, July 2012. DOI: 10.1109/TIE.2011.2167892

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 125-129 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A 3D Full Wave Inversion (FWI) Analysis for Handheld Ground Penetrating Radar (GPR)

Suki Dauda Sule*, Kevin Paulson

University of Hull, School of Engineering and Computer Science, HU6 7RX, United Kingdom

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
GPR
Handheld
Full wave inversion

ABSTRACT

This paper reports the results of an empirical, 3D full wave inversion (FWI) numerical analysis which provides an estimation of the performance of multi-static handheld ground penetrating radar (GPR,) compared to a bi-static system, for landmine detection using FWI imaging. The experiments are based on simulated data and provide a more realistic evaluation of the performance of multi-static handheld GPR for the landmine problem based on FWI, than previous studies based on a 2D analysis.. Furthermore, a novel method of estimating a parameter set that is closer to the global minimum for an iterative FWI optimization is introduced.

1. Introduction

This paper provides an extension of the work presented in [1], which is based on a sensitivity analysis of handheld GPR measurements, comparing bi-static and multi-static system performance using singular value decomposition (SVD). The results obtained confirmed the conclusions of a previous study conducted by Watson [2], which states that multi-static systems achieve greater subsurface information, for a landmine detection application using full wave inversion (FWI). The study in [2] also conducted a 2D FWI to verify the superior performance of multi-static arrays over a bi-static configuration, for handheld GPR. It concludes that the size of the multi-static array, or number of antenna elements, is insignificant. Furthermore, the acquisition system under test was simplistic and radio propagation properties such as the antenna cross-coupling, radiation pattern and geometry are neglected. Additionally, only a flat, homogenous domain was simulated for the 2D FWI analysis.

Nevertheless, a 2D FWI study is insufficient, as the landmine detection domain is a 3D, heterogeneous domain and a 3D analysis is required to produce a more realistic evaluation of the performance of bi-static and multi-static handheld GPR systems for the demining challenge. Here, an empirical FWI numerical analysis is reported for a 3D domain, which considers a homogenous as well as a heterogeneous or cluttered domain, using a derivative-free optimization algorithm. The bi-static and multi-static system performance are evaluated by comparing the estimated subsurface parameters obtained in each case with synthetic GPR data parameters. All the radio propagation effects are included with simulated experiments conducted in the CST STUDIO SUITE 3D electromagnetic (EM) environment. Finally, a novel method to determine an improved initial parameter set for the GPR FWI optimization, that is closer to the global minimum, is proposed. This is achieved by generating a database of prior forward problem solutions. The database is used with each measured GPR data to determine the least L2 norm objective function, whose parameters are considered as the initial parameter set for the actual FWI optimization.

*Corresponding Author: Suki Dauda Sule, University of Hull, +447443143897
Email: s.d.sule@2008.hull.ac.uk

2. FWI for Multi-static Handheld GPR

2.1. Gradient Based Method

The GPR FWI problem for landmine detection may be posed as a regularised least squares (LSQ) non-linear optimisation problem given by [2]

$$\mathbf{X}_{im} = \arg\min_{\mathbf{x}} \frac{1}{2}\left\|GPR(\mathbf{X}) - \mathbf{d}\right\|_2^2 + \lambda T(\mathbf{X}) \qquad (1)$$

where $\mathbf{X}_{im}$ is a vector of geometric and electrical parameters describing the ground, $\mathbf{d}$ is the GPR measured data, $GPR(\mathbf{X})$ is a forward model that returns the GPR measurement that would be made for a subsurface with parameter vector $\mathbf{X}$. The GPR inverse problem is ill-posed as arbitrarily large changes in $\mathbf{X}$ can have negligible effect on the error $\left\|GPR(\mathbf{X}) - \mathbf{d}\right\|_2^2$. The regularisation term

is required to control the size of components of $\mathbf{X}$ with little or no effects on GPR data. The function $T(\mathbf{X})$ introduces a penalty based on the size of these components and is a way to introduce prior information. For Tikhonov regularisation $T(\mathbf{X})=\|\mathbf{\Gamma X}\|^2$ and often $\mathbf{\Gamma}$ is chosen to be the identity matrix. The regularisation or Tikhonov factor $\lambda \geq 0$ is often adjusted dynamically during the iterative optimisation process to control convergence. Optimisation is posed as a LSQ problem as the error functional uses the L2 norm.

The GPR forward model is non-linear and so (1) is often solved by iterative linearization. Watson used an iterative, quasi Newton method called the limited memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) nonlinear optimisation algorithm [3]. The solution requires a calculation of the gradients of the LSQ error function $\|GPR(\mathbf{X})-\mathbf{d}\|_2^2$. Due to the special form of the LSQ error function, these derivatives can be directly related to the derivatives of the forward model.

2.2. Derivative-free Method

The derivative-free method for solving the FWI problem is used because it is computationally expensive to estimate the derivatives required for derivative based methods. Derivative-free methods require less computation per iteration and are suitable for a limited number of variables, but may require more iterations. A range of non-derivative nonlinear optimization algorithms may be used to solve the FWI optimisation problem when gradients do not exist or are expensive to compute. Local direct search methods may be used when there are a small number of variables and the objective function is computationally cheap to evaluate. The Nelder-Mead simplex algorithm [4] is the most cited, robust and efficient of local direct search methods. The algorithm fundamentally relies on an initial number of points that create a simplex i.e. a set of points spanning a volume in the variable dimensionality considered. The objective function is evaluated at the vertices of the simplex during each iteration to determine the highest objective value, which is used to redefine another vertex that produces a new simplex. Additional new points are produced by moving the vertex with the highest objective value through a series of transformations using the centroid of the associated simplex that include reflection, expansion, internal and external contractions [5]. These steps are iterated until convergence is achieved. Convergence can occur for non-smooth objective functions even when the second derivative of the function is unobtainable [6]. The Nelder Mead simplex algorithm may be applied to the FWI optimisation problem, to estimate the uncertainty in parameter sensitivity of the handheld GPR bi-static and multi-static systems. The Nelder Mead Simplex algorithm is embedded in the CST STUDIO SUITE environment and so the 3D GPR forward model may be integrated with the optimization process on the same platform.

3. Modelling and Simulation

3.1. The GPR System Models and Experiments

The GPR system models are developed with linear antenna arrays positioned with dipoles at a fixed distance above the ground surface. The configurations are for a bi-static system, and multi-static system with two and four receive elements, driven in a single input multiple output (SIMO) sequence. The antennas are end-fed (coaxial) dipole antennas designed and optimised using the Antenna Magus software for a centre frequency of 2.5 GHz and frequency range of 1.75-3.25 GHz. Assuming the same offset of 20 cm used by Watson, an antenna element spacing of 5 cm was used in all configurations. Antenna cross-coupling losses are significant, given this spacing which is less than a wavelength. The antennas are placed initially at a height of approximately 3.76 cm above the ground surface. The transmitting element is at the extreme left end of the array and the time series measured at each receive antenna are concatenated into a single data vector **d**. The target object is placed under the centre of the array, based on the assumption that a metal detector (MD) has located a conducting part of the device.

To reduce the computational cost, the ground size for this study is reduced to a 31 cm by 29 cm box with a depth of 9cm. The subsurface parameters are the relative permittivity $\varepsilon_r^{soil}=2.53$ and loss tangent $\tan\sigma=0.0036$. The target is a single AP mine which is modelled as a typical plastic cylinder with relative permittivity $\varepsilon_r^{mine}=2.8$ (typical US M14 mine). The diameter and height are 7cm and 6cm respectively, closer to a Colombian military MN-MAP-1 mine [7]. The mine also contains a tetryl charge (US M14) with relative permittivity $\varepsilon_r^{tetryl}=2.163\ \left(1g/cm^3\right)$ and an air void of free space relative permittivity of one. Instead of ABC boundary conditions (used by Watson), perfectly matched layer (PML) boundary conditions are applied on all faces of the box (ground) for all models in this study with added space above the antennas on the top of box to simulate the antenna to antenna and antenna to ground propagation. ABC boundary conditions have been studied longer and are generally easier to implement, but the PML region achieves less boundary reflections [8].

For a heterogeneous ground (cluttered domain), we introduce a rough surface to the flat ground (box) surface. The ground surface roughness height is modelled in the form of a Gaussian distribution [9]. Therefore, the box planar surface is modelled with depressions and protrusions to simulate a ground surface with a random height, assuming a Gaussian distribution, normally distributed white noise and a mean value of zero [10]. Additionally, subsurface clutter sources are modelled as several 3D rectangular blocks, grouped into two clusters, with each set having a different relative permittivity. One has a relative permittivity, $\varepsilon_r=5$ and the other with a relative permittivity, $\varepsilon_r=4$. An example bi-static dipole system and target subspace is shown in Figure 1. A LSQ FWI regularised optimisation problem, as in (1), is the test problem. The goal value to be evaluated is the sum squared difference of the total simulated GPR data (time-series) and the total forward model data (time-series), which is given by $\|GPR_{meas}-GPR(\mathbf{X})\|^2$. Watson performed a similar study for dipole antennas in a simplified 2D, homogenous domain only. Here both 3D homogenous and heterogeneous domains are investigated for the FWI analysis.

The optimization is set to a maximum of 20 iterations, due to computational constraints. The hexahedral mesh used to produce

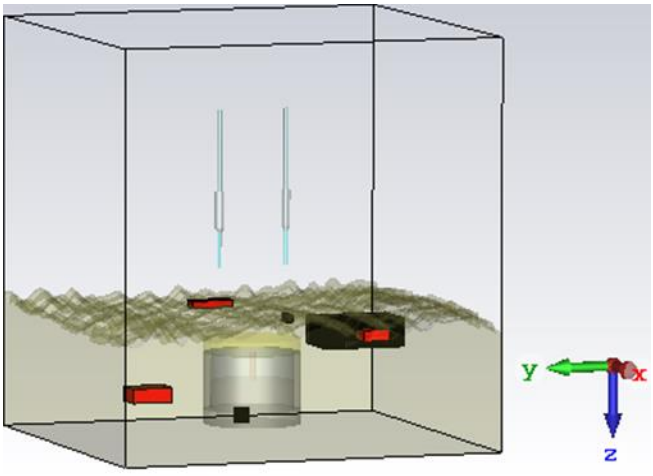


Figure 1: Bi-static dipole system for a heterogeneous ground

synthetic GPR measurements is different from the mesh used in the forward model within the FWI objective function. This avoids the spuriously good results that can be the inverse crime artefact [11] due to using the same mesh for both operations.

The parameter vector for the synthetic GPR data for the homogenous ground is given by

$$GPR_{initial} = [\varepsilon^T, \tan\sigma, \varepsilon^0, \varepsilon_r^m, \varepsilon_r^s] = [2.163, 0.0036, 1, 2.8, 2.53] \quad (2)$$

where ε^T = relative permittivity of tetryl charge

$\tan\sigma$ = loss tangent

ε_0 = relative permittivity of free space

ε_r^m = relative permittivity of plastic mine

ε_r^s = relative permittivity of dry sandy soil

Whereas the parameter vector for the synthetic GPR data for the heterogeneous ground is given by

$$GPR_{initial} = [\varepsilon^T, \varepsilon^{C1}, \varepsilon^{C2}, \tan\sigma, \varepsilon^0, \varepsilon_r^m, \varepsilon_r^s] = [2.163, 3.75, 6, 0.0036, 1, 2.8, 2.53] \quad (3)$$

where ε^{C1} = relative permittivity of first clutter source

ε^{C2} = relative permittivity of second clutter source

ε_0 = relative permittivity of free space

ε_r^m = relative permittivity of plastic mine

ε_r^s = relative permittivity of dry sandy soil

The initial subsurface parameter vector for the homogenous ground FWI is given by

$$GPR_{initial} = [\varepsilon^T, \tan\sigma, \varepsilon^0, \varepsilon_r^m, \varepsilon_r^s] = [2.363, 0.0046, 1.2, 3.01, 2.23] \quad (4)$$

Whereas the initial vector for the heterogeneous ground FWI is given by

$$GPR_{initial} = [\varepsilon^T, \varepsilon^{C1}, \varepsilon^{C2}, \tan\sigma, \varepsilon^0, \varepsilon_r^m, \varepsilon_r^s] = [2.363, 5, 4, 0.0046, 1.2, 3.01, 2.23] \quad (5)$$

3.2. FWI Numerical Analysis Results

To estimate the performance of the different antenna systems, we compare and consider the estimated parameter values approached by the FWI solutions for both models with their respective synthetic GPR data true parameter values. The error in the subsurface parameters for each system in the different ground conditions is estimated by determining the total sum squared difference between the GPR data vector parameters and the FWI solution vector parameters. The total estimated parameter errors are indicated in Tables 1 and 2 for the homogenous ground and heterogeneous ground respectively.

The 3D FWI data shows that for both ground domains, the conclusion that multi-static systems can achieve more subsurface information is achieved as the smallest subsurface error in both cases is recorded with a multi-static system. At this stage, we can confirm that the multi-static system performs better than the bi-static system in general. However, the subsurface parameter error is not monotonic for the number of antenna elements whereas it was found to be linear for a 2D FWI analysis. These results based on a 3D analysis are more realistic as the scattering on the soil surface and cylindrical mine introduce more degrees of freedom and complexity than a 2D numerical analysis.

For the homogeneous soil, the 4 receiver (RX) system FWI achieves a better performance than the two other systems though the difference between the bi-static system and 2 RX multi-static system is marginal. However, in the more realistic heterogeneous domain the 4 RX multi-static system yields the largest parameter uncertainty. The scattering and clutter signals are observed to be larger in the heterogeneous domain than the homogenous domain based on the larger subsurface parameter error and uncertainty figures that are estimated in the former. This is due to greater air-ground reflections from the rough ground surface and scattering from the buried clutter sources. In this case the clutter signal is much larger than the mine signal. Nevertheless, for both soils, particularly the heterogeneous soil, the parameterisation does not sufficiently describe the clutter and so the optimisation easily converges to the wrong solution. Some method is required to reduce the effects of the clutter signal. Therefore, better imaging with multi-static systems for a real GPR system and demining operation based on FWI is predicated on an optimised antenna design as well as clutter reduction.

Conversely, this analysis also shows that the multi-static system achieves only a small improvement over the bi-static system. A FWI optimization with actual GPR measurements or field evaluation data is expected to exhibit greater complexity and hence the improvement may only be marginal. This study has been limited to synthetic data and future studies may include further

validation of the results obtained with the use of measured GPR data.

Table 1: Summary of GPR and FWI solution parameter values for homogenous ground

Subsurface Parameters	GPR Data Parameter Values	FWI Solution Estimated Parameter Values		
		Bi-static	Multi-static (2RXs)	Multi-static (4RXs)
Charge relative permittivity	2.163	2.292	2.268	2.3
Loss tangent	0.0036	0.0026	0.0034	0.0069
Air void relative permittivity	1.0	1.49	1.5	1.45
Mine relative permittivity	2.8	3.069	3.185	3.032
Soil relative permittivity	2.53	2.2	2.2	2.6
Estimated parameter error	**-**	**1.219**	**1.320**	**0.892**

Table 2: Summary of GPR and FWI solution parameter values for heterogeneous ground

Subsurface Parameters	GPR Data Parameter Values	FWI Solution Estimated Parameter Values		
		Bi-static	Multi-static (2RXs)	Multi-static (4RXs)
Charge relative permittivity	2.163	2.4	2.388	2.314
Clutter1 relative permittivity	3.75	2.25	2.25	2.25
Clutter2 relative permittivity	6.0	7.28	6.67	8.0
Loss tangent	0.0036	0.0058	0.0066	0.0068
Air void relative permittivity	1.0	1.35	1.29	1.39
Mine relative permittivity	2.8	3.04	2.96	3.16
Soil relative permittivity	2.53	2.27	2.33	2.2
Estimated parameter error	**-**	**3.869**	**3.048**	**4.734**

4. Improvement of the FWI Initial Parameter Set

The steepest descent and direct search methods for iterative nonlinear optimisation all require an initial parameter set which is updated iteratively according to the chosen method. The GPR FWI problem is also a local minimisation problem that is nonlinear as well as ill-posed. A good initial parameter set that is as close as possible to the true solution is desirable to ensure convergence to the global minimum and less computational expense. A more general approach would be to employ global optimization techniques prior to the local optimization. However, for the GPR problem this would be computationally prohibitive. We propose a compromise approach that involves solving the forward problem for several parameter sets within the bounded local domain or search space and using an L2 norm objective with the true solution or measured data, as the goal value. The forward model meshing or grid would be set to the lowest tolerance level as only a coarse analysis is required to avoid a high computational expense. A higher tolerance is set for the forward model grid for the actual optimization.

A database of the forward model solutions for a specific operation is generated and can be evaluated with data from each GPR measurement. The initial parameter vector for the FWI solution is chosen by interpolation from the database of parameter vectors and measured time-series. Hence a single database is generated during a single campaign but can be used repeatedly for GPR data from the same source environment or location. The database campaign could be done during the training phase of a demining operation prior to the actual clearance operation. The database can be generated for any chosen number of parameter combinations and sample space or bounded conditions. More forward model solutions would be expected to increase the probability of a better initial estimate of the parameter set. For this experiment, the 4 RX dipole system GPR model for a heterogeneous domain is utilised. Eleven forward model solutions are generated by arbitrarily choosing eleven different parameter sets for values between a minimum and maximum bound. This yields a database of eleven sets of A-scan data and the L2 norm objective value of each of these data for any measured GPR is determined. The parameter set for the time-series (A-scan) that achieves the lowest objective function value would be considered as the closest to the true solution or global minimum and selected as the initial optimization parameter set. Figure 2 presents the objective function value for all eleven forward problem solutions.

It can be determined from Figure 2 that the lowest objective function value is obtained at simulation run eight which corresponds to a value of 0.001708. The parameter set for this forward model measurement, GPR_F is given by

$$\begin{aligned} GPR_F &= [\,\varepsilon^T, \varepsilon^{C1}, \varepsilon^{C2}, \tan\sigma, \varepsilon^0, \varepsilon_r^m, \varepsilon_r^s\,] \\ &= [2.25, 5.75, 6.50, 0.0061, 1.35, 3.05, 2.55] \end{aligned} \quad (6)$$

Therefore, this parameter set given in (6) is selected as the initial parameter set for the iterative FWI solution, for the heterogeneous domain under test. The FWI optimization solution for this parameter set is then obtained for 20 iterations, due to computational constraints, and compared directly with the FWI solution for the initial parameter set in (3). The result is shown in Figure 3.

The comparison of the FWI solution for the original initial parameter set and the one derived from the database generation (Figure 3) shows that the latter does not achieve an improvement in the accuracy of convergence as the absolute error is marginally less than the original solution. However, the result does show that the database generated initial parameter FWI solution is closer to the true solution. This could potentially lead to a more efficient optimization using a more powerful algorithm. Derivative based methods would benefit from this improvement and achieve convergence with less computational expense and fewer iterations as a local minimum would be realised more efficiently. The procedure achieves the primary goal of improving the initial parameter estimation or guess. The CST integrated Nelder-Mead optimisation does not allow the user to specify the entire simplex and objective function values. If this had been the case, the initial 10 iterations exploring the simplex could have been entirely

avoided by selecting all the simplex points from the database. The performance of gradient based algorithms are expected to benefit from the database generated improved initial parameter set.

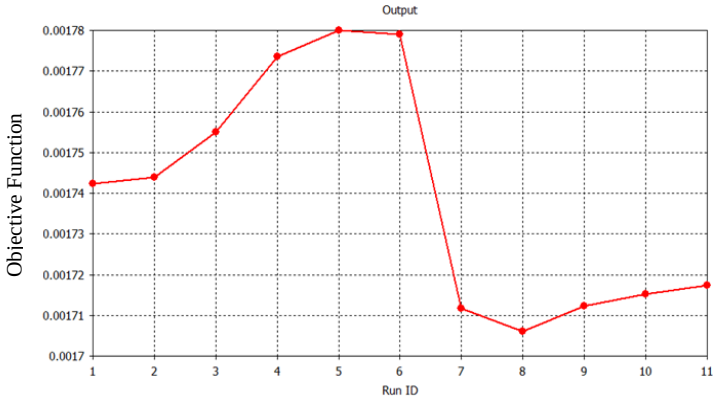


Figure 2: Objective function values for eleven forward problem solutions

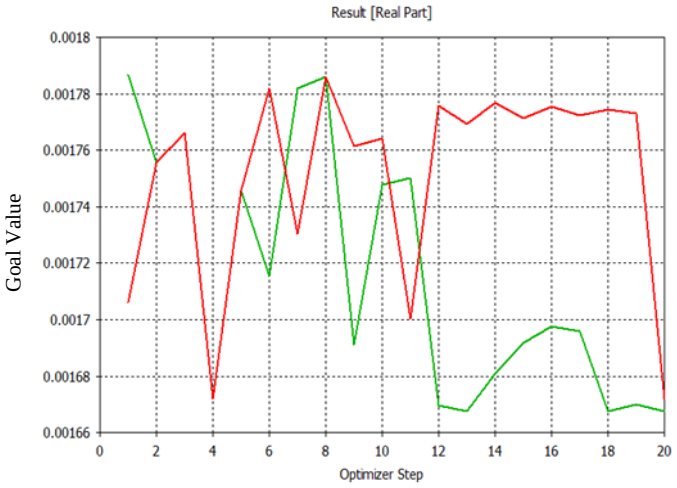


Figure 3: FWI solution result for 4 RX dipole a. original initial parameter set (GREEN) versus database selected parameter set (RED)

5. Conclusion

An empirical study has been undertaken to compare 3D FWI imaging using multi-static and bi-static systems in homogeneous and heterogeneous media. The results verify the possibility of multi-static systems to achieve greater subsurface parameter sensitivity and hence reliability of target detection than bi-static systems. In 2D, all multi-static systems outperformed the bi-static systems. However, a more realistic 3D analysis shows that the improvement in performance with increasing numbers of antennas is not simple or monotonic due to cross-coupling and antenna patterns. This underlines the need for optimisation of the antenna system configuration and size (number of elements), to achieve better performance than a bi-static system. Additionally, the effect of clutter significantly limits the accuracy of parameter estimation. Finally, a novel procedure has been proposed to determine the initial parameter vector for the FWI solution which yields an initial forward problem solution that is closer to the true GPR data solution. The procedure requires numerous forward problem solutions stored prior to a deeming campaign but has the potential to significantly reduce the computational expense of the FWI as well as the accuracy for an ideal local minimisation.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to acknowledge the Petroleum Technology Development Fund (PTDF), Abuja, Nigeria.

References

[1] Sule, S.D. and Paulson, K.S., "A comparison of bistatic and multistatic handheld ground penetrating radar (GPR) antenna performance for landmine detection", IEEE Radar Conference (RadarConf), 2017, pp. 1211-1215. **DOI:** 10.1109/RADAR.2017.7944389

[2] Watson, F. and Lionheart, W., "SVD analysis of GPR full-wave inversion", 15th International Conference on Ground Penetrating Radar (GPR), 2014, IEEE, pp. 484-490. **DOI:** 10.1109/ICGPR.2014.6970472

[3] Nocedal, J. and Wright, S., Numerical optimization. Springer Science & Business Media, New York, 2006.

[4] Nelder, J.A. and Mead, R., "A simplex method for function minimization", The computer journal, **7**(4), pp. 308-313, 1965. https://doi.org/10.1093/comjnl/7.4.308

[5] Rios, L.M. and Sahinidis, N.V., "Derivative-free optimization: a review of algorithms and comparison of software implementations", Journal of Global Optimization, pp. 1-47, 2013. doi:10.1007/s10898-012-9951-y

[6] Mckinnon, K.I., "Convergence of the Nelder--Mead Simplex Method to a Nonstationary Point", SIAM Journal on Optimization, **9**(1), pp. 148-158, 1998. https://doi.org/10.1137/S1052623496303482

[7] Lopera, O. and Milisavljevic, N., "Prediction of the effects of soil and target properties on the antipersonnel landmine detection performance of ground-penetrating radar: A Colombian case study", Journal of Applied Geophysics, **63**(1), pp. 13-23. https://doi.org/10.1016/j.jappgeo.2007.02.002

[8] Nataf, F., "Absorbing boundary conditions and perfectly matched layers in wave propagation problems, Direct and Inverse problems in Wave Propagation and Applications", 14, de Gruyter, Radon Ser. Comput. Appl. Math., pp.219-231, 2013, 978-3-11-028228-3. https://hal.archives-ouvertes.fr/hal-00799759

[9] Tajdini, M.M., Gonzalez-Valdes, B., Martinez-Lorenzo, J.A., Morgenthaler, A.W. and Rappaport, C.M., "Efficient 3D forward modeling of GPR scattering from rough ground", IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting, pp. 1686-1687, 2015. **DOI:** 10.1109/APS.2015.7305232

[10] Daniels, D.J., "The impact of antenna design on short range radar performance", IEEE Conference on Antenna Measurements & Applications (CAMA), pp. 1-4, 2014. **DOI:** 10.1109/CAMA.2014.7003338

[11] Kaipio, J. and Somersalo, E., "Statistical inverse problems: discretization, model reduction and inverse crimes", Journal of Computational and Applied Mathematics, **198**(2), pp. 493-504, 2007. https://doi.org/10.1016/j.cam.2005.09.027

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 128-137 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Mission-Critical Systems Design Framework

Kyriakos Houliotis[*,1], Panagiotis Oikonomidis[1], Periklis Charchalakis[2], Elias Stipidis[3]

[1]*Research Fellow, Vetronics Research Centre, University of Brighton, BN2 4GJ, United Kingdom*

[2]*Principle Research Fellow, Vetronics Research Centre, University of Brighton, BN2 4GJ, United Kingdom*

[3]*Professor, Vetronics Research Centre, University of Brighton, BN2 4GJ, United Kingdom*

ARTICLE INFO

Article history:
Received: 16 December, 2017
Accepted: 01 February, 2018
Online: 12 March, 2018

Keywords:
Vetronics
Mission-Critical systems
Model driven architecture
NGVA
Threat estimation

ABSTRACT

Safety-critical systems are well documented and standardized (e.g. IEC 61508, RTCA DO-178B) within system design cycles. However in Defence and Security, systems that are critical to the success of a Mission are not defined within the literature nor are there any guidelines in defining criticality in their design or operational capabilities. When it comes to Vetronics (Vehicle Electronics), a mission-critical system, is a system with much complexity and mixed criticality levels that is a part of the overall platform (military vehicle) offering integrated system capabilities. In this paper, a framework is presented, providing guidelines in designing efficiently and effectively mission-critical systems considering principles of Interoperable Open Architectures (IOA), mission-critical integrity levels and following new standardization activities such as NATO Generic Vehicle Architecture (NGVA). A Defensive Aid Suite (DAS) system is used as a case study to illustrate how this framework can be exploited. The indention of this extension is to provide an approach to precisely estimate threats in order to de-risk missions in the very early stages.

1. Introduction

Modern military vehicles rely on mission-critical systems that enhance and guarantee successful mission capabilities. Currently, these mission-critical systems come as black boxes that are installed and maintained by the same manufacturer through the vehicle's life-cycle. These black boxes are built on proprietary technology that only the manufacturer has access to, thus limiting the choices of maintenance and upgrades. Furthermore, existing mission-critical systems are limited to communicating with other on-board systems resulting to a vehicle having multiple instances of the same equipment (e.g. GPS sensor). This presents a number of issues including having network complexity and reduced flexibility in vehicle systems configuration depending on operational requirements.

For this reason, there is a need for an innovative architecture approach that allows components from different manufacturers to be integrated, paying particular attention to the system's mission, safety, and security. When building a mission-critical system, the system designer should have the freedom to choose components that fit appropriately to the intended use as well as enable integration to any legacy mission-critical system or sensors/actuators that exist on-board the vehicle.

In this study, a thorough investigation is conducted offering a new open modular architectural approach on mission-critical systems including a case study on Defensive Aid Suites (DAS), aided to extract the necessary technical and functional requirements directly related to the system. This presented an opportunity to research a conceptual approach to the bespoke system whereby a constructive framework could be established with firm recommendations on a high level (abstracted) design so that a target platform can be equipped with a tailored mission-critical system to meet its specific requirements.

The novel mission-critical system architecture for military platforms adopts an open and modular design approach offering flexibility in configuration, upgradability, and integration. This enables a better operational and functional understanding of the mission-critical system, increasing integrated survivability capabilities [1].

Following the presented framework of this work, qualitative and quantitative results are extracted in order to provide mission

[*]Kyriakos Houliotis, Vetronics Research Centre, University of Brighton, BN2 4GJ, United Kingdom, +44 (0) 1273 642251, k.houliotis@vetronics.org

functional concepts based on mission-critical systems and threats. Additionally from the results, an early de-risking estimation can be observed, that could be beneficial for stakeholders, systems engineers and architects to decide the appropriate elements for designing mission-critical systems in the very early stages of the overall system's life-cycle. The rest of the paper is structured as shown below:

- Section 2 presents some representative key questions that offer the direction of designing mission-critical systems.
- Sections 3 and 4 provides a definition for mission-critical systems in line with standardisation activities around military vehicle architecture design approach.
- Section 5 offers the Generic Architecture Framework.
- Section 6 identifies some key considerations on safety and security in aligning and defining mission criticality levels.
- Section 7 provides an approach of calculating threats using numerical values and mathematical equations.
- Sections 8, 9, 10 and 11 present a case study on how the framework can be used to define a DAS Architecture including qualitative and quantitative results.
- Section 12 concludes the paper with some indication of next steps to this research.

This paper is an extension of work originally presented in International Conference on Military Technologies (ICMT) 2017 [2].

2. Key Questions

2.1. Modularity and Openness

What approaches are needed for a mission-critical system to be feasible and extract advantages such as modularity and openness?

2.2. Construction, Maintenance and Safety Certification

What are the benefits of using modules sharing specific functionalities on a mission-critical system and how efficient and effective can become?

2.3. Alongside benefits from the framework

How the framework's modularity extracts through the mission-critical system development benefits such as safety cases and certifications?

2.4. Low-cost tools

What are the necessary tools needed to accomplish a rapid prototyping testing and permitting software functionality and operation of a mission-critical system using low-cost components?

2.5. Migration of low cost to a safety-critical performance verification testbed

How to achieve a transition from a low-cost functional testbed to a more elaborated safety-critical performance verification testbed?

3. Mission-Critical Systems

In general, the mission is the formal summary of the aims and values of an activity. The activity can be achieved with specific mission-critical elements. Those mission-critical elements are defined as vital to the functioning of an activity. Meaning that, a successful mission can be achieved when only the right mission-critical elements are applied. There are two attributes that make the specific mission-critical elements to be applied and to be right.

First, is usually when there is maturity in the applied mission-critical elements. The maturity must reach into a level that is satisfactory in each of the involved disciplines. When this level is reached, the expected outcome is sufficient and hence, the mission can be considered successful.

The second attribute is when enough knowledge is accumulated to allow for the prediction of a mission outcome to be more accurate. To achieve this, consideration needs to be given to all possible factors involved on the specific mission. Those factors are usually known or unknown and could be anything related to the mission. The knowledge can be gained when those factors are asked and answered using three main engineering questions; "What", "Why" and "How". Once these answers are mature enough and understandable the mission-critical elements can be referred to as vital and therefore, provide success to the mission.

Today, the technology has been developed in such a way that many missions could be successfully completed with the aid of systems. Those systems are referred to as mission-critical systems. A general definition of a mission-critical system is [3]:

"*A system that is essential to the survival of a service, and whose failure or interruption significantly impacts the mission*".

A mission-critical system for a typical land military vehicle is:

"*A system that is essential to complete the mission successfully*".

A mission-critical system in land military platform is composed of many discrete Vetronics (vehicle electronics) sub-systems and components including sensors, actuators, effectors, radars and processing resources. Each of these sub-systems may contain further sub-systems and components including mechanical parts.

In Vetronics the mission can be designed, described and/or accomplished either in simple or complex terms. This differentiation resulted from the characteristics that a Vetronics system has. A simple mission for a Vetronics system is when not many factors are involved. It is also simple, when a clear and an easy step-by-step procedure is provided. For instance, a mission-critical system has to transmit data from node A to node B. That can be described as a simple mission since there is only one task to be completed and if the right mission-critical elements are used.

However, in Vetronics, for a data to be transmitted from one node to another in reality it is more complicated than the previous example. What makes the mission more complicated in Vetronics mission-critical systems is when a number of multiple disciplines, such as safety, security and survivability, are involved to achieve the mission. A more desirable, refined and detailed mission procedure is required.

Assume each of the aforementioned disciplines require to complete a specific goal on the same mission. The safety prioritises the safety of people and environment; the security prioritises the protection of data from various threats; and survivability prioritise the whole mission envelop. This makes the mission more difficult

to accomplish if there is neither enough maturity nor confidence on the applied mission-critical elements.

In conclusion, "mission in military applications cannot be specified or narrowed down into a single element that easily". Therefore, an innovative unified framework is required to define and guide Mission-Critical Systems development so as to enhance mission success.

4. IOA International Activities

Today within a modern military platform, land, naval and air force, have adapted the principles of the Interoperable Open Architecture (IOA) in their system design to speed up acquisition and upgrading alongside with reducing life-cycle costs through data modelling. Below a selection of significant activities in the area of architectures and standardisation with IOA is presented.

4.1. Generic Vehicle Architecture (GVA)

The Generic Vehicle Architecture is an approach taken by the UK Ministry of Defence (MOD) to the design of electronic and power architectures for military vehicles. The approach is based on establishing system engineering principles to define a generic architecture that requires open implementation standards, Def-Stan 23-009, to support cost-effective integration of sub-systems on land platforms, electronically, electrically and physically. Any equipment shall be integrated with the GVA military land platforms must be designed in the Land Data Model (LDM) which is a (sub)-system standardisation process [4].

4.2. NATO Generic Vehicle Architecture (NGVA)

The NGVA is an approach to ensure interoperability among military land vehicles equipment. The NGVA follows a similar line to the GVA, incorporating a new method of verification and validation and by maturing the NGVA Data Model concepts, focus and implementation can be achieved [5], [6].

4.3. Future Airborne Capability Environment (FACE)

The FACE approach is an aviation US government-industry software standard and business strategy for acquisition of affordable software systems that promotes innovation and rapid integration of portable capabilities across global defence programs. The main objective of this approach is to make military operations more robust, interoperable and secure using open standards [7].

4.4. Vehicle Integration for C4ISR/EW Interoperability (VICTORY)

VICTORY is a US army vehicle's open standard for physical and logical interfaces between systems and C4ISR/EW components. The VICTORY architecture targets to provide a clear picture between to the users and the developers. Throughout the usage of an open architecture, the platforms can accept upgrades without a significant impact on the design.

5. Generic Architecture Approach

Model Driven Architecture (MDA) is an approach that is used in the system engineering domain to improve product development and delivery. The approach was initially launched in 2001 by the Object Management Group (OMG) to support the model-driven engineering of software systems. The main objective of the MDA is to provide a set of specifications of system's functionality and behaviour expressed in models. Instead of writing the code manually, the MDA approach with the help of a data modelling tool it is possible to regenerate automatically an application code. Additionally, this approach reduces implementation and integration risks when an activity is designed.

In Figure 1 an illustration of the MDA process is presented. Initially, the system requirements can be defined and specified into a Platform Independent Model (PIM) model. Using a specific standard and specification, the model can be constructed in a formal way such as the Unified Modelling Language (UML). The objective of the PIM model is to specify data, operations, functions and modes of a system, independent of the platform in which it may be integrated. In order to organise and standardise the data as well as facilitate a long-term improvement in interoperability and upgradability within the model, a data model is essential. A Platform Specific Model (PSM) model contains elements of a specific software platform. It can be generated from the PIM model either manually or automatically if appropriate tools are used. The PSM embeds the chosen software architecture strategies which refine the PIM model based on specifications. The Platform Specific Implementation (PSI) embeds the chosen middleware technology and explains the usage of a specific platform.

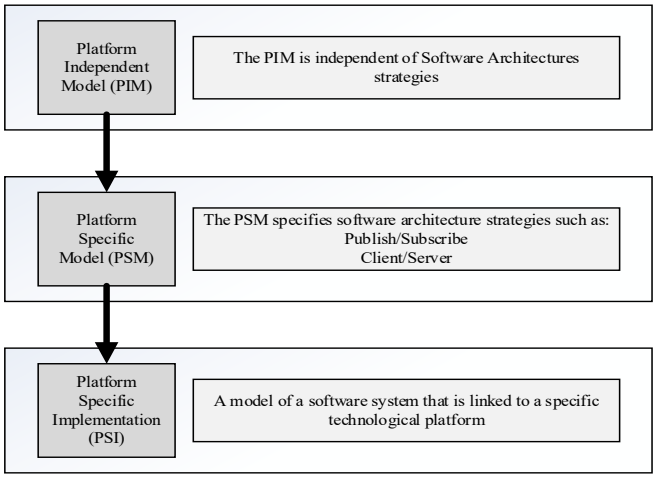


Figure 1. Model Driven Architecture Approach.

Each electronic device is developed to satisfy one or more specific task(s) thus, the devices can only generate or receive a set of specific data for their operation. Therefore, when these devices are integrated into an IOA architecture, the broadcasted data cannot be ensured if is critical or not. For instance, the real-time level cannot be defined just by the device itself but is needed to be declared from the designers. Below there is a brief explanation of the real-time levels.

Real-Time Level – In any electronic architecture there is a set of data or information it might be critical or noncritical. Real-time can enable non, soft and hard responsiveness depending on what level, prioritisation, and importance the data is designed for. In Vetronics different real-time levels are applied for satisfying different level processes or events. The definitions of real-time levels are:

- Non-Real-Time (**NRT**) - Best Effort Service with no time constraints.

- Soft Real-Time (**SRT**) - Relaxed time (latency) requirements.
- Hard Real-Time (**HRT**) - Fixed time requirements. [8]

Due to the different critical level of data in Vetronics, this paper proposes a novel framework to support the developers to design and decide whether the data is critical or noncritical. With the aid of a flowchart, Figure 2 demonstrates a modular framework that is aimed to accurately design a data for a mission-critical system. The designer must firstly define the data attribute and if is already existing on the model. When the data attribute is declared, the designer must decide in what level of real-time responsiveness the data is belonging to. The different levels the data can take are the Hard, Soft, and Non-Real-Time. This can be useful for the developers to choose the appropriate network communication technology. Next, the designer must declare the criticality level of the data by choosing between Mission, Safety, and Security Critical levels. The critical levels shown in the diagram are the most commonly used levels used in the Vetronics systems. The user has the freedom to add/remove other critical levels such as survivability or business critical.

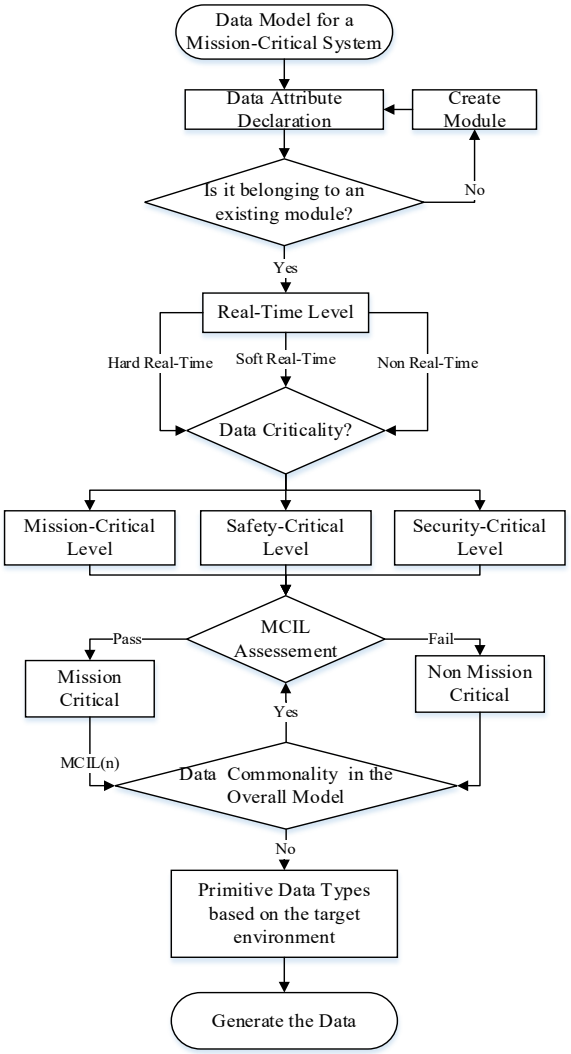


Figure 2. Mission-Critical Modelling Framework.

The next step is an assessment that is applied for the mission planning and goals based on risk assessments, a detailed explanation will be discussed in Sections 10 and 11. If the data fails the assessment is considered as non-mission-critical and if the data meets the requirements then is considered as mission-critical. In the mission-critical block, the data will be labelled with the Mission-Critical Integrity Level (MCIL). To ensure that the model is not polluted with data having similar attributes, a data commonality assessment is provided for addressing that issue. Finally, the data must have its own data type and must be checked whether the data type can be supported or not by the targeted programming language.

In a communication network, there are multiple connection points that are able to receive, store and send data across. These connection points are known as network nodes. Figure 3 depicts a generic network node which is divided into three elements, communication, processing, and application. The application element, is a computer program designed to perform a group of coordinated functions, tasks or activities for the benefit of the end user. The processing in a networked node, is a combination of machines, people, and processes that for a set of inputs produces a defined set of outputs. And the communication is the communication endpoint. The transmission of data from one computer to another is achieved by the communication device using various communication technologies [9].

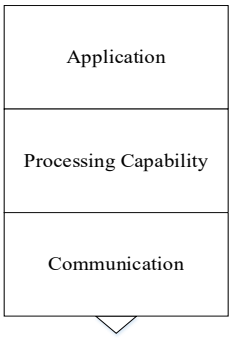


Figure 3. Network Node.

In the military vehicles, multiple (sub)-systems are integrated on a single platform. In a mission-critical system the most commonly used network nodes are sensors, actuators, effectors, processing nodes and so on. All the network nodes are interacting together through a virtual networking, as shown in Figure 4. A virtual network is capable of controlling one or more nodes over a logical or virtual networks that are decoupled from the underlying network hardware. This is used to ensure that the network nodes can efficiently integrate and perform on a single network. Common data sharing technologies used for military applications are the Data Distribution Service (DDS) and the Message Queuing Telemetry Transport (MQTT). The gateway block is integrated to support system legacy and allow different existing on-board systems to provide and receive services.

6. The need of Safety and Security

In land military platforms, there are other existing critical systems. These critical systems are for the safety and for the security of the platform and the crew. Safety critical systems are the systems whose failure may endanger human life, economics or the environment. Examples of safety critical systems in military

vehicles are the vehicle's steering and fire control. Security critical systems deal with the integrity and loss of sensitive data through theft or accidental loss [10].

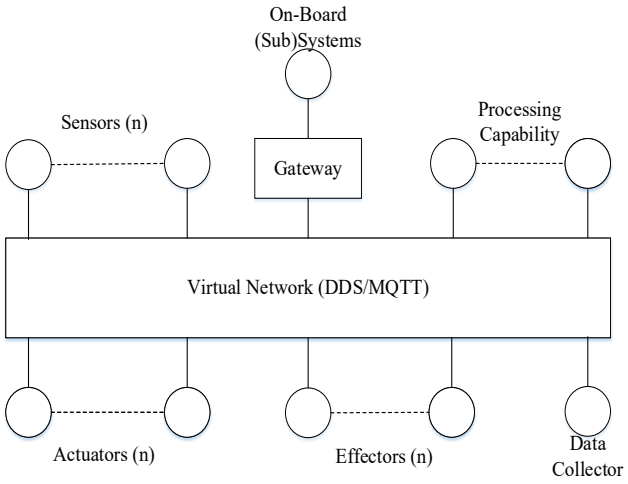


Figure 4. Virtual Network.

Safety and security critical systems can be used to extract logical capabilities, in which they can be mapped and used as the essentials of a successful mission-critical system. Considering safety and security capabilities on the development of a mission-critical system, it is possible to achieve a complete integrated survivability system [11]. It is vital to increase the dimensions and properties of safety and security in a mission-critical system to address issues such as intrusion detection, component fault/failure detection system behaviour, and restoring essential services in case of a security attack [12].

The developer must specify and approve the criticality level of each data. This can be varied depending on the application and technical specification related to user and system requirements. Although it is possible to declare and assess the data's mission-critical level using the Tables 1-4. The tables are essential to achieving a successful mission-critical system in which the mission goals are based on risk assessments. Risk assessments should be set and then that, the rigour of management and processes should be appropriate to meeting them.

Assume a civilian vehicle used for a military application; when the vehicle's passenger opens the vehicle's door some interior lights are switched on. The interior lights are designed to provide luminosity in the vehicle at night or in dark environments. During the day or at bright environments the light does not significantly impact any process thus the related data for the light can be categorised as negligible or MCIL4. During the night the light may provide luminosity to the passengers but it also indicates the vehicle's location. If the vehicle is used for a mission during night, it is likely the mission will fail, thus the corresponding data, for the light, can be categorised to as catastrophic or MCIL1. However, any data must be filtered and tested through the risk assessments before is integrated into the platform. Furthermore, the framework requires, that hazard and risk assessments be executed for the analysis of the likelihood of occurrence, consequences and detection levels provided by the tables below.

Table 1 Categories of likelihood of failure occurrence

Category	**Definition**	**Range (Mission Failure)**
Frequent	Many times in missions	$> 10^{-3}$
Probable	Several time in missions	10^{-3} to 10^{-4}
Occasional	Once in mission	10^{-4} to 10^{-5}
Remote	Unlikely in missions	10^{-5} to 10^{-6}
Improbable	Very Unlikely	10^{-6} to 10^{-7}
Incredible	Cannot believe that it could occur	$< 10^{-7}$

Table 2 Consequence categories

Category	**Definition**
Catastrophic	Complete mission failure
Critical	Impacts mission but not complete failure
Marginal	Major mission issues
Negligible	Minor mission issues

Table 3 Risk class matrix

Likelihood	**Consequence**			
	Catastrophic	***Critical***	***Marginal***	***Negligible***
Frequent	Class 1	1	1	2
Probable	1	1	2	3
Occasional	1	2	3	3
Remote	2	3	3	4
Improbable	3	3	4	4
Incredible	4	4	4	Class 4

The classification of the consequences are as follow:

- **Class 1**: Unacceptable in any circumstance
- **Class 2**: Undesirable: tolerable only if risk reduction is impracticable or if the costs are grossly disproportionate to the improvement gained.
- **Class 3**: Tolerable if the cost of risk reduction would exceed the improvement.
- **Class 4**: Acceptable as it stands, though it may need to be monitored.

Once the hazard and risk assessments are identified, each of the threats should also be assigned with a detection level, as given in Table 4, in order to provide a definition in which degree a threat can be detected.

Table 4 Detection Levels

Difficulty Level	Definition
No effort	Very likely to be detected
Very Easy	With almost no effort
Easy	Without great effort
Normal	Conforming to a standard
Hard	With a great deal or effort
Very Hard	Not likely to be detected

Mission-critical levels, see Table 5, offers the ability to attain in regards to mission-critical system development and related to the Classification Matrix given in Table 3. Throughout risk assessments the target MCIL can be identified and thus, it can be converted as a requirement for the mission-critical system. The derived requirements can provide an efficient data model development that can be used and ensure that the mission-critical system can succeed a mission.

Table 5 Mission-Critical Levels

Mission-Critical Integrity Level	Mission Failure Factor	Risk Classification
MCIL 4	100,000 to 10,000	Class 4
MCIL 3	10,000 to 1,000	Class 3
MCIL 2	1,000 to 100	Class 2
MCIL 1	100 to 10	Class 1

Safety (mission) critical applications, such as in [13] and [14], are using safety standards, such as the IEC 61508 and the RTCA DO-178B, in which are oriented for people's safety when a system is designed. The standards are intended to be a basic functional safety standard applicable to cover the safety management of electrical, electronic and programmable electronic systems throughout their lives. If the development of a mission-critical system involves human factors then the system should be considered as a safety critical system.

7. Threat Estimation

First, it is important to note that Tables 1, 2 and 4 are abbreviated as, Occurrence – O[n], Severity – SE[n], Detection – D[n] and threat estimation as T[n], with "[n]" representing a natural number [1] of each element or requirement. For example, if one threat is identified within the framework, then the threat will be assigned to as T[1]. If threat T[1] has sub-requirements, then it will be assigned to as T[1][n].

In order to evaluate or estimate the criticality of the identified threat T[n], the aforementioned tables and their elements must be assigned with numerical values, as depicted in Table 6. The idea behind the values is indicative (assumption), therefore, the maximum value of the threat T[n] can be roughly 99.9% and the lowest 0%. These values will indicate the probability of the threat affecting the mission.

Therefore, the threat level of the identified threat T[n], TL_T[n] can be calculated using the following expression,

$$TL_T[n] = O[n] + SE[n] + D[n] \quad (1)$$

Where,

TL_T[n]: The threat level of the identified threat.

O[n]: The occurrence value of the O[n].

SE[n]: The severity value of the SE[n].

D[n]: The detection value of the D[n].

In the event of having multiple threats or sub-threats, an estimation of an overall threat must be calculated in order to predict the probability of the mission success. A representation of this is shown in Table 7.

Where,

Req_core: The core requirement.

1st Sub: The first sub_core_requirement.

n Sub: Indicated the last sub-requirement.

n: Represents the real number.

i: Represents the sequential number of sub-requirements.

An approach on how to calculate the average value of two or more requirements of the same degree is as follows, (2),

Table 6 The Assigned Values for Calculating Threat T[n]

Occurrence – O[n]	Occurrence (%)	Severity – SE[n]	Severity (%)	Detection – D[n]	Detection (%)
Frequent	25	Catastrophic	≤50	Very Hard	25
Probable	20	Critical	~33.4[3.s.f]	Hard	20
Occasional	15	Marginal	~16.7[3.s.f]	Normal	15
Remote	10	Negligible	0	Easy	10
Improbable	5			Very Easy	5
Incredible	0			No Effort	0

Table 7 Requirement Sequence

Req	1st Sub	n Sub
	Req[1][1]	Req[1]…i…[1]
Req[1]	…	…
	Req[1][n]	Req[1]…i…[n]
	…	…
	Req[n][1]	Req[n]…i…[1]
Req[n]	…	…
	Req[n][n]	Req[n]…i…[n]

$$Req_core = \sum_{i=1}^{n} \frac{Req[i]}{n} \quad (2)$$

Where:

Req_core: The overall average value of the core requirement.

i: Lower limit number of requirement.

n: Upper limit number of requirement.

8. Case Study: Defensive Aid Suite (DAS) System

A most commonly used mission-critical system in the military platforms, is the Defensive Aid Suite (DAS) system. Is a survivability system, and can be used as a potential case study that addresses similar complexity issues to the aforementioned mission-critical systems and integration levels. A DAS is composed of sensors, effectors, algorithms and Human Machine Interfaces (HMI) that enhances the integrated survivability of a military vehicle. DAS also consists of different decoupling physical and logical capability networks applied for specific tasks or applications.

The DAS system can be a semi-autonomous or autonomous system that is capable of detecting, recognising and addressing threats. DAS elements are classified into two major categories Soft-Kill and Hard-Kill; other action categories are also possible where effective countermeasures can be deployed by other assets.

A **Hard-Kill** system engages and destroys threats. It creates an active fire zone of protection at a safe distance near the vehicle.

A **Soft-Kill** system is designed to avoid threats by confusing or re-directing the threats using jammers, decoys, and signature reduction measures.

An existing soft-kill DAS system from [14] has been selected for this case study, using existing DAS system components. In this DAS system various electronic components can be addressed, which are used in a Light Armed Vehicle (LAV) vehicle. A component such as the Long Range Passive Sensing (LRPS) sensor. The threat can be detected using optical systems with either Wide Field Of View (WFOV) or Narrowed Field Of View (NFOV) mounted on the platform. Each of the detected threats produces a signature identifying a potential threat that may be weapon systems such as guns and anti-tank rocket-propelled grenade launchers (i.e. M-712 and RPG-7). The detection range of the WFOV and NFOV optics are represented in Table 8. The table can be used as a message specification for a generic DAS system.

Table 8 Sensor Camera and Threats Attributes [15]

Anti-Armour Threats	***Threat, Calibre***	M-712, LSAH, 155mm	RPG-7, 80mm	Gun, 20mm, APDS
IR WFOV	***Distance, [m]***	400	470	5480
IR NFOV	***Distance, [m]***	3600	4200	340
LI/RG Camera	***Threat, [Pixels]***	1.3	42 x 42	0.8
	Target, [Pixels]	25x20	234 x 187	118 x 60
Threat Variables	***Dimensions, [m]***	0.155 dia.	0.18 dia.	2.1 dia.
	Range, [m]	14000	500	2000
	Velocity, [m/s]	255	255	815

9. Proposed DAS Architecture and Modules

This section presents a novel DAS system architecture that could be potentially used to satisfy the questions in Section II as presented earlier in this paper. For the system to facilitate a modular framework the Model Driven Architecture approach is used to construct modules that enable upgradability, maintainability and system legacy with technologies, devices and operational or functional capabilities. This proposed DAS architecture aims to fuse systems and software modelling and simulation capabilities, modular open system architectures and device integration techniques into a single package to enable rapid design, development, verification, certification and deployment of interoperable, platform portable and manoeuvre embedded mission criticality. The following sub-sections are classified as module models used for a DAS system and can also be applied for any mission-critical system.

9.1. Threats

Threats are all the known causes that can damage the platforms. Threats can be either external or internal; internal threats could be cyber-attacks, malfunction etc. External threats could be missiles or mines.

9.2. Sensors

Sensors module represents all the candidate sensors used for the DAS system. This module is for detecting and responding to any incoming threat from the physical environment. The specific input may be motion, heat, light or any of other environmental phenomena. The output is a signal that is converted into a format that is human readable or machine readable electronically transmitted over a network for exploitation and further processing.

The Sensors module can be directly connected to the DAS Sensor Processing Module to apply the safety, security, and performance on the indented design.

9.3. DAS Sensor Processing Module (DASSPM)

The DASSPM is the module that is capable of receiving the data/information from the DAS sensor. The DASSPM converts the data into a usable format that can classify, position and eventually eliminate or avoid the detected threat. The module will be able to detect also the attributes of the candidate DAS sensor in terms of type and capability if the appropriate message specification is constructed.

The DASSPM, therefore, receives the data/information from the candidate sensor and extra qualities can be added. Developers can customise the data and modify it into the corresponding real-time environment, criticality level, and mission-critical integrity level. The DASSPM will also communicate with the DASECM.

9.4. DAS Computer Module (DASCM)

The DASCM collects and processes the information from the DASSPM, to decide the best action for each available threat. This action may require user interaction or it can be fully automated. The modular design allows the system designer to use a single or multiple DASC that may deal with the same or different types of threats at the same time, providing fault tolerant and distributed design. Then the DASC communicated with the appropriate DAS Effector Control Module (DASECM), depending on the type of action decided. When there is no DASC available, the DASECM takes responsibility for the appropriate countermeasure action.

9.5. DAS Effector Control Module (DASECM)

The DASECM is responsible for controlling the DAS effectors according to the commands received from the DASCM. This module has many commonalities in data and functionality with the DASSPM. In the case of DASC failure, the DASECM can be actioned to directly deal with the threat. Additionally, it offers configuration options for the effector to achieve modularity and dynamic configuration.

9.6. Effectors

The Effectors Module is the representation of the all the candidate effectors. This module is constructed upon the available effectors needed to be installed on the platform. Each of the candidate effectors carries specific attribute thus, the module is using message specifications common to the overall architecture.

10. Top Level of a DAS Data Model

Due to the increased number of electronic components designed from different manufacturers who use different technologies, approaches, standards and architectures, the systems integration process becomes rather complex. A suggested approach to improve the vehicle's performance and survivability is to use, information management techniques. In this paper, the potential approach to such component's capability acquisition is proposed by structuring a set of message specifications and develop the DAS Data Model, represented in Figure 5.

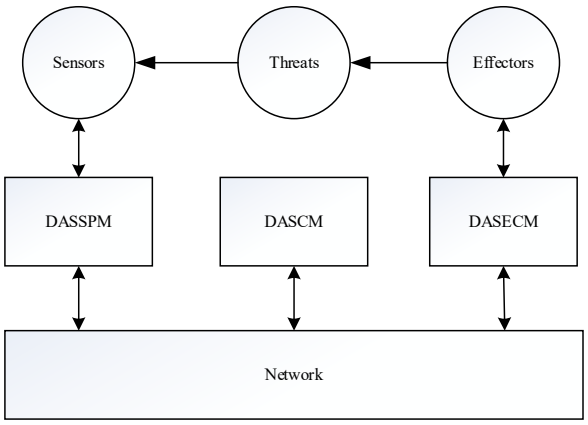


Figure 5. Top Level of DAS Architecture.

Following the proposed framework and the proposed DAS architecture and modules, the system will be able to use different battlefield scenarios, operational modes and different electronic devices from different manufacturers. With this development and the combination of the proposed approaches, the DAS Data Model will provide the different criticality levels and the different real-time environments that will be the essentials for a successful mission.

10.1. DAS PIM

Using the IBM Rational Rhapsody Developer for C++ tool, the message specification for the soft-kill DAS system sensors is represented in a UML model as shown in Figure 6. The tool is selected because it is the most commonly used for land vehicle electronic architectures, specifically in the GVA approach implementation [4]. This a PIM in which it has the ability to be redesigned when additional modifications are required during the development, or additional sensors are included in the DAS system. This PIM model is constructed using the attributes in Table 8 and with the usage of the proposed MCIL framework and DAS architecture, the mission-critical message specifications of a soft kill DAS can be created.

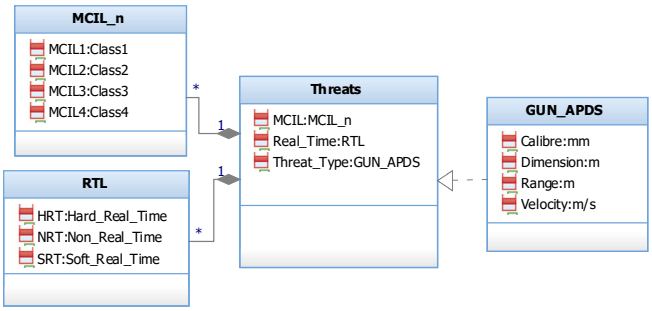


Figure 6. Threat PIM.

The message specification in Figure 6 has its own Mission-Critical Integrity Level (MCIL_n) and Real-Time (RTL) levels.

The developer will have the freedom to choose between the aforementioned levels when applying the paper's approaches. For instance, if the vehicle uses passive armour the specific threat is not lethal, therefore, the Mission-Critical Integrity level of the message can be classified as Class 4. If is a light armoured vehicle then the Mission Criticality Integrity level can be classified as Class 2. The same can be applied for the Real-Time responsiveness.

10.2. DAS PSM

Once the PIM model is designed and specified, the model can be translated into PSM for integration into a specific architecture. The middleware technologies that can be used are the Data Distribution Service (DDS) or Message Queue Telemetry Transport (MQTT). DDS is an OMG machine-to-machine middleware standard that offers scalability, real-time, Quality of Service (QoS), high performance and interoperability data exchange between data publishers and subscribers. Publish/Subscribe message patterns are used for sharing the DAS system data in order to minimise the impact of adding new sub-systems [16]. MQTT is a lightweight messaging protocol that offers bi-directional communication to nodes. Its design has been created to minimise network bandwidth that uses the messages in a reliable degree of delivery.

When each attribute in the PSM model has its own data type, it must be specified and validated from the supported primitive data types of each target environment of DDS. The message specification file is used to describe the software component's application in order to enable communication between software components from different programming languages. However, the PSM transformation can be translated into a PSI model for the simulations and apply case studies of any mission-critical environments.

11. Qualitative and Quantitative Results

Using the example in Figure 6 (Section 10.1) and applying the threat estimation procedure discussed in Section 7, the threat can be calculated and then estimate the effect level of the mission. The example stated the following, "if the vehicle uses passive armour the specific threat (see Figure 6) is not lethal, therefore, the Mission-Critical Integrity level of the message can be classified as Class 4. For a light armoured vehicle then the Mission Criticality Integrity level can be classified as Class 2". In this section an estimation of the effect level of the stated threat is presented using anticipated values for each element.

11.1. Case 1: Passive Armour Vehicle

Case 1: If the vehicle uses passive armour then threat is as depicted in Figure 7.

The occurrence O[1] of the potential cause can be occasional. Therefore, the occurrence selected from Tables 1 and 6 is,

Occurrence: **O[1]** – Occasional (15%)

Using Table 6, the O[1] is 15%.

Considering that the threat does not significantly impact the mission or the vehicle, the severity of the specified threat can be negligible. Using Tables 2 and 6,

Severity: **SE[1]** – Negligible (0%)

Assuming that the vehicle uses passive armour instead of active, the threat T[1] will be detected from the crew. Therefore, the detection might be considered to as hard to detect, and using Tables 4 and 6,

Detection: **D[1]** – Hard (20%)

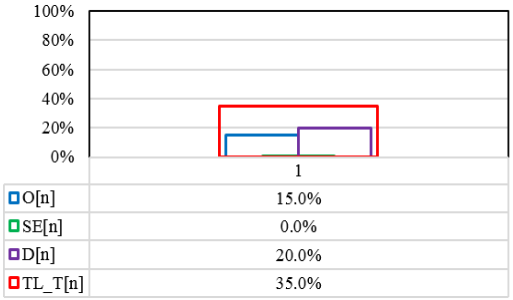


Figure 7 Case 1: T[n] Level

TL_T[1]: [(O[1]:15%)+(SE[1]:0%)+(D[1]:20%)= 35%

After the threat analysis of the specific threat for this case study, it has been identified to be 35% hazardous against the mission.

11.2. Case 2: Light Armoured Vehicle

Case 2: If the vehicle is a light armoured vehicle then the threat will be as in Figure 8. The occurrence O[1] of the potential cause can be occasional similar to the Case 1. Therefore, the occurrence selected from Tables 1 and 6 is,

Occurrence: **O[1]** – Occasional (15%)

Using Table 6, the O[1] is 15%.

Considering that the threat significantly impacts the mission or the vehicle then the severity of the specified threat can be catastrophic. Using the table 2 and 6,

Severity: **SE[1]** – Negligible (50%)

As aforementioned, the vehicle is a light armoured vehicle and assume that it has an integrated DAS system. Therefore, the threat T[1] to be detected can be categorised as easy. Using the Tables 4 and 6, detection is,

Detection: **D[1]** – Easy (10%)

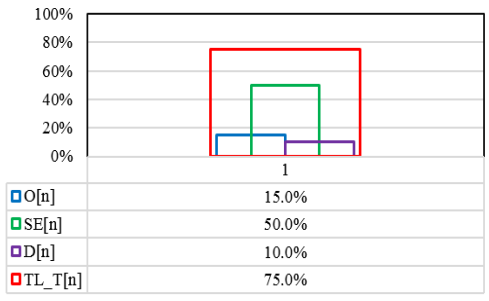


Figure 8 Case 2: T[n] Level

TL_T[1]: [(O[1]:15%)+(SE[1]:50%)+(D[1]:10%)= 75%.

The same threat as identified in Case 1 has being calculated to be 75% within this case study (2); whilst the same threat for a different mission scenario has increased in terms of threatening level. This in turn, elements that can be identified for a mission shall be re-used and analysed accordingly using this preliminary mission analysis at this early stage of a mission-critical system.

12. Conclusion and Future Work

At present there are no current development activities employed for mission-critical systems. In this paper, the importance of those activities is discussed. The objective of the proposed framework, is to enable system engineers achieve the mission development of any critical system's life-cycle, efficiently and effectively. This is achieved through a single package, using the three following main capabilities.

Firstly, a basic functional "mission standard" that enables the mission life-cycle development in a critical system; Secondly, the ability of mission interoperability for services and functionalities between systems and sub-systems built and procured in different times; and finally, the process that can support to define and analyse the mission data requirements of various critical systems.

The paper covered only the primary mission-critical attributes of a critical system, extracting the basic functional requirements. The scope of those requirements is paying particular attention on the mission, safety and security-critical attributes.

Furthermore, the extended part of this work, proposed an approach to estimate threats and their impact on a mission so as to early de-risk missions and systems (mission-critical) whilst in their design stages.

The next step is to further develop and refine the proposed using detail low level design case studies along with measures of performance.

References

[1] C. Ponsard, P. Massonet, A. Rifaut and J. F. Molderez, "Early Verification and Validation of Mission Critical Systems," Electronic Notes in Theoretical Computer Science 133, pp. 237–254, 2005

[2] K. Houliotis, P. Oikonomidis, P.Charchalakis, E. Stipidis, "An Efficient Approach to Designing Mission-Critical Systems, Case Study: Defensive Aid Suite (DAS) Systems", 2017 International Conference on Military Technologies (ICMT), Brno, Czech Republic May 31 – June 2, pp 402-409, 2017

[3] C. Ponsard, P. Massonet, A. Rifaut and J. F. Molderez, "Early Verification and Validation of Mission Critical Systems," Electronic Notes in Theoretical Computer Science 133, pp. 237–254, 2005

[4] F. Ciccozzi, I. Crnkovic, D. Di Ruscio, I. Malavolta, P. Pelliccione, R. Spalazzese, "Model-Driven Engineering for Mission-Critical IoT Systems." in IEEE Software, vol. 34, no. 1, pp, 46-53, Jan.-Feb.2017

[5] UK Ministry of Defence (MOD), "Generic Vehicle Architecture (GVA)," 2010

[6] NATO, STANAG 4754, "NATO Generic Systems Architecture (NGVA) for Land Systems," Edition 1, Ratification draft, August 2015

[7] M. Pradhan and D. Ota, "An adaptable multimodal crew assistance system for NATO generic vehicle architecture," 2016 International Conference on Military Communications and Information Systems (ICMCIS), Brussels, pp. 1-8, 2016

[8] M. Williamson, "Future Airborne Capability Environment (FACE),"2010, from"http://www.defensedaily.com/Assets/Williamson%20Panel%204%20 .pptx

[9] S. A. Brandt, S. Banachowski, C. Lin, T. Bisson, "Dynamic Integrated Scheduling of Hard Real-Time, Soft Real-Time and Non-Real-Time Processes", Proceedings of the 24th IEEE International Real-Time Systems Symposium (RTSS'03), 2003

[10] R. M. Connor, "VSI Vetronics Standards and Guidelines", QINETIQ/EMEA/TS/CR0702540 Issue 3, June 2009

[11] J. P. Lobo, P. Charchalakis, and E. Stipidis, "Safety and security aware framework for the development of feedback control systems," 10th IET System Safety and Cyber-Security Conference 2015, 2015

[12] A. Deshpande, O. Obi, E. Stipidis, and P. Charchalakis, "Integrated vetronics survivability : Requirements for vetronics survivability strategies," 6th IET International Conference on System Safety, Birmingham, 2011

[13] O. Obi, a Deshpande, E. Stipidis, and P. Charchalakis, "Intrusion Tolerant System for Integrated Vetronics Survivability Strategy," 8th IET International Safety Conference incorporating the Cyber Security Conference, Cardiff, 2013

[14] A. Larrucea, J. Perez, and R. Obermaisser, "A Modular Safety Case for an IEC-61508 Compliant Generic COTS Processor," IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, pp. 1788–1795, 2015

[15] J. Kong and H. Yan, "Comparisons and analyses between RTCA DO-178B and GJB5000A, and integration of software process control," 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), Chengdu, 2010

[16] [15] J. L. Rapanotti, "Developing soft-kill capability for light armoured vehicles through battlefield simulations," Defence R&D Canada – Valcartier Technical Memorandum DRDC Valcartier TM 2003-276, 2007

[17] USA Department of Defence (DOD), "The Data Distribution Service Reducing Cost through Agile Integration", 2011, from "http://www.twinoakscomputing.com/wp/DDS_Exec_Brief_v20l-public.pdf

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 130-135 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Leaf-shaped solar cell antenna for Energy Harvesting and RF Transmission in ku-band

Chokri Baccouch*, Hedi Sakli, Dhaou Bouchouicha, Taoufik Aguili

National Engineering School of Tunis, Tunis El Manar University, B.P. 37, Le Belvédère, 1002 SYS'COM Laboratory, Tunis, Tunisia.

ARTICLE INFO

Article history:
Received: 31 August, 2017
Accepted: 13 November, 2017
Online: 10 December, 2017

Keywords:
Antenna
Solar cell
Communication systems

ABSTRACT

In this paper, a hybrid system solar cell antenna was studied. Our contribution lies in the combination of both optical and radio-frequency signals in a solar cell antenna. Hence, we propose a new method to merge antenna and solar cell. The leaf- shaped patch solar cell antenna was dedicated at a time to the energy harvesting and the RF transmission. A performance analysis of the solar cell antenna was conducted using Advanced Design System (ADS) software. Simulation results showed a resonance at a frequency of 15.77 GHz with an effective return loss of -27.62dB and a gain of 5.77dBi.

1. Introduction

Solar energy was considered as one of the most renewable energy resources that could fill the shortage of energy needs. In fact, human beings used the solar energy to produce electricity. Although the photovoltaic (PV) cell was first employed to produce it, its efficiency remained limited [1]. Solar energy has become an attractive alternative for powering autonomous communication systems. Many communication systems, such as standalone systems like environmental monitoring system, vehicular communication and satellite systems [2] require electric energy for their operation and use. Solar powered communication systems have received considerable attention due to their ability to operate without the necessity of being connected to an electricity grid. This has become a significant challenge when it comes to powering communication systems in remote places where the electricity grid is not available. In order to address this challenge, the use of photovoltaic in communication systems has recently been the subject of many researches. The photovoltaic systems of power generation when combined with communication systems can provide compact and reliable autonomous communication systems, which can be used for many applications.

Traditionally, telecommunication antennas and solar cells do not mix. They must operate independently so as not to interfere. They compete for the available space on mobile and standalone systems which are generally limited in size. Furthermore they may be bulky and expensive and they limit the capabilities of product designs. In space applications for instance, both solar panels and communication system are major contributors to the overall size and weight of the satellites and combining these two systems could save real estate and cost. In urban areas, wireless communication infrastructures could be integrated into solar panels. Finally the integration of antennas and solar cell panels can be essential for the transportability of emergency autonomous communication stations [3-4] and [5-6].

In this paper, we describe the solar cell antenna in the second section; we detail the proposed approach as well as the effects of optical transparency on the antenna parameters. In the third section we explain the proposed solar cell antenna structure by giving a model of the mesh patch with optimized geometry. The simulations results of the solar cell antenna are presented. A solar cell antenna array has also been studied in this section by giving their parameters. The solar cell antenna equivalent circuit as well as the RF / DC decoupling circuit have been studied in order to eliminate the influence of the RF signal on the DC signal and to show that this hybrid system will be both for the recovery of a solar cell, energy and RF transmission. Finally, we conclude the paper.

2. Solar cell antenna

2.1. Approach

Several attempts to integrate solar cells with patch antennas have been studied, but they did not pay attention to the antenna patch when it is placed above a solar cell. To remedy the drawbacks, especially of limited surface area, encountered during the integration of patch antenna with a photovoltaic cell, we

* Chokri Baccouch, National Engineering School of Tunis, Tunis El Manar University, B.P. 37, Le Belvédère, 1002 SYS'COM Laboratory, Tunis, Tunisia,

propose here new types of antennas based on optically transparent photovoltaic cells with mesh patches.

Optically transparent mesh patch antennas are antennas that have a certain level of optical transparency. These transparent antennas, for example, are potentially suitable for integration with solar cells for small satellites. Traditional patch antennas used on small satellites compete with solar cells for the surface. However, a mesh patch antenna can be placed directly on the solar cells and solve the surface limitation problem. For such integration, a high optical transparency of the patch antenna is required from the solar cells point of view, since solar cells require sufficient sunlight to generate adequate electrical power. On the other hand, the antenna must have at least acceptable electrical properties at the same time so that it can radiate correctly and efficiently [7-8] and [9-11].

The hybrid system studied in this paper was devoted at a time to the energy harvesting and the RF transmission for satellite communication. This structure was made of a photovoltaic cell in which the front grid was designed to have a miniature antenna suited to the band transmission and to minimize the power loss of the cell devoted to energy conversion. The received electrical energy was used to operate the complete system.

2.2. Optical transparency effects on the antenna parameters

The openings in the conductor make it possible to pass the light while the mesh can still be conceived as an efficient radiator. Mesh patch antennas, compared to their solid counterparts, have a higher surface resistance due to much lower conductor coverage. Consequently, it is normal, when the optical transparency increases, to have a low performance of the antenna.

The antenna optical transparency is given most often by the ratio of the mesh patch surface to the patch total area (eq.1).

$$T = \left[1 - \frac{A_{conductor}}{A_{patch}}\right].100\% \qquad (1)$$

It has been demonstrated that, when the line width is fixed, we notice that during the patch transparency increase the frequency of resonance decreases. Since t the patch physical length is the main determinant of the resonance frequency as noted by Clasen [7], the increase in transparency introduces the possibility of minimizing the antenna design. Using the mesh patch increases the fringing fields; therefore, increase the antenna electrical length [12]. Also from Figure 1 one is able to see that increasing the transparency decreases the bandwidth. Bandwidth is defined by the following equation:

$$BW = \frac{f_h - f_l}{f_r} \qquad (2)$$

With f_h the highest frequency at which the voltage standing wave ratio (VSWR) is below 2, f_l the lowest frequency at which the VSWR is below 2, and f_r is the frequency of the minimum VSWR.

Also as seen in Figures.2, 3 and 4, the increase in transparency decreases the antenna efficiency, gain and directivity.

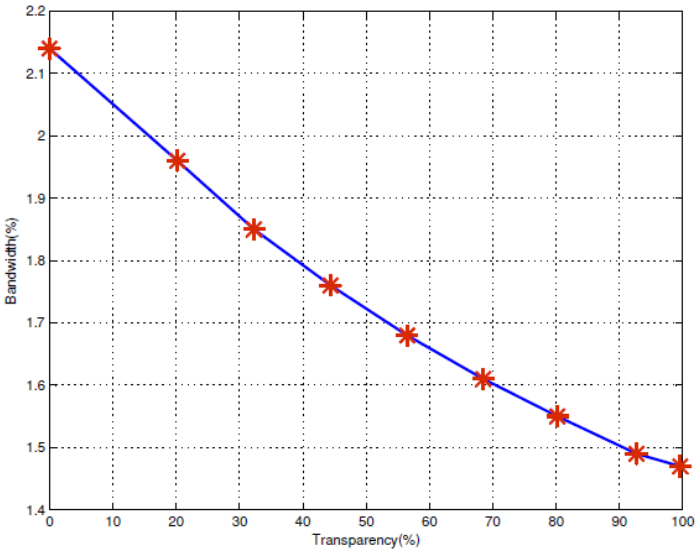


Figure 1: Bandwidth in terms of transparency

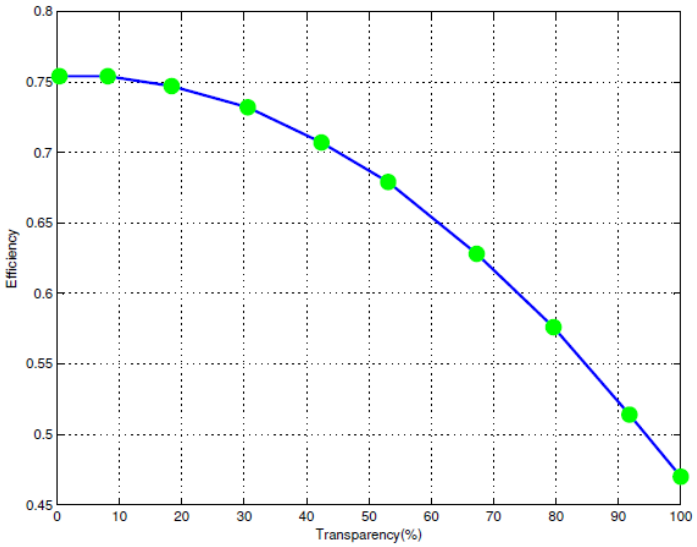


Figure 2: Efficiency in terms of transparency

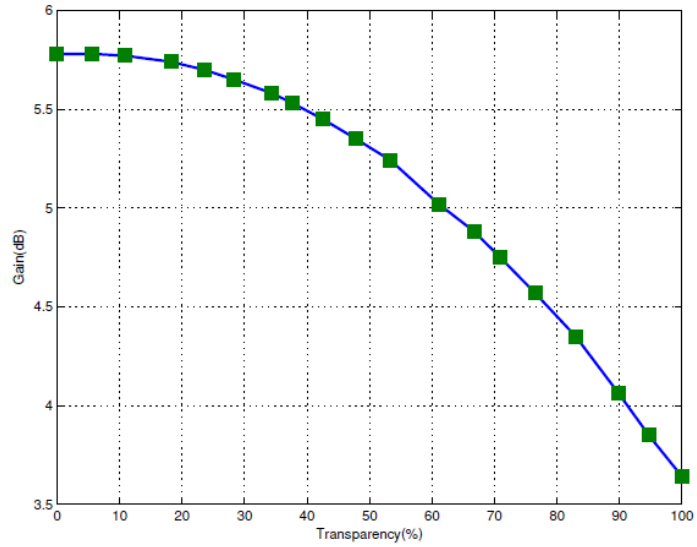


Figure 3: Gain in terms of transparency

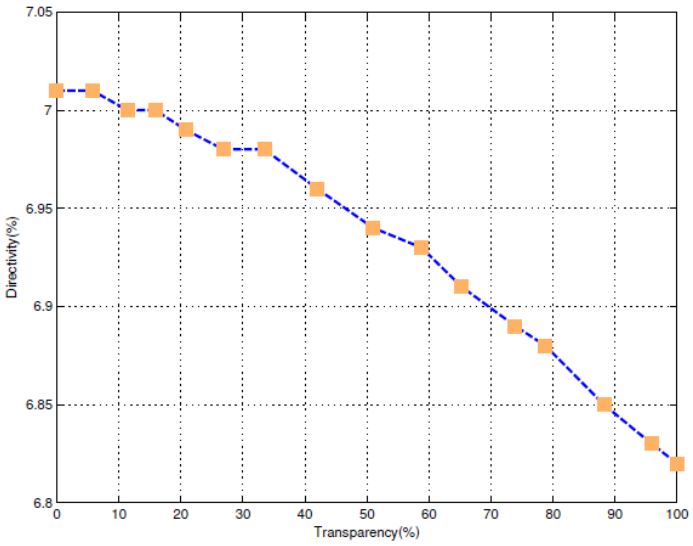


Figure 4: Directivity in terms of transparency

2.3. *Effect of the line width*

The study of optically transparent antennas is based essentially on the optimization of the mesh patch dimensions. In this paper, antennas with an optical transparency of 70%, 80% and 90% are generated. Each transparency depended of the line width and the number of mesh lines. Simulation results present the efficiency of the mesh antenna as a function of the line width for each transparency is plotted in figure 5. We interpret that for a given transparency, efficiency of a mesh antenna can be improved by refining the line width. From fabrication standpoint of view, the line width realization of 0.1 mm or more is simple. The efficiency and transparency are adequate for this hybrid system of solar cell antenna for satellite and terrestrial communications. From the simulation results (see Figure 5), we conclude that with a line width of 0.1 mm, it is possible to achieve more than 60% efficiency with an antenna of 90% transparency, 63% efficiency with an antenna of 80% transparency and 67% efficiency with an antenna of 70% transparency.

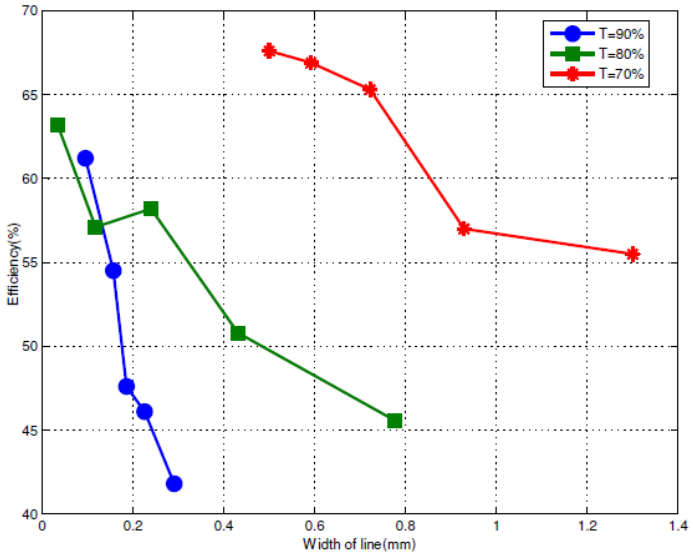


Figure 5: Efficiency in terms of width of line

Efficiency is inversely proportional with optical transparency, this is true because traditional patch antennas are more efficient but when they are integrated with solar cells they screen these and can reduce the collected electrical power. Our method allows us to propose a hybrid solar cell antenna system for energy harvesting and RF transmission for satellite and terrestrial communications.

3. Simulations results and discussions

3.1. *Structure of solar cell antenna*

We propose a solar cell antenna with mesh patch. A mathematical model that we have already studied to minimize the power losses of the solar cell antenna and improve the conversion efficiency as a solar cell [13-14]. Optimization of the maximum electrical power collected as a function of finger width was determined. The optimal width W_f has been used for the mesh patch design or front face collection grid.

Improving the performance of a solar cell depends not only on the materials and structure but also on the design of the metal grid front face.

With this mesh patch structure two types of waves can exist. The optical waves that will be absorbed by the silicon (semiconductor) and the RF waves that will be collected by the mesh patch metal whose width of fingers W_f is optimized. The structure designed was a solar cell antenna printed on a multi-layered substrate Figure 6.

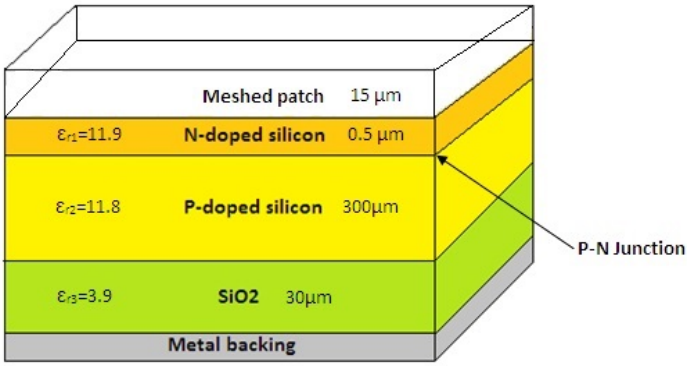


Figure 6: Multi-layered substrate of solar cell antenna

The silicon on an insulating layer SiO_2 confers to the components that are realized, a higher operating frequency, an ability to operate at low voltage and low power consumption and an insensitivity to the effects of ionizing radiation. When using a thin-film substrate we observe that the simulations are complex and time consuming. The meshed patch of solar cell antenna proposed is given in Figure 7. The dimensions of the proposed patch of solar cell antenna are given in the following table (Table 1).

Table 1: Dimensions of solar cell antenna

Parameters	Values
Width of busbars	W_b=0.5mm
Width of finger	W_f=60 µm
Length of feedline	L= 9.1 mm
Width of feedline	W=0.235 mm

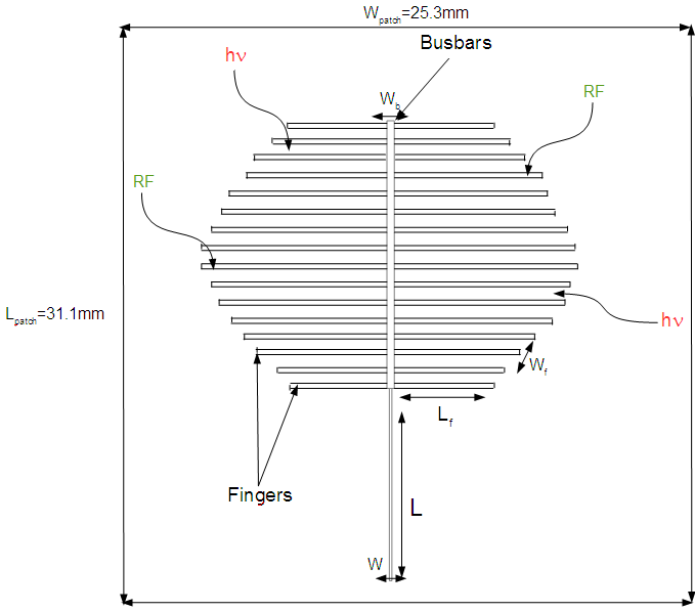


Figure 7: Meshed patch of solar cell antenna proposed

This antenna was excited by a microstrip line of characteristic impedance equal to 50 Ω. In this paper, the transparency of a patch is defined by the percentage of see-though area on the patch as shown below (see equation 3):

$$T = \left[1 - \frac{A_{conductor}}{A_{patch}}\right].100\% = \left[1 - \frac{W_b L_b + WL + \sum_{1}^{32} W_f L_f}{W_{patch} L_{patch}}\right].100\% \quad (3)$$

where W_f is the mesh line width, L_f length of each mesh line, W_b is the busbars line width, L_b the busbars length, W is the InsetFeed width, L is the InsetFeed length, W_{patch} is the patch width and L_{patch} is the patch length (see Figure7).

3.2. Antenna Parameters

As an antenna, the electromagnetic performances of the solar cell antenna were simulated by using the ADS software. The simulated S11 parameter was presented in Figure 8. An antenna should be perfect radiator, rather than perfect absorber. The amount of radiated power returned back through the port can be calculated for finding return loss at that resonating frequency. For the resonant frequencies the return loss should be less than -10dB i.e. S11<-10 dB. Simulation results show that the designed antenna can be used as a frequency antenna with an effective return loss of -27.62dB at 15.77GHz. To the normal scale $S11=10-2.762, ie only 0.17% of the emitted power will be reflected as well as the impedance Z_a=49.82 Ω which confirms the adaptation of the antenna. The adaptation of our antenna is substantiated by the return loss low value S11. Similarly, its operation can be assessed by the VSWR curve which is another performance measurement parameter at the input. It indicates the signal reflection due to a mismatch between the impedance of the antenna and that of the source. The reflected signal leads to the appearance of the standing wave voltage in the feed line which, in turn, destroys the reliability of the antenna. The ideal value of VSWR is 1 which means that 100% of the delivered power is accepted (no reflection). But, practically speaking, it is tolerable up to 2 [15]. The VSWR obtained for this antenna equal to 1.0036.

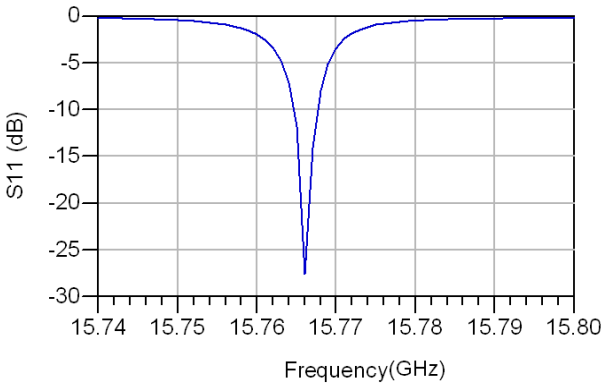


Figure 8: Reflection coefficient S11

The radiation of our antenna is gathered in the main lobes which have a wide beam width at -3 dB and whose amplitude is much higher than that of the side lobes. This reflects a very directive radiation by offering a total directivity equal to 7.26 dB at 15.77 GHz. The radiation pattern of this antenna at a frequency of 15.77GHz is shown in Figure 9. The polarization of the radiated field was linear.

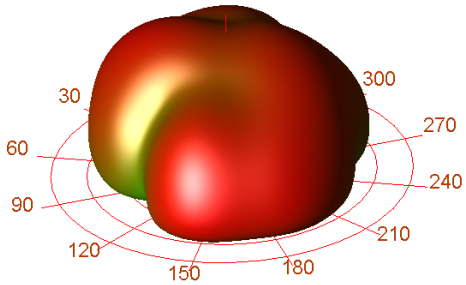


Figure.9: Radiation pattern

When designing an antenna, the gain must be taken into consideration as it is an important metric. The good values of S11 and VSWR are not enough to confirm a good radiation. Antenna gain describes how much power is transmitted in the direction of peak radiation to that of an isotropic source. The gain of this antenna was 5.77dBi.

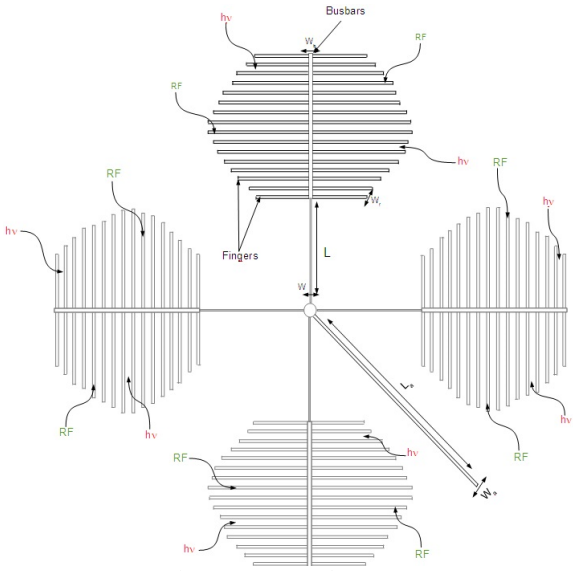


Figure 10: Array of antennas

3.3. Array of antennas

In order to increase the gain, we proposed an array of antennas shown in Figure 10. Simulation results of S11 parameter and radiation pattern are given, respectively, in Figures 11 and 12.

This array of antennas was excited by a microstrip line of La=16.82 mm in length and W_a=0.228 mm in width and of characteristic impedance equal to 50 Ω. S11 parameter shown in Figure 11 of a dual resonant frequency at 15.72GHz and 15.77 GHz with an effective return loss respectively of -20.18 dB and -22.81dB.

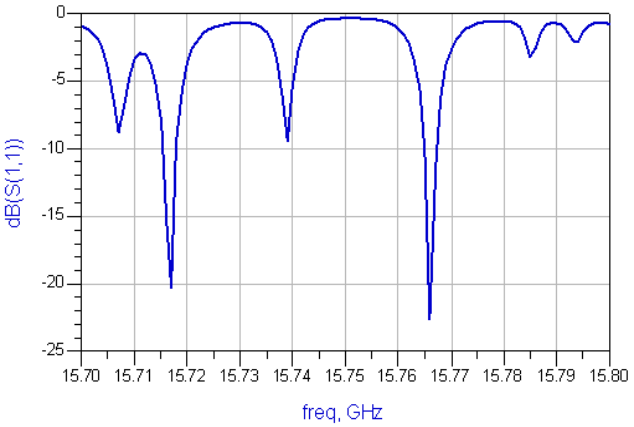


Figure 11: Reflection parameter S11 of array antennas

The radiation pattern of this array of antennas is shown in figure 12. It is gathered in the main four lobes. The gain at the resonance frequency of 15.77GHz is 9.76 dB and the directivity is 11.24 dB.

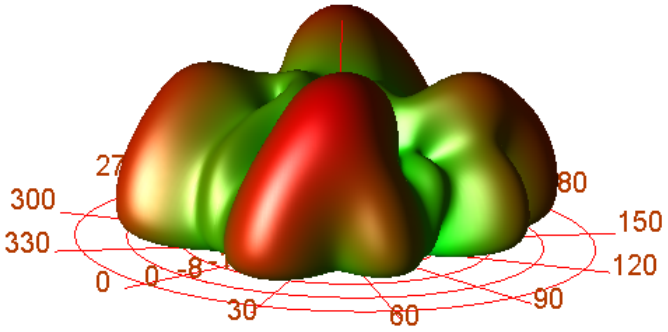


Figure 12: Radiation pattern of array antennas

The designed array antennas are showing excellent gain and directivity at the resonating frequencies with high radiation efficiency.

3.4. Equivalent circuit of solar cell antenna

Solar cell equivalent circuit is specially simulated in Advanced Design System for measuring power (P), current (I) and voltage (V). Microstrip antenna equivalent circuit is separately simulated and the same output is measured. Finally both microstrip and solar cell circuits are combined and output is measured. While comparing both outputs, combination of microstrip and solar cell equivalent circuit produces high voltage (V), current (I) and power (P). The simulation of SOLAN in ADS provides the following values of current (I) =872.8mA Voltage (V) =8.728mV and Power (P) =7.6177984mW were measured as shown in Figure 13.

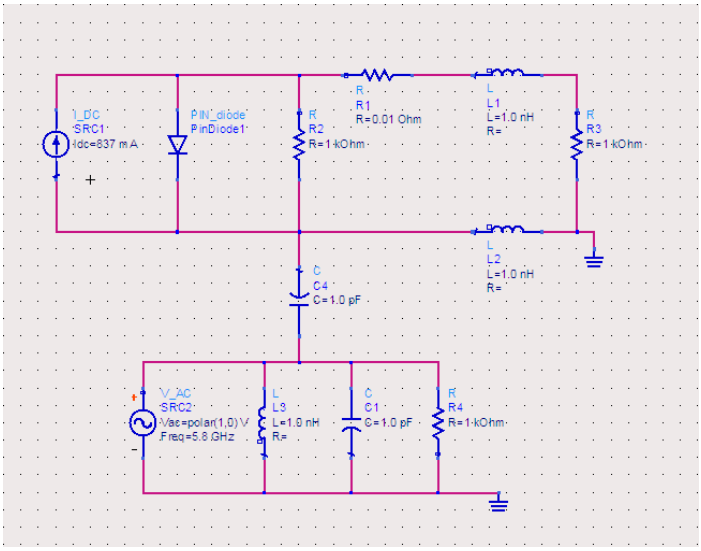


Figure 13: Equivalent circuit of solar cell antenna

3.5. Decoupling RF/DC in solar cell antenna proposed

Since the solar cell has a DC circuit, the direct current path must be decoupled from the RF signal path in such a way that the DC load has no influence on the RF properties of the antenna. The decoupling can be realized by means of concentrated reactive elements and distributed elements, respectively. In this case, we propose an inductive and capacitive RF/DC decoupling circuit shown in Figure 14.

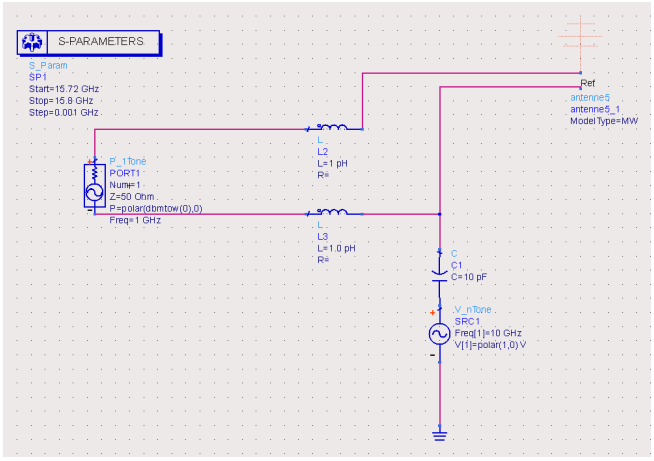


Figure 14: RF/DC decoupling circuit

The decoupling circuit was simulated (see Figure 15) and values are compared to the S11 parameter shown in Figure 8. The simulation of the reflection parameter S11 of decoupling circuit provides the same result of resonant frequency of the proposed solar cell antenna in Layout with a slight increase of return loss.

After RF/DC decoupling, this solar cell antenna can meet the purpose for which it is being studied energy harvesting for the relevant telecommunication system power supply and RF

transmission. Based on the simulations results obtained, the solar cell antenna proposed here can be used for ku band satellite communication. Figure 16 explains well the operation of this solar cell antenna.

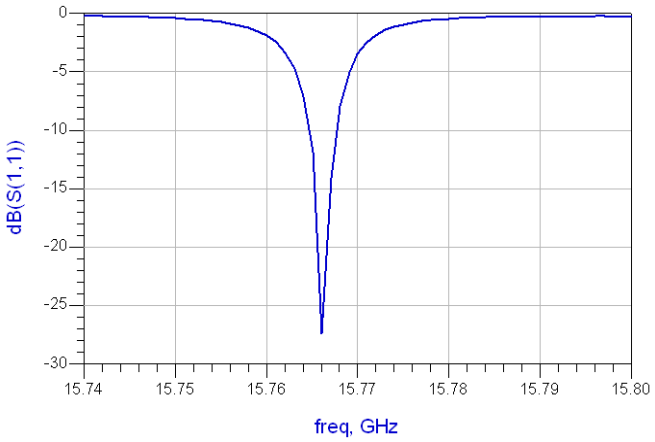


Figure 15: Simulation of RF/DC decoupling circuit

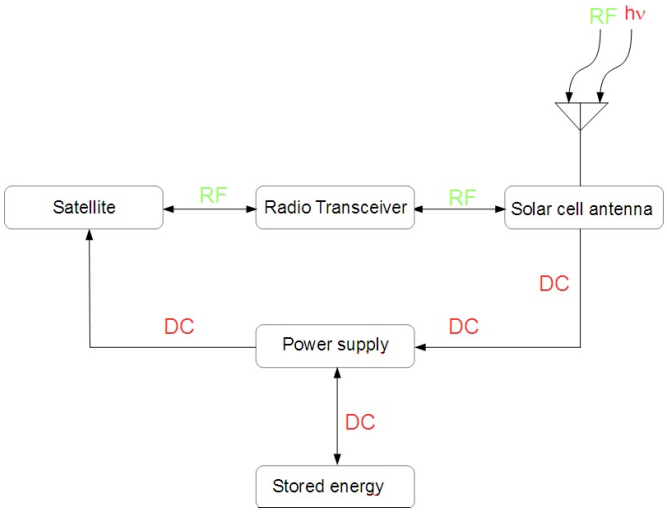


Figure 16: Self-contained wireless system with solar cell antenna

4. Conclusion

In this paper, we presented a solar cell antenna with leaf-shaped patch. This optically transparent antenna is particularly well adapted for energy harvesting and for the RF transmission in Ku band with a large gain. Good performances are obtained for this solar cell antenna, as a PV cell with its collected electrical power and as an antenna with its gain 5.77dBi and 9.76 dBi of array of antennas, directivity and efficiency. We conclude that the hybridized method where the patch element of a patch antenna has been replaced by a solar cell is more efficient than a simple integration where the patch antenna was placed directly on the solar cell is what which limited the available area of the cell and consequently a loss of electrical energy collected.

References

[1] Maharaja M., Kalaiselvan C, "Integration of Antenna and Solar Cell for Satellite and Terrestrial Communication". International Journal of Scientific and Research Publications, **3**(5), 2013.

[2] Danesh, M. and J. R. Long, "An autonomous wireless sensor node incorporating a solar cell antenna for energy harvesting", IEEE Trans. Microw. Theory Tech, **59** (12), 3546--3555, 2011. DOI: https://10.1109/TMTT.2011.2171043.

[3] Shynu, S. V., M. J. R. Ons,P. McEvoy,M. J. Ammann,S. J. McCormack, B. Norton, "Integration of microstrip patch antenna with polycrystalline silicon solar cell". IEEE Trans. Antennas Propag, **57** (12), 3969--3972, 2009. DOI: https://10.1109/TAP.2009.2026438.

[4] A. Suresh Kumar, S. Sundaravadivelu, "Performance analysis of solar cell antenna with hybrid mesh and aght-8 material". Scholarly Journal of Scientific Research and Essay (SJSRE), **3**(4), 51-55, 2014. DOI: http://dx.doi.org/10.1109/ghtc.2011.14.

[5] Yurduseven, O., Smith, D., Pearsall, N., Forbes, I, "A solar cell stacked slot-loases suspended microstrip patch antenna with multiband resonance characteristics for WLAN and WIMAX SYSTEMS", Progress In Electromagnetics Research, 2013, 142, 321--332. DOI: http://dx.doi.org/10.2528/pier13081502.

[6] Ons, M.J.R., Shynu, S.V., Ammann, M.J., McCormack, S., Norton, B, "Investigation on Proximity-Coupled Microstrip Integrated PV Antenna", IEEE Antennas and Propagation,EuCAP 2007, 3, 1--3. DOI: http://dx.doi.org/10.1049/ic.2007.1374.

[7] G. Clasen and R. Langley, "Meshed patch antennas," IEEE Transactions on Antennas and Propagation, vol. 52, pp. 1412-1416, 2004.

[8] C. Baccouch, D. Bouchouicha, H. Sakli, T. Aguili, "Patch Antenna on a Solar Cell for Satellite Communications". International Journal on Communications Antenna and Propagation, **6**(6), December 2016. DOI: https://doi.org/10.15866/irecap.v6i6.9807.

[9] C. Baccouch, H. Sakli, D. Bouchouicha, T. Aguili, "Patch antenna based on a photovoltaic solar cell grid collection". 2016 Progress in Electromagnetic Research Symposium (PIERS). DOI: https:// 10.1109/PIERS.2016.7734278.

[10] C. Baccouch, H. Sakli, D. Bouchouicha, T. Aguili, "Patch Antenna based on a Photovoltaic Cell with a Dual resonance Frequency". ADVANCED ELECTROMAGNETICS, VOL. 5, NO. 3, November 2016. DOI: https://10.7716/aem.v5i3.425.

[11] Turpin, T. W. and R. Baktur, "Meshed patch antennas integrated on solar cells. IEEE Antennas Wireless Propag. Lett, 2009, Vol. 8, 693– 696. DOI: https://10.1109/LAWP.2009.2025522.

[12] C. A. Balanis, Antenna Theory Analysis and Design. Hoboken, NJ: John Wiley and Sons, 2005.

[13] C. Baccouch, D. Bouchouicha, H. Sakli and T. Aguili, "Optimization of the Collecting Grid Front Side of a Photovoltaic Cell Dedicated to the RF Transmission", 2nd International Conference on Automation, Control, Engineering and Computer Science ACECS, 22- 24 March 2015 – Sousse, Tunisia.

[14] C. Baccouch, D. Bouchouicha, H. Sakli, T. Aguili, "Patch Antenna on a Solar Cell for Satellite Communications". International Journal on Communications Antenna and Propagation, 6(6), December 2016. DOI: https://10.15866/irecap.v6i6.9807.

[15] M. Titaouine, "Analysis of Antennas Micro ruban by the model of the cavity, the model of the transmission line and the method of moments", Magister thesis University Ferhat Abbas, Sétif, 1998.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 130-134 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Modelling of the resistance heating of the moving molybdenum sheet

Miroslav Pavelek[*,1], Michal Frivaldsky[1], Pavol Spanik[1], Tibor donič[2]

[1]*Department of mechatronics and electronics, University of Zilina, 01026, Slovakia*

[2]*Research Centre of University of Zilina, University of Zilina,01026, Slovakia*

ARTICLE INFO

Article history:
Received: 02 January, 2018
Accepted: 17 January, 2018
Online: 30 January, 2018

Keywords :
Resistance heating
Molybdenum sheet
Finite element method

ABSTRACT

This article describes the modeling procedure of a direct resistive heat model of the molybdenum sheet and its simulation with electrodes made from various materials. The main characteristic of the proposed model is its dynamic behavior, because model considers the movement of the molybdenum sheet. The simulation results are mutually compared and the optimal material parameters are selected. The development of the model, in which the movement of molybdenum sheet is considered is used for determination of optimal movement speed of the molybdenum sheet in order to achieve requested temperature. The presented model can be also used for determination of optimal electrode materials and its geometrical properties/shape.

1. Introduction

This paper is an extension of work originally presented in Electrical Power engineering 2017 [1]. This article discusses the usage of finite element method for the development of a resistive heating model of the molybdenum sheet. The model is designed for the needs of the development of the equipment for the heating and shaping the molybdenum sheets into molds. The mentioned molds are used as containers for horizontal crystallization of the sapphire single crystal. Molybdenum is one of the few materials suitable for this purpose in a view of the long-term high-temperature load of the mold. The mold is exposed to the temperature gradient of up to 100 °C./ mm at the maximum temperature of 2150 °C. in the longitudinal direction of the container, where in one part is the melt and in the adjacent part is the corundum as an already growth product, i.e. single sapphire crystal. Based on the extreme conditions during the process, it is crucial to not decrease the mechanical abilities during the shaping process of the mold. For that purpose, the structural changes and possibilities of avoiding them by properly set pre-heating respectively heating/cooling system of Mo sheets need to be predicted. Currently, the molds are made of molybdenum sheet having a thickness of 0,5 mm, made by powder metallurgy technology, i.e. by plastic deformation and sintering procedure. Specifically, the molybdenum sheet Grade M1 corresponding to the American standard ASTM B386 or GB 3877, where the chemical composition is approximately the same. [2-5]

[*]Corresponding Author: Miroslav Pavelek, University of Zilina
Email: miroslav.pavelek@fel.uniza.sk

2. Basic simulation model

The simulation model is developed in COMSOL environment as a 3D model with the options for reconfiguration the mutual position of electrodes, the electrode material and the pressure on the electrodes. The model is composed of the molybdenum sheet and two electrodes. The molybdenum sheet is modeled as a block (domain) with a wanted width(a), depth(b) and thickness(c) and electrodes are modeled as cylinders with a wanted diameter(d), height(h) and distance(x) between them. The geometry (Fig.1.) is then complemented by physic setup. The Multiphysics "Joule heating (JH)" is used for the model. This Multiphysics is formed by the connection of "Electrical circuit (EC)" physic and "Heat transfer in solid (HTS)" physic. The physic "Electrical circuit" is used to set up the value of input current trough top boundary of one electrode and grounding through the top boundary of the other electrode. Within the EC module, the required contact pressure applied to the electrodes and the roughness of the contact between the electrode and Mo sheet can be adjusted. The HTS is used to set up parameters as initial temperature of the system and the method in which the temperature is transferred to the surrounding area. The proper function of the model is achieved by setting the spherical domain with a diameter of one meter, which is filled with air. All simulations are the set as time domain simulation, so we can determine the influence of studied materials on the speed of the heating of the Mo sheet. [6-9]

As it has been already mentioned, the model is reconfigurable, but for means of this paper the following geometrical parameters are set: Parameters of molybdenum sheet (based on manufactured molybdenum sheets [10,11]): **1000/300/1 (a/b/c) mm,** parameters

of electrodes: **20/50 (d/h) mm**, the electrodes distance: **100 (x) mm**. The electrodes are pressed to the Mo sheet with a pressure of **0.3 MPa.** For the electrode's material study, three classes of materials based on the electrical conductivity of the electrodes are defined (lower than Mo, Mo and higher than Mo). The electro-thermal parameters of the basic model are defined in TABLE I. The model needs to be particularly accurate in the molybdenum sheet domain. For that purpose, the temperature dependencies of electro-thermal parameters of molybdenum are implemented, specified by the manufacturer "PLANSEE" (Fig.2-4.).

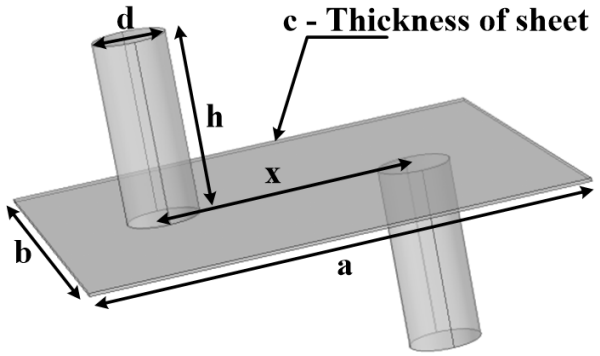


Fig. 1. Representation and definition of Mo sheet model

Table 1: Electro-thermal parameters used in simulations

Material	Parameter				
	Electrical cond. [S/m]	*Relative perm. [-]*	*Thermal cond. [W/(mK)]*	*Specific heat capacity [J/(kgK)]*	*Density [kg/m^3]*
Molybdenum (20 °C) (293.15 K)	17.9e6	1	142	254	10280
Class I. Lower El. Con.	$\sigma_{Mo}/10$	1	400	385	8700
Class II. Equivalent El. Con.	σ_{Mo}	1	400	385	8700
Class III. Higher El. Con.	$\sigma_{Mo}*100$	1	400	385	8700

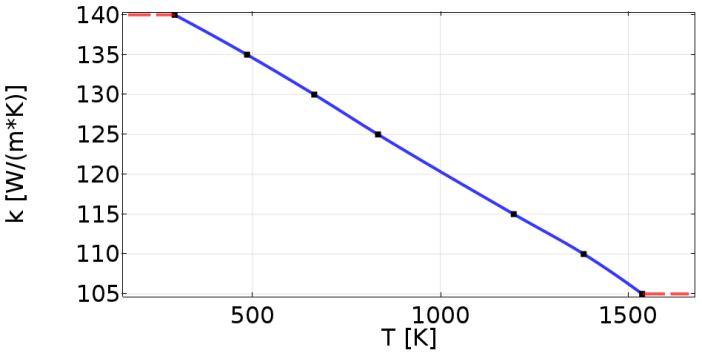


Fig. 2. Thermal dependence of thermal conductivity of molybdenum used in simulations

The mentioned JH module of the COMSOL environment uses well-known equations for modeling a heat transfer in solid materials (1).

$$\rho C_P \frac{\partial T}{\partial t} - \nabla \cdot (k \nabla T) = Q \qquad (1)$$

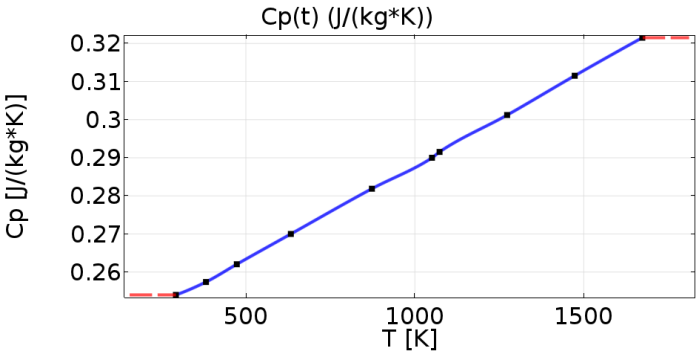


Fig. 3. Thermal dependence of specific heat capacity of molybdenum used in simulations

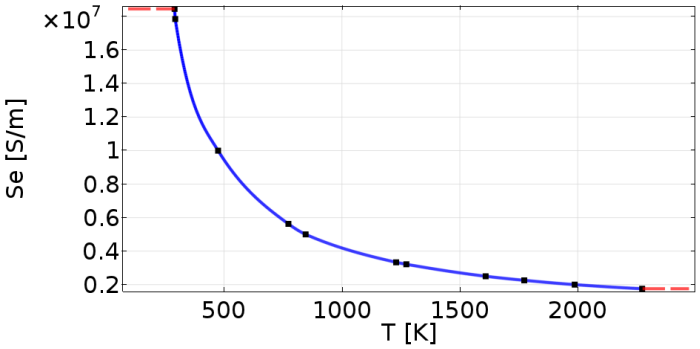


Fig. 4. Thermal dependence of electrical conductivity of molybdenum used in simulations

Where "ρ" is the density of the modeled material, "C_P" is specific heat capacity of the material, "k" is thermal conductivity of the material and "Q" is a heat source.

In this case, the heat source "Q" is determined by Joule-Lenz law of resistive loses in material structure due to the current flow through this structure.

3. Simulation results I.

Based on [6] the wanted temperature of Mo sheet is at a level between 1000 K to 1200 K. At those levels, the Mo sheet can be easily bendable. The simulation is done for before mentioned electrode's material classes (TABLE I.) not for specific material. The input current of the system is selected within the interval from 1000 A to 1400 A. The results are compared in a way of temperature distribution within Mo sheet domain and speed of the heating process. Fig.5. shows actual simulation model with geometrical parameters defined in the previous paragraph.

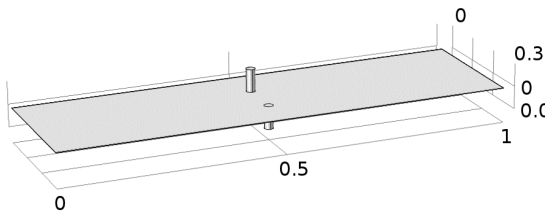


Fig. 5. Model of Mo heating with given geometrical parameters

The simulation is done with electrodes placed in the middle of the longer side of the sheet while the distance between electrodes is set to 100 mm. The following results are displayed as the temperature in the point in the middle of the sheet's volume and temperature distribution in the cutline between electrodes (middle of the Mo sheet) as can be seen at Fig.6.

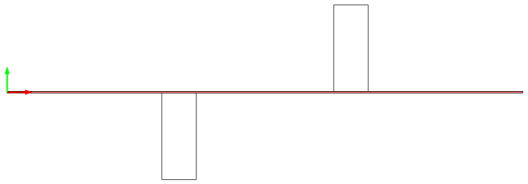


Fig. 6. Line for result extraction (Temperature distribution)

Fig.7. shows mentioned results for the first class of electrode's material, which has lower electrical conductivity than Molybdenum. The wanted temperature of 1000 K (Fig.7. up) was achieved in 75 s for lowest current (1000 A) and in 4 s for highest simulated current (1400 A). The wanted temperature interval is marked as black dashed lines in Fig.7,8,9 (up). The second part of the graph (Fig.7.middle) shows temperature distribution (based on Fig.6.) for different current values, while the time when the desired temperature value was achieved (1000 K) is considered. The last part of the graph shows temperature distribution within whole Mo sheet with the chosen current flow. Fig.8. and Fig.9. have the same purpose as Fig.7., but other material types are considered. For second electrode's material (Se=17.9e6 S/m) can be seen that wanted temperature was achieved in 95 s for current value of 1400 A.

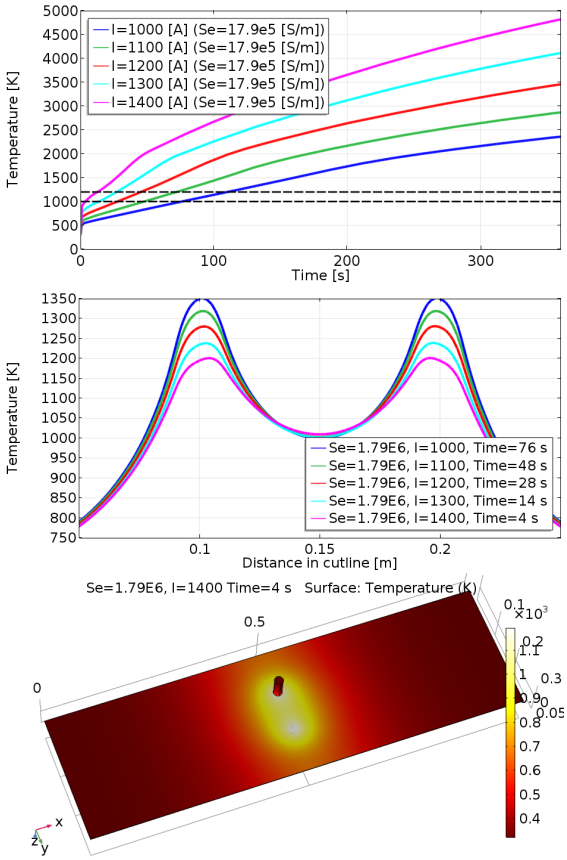


Fig. 7. Results for first electrodes material type (Se = 17.9e5 [S/m])

In case of third material the heating process was much slower (the major part of heat is produced through the direct resistive heating, while in first two cases there was also considerable influence of indirect resistive heating), wanted temperature was achieved in 164s for current 1400 A and it can be said that minimal current needed to achieve wanted temperature in given time is 1200 A.

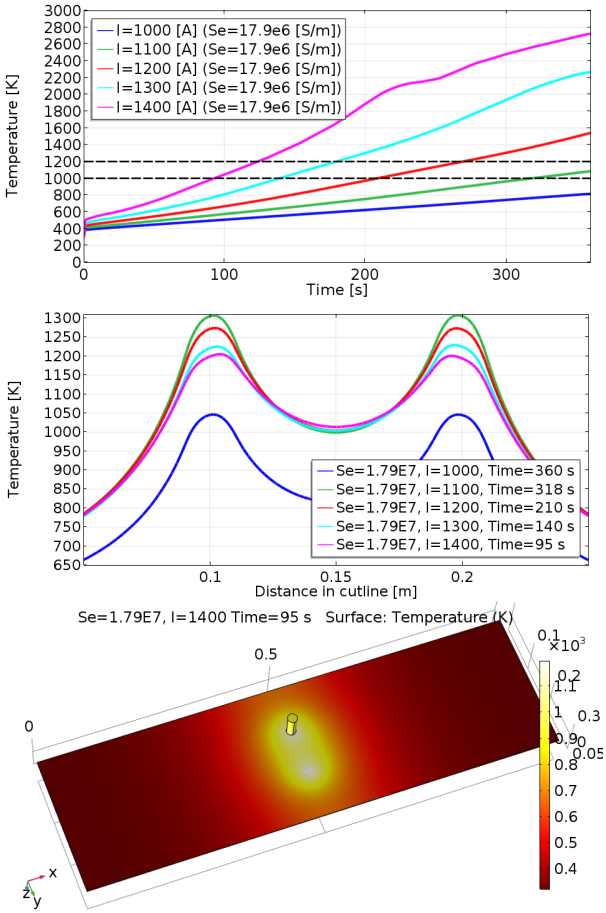


Fig. 8. Results for second electrede's material type (Se=17.9e6 [S/m])

4. Model of sliding molybdenum plate

For simulation of sliding molybdenum sheet the same settings of physics and geometry are used as within the basic model (previous part), but for the shift of the Mo sheet a special script in Matlab language is developed, which basically simulates the heating in a loop with the different initial position of electrodes. For the required temperature distribution, which is different for each step of the simulation, the molybdenum sheet domain is divided into subdomains (blocks) (Fig.10.). Within individual subdomains, the average temperature is defined (volumetrically) from previous simulation step and next simulation step is simulated with given electrodes shift. Each domain is within script defined as a 3D object with given length, width and height. These geometrical parameters are computed from geometrical parameters of whole Mo plate and from a wanted number of the subdomain in every direction of Mo plate. Mathematically, this model is described as the basic simulation model with equation 1 and by Joul-Lenz law for heat source of the modeled physic.

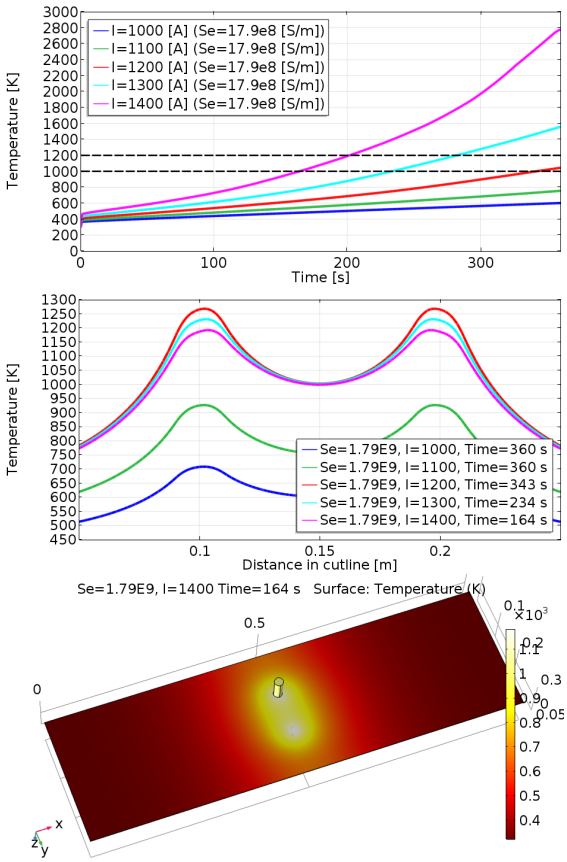


Fig. 9. Results for third electrode's material type (Se=17.9e8 [S/m])

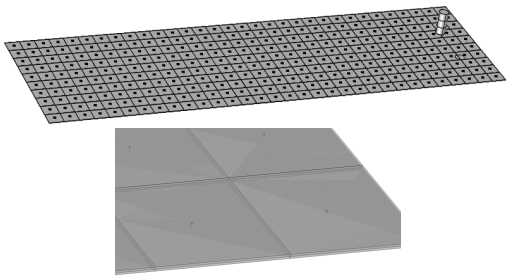


Fig. 10. Model of Mo heating with subdomains

5. Simulation results II.

The simulation is done for the second class of materials with electrical conductivity value close to Mo sheet, in which the temperature distribution was most uniform. Based on the optimal speed of the heating process and distribution of heat within Mo sheet, the input current of 1400 A is applied. It needs to be mentioned that this part of the model was created purely for verification of the proposed modeling process and thus, it does not consist any actual comparison with measurement. For that reason, a quite large (120 mm) electrodes shift is chosen for simulation. Model of the Mo sheet is composed of (32x10x2) subdomains and it is simulated in 5 steps with total electrodes shift of 600 mm in 100 s. The results shown in fig. 11 demonstrates the functionality of the proposed model (chapter 4), while each step of simulation is defined two times. The initial time (0 s) is representing initial temperature setting within each subdomain with new electrode position. The secondary time (20 s) is representing end time for each step and for the temperature distribution needed what is needed for next step's initial condition. Based on the results of the individual simulation steps, it can be said that the modeling process is suitable for this application and may be applied also in other cases/studies.

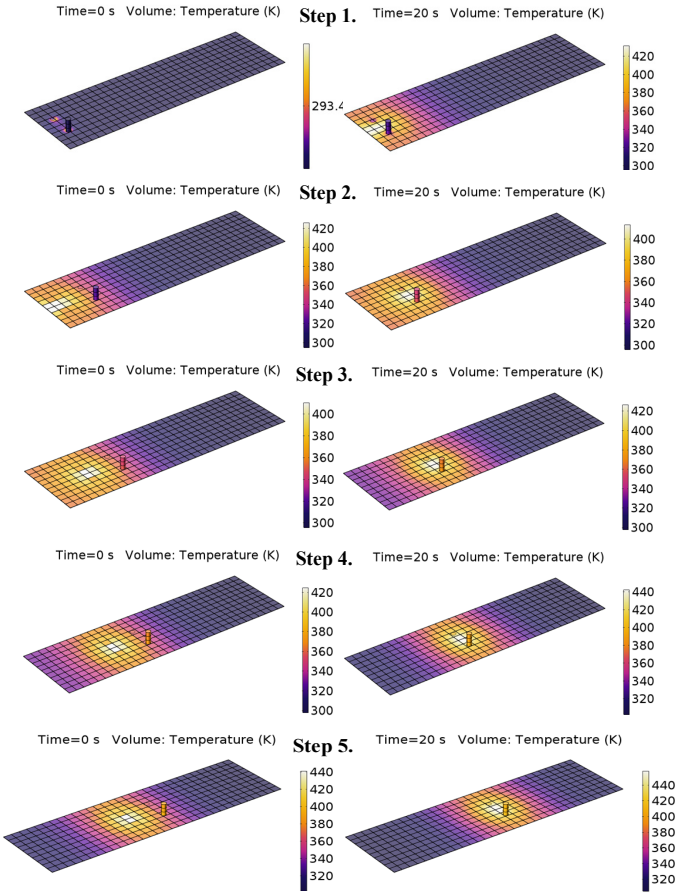


Fig. 11. Simulation results for given parameters

6. Conclusion

In this paper, the design of thermal finite element simulation model of Mo sheet was described. The main aim was to target exact modeling of thermal field in the structure of Mo sheet, while multiple material properties of heating electrodes where considered because resistive heating is considered within the proposed paper.

The presented model had the possibility of the reconfiguration of the electrode material, the electrodes distance, the electrode geometry as well as Mo sheet geometry. Also, Multiphysics is considered, whereby various values of electric current have been applied during experiments. For these purposes, several types of simulation models were developed. The main problem during the investigation is the proper definition of the material properties of Mo sheets. Each manufacturer has different material composition. Therefore, the simulation model has to have the possibility of reconfiguration in order to meet experimental results, which have been made within previous researchers. Consequently, the presented model shall serve for optimization procedures, because experimental investigation mostly acts as time-consuming. Based on investigated configuration it can be claimed that there are possibilities of improvement in speed of heating (point-heating)

as well as possibility of pre-heating (line-heating) by using electrodes from different materials.

Second approach o presented paper was focused on the design of the model, which will consider a dynamic change of the heated place within Mo sheet, i.e. translation move was applied. Various software products have limitations regarding the dynamic movement of the investigated sample, therefore special script in MATLAB environment was developed. With the use of this approach, each iteration of the simulation accepts previous results, which are repeatedly implemented within the computation solution.

Given proposal of the solution was required due to future works, which will be focused on the design of multi-physics simulation model of Mo sheet heating and molding system, where high-validity simulation models are expected to be used namely for further investigations of mechanical parameters and other molybdenum restrictions.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors wish to thank Slovak grant agency APVV for the project no. 0396-15 - Research of perspective high-frequency converter systems with GaN technology and for the project no. 14-0284 - Study of Utility Properties of Formed Molybdenum Sheets Applicable for Horizontal Crystallization of Sapphire Single Crystal.

References

[1] M. Pavelek, M. Frivaldsky, P. Spanik, T. Donic, “Electrothermal model of resistance heating of molybdenum sheet” in EPE 2017, Kouty nad Desnou 2017. https://doi.org/10.1109/EPE.2017.7967262

[2] L. Hargaš, M. Hrianka, J. Lakatos, D. Koniar, “Heat Fields Modeling and Verification of Electronic Parts of Mechatronics Systems” Metalurgija **49**(2), 268 – 272, 2012. https://doi.org/10.5755/j01.eee.123.7.2371

[3] P.Karban, F.Mach, I.Doležel, “Hard-couple model of local direct resistance heating of thin sheets” Journal of Computational and Applied Mathematics **236**(18), 4725-4731, 2012. https://doi.org/10.1016/j.cam.2012.02.036

[4] F.Dughiero, M.Forzan, C.Pozza, E. Sieni, “A translational couple electromagnetic and thermal innovative model for induction welding of tubes” IEEE Transactions on Magnetics **48**(2), 483-486, 2012. https://doi.org/10.1109/TMAG.2011.2174972

[5] K.Mori, S. Maki, Y.Tanaka, “Warm and Hot Stamping of ultra-high tensile strenght sheets using resistance heating” CIRP Annals-Manufacturing Technology **54**, 209-212, 2005. https://doi.org/10.1016/S0007-8506(07)60085-7

[6] S. Maki, Y. Harada, K.I. Mori, H. Makino, “Application of resistance heating technique to mushy state forming of aluminium alloy” Journal of materials processing technology **125**, 477 – 482, 2002.

[7] J. Cuntala, A. Kondelova, O. Hock, M. Pridala, “Electro-Thermal Modeling of Power LED Using COMSOL Environment” in ELEKTRO 11TH INTERNATIONAL CONFERENCE, Slovakia, 2016.

[8] M. Hrianka, L. Hargaš, J. Lakatš, D. Koniar, “Modeling, Simulation, and Verification of Heat Transfer in Power Transistor Cooler” Metalurgija 2 **49**(), 283 – 287, 2012.

[9] K.Mori, S. Maki, Y.Tanaka, “Warm and Hot Stamping of ultra-high tensile strenght sheets using resistance heating” CIRP Annals-Manufacturing Technology, **54**(), 209-212, 2005.

[10] Plansee Group, https://www.plansee.com/en/index.html [10.11.2017]

[11] ED FAGAN INC., MACHINING GUIDE, http://www.edfagan.com/litPDF/Machining_Guide_All_Materials.pdf [10.11.2017]

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 135-147 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Algorithms for Technical Integration of Virtual Power Plants into German System Operation

André Richter*, Ines Hauer, Martin Wolter

Otto-von-Guericke University, Institute of Electric Power Systems, Chair Electric Power Systems and Renewable Energy Sources, 39106 Magdeburg, Germany

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 20 December, 2017
Online: 30 January, 2018

Keywords :
Virtual Power Plants
Operational concepts
Power system
Renewable Energy Sources
Optimization
Case study

ABSTRACT

This paper critically evaluates the operational perspective of Virtual Power Plants (VPP) in Germany by analyzing key factors to replace conventional power plants in the future power system. Therefore, its necessity for the secure operation as well as the technical and economic benefits for the German power system are pointed out. The single sections describe in detail how the requirements on technical functions and standardized communication can be reconciled with the increasing challenges on volatile generation. Furthermore, different operation concepts (profit maximization, intra-day schedule loyalty and system services maximization) are described with respect to their mathematical algorithms and their practical feasibility under consideration of given circumstances and future developments. The impact on the power system by the different possible VPP-operational concepts are exemplary pointed out in a case study with use of a medium voltage Cigre-Benchmark Network. The results indicate a high impact on hand-over points by VPP operational concepts.

1. Introduction

This paper is an extension of work originally presented in 14th International Conference on the European Energy Market (EEM) [1]. The goal of the German government to reduce CO2 emissions about 40 % until 2020 leads to a rising energy infeed fluctuation evoked by renewable energy sources in the German power system[2]. Due to a high percentage of volatile power infeed, the share of gross electricity generation in 2016 increased to 12.3 % from wind, 7.0 % from bio power plants and 5.9 % from photovoltaic (PV) [3]. The future share of renewable energy sources should rise up much more according to the German government. 45 % of the renewable power was installed in medium and low voltage level [4], so that a large percentage of uncontrolled generation is located in this area. Therefore, the electricity system requires intelligent power networks to secure a stable balance between generation and demand [5]. One part of this intelligent power network is a Virtual Power Plant (VPP) that provides centralized control for aggregated units. The future system shall provide advanced functionalities similar to real power plants aiming on economically sensible generation and increased reliability [6, 7]. Therefore, the individual functionalities of decentralized units are bundled up using suitable communication and control systems to entirely replace conventional power plants [7]. However, bundling up single functionalities of single generation unit's does not directly secure similar operation concepts to conventional power plants.

Therefore, this paper deals with the question how VPP can be integrated into the current German power system exploiting their advantageous characteristics like centralized control of small volatile renewable generation units. Included in this, current available technical VPP properties as well as available and economic reasonable operation concepts are checked under consideration of current market circumstances. For example, so far there is no operation concept for VPP that supports the elimination of critical situations in TSO and DSO voltage levels.

[6 – 8] present some investigations already dealt with corresponding optimization methods for operational concepts, but none of them addresses the overall system requirements in its entirety. For that reason, Figure 1 illustrates the system design of a VPP as well as the interaction of market and power system participants and components. All units are remotely accessible through a central energy management and control center, which calculates the optimal schedule by taking into account information from the electricity market and forecast data of the individual generation and load units.

The overall flexibility of decentralized power units could be improved by applying the concept of VPP's [6, 7, 9] but the German electricity market together with applicable regulations still

*Corresponding Author: André Richter, Email: andre.richter@ovgu.de

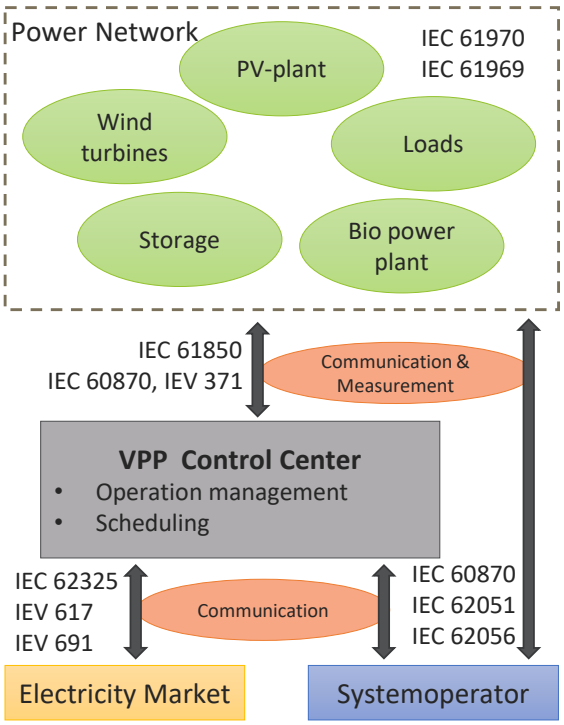


Figure 1. Components of the VPP and influencing factors

not provide the required incentives for a sustainable operation and consequently for a large-scale roll-out. Based on the identification of basic requirements, this paper points out different necessary presupposes for the operation of VPP's similar to conventional power plants. This includes different aspects like communication, generation and demand forecast as well as the German electricity market behavior and costs (*c*) for generation (**Error! Reference source not found.**).

Table 1: Costs of renewable generation without investment costs

	Variable costs in € /kWh	**Annual fix costs in € /kW**
Wind turbines [10]	0.0241 – 0.0268	56 – 61
PV plants [11, 12]	0.0000	13.00 – 35.00
Biopower plants [12]	0.0325	175.00

Furthermore, these presupposes are necessary to realize the new VPP operational concepts for taking part in network congestions elimination or control reserve contribution as well as intra-day schedule loyalty concerning the day-ahead planned generation unit operation. Therefore, not only the benefit for VPP owners and direct marketers is identified, but mainly the operational concpets are mathematically described. The paper presents an ongoing case study, where operational concepts are tested and evaluated according to the impact in a medium voltage CIGRE benchmark system. Thereby, the basic preconditions for the operation of Virtual Power Plants according to the German Market system are considered. Finally, the results and conclusion will indicate the high impact of small centralized controlled generation units as well as use cases where only small incentives are needed for the realization of the system supporting technical concpets. This paper differs strongly from the original paper in [1] by an extended definition of market restriction, in detail concerning the control reserve market and by presenting defined operational concepts and algorithms.

2. Basic Requirements of VPP's

2.1. Communication

Communication is a basic requirement for the operation of VPP as well of intelligent power networks in general. Therefore, it is necessary that generation units, controllable loads and all players offering power flow flexibility and aiming on benefits from variable tariffs can use a standardized and reliable communication system [5]. Based on this, there is a need for a standardized communication language and easy data exchange between all players [13]. Table II lists typical standards for communication in electric power systems, which can be applied in German power system. Furthermore all these standards are necessary for the operation of VPP and the communication with measurement systems, market and the system operator.

Table 2: Communication standards for intelligent power networks [14, 15]

Standard content	**Standard**
Telecontrol	IEC 60870, IEV 371
Communication systems in substations	IEC 61850
Interface for distribution management	IEC 61968
Energy market communication	IEC 62325
Data exchange with metering equipment / Electricity Metering	IEC 62051 /IEC 62056
Application in energy management system	IEC 61970 / 61969

Table II shows that various communication standards are applied to VPPs [16, 17]. According to [17] the IEC 61850 represents one of the most important communication standards for the operation and control of VPP with the integrated intelligent networks and distributed units. The innovative standard enables a secure and reliable communication and control using intelligent measurement systems like Phasor Measurement Units (PMUs) and control units like Remote Terminal Units (RTUs) [17 – 19]. Its necessity is based on the requirement to standardize data-models and furthermore to describe communication mechanisms and system management aspects. This standard is an extension to IEC 60870 resulting from adapted power system requirements. IEC 61850 has the advantage of defining data units through logical nodes with defined data structures (common data classes) [20]. The adressed protocol enables the control center to have direct data access and direct control to any physical unit of the VPP in the power network. Furthermore, the application of this standard combined with direct communication supplies the VPP operator with information about the physical location so that he is able to separate physical units according to accounting areas for instance.

2.2. Forecast-oriented power unit scheduling

For setting up an optimal power unit schedule, the VPP operator needs to distinguish the individual characteristics of the available resources. This information is not only necessary for an

optimal market participation, but also for enabling the system operator to make day-ahead and intra-day forecasts as well as to know the maximum controllable power for critical situations. The demand forecast for high numbers of small consumers can be realized using standardized load profiles. Bigger customers, like industry companies with bilateral contracts, require separate measurement systems in order to meet the requirements for providing dedicated demand side potential. In contrast, a larger percentage of units in the VPP are generation units and especially volatile ones like PV plants and wind turbines. Consequently, an exact forecast for those units is indispensable from the operator‘s point of view, e.g. for scheduling the secondary control reserve [21].

The forecast for wind and PV generation can be realized with different methods presented in Figure 2 and Figure 3. Thereby, the volatile factors are wind speed and solar radiation [19].

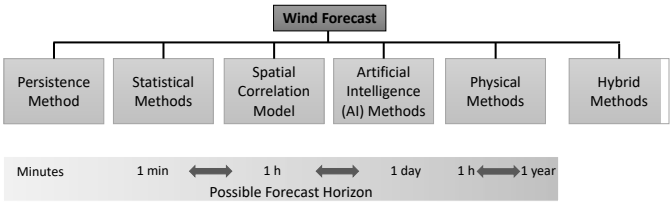


Figure 2. Overview of wind forecast methods and corresponding mathematical algorithms, [21 – 23]

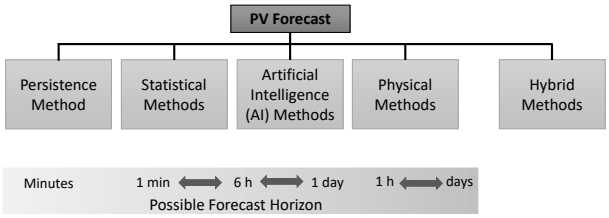


Figure 3. Overview of PV forecast methods and corresponding mathematical algorithms [25]

Wind forecast methods can be divided in six groups depending on its corresponding mathematical algorithms. Four methods (Persistence-, Statistical-, Artificial Intelligence method and Spatial Correlation model) are based on current wind measurement and statistical data as well as on learning approaches to predict wind generation. Depending on the applied method, the forecast horizon varies between a few minutes up to one day. To achieve an increase in accuracy and time horizon, it is necessary to use a physical or hybrid method. Both methods are based on meteorological and climate data and use the numerical weather prediction (NWP) for that purpose, but under the constraint of high computational effort. Regarding the methods and the forecast horizon the best accuracy (mean absolute error) varies between 5 % and 7 % for one to two hours forecast and increases for day-ahead forecast [21 - 24]. According to Figure 3, three of the shown methods (Persistence-, Statistical- and Artificial Intelligence Method) process current solar radiation and statistical data. This way utilizable data can be provided for a forecast horizon of a few minutes until one day. However, similar to wind forecast the accuracy rapidly decreases with the forecast horizon. Compared to wind forecast, PV forecast depends even more on the availability of information from weather stations and satellite data to detect clouds, high fog and other phenomena that effect solar radiation. Therefore, the physical method and the hybrid methods use NWP and Cloud Imagery for prediction with subsequent data post processing [25].

In consequence to the described uncertainties in generation forecast, the VPP operator has to deal with different impacts. In case of direct market participation, he has to expect sanctions from the power system operator or to participate in the Intra-Day market accepting possible high costs [26]. Besides using storage systems, there is a solution by using existing flexibilities from wind turbines pv, biomass plants and other technical solutions as in **Error! Reference source not found.**. Those technologies can provide power reserve if they are not operated at their maximum power output limit. However, since storage systems are still expensive and an appropriate market model involving positive and negative power capacity of wind and bio generation is not provided, more profit can be achieved by operating the units at its maximum power output level.

2.3. German Electricity Market

In this section, the German electricity market will be characterized focusing on VPP to identify possible operation behavior which influence system operation. Based on this, there are three possibilities to sell energy from renewable energy sources. The first possibility is to receive fixed payment, which is determined based on the German Renewable Energy Act (EEG). The other two possibilities are to participate in the Energy-Only-Market and to sell control reserves at the balancing market [26].

The electricity is traded in Germany either at the energy exchange or bilateral at the Over-the-Counter market. The European Energy Exchange (EEX) in Leipzig is the biggest market place for electricity produced in Germany. Electricity on the EEX can be traded either on the futures market or on the spot market (EPEX SPOT in Paris). The main difference between these markets is the electricity delivery time. Thus, the future markets provides contracts for long-term supply of electricity up to six years in the future [26]. The spot market is used as a trading platform for the short-term supply of electricity within 1-2 days, namely Day-Ahead and Intra-Day market [26, 28]. Thus, the day-ahead market is used to trade the electricity, that will be generated and consumed the next day. Because of the fact, that the day-ahead market is based on forecasted profiles (generation and consumption), an Intra-Day market is necessary to balance the forecasted errors. The contracts on the Spot market are mostly carried out physically, that means the physical electricity delivery to the fixed date [29]. According to (§11 EEG 2014), a system operator is obligated to transmit the renewable electricity and to sell it at the exchange. The operators of renewable power plants receive a fixed remuneration, which is supported by the EEG-apportionment. As a result, the renewable energy sources are always first at the merit-order, independently of the operation costs. Nuclear and lignite power plants follow, since they are usually used to cover the base load due to technical and economic advantages. Hard coal, gas and oil power plants are usually more expensive. Thus, renewable electricity has a great influence on the electricity price, because of its prior position. The operator of renewable energy sources can freely choose between a fixed EEG remuneration and the direct marketing at the beginning of each calendar year, §20 (1) EEG 2014. In addition, all systems with an installed power of more than 100 kW, which are set to operation from 1 January 2016, must be marketed directly (§ 21 EEG 2017). The benefit of direct trading in contrast to the fixed remuneration is that trader get the chance of higher profits by receiving market earnings plus a market premium. The market premium is lower

than the fixed remuneration but another governmental incentive to get a higher percentage of renewables in the market system. However, there are some technical limitations which restrict the operation of VPP at the market and in the power system. These are for example the lowest bid of 0.1 MW to participate in the Day-ahead and in the continous Intra-Day market. Furthermore, there are finacial requirements to participate in the spot market, in Germany [30]. The participation fee for Day-ahead market is about 10000 €/ a with additional 0.04 €/MWh (Day-Ahead auction) and 0.08 €/MWh (intra-day auction). Aditionally, there is a one-time participation fee to get into the spot market of about 25000 € for Day-Ahead market. These are costs that have to be covered by an optimal operational concept or new regulatory incentives.

The control reserve market in Germany is characterized by the fact that the German transmission system operator are responsible to keep the power system stable and the rated frequency of 50 Hz. There are 3 types of control reserves (primary control reserve, secondary control reserve as well as minute reserve). These types are differentiated according to activation speed and duration. The procurement of these control reserves is organized in a tender auction with participation of power plant operators and customers. One main fact is that pooling is still permitted to reach the market entry boundaries [31]. Main characteristics of the three control reserves are explained subsequently according to [32]:

Primary control reserve (PCR):

- full activation of PCR within 30 seconds
- incident period that have to be covered 0 <t <15 min
- automatic activation
- symmetrical (positive and negative PCR is not separated) bid of minimum +/- 1 MW
- the tender submission period is one week
- fixed power price

Secondary control reserve (SCR):

- energy balance of the control area and frequency control
- immediate automatic activation by the concerned TSO
- complete activation within 5 minutes
- automatic activation based on Merit-Order-List
- separated bid for negative and positive SCR is possible
- separation of SCR positive and negative in high tariff (HT, on Monday till Friday, 8 am til 8 pm) and low tariff (LT, weekend, national holidays and 8 pm til 8 am)
- minimum bid of 5 MW and an increment of 1 MW
- the tender submission period is one week
- fixed power price and energy price on call

Minute reserve (MR):

- automatic activation is based on a Merit-Order-List
- complete activation within 15 minutes
- incident period that have to be covered are t > 15 min to 4 quarter hours or up to severalhours in case several incidents
- separated bid for negative and positive MR is possible
- positive and negative MR, always with six 4-hour time slices
- minimum bid of 5 MW
- the tender submission period is one day
- fixed power price and energy price on call

Finally, the market regulatories and EEG incentives for the operation of single generation units are nowadays much higher, instead of VPP-operation with centralized controlled of small generation units. But there are ongoing changes in the market system which enable VPP for offering more system supporting functionalities. In [33] a change in the control reserve market system is already anounced. For instance, SCR and MR are intended for a daily (calendar) tender submission period with 4-hour time slices. Furthermore, exceptional rules that allow for minimum bids below 5 MW are introduced.

In conclusion, there are still no high financial and regulatory incentives to operate VPP in a way, which serves both areas – market and technical operation – at the moment. Because of this physical delivery of electricity remains uncertain with respect to a future power system including a higher share of renewables. Consequently, it will be indispensable to use the VPP benefits of generation unit aggregation and its aggregated central controlled flexibilities to have minimum system effecting controllable power. Therefore, some assumptions are made for the ongoing analyzes for a medium-term future VPP.

3. VPP -Operational concepts

One of the main challenges of VPP operators is the setup of the internal merit order, that defines which power unit shall be used to satisfy a certain power demand or infeed. This has to be done under consideration of type of generation unit, its technical capability and regulatory as well as economical restrictions. However, right now VPP operators maximize their profits by offering as much energy as possible in the markets aiming on the highest possible profit. As one result, VPP functionalities remain unused due to missing monetary or regulatory incentives, in particular the centralized control of aggregated small generation units. This causes a problem for the system operator: VPP operators sell energy based on the market premium at every point of time, even in case of congestions in the DSO or TSO network.

Therefore, this paper offers two main system supporting operational concepts and additonally an economical concept (see Figure 4). Each concept is explained in detail in the ongoing subsections 3.2 and 3.3.

1) Economical oriented (green): porift maximization, by optimal energy selling on the electricity market.
2) Technical oriented (orange): intra-day schedule loyalty and system service maximization including centriole reserve maximization and redispatch maximization.

3.1. Mathematical optimization problem

The operational concepts can be mathematically involved and expressed in terms of a maximization or minimization problem (objective function) with a given number of restricted goods (constraints). Therefore, this paper presents options to solve the optimization problems in MATLAB for day-ahead scheduling and intra-day schedule loyalty in intervals of 15 minutes. The development of schedules in day-ahead and intra-day planning rely on an optimization problem with an application-specific objective function. The restrictions are defined for technical characteristics of the system, electricity market restrictions and for current as well as predicted system states. Linear optimization method based on simplex algorithm is applied to solve the optimization problem. Linear optimization either minimizes or maximizes a linear

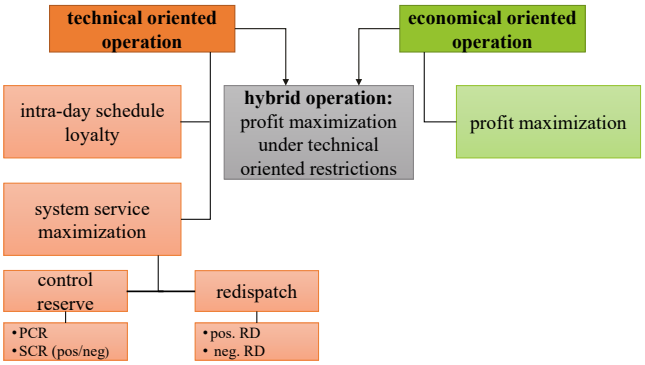


Figure 4. VPP operational concepts

objective function, taking into account linear constraints. The linear optimization contains only continuous variables, whereby benefitery short calculation times are realized. From the linear optimization, the mixed integer linear optimization can be distinguished. Mixed integer linear optimization (MILP) is the basis of the optimization problem for day-ahead and intraday optimization implemented in this paper. Unlike linear optimization, variables can be either continuous or integer or binary. Especially with binary variables it is possible to model logical conditions. The degree of detail of the optimization can be increased, but this is accompanied by an increase in the computational effort for the determination of the optimal solution. [34] The solver intlinprog is used to implement the scheduling problem in MATLAB. Firstly, the objective function is defined and implemented according profit maximization. Ongoing the different technical and economical upper and lower bounds as well as linear constraints were defined and implemented. Finally, the integer constraints have to be implemented. The solver finds a first less constraint starting solution by using linear optimization without integer constraints. The starting solution serves as a starting point for new solutions for the method branch and bound. The method branch and bound takes all constraints and inequaltities into account by generating subproblems and finally finds the best solution, if this solution is within a defined tolerance gap according to the other solutions.

3.2. Economical oriented operation

The main objective of a power plant operator and thus a VPP operator is always to maximize its profits, regardless of the system situation. The regulatory framework and the optimization skills always determine whether or in what amount the operator generates profit. Consequently, the composition of the VPP has a significant impact on potential profits. In the analysis of this concept for resource planning, an optimization based on today's market data for control reserve markets ($p_{\mathrm{PCR/\,SCR/\,MR}}$ / $P_{\mathrm{PCR/\,SCR/\,MR}}$), Day-Ahead auction (p_{auction} / $P_{\mathrm{wind/pv/bio,auc}}$) and continuous Intra-Day market ($p_{\mathrm{continous,IH,T}}$ / $P_{\mathrm{wind/pv/bio,continous,IH,T}}$) is made. The analysis assumes that both a fixed EEG allowance and a market premium for market participation through EEG plants are not paid out. Aim of this assumption is to analyze the VPP operation under perfect market conditions without any subsidies. This leads to the objective function where profit (p) should be maximized by optimal day-ahead generation unit schedule(P_{wind}, P_{pv} and P_{bio}). The optimal distribution of power to different selling options is limited to the forecast of renewables ($P_{\mathrm{wind/pv,\,forecast}}$), the installed power, various constraints and other secondary conditions which are described in (2) – (34).

$$\begin{aligned}\max f(p) = {} & p_{\mathrm{wind}}(t)\cdot P_{\mathrm{wind}}(t) + p_{\mathrm{pv}}(t)\cdot P_{\mathrm{pv}}(t) \\ & + p_{\mathrm{bio}}(t)\cdot P_{\mathrm{bio}}(t) \\ & + p_{\mathrm{auction}}(t)\cdot P_{\mathrm{wind,auc}}(t) \\ & + p_{\mathrm{auction}}(t)\cdot P_{\mathrm{pv,auc}}(t) \\ & + p_{\mathrm{auction}}(t)\cdot P_{\mathrm{bio,auc}}(t) \\ & + p_{\mathrm{continous,IH,T}}(t)\cdot P_{\mathrm{wind,continous,IH,T}}(t) \\ & + p_{\mathrm{continous,IH,T}}(t)\cdot P_{\mathrm{pv,continous,IH,T}}(t) \\ & + p_{\mathrm{continous,IH,T}}(t)\cdot P_{\mathrm{bio,continous,IH,T}}(t) \\ & + p_{\mathrm{PCR}}(t)\cdot P_{\mathrm{PCR}}(t) \\ & + p_{\mathrm{SCR,pos}}(t)\cdot P_{\mathrm{SCR,pos}}(t) \\ & + p_{\mathrm{SCR,neg}}(t)\cdot P_{\mathrm{SCR,neg}}(t) \\ & + p_{\mathrm{MCR,pos}}(t)\cdot P_{\mathrm{MCR,pos}}(t) \\ & + p_{\mathrm{MCR,neg}}(t)\cdot P_{\mathrm{MCR,neg}}(t)\end{aligned} \tag{1}$$

During the analyzes several constraints are considered. The main constraints are:

- The profit calculation based on revenue (r) is depending on the market selling option and the costs (c) (energy generation costs, spot-market selling).

$$p_{\mathrm{wind}}(t) = -c_{\mathrm{wind}} \tag{2}$$

$$p_{\mathrm{pv}}(t) = -c_{\mathrm{pv}} \tag{3}$$

$$p_{\mathrm{bio}}(t) = -c_{\mathrm{bio}} \tag{4}$$

$$p_{\mathrm{auction}}(t) = r_{\mathrm{auction}}(t) - c_{\mathrm{auction}} \tag{5}$$

$$p_{\mathrm{continous.IH,T}}(t) = r_{\mathrm{continous.IH,T}}(t) - c_{\mathrm{continous.IH}} \tag{6}$$

- Minimum possible generation for wind, PV and bio generation that can be used for optimization: The factors $f_{\mathrm{wind,min}}$ $f_{\mathrm{pv,min}}$ and $f_{\mathrm{bio,min}}$ set the lower bounds of minimum generation output. They define pre-process depending generation output especially for bio generation. The minimum for wind and PV is zero by default.

$$f_{\mathrm{wind,min}} \cdot P_{\mathrm{wind,forecast}}(t) \le P_{\mathrm{wind}}(t) \tag{7}$$

$$f_{\mathrm{pv,min}} \cdot P_{\mathrm{pv,forecast}}(t) \le P_{\mathrm{pv}}(t) \tag{8}$$

$$f_{\mathrm{bio,min}} \cdot P_{\mathrm{bio,inst}} \le P_{\mathrm{bio}}(t) \tag{9}$$

- Bio generation power ramping: The maximum rate of change of power ($f_{\Delta P\mathrm{bio,pos}}$, $f_{\Delta P\mathrm{bio,neg}}$) is limited, comparable with conventional power plants.

increase of power for t = 1:

$$P_{\text{bio}}(t) - f_{\text{bio,Start}} \cdot P_{\text{bio,inst}} \leq f_{\Delta P\text{bio,pos}} \cdot P_{\text{bio,inst}} \tag{10}$$

increase of power for t ≥ 2:

$$P_{\text{bio}}(t) - P_{\text{bio}}(t-1) \leq f_{\Delta P\text{bio,pos}} \cdot P_{\text{bio,inst}} \tag{11}$$

decrease of power for t =1:

$$f_{\text{bio,Start}} \cdot P_{\text{bio,inst}} - P_{\text{bio}}(t) \leq f_{\Delta P\text{bio,neg}} \cdot P_{\text{bio,inst}} \tag{12}$$

decrease of power for t ≥ 2:

$$P_{\text{bio}}(t-1) - P_{\text{bio}}(t) - \leq f_{\Delta P\text{bio,neg}} \cdot P_{\text{bio,inst}} \tag{13}$$

- Limitation of provision of control reserve (CR) by bio generation: Not every bio power plant is able to fulfill the technical requirements of PCR provision. Therefore, the proportion of bio generation units that can provide PCR can be limited with the factor $f_{\text{Bio,PCR}}$. Furthermore, the contribution of PCR and SCR has to be defined in constraints according to the technical requirements that PCR have to be provided in at least 30 s and SCR in at least 5 min. $f_{\text{d}P\text{bio,pos}}$ and $f_{\text{d}P\text{bio,neg}}$ describe the maximum possible rate of change of power in one minute so that the constraints for PCR and SCR are described as follows:

PCR contribution within 30 s, positiv and negative:

$$\text{d}P_{\text{bio,PCR,pos}}(t) \leq \frac{f_{\text{d}P\text{bio,pos}}}{2} \cdot f_{\text{bio,PCR}} \cdot P_{\text{bio,inst}} \tag{14}$$

$$\text{d}P_{\text{bio,PCR,neg}}(t) \leq \frac{f_{\text{d}P\text{bio,neg}}}{2} \cdot f_{\text{bio,PCR}} \cdot P_{\text{bio,inst}} \tag{15}$$

SCR contribution within 5 min, positiv and negative:

$$\text{d}P_{\text{bio,SCR,pos}}(t) \leq 5 \cdot f_{\text{d}P\text{bio,Pos}} \cdot P_{\text{bio,inst}} \tag{16}$$

$$\text{d}P_{\text{bio,SCR,neg}}(t) \leq 5 \cdot f_{\text{d}P\text{bio,neg}} \cdot P_{\text{bio,inst}} \tag{17}$$

- Maximum possible control reserve. The provision of control reserve is limited to the requirements described in section 2.3. These are the continuous availability of CR according to the defined periods (PCR-continuous over the day, SCR-LT, SCR-HT, MR-4 hour intervals) and under consideration of wind and PV forecast and the limited contribution of bio generation, respectively.

PCR positive and negative:

$$\begin{aligned} P_{\text{PCR}}(t) \leq\ & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + \text{d}P_{\text{bio,PCR,pos}}(t) \end{aligned} \tag{18}$$

$$\begin{aligned} P_{\text{PCR}}(t) \leq\ & P_{\text{wind}}(t) - f_{\text{wind,min}} \cdot P_{\text{wind,forecast}}(t) \\ & + P_{\text{pv}}(t) - f_{\text{pv,min}} \cdot P_{\text{pv,forecast}}(t) \\ & + \text{d}P_{\text{bio,PCR,neg}}(t) \end{aligned} \tag{19}$$

SCR positive and negative:

$$\begin{aligned} P_{\text{SCR,pos,HT}}(t) + P_{\text{SCR,pos,NT}}(t) \leq\ & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + \text{d}P_{\text{bio,SCR,pos}}(t) \end{aligned} \tag{20}$$

$$\begin{aligned} P_{\text{SCR,neg,HT}}(t) + P_{\text{SCR,neg,NT}}(t) \leq\ & P_{\text{wind}}(t) \\ & - f_{\text{wind,min}} \cdot P_{\text{wind,forecast}}(t) \\ & + P_{\text{pv}}(t) \\ & - f_{\text{pv,min}} \cdot P_{\text{pv,forecast}}(t) \\ & + dP_{\text{bio,SCR,neg}}(t) \end{aligned} \tag{21}$$

MR positive and negative:

$$\begin{aligned} P_{\text{MCR}}(t) \leq\ & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + P_{\text{bio,inst}} - P_{\text{bio}}(t) \end{aligned} \tag{22}$$

$$\begin{aligned} P_{\text{MCR}}(t) \leq\ & P_{\text{wind}}(t) - f_{\text{wind,min}} \cdot P_{\text{wind,forecast}}(t) \\ & + P_{\text{pv}}(t) - f_{\text{pv,min}} \cdot P_{\text{pv,forecast}}(t) \\ & + P_{\text{bio}}(t) - f_{\text{bio,min}} \cdot P_{\text{bio,inst}} \end{aligned} \tag{23}$$

- Avoidance of multiple use of power for control reserve provision: This constraint defines an upper limit for the sum of CR not exceeding the amount of available power. This is divided into positive and negative CR.

$$\begin{aligned} CR_{\text{pos}}(t) = & P_{\text{PCR}}(t) + P_{\text{SCR,pos,HT}}(t) \\ & + P_{\text{SCR,pos,NT}}(t) + P_{\text{MR,pos,4 h-slice}}(t) \end{aligned} \tag{24}$$

$$\begin{aligned} CR_{\text{pos}}(t) \leq\ & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + \text{d}P_{\text{bio,PCR,pos}}(t) \end{aligned} \tag{25}$$

$$\begin{aligned} CR_{\text{pos}}(t) \leq\ & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + P_{\text{bio,inst}} - P_{\text{bio}}(t) \end{aligned} \tag{26}$$

$$\begin{aligned} CR_{\text{neg}}(t) = & P_{\text{PCR}}(t) + P_{\text{SCR,neg,HT}}(t) \\ & + P_{\text{SCR,neg,NT}}(t) + P_{\text{MR,neg,4 h-slice}}(t) \end{aligned} \tag{27}$$

$$\begin{aligned} CR_{\text{neg}}(t) \leq\ & P_{\text{wind}}(t) - f_{\text{wind,min}} \cdot P_{\text{wind,forecast}}(t) \\ & + P_{\text{pv}}(t) - f_{\text{pv,min}} \cdot P_{\text{pv,forecast}}(t) \\ & + \text{d}P_{\text{bio,SCR,neg}}(t) \end{aligned} \tag{28}$$

$$\begin{aligned} CR_{\text{neg}}(t) \leq\ & P_{\text{wind}}(t) - f_{\text{wind,min}} \cdot P_{\text{wind,forecast}}(t) \\ & + P_{\text{pv}}(t) - f_{\text{pv,min}} \cdot P_{\text{pv,forecast}}(t) \\ & + P_{\text{bio}}(t) - f_{\text{bio,min}} \cdot P_{\text{bio,inst}} \end{aligned} \tag{29}$$

- Minimum bid and increment for control reserve provision, according to the market restrictions: The control reserve amount can be zero or or an integer

multiple of the increments, depending on size of minimum bid and minimum increment. The restriction is presented for PCR with analog constraints for SCR and MR. The factor $y(t)$ is a binary variable for the proof, if enough power is available for CR provision.

$$P_{\mathrm{PCR}}(t) - y_{\mathrm{PCR}}(t) \cdot P_{\mathrm{PCR,min}} \geq 0 \tag{30}$$

- Spot-market restriction: According to the spot-market options regarded in this paper, there are the possibilities to sell the energy in the Day-Ahead auction offered by hour or in the continuous Intra-Day market offered by 15 minutes. Because of the fact that this is a day-ahead planning process, Intra-Day market prices have to be assumed as forecasted values. A further constraint is that the available energy should be sold out under profitable conditions.

$$P_{\mathrm{wind}}(t) = P_{\mathrm{wind,auc}}(t) + P_{\mathrm{wind,continous,IH,T}}(t) \tag{31}$$

$$P_{\mathrm{pv}}(t) = P_{\mathrm{pv,auc}}(t) + P_{\mathrm{pv,continous,IH,T}}(t) \tag{32}$$

$$P_{\mathrm{bio}}(t) = P_{\mathrm{bio,auc}}(t) + P_{\mathrm{bio,continous,IH,T}}(t) \tag{33}$$

$$\begin{aligned} & P_{\mathrm{wind}}(t) + P_{\mathrm{pv}}(t) + P_{\mathrm{bio}}(t) \\ & = P_{\mathrm{wind,auc}}(t) + P_{\mathrm{wind,continous,IH,T}}(t) \\ & + P_{\mathrm{pv,auc}}(t) + P_{\mathrm{pv,continous,IH,T}}(t) \\ & + P_{\mathrm{bio,auc}}(t) + P_{\mathrm{bio,continous,IH,T}}(t) \end{aligned} \tag{34}$$

As a result of the economic optimization schedules are defined for wind, PV and bio generation. All presented constraints consider the figures presented in section 2. This methodology allows for an economically optimal day-ahead schedule, regardless of the size of the VPP (available generation units) and the impact on the electrical power system. In future systems with high penetration of RES, this can lead to unstabelizing effects and to a lack of system services. Therefore, the system itself and system operator need system supporting services to handle exemplary high volatility generation and forecast errors.

3.3. Technical oriented operation

Within this section, various new network-relevant VPP operating concepts are to be examined, which are intended to ensure stable grid operation in the future and also to minimize the effort of operational management for directly and indirectly connected grid operators. These different concepts thus represent the integration of network operator requirements into the operational planning. The new technical concepts meet the request of reliable schedules contributed by VPP and by renewable generation units in general. Furthermore, the presented concepts will enable a maximum of PCR and SCR contribution as well as redispatch. Therefore, the day-ahead maximization function in (1) with its constraints in (2) – (34) are applied. Changed or added objective functions and constraints for optimal day-Ahead scheduling will be presented according to the concepts. The intra-day schedule loyalty represents a further concept which was presented in general in [35].

3.3.1. Intra-day schedule loyalty

The objective of intra-day schedule loyalty is to minimize the schedule deviation, which is mainly based on forecasting errors related to renewable energy source (RES) infeed in the operational planning of the VPP control center. In this case, the power consumption or the power recovery from or into the higher-level network should be reduced, so that the higher-level network operator or the accounting area manager only has to make few or no interventions. If all VPP operators were capable of doing this in the future, day-ahead planning process of system operators can be optimized and network congestions may be avoided.

However, this requires a defined level of flexibility within the VPP and can also lead to curtailments of generation units. The intra-day loyalty operation must be preceded by a day-ahead optimization in which capacities are blocked in order to compensate intra-day schedule deviations. The consumption of control energy should be avoided, if possible. For this purpose, a high fictitious penalty of 1000 €/MW is initially assumed for the balancing energy consumption when positive balancing power ($p_{\mathrm{reBAP,pos}}(t)$) or negative balancing power ($p_{\mathrm{reBAP,pos}}(t)$) is consumed. If the day-ahead schedule can not be met without consumption of balancing energy, balancing energy can be received ($P_{\mathrm{reBAP,pos}}(t)$ and $P_{\mathrm{reBAP,neg}}(t)$). Therefore, the term of the balancing energy represents a measure of the deviation from schedule. Thereby, a main constraint is that a control reserve deviation ($P_{\mathrm{deviation,CR,pos}}(t)$ and $P_{\mathrm{deviation,CR,neg}}(t)$) is prohibited by use of fictitious penalty ($p_{\mathrm{deviation,CR}}$) about 10000 €/MW. The objective function is presented in (35) with the objective to minimize the deviation according to the planned market schedule.

$$\begin{aligned} \max f(p,P) = & \; p_{\mathrm{wind,intraday}}(t) \cdot P_{\mathrm{wind,intraday}}(t) \\ & + p_{\mathrm{pv,intraday}}(t) \cdot P_{\mathrm{pv,intraday}}(t) \\ & + p_{\mathrm{bio,intraday}}(t) \cdot P_{\mathrm{bio,intraday}}(t) \\ & + p_{\mathrm{reBAP,pos}}(t) \cdot P_{\mathrm{reBAP,pos}}(t) \\ & + p_{\mathrm{reBAP,neg}}(t) \cdot P_{\mathrm{reBAP,neg}}(t) \\ & + p_{\mathrm{deviation,CR}} \cdot P_{\mathrm{deviation,CR,pos}}(t) \\ & + p_{\mathrm{deviation,CR}} \cdot P_{\mathrm{deviation,CR,neg}}(t) \end{aligned} \tag{35}$$

This operational concept requires similar constraints, like fictitious penalties for CR deviation and the possible use of balancing power. But the most important point for this operational concept is the flexibility, which is used to fulfill the schedule. Therefore, it is needed to reduce the maximum available power output from bio-generation in the day-ahead planning process by a defined value (e.g. 5 %). This procedure has to be applied for wind and PV forecast as well. The day-ahead forecast has to be reduced by a factor to ensure intra-day flexibilities and intra-day schedule loyalty. The factor has to be optimized depending on forecast precision and VPP generation unit composition. Two further constraints are loyalty of generation units and observance of day-ahead and intra-day schedule. The constraint loyalty of generation units has the aim to consider optimized day-ahead generation unit schedules for wind, PV and bio under the changed intra-day

conditions for renewable infeed ($P_{\text{wind,intraday}}(t)$ and $P_{\text{pv,intraday}}(t)$). If this will not be possible, energy has to be bought according to positive balancing power ($P_{\text{reBAP,pos}}(t)$) or negative balancing power ($P_{\text{reBAP,pos}}(t)$). The constraint observance of day-ahead and intra-day schedule ensure that the schedules of generation units can be changed according to the failure in forecast of renewables. But the sum of generation and the planned amount of cumulated power in each time step for trading on Day-Ahead and Intra-Day market is as much as possible prohibited to change. Both constraints can be expressed as:

$$\begin{aligned} P_{\text{wind}}(t) + P_{\text{pv}}(t) + P_{\text{bio}}(t) = {} & P_{\text{wind,intraday}}(t) \\ & + P_{\text{pv,intraday}}(t) \\ & + P_{\text{bio,intraday}}(t) \\ & + P_{\text{reBAP,pos}}(t) - P_{\text{reBAP,neg}}(t) \end{aligned} \tag{36}$$

The result of this operational concept is a reliable schedule planning which supports system operator in systems with a high amount of RES to keep a stable system and avoid redispatch. The changed generation unit schedule depends on the quality of forecast and reserved capacity for intra-day flexibility and thus indirectly depending on the quality of the forecast.

Furthermore, this concept only needs small monetary or regulatory incentives, because the needed flexibility reserve not leads to high loss in profit. But there is a need of incentives for this operational concept.

3.3.2. Redispatch maximization

The concept of redispatch maximization is a mainly technical triggered concept, which currently has no regulatory or monetary incentives. But in medium-term and future systems there will be an incredible high need for system operators to handle congestions in a system with less controllable conventional power plants. VPP can provide those functions utilizing the concept of centralized control of RES. It has to be assumend that the VPP is either small, so that it feed in a defined network section like in one medium voltage network, or that the schedule for this concept is adaptable for defined network sections. This assumption has to be done to ensure the effectiveness of redispatch in the system. The objective function (37) is focused on the maximization of negative ($P_{\text{RD,neg}}(t)$) or positive ($P_{\text{RD,pos}}(t)$) available power for redispatch. Therefore, the monetary objective function (1) is extended with fictious positive ($p_{\text{RD,pos}}(t)$) and negative ($p_{\text{RD,neg}}(t)$) profit for redispatch. This profit has to be given a high constant value only. A benefit of the methodology of fictious profits to maximize technical needs is that minimum monetary incentives can be identified under given market circumstances, if there is an ambition to create and examine new market concepts for technical needs.

With the integration of redispatch, additional constraints have to be defined. These are the definition of lower bounds for available power, exceeding a defined minimum in (38) – (39) and the calculation of available power for redispatch in (40) – (41).

$$\begin{aligned} \max f(P_{\text{RD}}) = {} & p_{\text{Wind}}(t) \cdot P_{\text{Wind}}(t) + p_{\text{PV}}(t) \cdot P_{\text{PV}}(t) \\ & + p_{\text{Bio}}(t) \cdot P_{\text{Bio}}(t) \\ & + p_{\text{auction}}(t) \cdot P_{\text{Wind,auc}}(t) \\ & + p_{\text{auction}}(t) \cdot P_{\text{PV,auc}}(t) \\ & + p_{\text{auction}}(t) \cdot P_{\text{Bio,auc}}(t) \\ & + p_{\text{continous,IH,T}}(t) \cdot P_{\text{Wind,continous,IH,T}}(t) \\ & + p_{\text{continous,IH,T}}(t) \cdot P_{\text{PV,continous,IH,T}}(t) \\ & + p_{\text{continous,IH,T}}(t) \cdot P_{\text{Bio,continous,IH,T}}(t) \\ & + p_{\text{PCR}}(t) \cdot P_{\text{PCR}}(t) \\ & + p_{\text{SCR,pos}}(t) \cdot P_{\text{SCR,pos}}(t) \\ & + p_{\text{SCR,neg}}(t) \cdot P_{\text{SCR,neg}}(t) \\ & + p_{\text{MCR,pos}}(t) \cdot P_{\text{MCR,pos}}(t) \\ & + p_{\text{MCR,neg}}(t) \cdot P_{\text{MCR,neg}}(t) \\ & + p_{\text{RD,pos}}(t) \cdot P_{\text{RD,pos}}(t) \\ & + p_{\text{RD,neg}}(t) \cdot P_{\text{RD,neg}}(t) \end{aligned} \tag{37}$$

$$P_{\text{RD,pos}}(t) \geq P_{\text{min,RD,pos}}(t) \tag{38}$$

$$P_{\text{RD,neg}}(t) \geq P_{\text{min,RD,neg}}(t) \tag{39}$$

$$\begin{aligned} P_{\text{RD,pos}}(t) = {} & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + P_{\text{bio,inst}} - P_{\text{bio}}(t) \end{aligned} \tag{40}$$

$$\begin{aligned} P_{\text{RD,neg}}(t) = {} & P_{\text{wind}}(t) + P_{\text{pv}}(t) \\ & + P_{\text{bio}}(t) - f_{\text{bio,min}} \cdot P_{\text{bio,inst}} \end{aligned} \tag{41}$$

3.3.3. Control reserve maximization

Control reserve is one of the most needed products in systems with high volatile generation. CR is important for a global point of view for systems with increased percentage of renewables and decreased number of CR providing conventional power plants. VPP offers the possibility to fulfill market restriction for control reserve market by pooling small decentralized generation units. Depending on the size (amount of installed renewable capacity), VPP can provide only CR products most of the time. Therefore, this optimization maximizes the contribution of positive or negative SCR or the contribution of PCR. The only limiting factors are market restriction concerning minimum bid size and minimum increment, so that CR is maximized first and remaining energy is traded according to profit maximization. This optimization is again based on (1) extended by fictitious high profit for CR. The fictitious profit is chosen very high, depending on the maximization problem (PCR, pos. SCR and neg. SCR).

$$\begin{aligned}\max f(P_{CR}) = {} & p_{Wind}(t)\cdot P_{Wind}(t) + p_{PV}(t)\cdot P_{PV}(t) \\ & + p_{Bio}(t)\cdot P_{Bio}(t) \\ & + p_{auction}(t)\cdot P_{Wind,auc}(t) \\ & + p_{auction}(t)\cdot P_{PV,auc}(t) \\ & + p_{auction}(t)\cdot P_{Bio,auc}(t) \\ & + p_{continous,IH,T}(t)\cdot P_{Wind,continous,IH,T}(t) \\ & + p_{continous,IH,T}(t)\cdot P_{PV,continous,IH,T}(t) \\ & + p_{continous,IH,T}(t)\cdot P_{Bio,continous,IH,T}(t) \\ & + p_{PCR}(t)\cdot P_{PCR}(t) \\ & + p_{SCR,pos}(t)\cdot P_{SCR,pos}(t) \\ & + p_{SCR,neg}(t)\cdot P_{SCR,neg}(t) \\ & + p_{MCR,pos}(t)\cdot P_{MCR,pos}(t) \\ & + p_{MCR,neg}(t)\cdot P_{MCR,neg}(t)\end{aligned} \quad (42)$$

The only constraints which have to be added according to the day-ahead optimization are the definition of lower bounds (P_{min}) for minimum provision of control reserve. Further restrictions are still defined in section 3.2. The constraints are:

$$P_{PCR}(t) \geq P_{min,PCR} \quad (43)$$

$$P_{SCR,pos,HT}(t) \geq P_{min,SCR,pos,HT} \quad (44)$$

$$P_{SCR,pos,NT}(t) \geq P_{min,SCR,pos,NT} \quad (45)$$

$$P_{SCR,neg,HT}(t) \geq P_{min,SCR,neg,HT} \quad (46)$$

$$P_{SCR,neg,NT}(t) \geq P_{min,SCR,neg,NT} \quad (47)$$

The result of the control reserve maximization is a scheduling concept to provide system services and enables the transmission system operator to have stable and effective control. Furthermore, if the VPP is located in lower voltage levels, this concept supports the appropriate connection system operator with defined schedule information concerning possible control reserve activation for distribution system operator. This is a necessary information during the day-ahead planning process in case of expected congestions in the distribution system.

4. Case Study

The case study demonstrates the application of the presented concepts under consideration of a test network. The impact of the operational concepts on the power system are highlighted according to the case study. The test nework in Figure 6 is based on the CIGRE Benchmark System. The used medium voltage network represents areas of central city, suburb and country grid according to [36 – 39]. Furthermore, there is only a single hand-over point to the overlay 110 kV network. Renewable generation from PV, wind and bio mass are collected in the voltage levels HV/MV, MV, MV/LV and LV in this test network.

Based on the investigation in [40] a defined amount of load (9815 kW) and installed renewable generation were randomly distributed. The main amount of RES is installed in country and suburb area and the main installed load is in the central city. Load profiles are modelled with standardized load profiles for households, business and agriculture. The installed renewable power for the VPP operation is considered to be 339.25 kW bio generation, 2126.56 kW wind generation and 7279.39 kW photovoltaic (pv, mainly in low voltage). The case study represents a scenario with high wind and photovoltaic infeed time-series for one day in 15 minute values and the operational starting point for bio power plants were expected to be 50 % of nominal power output. Figure 5 shows the uncontrolled renewable energy infeed forecast in the VPP area.

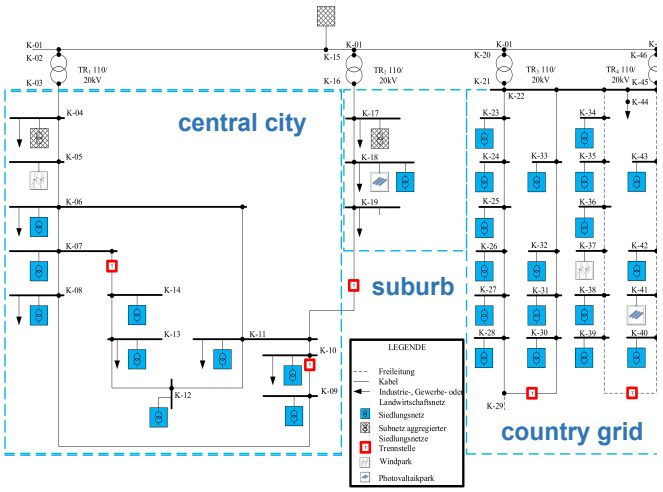


Figure 6. Medium voltage test network to [36 – 39]

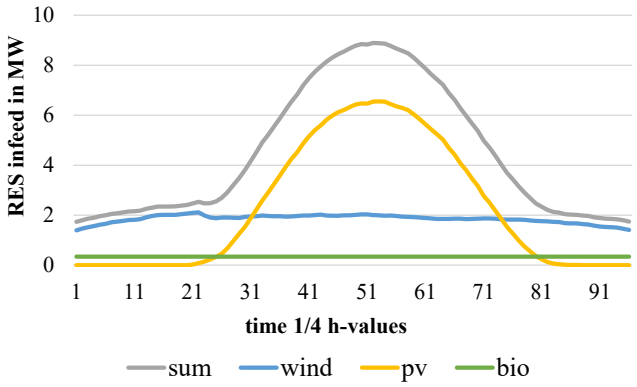


Figure 5. VPP generation day-ahead forecast

In the next step, the impact of the optimization is presented for the day-ahead planning process, with change of generation unit schedule in comparison to the uncontrolled renewable energy infeed (forecast). The test network with the implementation of load and generation as well as the according profiles offers the possibility to highlight the qualitative impact of the operational concepts on the 110 kV hand-over point exemplary. The impacts demonstrate the need of schedules and the impact of centralized controlled RES in low voltage levels. It is assumed that minimum bid and increment for control reserve is below 0.1 MW, because the case study only considers a small part of a network. The VPP itself can operate with more installed capacity of RES that is distributed in a bigger area and fulfill CR market requirements.

Furthermore, the technical oriented concepts will be exemplary represented by negative SCR contribution and negative redispatch contribution. Generation unit schedule is close to zero in case of maximum positive power provision for SCR and RD.

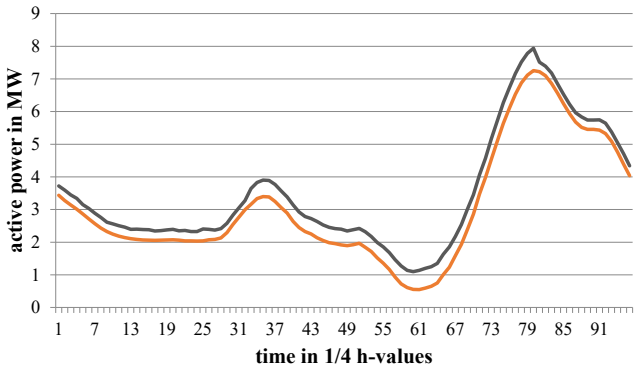


Figure 7. Active power demand of the medium voltage network with (grey) and without (orange) profit maximization

4.1. Economical oriented operation

The results of profit maximization are based on average prices for Day-Ahead market (blue) and Intra-Day market (green) as well as for CR participation (Figure 7). The revenue for PCR participation was assumed with 3646 €/MW and for SCR: 357 €/MW (pos. HT), 605 €/MW (pos. LT), 122 €/MW (neg. HT), 318 €/MW (neg. LT).

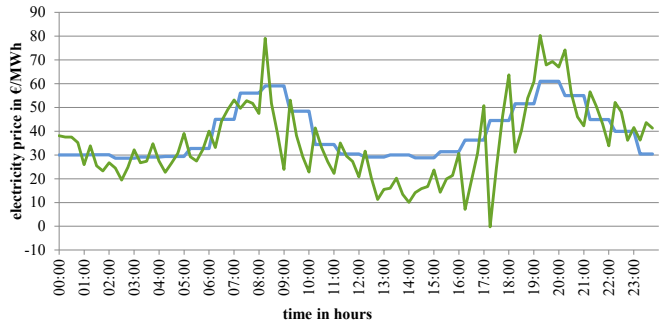


Figure 8. VPP generation unit scheduling for profit maximization

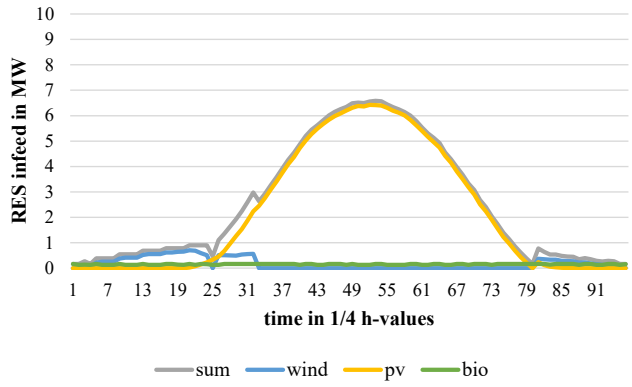


Figure 9. VPP generation unit scheduling for profit maximization

The overall expected profit for average spot-market prices of CR-market, Day-Ahead and Intra-Day market is about 1443.24 €. Most of the energy is planned to be traded on the Intra-Day market. Only less power is provided for positive SCR in low tariff with 1.563 MW and 2.297 MW in high tariff. The optimized generation unit schedule is reduced according to SCR provision (Figure 9).

Figure 7 demonstrate the impact of the profit maximization at the 110 kV hand-over point before profit maximization (orange) and after (grey). Due to the CR provision, the demand in this network area increases. It can be stated that the power demand of the network section is strongly reduced during the middle of the day, caused by a high PV infeed.

4.2. Technical oriented operation

The schedule loyalty is one of the basic needs for future system operation with high percentage of RES. The starting point for this concept is close to profit maximization, but the capacity for flexibilities has to be defined. Therefore, the day-ahead planning process is optimizing with only 95 % of forecast and a maximum of 90 % for bio generation. Therefore, the day-ahead schedule is set according to Figure 10. Furthermore, a random failure of maximum 5 % is assumed in each time step for wind and PV forecast. The Day-Ahead schedule expects a profit of about 1337.21 €, which lower compared to profit maximization. Most of the power is planned to be traded again in the Intra-Day market, but less power is provided for positive SCR in low tariff with 1.493 MW and 2.191 MW in high tariff.

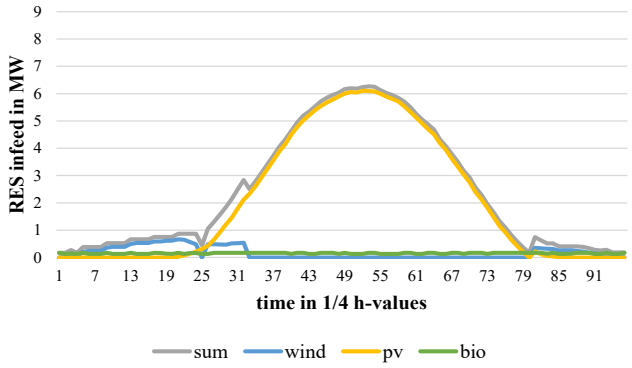


Figure 10. VPP day-ahead generation unit scheduling for schedule loyalty optimization

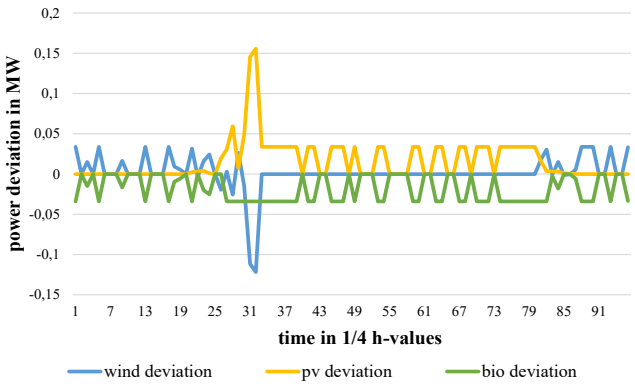


Figure 11. VPP intra-day deviation of generation in comparison to day-ahead schedule

The planned schedule can be fullfilled during the intra-day operation without any CR constraint violation and without any balancing power consumption. The intraday difference in generation unit operation is shown in Figure 11. The results demonstrate the functionality of this optimization concept. Furthermore, it was shown that the reduction of profit is relatively small in comparison to profit maximization, but the system benefit is enormous.

Eventually, the results for negative secondary control reserve and negative redispatch maximization are presented exemplary for all concepts of system service maximization. Both concepts represent the objective to maximize the power in within the planning horizon to provide maximum power that can be reduced by negative RD or negative SCR activation. Though results in generation unit schedule are the same, the reason of use is different. Figure 12 presents the day-ahead generation unit schedule. By considering the sum of generation, power output is maximized with repect to planned negative SCR and RD. In case of call of power reduction (RD) or negative SCR the maximum power is reduced. In particular at time step 25, there is a small reduction of wind caused by local weather conditions. It has a high impact on the maximization because PV infeed is low in the morning and the share of bio generation is small

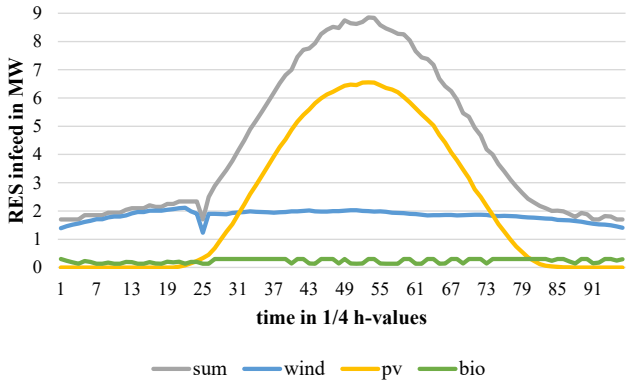


Figure 12. VPP generation unit schedule for maximization of negative SCR and negative RD

One main challenge and requirement for these concepts is the embodiment of incentives to support the system with these functionalities. On the other hand, is the concept of intra-day schedule loyalty. Schedule loyality is one of the best circumstances for system operatores point of view in terms of high voltilty during system operation. The concept of intra-day schedule loyalty needs only small incentives to be integrated in the german power system and leads only to small losses in profit. This concept can optimally be applied with controllable, forecast independent, generation or storage units. Otherwise, with a high amount of pv and wind it is necessary to make a day-ahead schedule with reduced forecast. But in the end this concept is main requirement for future system with high amount of RES and less conventionals to keep the system stable.

Figure 13 illustrate the change of power demand of the network at 110 kV hand-over point in case of call of negative SCR or negative RD (orange). The demand of active power before activation (grey) indicates a low load in comparison to the installed RES with times of power infeed in the 110 kV network. The activation of SCR or RD can almost double the demand during the middle of the day, depending on the high PV infeed.

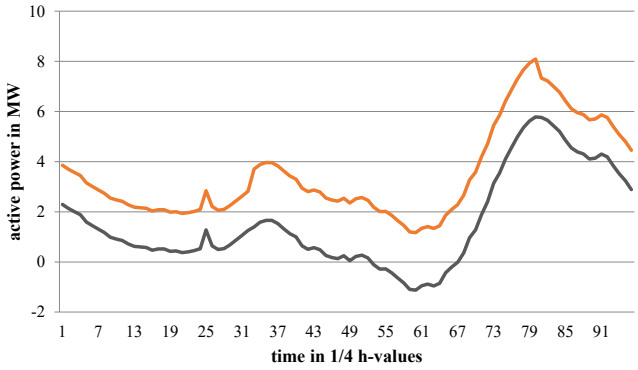


Figure 13. VPP intra-day deviation of generation in comparison to day-ahead schedule

5. Summary

The ongoing penetration of RES, especially in lower voltage levels, lead to a dramatically lack of needed system operation control funcitionalities as well as to enourmous increase of volatile generation. Therefore, this paper adresses their synergetic use in VPPs and their operational concepts. One basic advantage is the available schedule information for the system operator due to the centralized control for a defined network area. This can immensely optimize the DSO's and TSO's planning and operation process and enable generation process interventions (active and reactive power control) by involving the VPP control center, without having expensiv direct communication excess to each small unit. On top of that, it has been shown that the technical preconditions for the implementation of VPP in Germany are provided. Several suitable communication standards that meet the requirements of an intelligent power network and its decentralized generation units are already in use.Furthermore, the profit maximization concept indicates the generation of profit under currently given market conditions and without any renewable remuneration in the case study. But this result is without regarding costs for the operation of the VPP and this concept requires a sensitivity analysis to identify the minimum optimal number and kinds of generatin units to generate profit. However, there are many technical flexibilities like active and reactive power adjustments of decentralized generation units as well as storage systems, which can partly solve the uncertainty problem of demand and volatile generation forecast in particular as well as the contribution of system services. For this purpose, the concepts of intra-day schedule loyalty, control reserve maximization and redispatch maximization have been presented. The different concepts can be used in order to improve a stable and reliable system operation today and in the future. The presented technical concepts of CR and RD maximization algorithms are based on the profit maximization extended by fictious profits. This leads to the advantage that the algorithms enable the definition of needed monetary incentives to fulfill technical oriented operational concepts by optimal identification of fictious prices under changing market conditions. Furthermore, the case study demonstrates the importance of the information about location of generation units and day-ahead scheduling for CR and RD

maximization, where the call of system services can dramatically change the power demand at the hand-over point.

Eventually, the economical aspect is the most important one, especially from a VPP operator's point of view. As long as there is no sufficient level of monetary incentives, a reasonable support for the power system will not be provided. Under the given market conditions and the validity of the German Renewable Energy Act the profit for a single generation unit is much higher as for a VPP system compound, especially with regard to installation costs for a communication system and control center. Nevertheless, the future market will be dominated by renewable energy sources, so that the VPP concept will become more and more important for maintaining and improving a reliable and secure power system operation.

Limitations of the presented methods might be the computational time in case of increasing the number of constraints and by using MATLAB as the solver. For example, by defining individual technical and economical constraints for each generation unit in a VPP, the optimization complexity will increase dramatically, so that a more professional mathematical solver (e.g. Gurobi optimizer) is needed. Furthermore, the presented methods are day-ahead planning methods for a complete schedule over 96 time steps without any interruption in between two time steps. A further challenge can be to transfer the concepts of CR and RD into intra-day operation by taking system states of network frequency, line congestions and voltage stability into account [41], under consideration to calculate a new optimal solution for each next time step. By taking this aspects into account, the mathematical problem will rise strongly depending on the size of the network, size of VPP and the assumptions made.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper has been worked out as part of the project "REGEES" funded by the German Federal Ministry for Economic Affairs and Energy. The authors are grateful for funding their research.

References

[1] A. Richter; N. Moskalenko, I. Hauer, T. Schröter and M. Wolter, "Technical integration of virtual power plants into German system operation", 14th International Conference on the European Energy Market (EEM): 6-9 June 2017, Dresden, Germany, DOI: 10.1109/EEM.2017.7981876

[2] Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit, Nationale Klimapolitik, http://www.bmub.bund.de/themen/klima-energie/klimaschutz/ nationale-klimapolitik, 14.09.2015

[3] Statistisches Bundesamt, Bruttostromerzeugung 2016: 30 % stammten aus erneuerbaren Energien, 2017

[4] Informationsplattform der Deutschen Übertragungsnetzbetreiber, EEG-Anlagenstammdaten, Stand 31.12.2013

[5] Bundesministerium für Wirtschaft und Energie, Bausteine für die Energiewende: 7Eckpunkte für das 'Verordnungspaket Intelligente Netze', 2014

[6] PwC, Virtuelle Kraftwerke als wirkungsvolles Instrument für die Energiewende, PricewaterhouseCoopers AG (pwc), 2012

[7] B. Droste-Franke et al., Brennstoffzellen und Virtuelle Kraftwerke – Energie-, umwelt- und technologiepolitische Aspekte einer effizienten Hausenergieversorgung, Springer-Verlag Berlin Heidelerg, 2009

[8] Energietechnische Gessellschaft im VDE (ETG), Smart Distribution 2020 – Virtuelle Kraftwerke in Verteilungsnetzen, VDE, Frankfurt a. M., 2008

[9] Next Kraftwerke GmbH, Praxisbeispiel Virtuelles Kraftwerk: Flexibilität aus erneuerbaren Energien und Industrieprozessen, 2014

[10] Deutsche Windguard, Kostensituation der Windenergie an Land in Deutschland, Varel, Dezember 2015

[11] Fraunhofer ISE, Aktuelle Fakten zur Photovoltaik in Deutschland, Freiburg, Fassung vom 22.04.2016

[12] Fraunhofer ISE, Stromgestehungskosten Erneuerbare Energien, Freiburg, November 2013

[13] B.M. Buchholz, C. Brunner, A. Naumann and A. Styczynski, "Applying IEC standards for communication and data management as the backbone of smart distribution", Power and Energy Society General Meeting, 2012 IEEE, 10.1109/PESGM.2012.6345011

[14] CEN/CENELEC/ETSI Joint Working Group, Standards for Smart Grids, Final Report, 2011

[15] Dilan Sahin, et al., Smart Grid Technologies: Communication 1st ELECON Workshop 41 Technologies and Standards Transactions on Industrial Informatics, vol. 7, no. 4, pp. 529-540, November 2011.

[16] International Electrtechnical Commission (IEC), Electropedia: the World's Online Electrotechnical Vocabulary, http://www.electropedia.org/

[17] BDEW (Bundesverband der Energie und Wasserwirtschaft) and ZVEI (Zentralverband Elektrotechnik- und Elektroindustrie e.V.), Smart Grids in Deutschland – Handlungsfelder für Verteilnetzbetreiber auf dem Weg zu intelligenten Netzen, 2012

[18] M. Richter, "PMU-basierte Zustandsabschätzung in Smart Distribution", Dissertation, Otto-von-Guericke-Universität Magdeburg, Germany, 2016

[19] N. Moskalenko, "Optimal Dynamic Energy Management Systems in Smart homes", MaFo, vol. 59, ISBN 978-3-944722-16-0, Magdeburg, 2014

[20] A. Naumann, "Leitwarte im Smart Grid", MaFo, vol. 47, ISBN 978-3-940961-81-5, Magdeburg, 2012

[21] Wen-Yeau Chang, "A Literature Review of Wind Forecasting Methods", Journal of Power and Energy Engineering, 2013

[22] C. Skittides and W-G. Früh, "Wind forecasting using Principal Component Analysis", Journal Elsevier, Edinburgh, 2014

[23] Aoife M. Foley et al., "Current methods and advances in forecasting of wind power generation", Journal Elsevier, 2011

[24] National Renewable Energy Laboratory (NREL), Solar and Wind Forecasting, http://www.nrel.gov/electricity/transmission/resource_forecasting.html, 2015

[25] International Energy Agency (IEA), Photovoltaic and Solar Forecasting: State of the Art, Report IEA PVPS, Canada, 2013

[26] Next Kraftwerke GmbH: Strommarkt, [Online], Available: https://www.next-kraftwerke.de/wissen/strommarkt.

[27] Tariq Kamal, Syed Zulqadar Hassan, Muhammad Hussnain Riaz, Hui Li, Muhammad Sarmad and Gussan Maaz Mufti, "Design and Control of Photovoltaic/Microturbine/Super-Capacitor based Microgrid System", 20th IEEE International Multitopic Conference 2017 (INMIC 2017) at National University of Computer and Emerging Sciences (FAST-NUCES) Lahore campus, Pakistan from November 24 – 26, 2017.

[28] European Energy Exchange AG (EEX), EEX Produktbroschüre Strom, 2012. [Online]. Available: http://eex.com/blob/66450/6aee0902c1f341968d1f1b948f3b1a5b/konzept-strom-release-03a-deutsch-pdf-data.pdf.

[29] P. Konstantin, Praxisbuch Energiewirtschaft, 3 Hrsg., Berlin, Heidelberg: Springer Vieweg, 2013

[30] EPEX SPOT SE. Price List: Valid as of 1st January 2016. https://www.epexspot.com/document/34180/EPEX%20SPOT_Price%20List%20January%202016.pdf

[31] 4 German TSO, Market for control reserve in Germany, [Online] Available: https://www.regelleistung.net/ext/static/market-information

[32] 4 German TSO, About control reserve and Tender details, [Online] Available: https://www.regelleistung.net/ext/

[33] Bundesnetzagentur, Bundesnetzagentur verbessert die Bedingungen zur Teilnahme an den Regelenergiemärkten Strom, Press announcement, 28.06.2017

[34] W. Domschke, Einführung ins Operation Research. 9. Auflage, Berlin: Springer-Verlag, 2015

[35] M. Tröschel, "Aktive Einsatzplanung in holonischen virtuellen Kraftwerken", Dissertation, 2010

[36] CIGRE Task Force C6.04, Benchmark Systems for Network Integration of Renewable and Distributed Energy Resources, ISBN: 978-285-873-270-8, 2013

[37] Netzentwicklungsplan Strom 2025, Version 2015: Zweiterer Entwurf der Übertragungsnetzbetreiber, Feb. 2016, [Online]. Available: http://www.netzentwicklungsplan.de/_NEP_file_transfer/NEP_2025_2_Entwurf_Teil1.pdf

[38] Thüringer Energie- und GreenTech-Agentur, Studie zu lokalen Energiespeicherbedarfen: Analyse und Prognose der Technologien und Anwendungsfelder thermischer und elektrischer Energiespeicher auf Nieder- und Mittelspannungsebene, Juni 2014, Erfurt
[39] BMWi, Studie: Moderne Verteilernetze für Deutschland, Abschlussbericht, September 2014
[40] A. Richter, Chr. Ziegler, N. Moskalenko and M. Wolter, Das Virtuelle Kraftwerk als gewinnorientierter Akteur mit verteilnetzunterstützendem Potential, im Tagungsband 4. Konferenz Zukünftige Stromnetze für Erneuerbare Energien, Berlin, Germany, 2017
[41] D.P. Kothari and J. S. Dhillon, Power System Optimization, Prentice-Hall of India, 2004, 2nd Edition, 2011

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 136-144 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Distributing the computation in combinatorial optimization experiments over the cloud

Mario Brcic*, Nikica Hlupic, Nenad Katanic

University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, 10000, Croatia

ARTICLEINFO

Article history:
Received: 26 October, 2017
Accepted: 12 November, 2017
Online: 10 December, 2017

Keywords:
Combinatorial optimization
Computational experiments
Cloud computing

ABSTRACT

Combinatorial optimization is an area of great importance since many of the real-world problems have discrete parameters which are part of the objective function to be optimized. Development of combinatorial optimization algorithms is guided by the empirical study of the candidate ideas and their performance over a wide range of settings or scenarios to infer general conclusions. Number of scenarios can be overwhelming, especially when modeling uncertainty in some of the problem's parameters. Since the process is also iterative and many ideas and hypotheses may be tested, execution time of each experiment has an important role in the efficiency and successfulness. Structure of such experiments allows for significant execution time improvement by distributing the computation. We focus on the cloud computing as a cost-efficient solution in these circumstances. In this paper we present a system for validating and comparing stochastic combinatorial optimization algorithms. The system also deals with selection of the optimal settings for computational nodes and number of nodes in terms of performance-cost tradeoff. We present applications of the system on a new class of project scheduling problem. We show that we can optimize the selection over cloud service providers as one of the settings and, according to the model, it resulted in a substantial cost-savings while meeting the deadline.

1. Introduction

This paper is an extension of work originally presented in conference MIPRO 2017 [1].

Combinatorial optimization (CO) is a research field with many important real-world applications. Scheduling [2], auctions [3], and vehicle routing [4] are just a few notable examples. Combinatorial optimization is a subfield of mathematical optimization. It deals with problems where optimal selection needs to be done from a discrete feasible set. Exhaustive search evaluates all possible solutions before selecting the best one which is infeasible for realistic problem sizes. There are special classes of CO problems that can be solved with polynomial-time algorithms such as shortest paths, flows, spanning trees, and matching. However, many interesting problems are NP-complete and for these problems, unless P=NP, there are no computationally efficient solving algorithms. For such problems, different search or metaheuristic algorithms are created in order to get as good as possible performance in a realistic amount of time. Design of such algorithms is an intrinsically empirical process, guided by the experiments while the ranking of different design choices, hyperparameter values and algorithms depends on the results from experimental runs, often performed on benchmark test sets.

Each experiment consists of experiment units which denote the smallest indivisible executable unit. Experiment units are in the focus of this paper as they tend to be independent during the execution, which enables a high degree of parallelization. In deterministic CO problems all the parameters are deterministic. For them, each sampled instance of CO problem comprises an experimental unit. In stochastic and robust CO problems, some of the parameters are uncertain or unknown. In that case, each combination of sampled problem instance with its sampled parameter scenarios makes one experimental unit. For that reason, the number of experimental units in stochastic and robust problems can grow exponentially in the number of uncertain parameters. Additionally, the aforementioned set of experimental units is increased in Cartesian product with other experimental factors, as shown in Figure 1. Such factors include hyperparameter values, used algorithm, and algorithm design choices.

*Corresponding Author: Mario Brcic, Unska 3, Zagreb 10000, Croatia, +38516129951, Email: mario.brcic@fer.hr

https://dx.doi.org/10.25046/aj020617

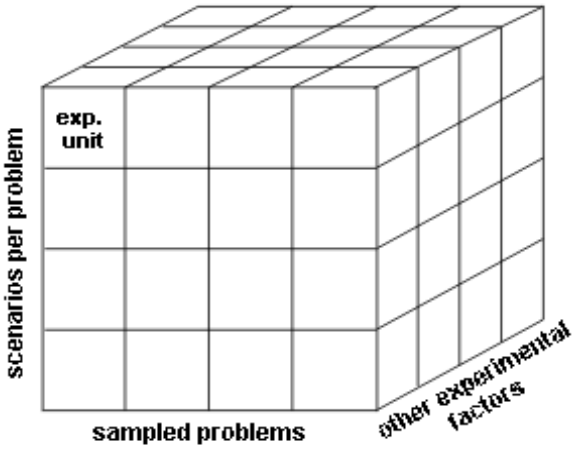


Figure 1. Exponential growth of number of experimental units with added factors which constitute independent dimensions

The experiment requires execution of all experimental units and empirical data must be recorded over those runs. This can be expensive operation, depending on the number of experimental units and the complexity of the underlying problem in each unit. However, as already mentioned, experimental units tend to be computationally independent and the process of experiment execution is inherently parallelizable over the units. In the current state-of-the-art combinatorial optimization, experimentation is mostly done on single computer. This was the case in stochastic variant of vehicle routing problem in [5] and project scheduling in [6]. There were applications where the problem was of such complexity that computer cluster had to be employed in order to make the algorithm practically usable. This was the case for fleet optimization in [7].

In this paper we present an architecture of a system for distributing extensive computational experiments over the cloud that takes advantage of the independence between experimental units to achieve inexpensive scalability in the cloud. This architecture can be implemented as a standalone system or it can use some of the frameworks and systems described in section 2 if the amount of reuse and alignment of the blueprint with the implemented functionalities in those systems can be achieved. The architecture was designed in a way to reduce the amount of communication to a minimum level while still enabling efficient load balancing. We explain the distributed design of data storage for the experimental results. The distributed design reduces the communication overhead which is a problematic aspect for high-performance computing (HPC) in the cloud. Storing rich experimental data is important for further analysis and also important as a mean of scientific scrutiny by enabling efficient sampling-based reproducibility. An optimization model is presented that describes the best choice of settings for running specific experiments. The model uses data from pilot runs which execute sampled subset of experiment units. Finally, the proposed architecture was applied on a specific problem of developing and validating algorithms for a new type of complex stochastic combinatorial optimization problem in project scheduling.

The paper is organized as follows: in section 2 we present the idea of high-performance computing in the cloud as an alternative to in-house grids. Section 3 lists the related work. Section 4 presents the general architecture of our system for distributed experimental runs. In section 5, a distributed design of the storage is described. Applications of the architecture on a real research problem are described in section 6. Section 7 drafts future research ideas and section 8 offers concluding remarks.

2. High-performance computing in the cloud

Distributed computing is unavoidable in high-performance computing (HPC) where job is divided between many available processors in order to significantly reduce the runtime with currently available hardware. Traditionally, dedicated in-house grids (super-computers) are used. They are difficult to setup, maintain and operate [8]. In this paper we shall deal with cloud computing as a flexible and cheap resource alternative that can be rented on demand instead of being owned the whole time – for a lower overall cost and a relatively small performance penalty. Studies have been conducted on the matter of using the cloud for high performance computing. In [9] and [10] authors concluded that there is a limit on a number of used computational nodes where coupled applications are competitive with in-house grids. Beyond that limit, overheads become overwhelming performance detractors. In [11] virtualization, latency and system noise are singled out as the biggest issues in comparison to dedicated supercomputers. They found that research groups with limited access to supercomputer resources and with varying demand might find cloud computing to be a beneficial choice. Latency is problematic due to used commodity equipment in most of the cloud infrastructure and network virtualization. Virtualization introduces performance penalties through network virtualization and other virtualization overheads while accessing physical resources. In [11], network virtualization was found to be the primary bottleneck of the cloud that increases latency, reduces bandwidth and interferes with processes. System noise affects the performance due to multi-tenancy which introduces resource sharing with virtual machines deployed on the same physical hardware. Service providers can manipulate the degree of multi-tenancy, which enables greater profit by overallocating resources to the users. The problems with HPC in the cloud were reiterated in [12] where authors have put the focus on necessary reductions in communication overhead and virtualization. For the former, they propose the implementation of better load balancing, and using bare-metal containers for the latter. Comparative study in [13] confirmed raw performance superiority of in-house grids to Amazon's Elastic Compute Cloud (EC2) cluster. However, waiting time in queue on HPC clusters plays a significant role when taking turnaround time into account . In such circumstances EC2 cluster could produce better turnaround times. The cost-effectiveness of running HPC application was observed as dependable on raw performance and scalability. Cloud computing enables utilizing available monetary resources to rent practically as many as possible identical processing instances. This identicality of processing instances is desirable in running experiments as it sets all runs in the identical environment. This keeps most of the variance in measurements related to designed experimental factors. The effect of system noise on experiment results can be reduced using effective randomization in the job balancing.

3. Related work

In the last decade, with advent of Big Data, usage of cloud computing became all-pervasive. Related to scientific

applications, authors in [14] presented a reproducible genome sequencing task that was run in the cloud for a small cost with a near linear scalability. In [15], a new artificial intelligence algorithm for complex control tasks has been created. It features linear speedups with over a thousand workers on a public cloud service to cut down on total execution time in comparison to other algorithms. Authors in [16] have created distributed architecture for deep neural networks in the cloud for sensor fusion and inference based on the data from multitude of end devices. The communication cost was reduced by a factor of over 20x compared to the alternative. All of the aforementioned applications intentionally achieved low communication requirements between subtasks, hence avoiding or mitigating problems related to the communication overhead in cloud computing. Distributed execution engines have been created to simplify parallelization of computationally intensive tasks. Ray [17] is a python-based example of such system which enables computations.

Existing tools that provide support in design and comparison of optimization algorithms are listed in this paragraph. Comparing Continuous Optimizers (COCO) [18] is a platform for continuous optimization, hence it does not cover combinatorial optimization. Nevertheless, it has many of the features needed in a tool for our needs. It has a library of standard benchmark problems on which optimizers can be compared. Experiments can utilize Shared Memory Parallelism (SMP), but grid computing is not utilized. For that reason, big-scale execution in the cloud is not standard feature of that platform. ParadisEO [19] is a white-box C++ framework for reusable design of parallel and distributed metaheuristics. It has features and components helpful for creating new algorithms. The intent of this framework is to simplify the design of topologies within a single running system, i.e. a single optimizer that can be distributed. However, it does not specify the efficient way of executing distributed experiments. Java Evolutionary Computation Toolkit (ECJ) [20] is an option similar to ParadisEO. Multi-Objective Evolutionary Algorithm (MOEA) [21] Java-based Framework deals with multi-objective optimization by combining the features of COCO, ParadisEO, and ECJ. We have pointed out that COCO covers only continuous optimizers, while other tools enable easier and faster algorithm design. The latter is achieved through reusability of common algorithmic components when the optimization problem and algorithm design have favorable features. These tools do not specify guidelines for distributing extensive computational experiments over the cloud.

Performance of executing the experiment in the cloud is an important issue. Predictions can be used in scheduling as well as in finding optimal settings of computational nodes. The focus of [22] is on comparing public cloud providers using measurements on specific applications. These measurements can inform the processes of provider selection and performance prediction. Performance prediction using machine learning for improving the quality of system management decisions has been investigated in [23]. Authors in [24] used machine learning to predict the execution time of computational fluid dynamics applications in the cloud. These predictions were used in scheduling algorithms. A system for efficient performance prediction for large-scale analytics on EC2 cloud has been created in [25]. It utilizes optimal experimental design in order to minimize the resources in building the model. The system is used to find the optimal configuration in number of instances. Our architecture uses simple statistical model for performance prediction in order to calculate cost-optimal instance-type and number of necessary instances in order to satisfy desired probabilistic level of satisfying the deadline. In our case, settings also include the cloud provider, hence combining the intents of aforementioned works: performance prediction and optimization of node-selection that takes into account the cloud provider as well.

4. Architecture

The intended use-case is inherently parallelizable task. This architecture is designed in a way to use those favorable features of the task in order to achieve low communication between the computational nodes. In that way, latency is not an issue and there is only occasional communication where bandwidth plays the main role. Communication between the nodes is necessary for creating the computational nodes, sending instructions for job chunks (that is, batches) to them, and during migration of the final results. Instructions for job chunks contain small amount of information. Chunks are sized in a way to keep the nodes occupied for some time. There is a tradeoff between achieving good job balancing and reducing communication overheads in relation to the amount of computation done on the node. Computational nodes keep all generated and logged experimental data locally in their part of distributed database. The results, raw or processed, can be pooled periodically, upon finalization of the assigned experiment chunk or at the end of node's part in the experiment. Aforementioned features make the problem suitable for cloud deployment as we can avoid communication-related detrimental effects on performance.

Our pipeline architecture includes five general stages in the experimental process:

1. Creation of experimental data, that is the data about experiment units in initial seed database.

2. Creation of node images populated with all the necessary data and optimized code for executing the experiment units, logging and storing the results.

3. Optimizing the settings of execution environment. Pilot runs are used on a small subset of experiment units for different settings of the nodes, including a service provider to inform the optimization procedure.

4. Executing the experiment on pre-calculated number of identical nodes with settings selected in the previous stage.

5. Collecting experimental data from computational nodes, conducting analysis and getting the results.

In the rest of this section, we shall describe each of the stages in more details. All steps are enumerated according to their corresponding figures.

4.1. Creation of the experimental data

The data needed for running an experiment is created in the first phase, which is depicted in Figure 2. This means creating all the data that sufficiently describes experiment units so that they can be created and executed.

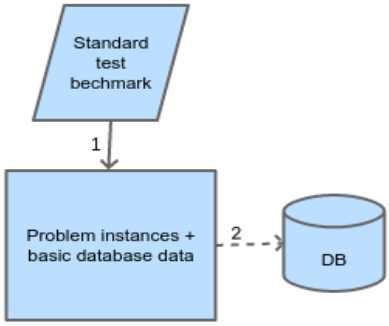


Figure 2. Preparing the seed database

The first step in this stage is optional. It uses standard test set as a source for creating problem instances. For many well-researched CO problems, such sets are shared within the community of the researchers to enable consistent ranking of algorithms. In case of new CO problems, maybe a standard test benchmark related to the problem can be found and used for creation of problem instances. Also, problem-instance generators are available for some problems and custom test set can be created. In this step, sampling and transformations of problem instances can be done to create our own set of problem instances that are going into an experiment.

Actually used test set is created at this point. The "seed" database is created and populated with the metadata necessary for experiment execution. Also, it is filled with all the data about experiment units needed for execution and result logging. Part of the data can be in the form of external files if that is more appropriate, but they must be linked from a database. In order to enable distributed execution, a database (and external data) is distributed over the nodes in such a way that each node has its independent, unsynchronized version of a database. Each such database is initiated from a singular seed database. Horizontal fragmentation of writeable relations is employed as a mean of data distribution over the nodes. Details of a database are given in section 5.

4.2. Creation of node images

One or more virtual machine (VM) images need to be created at this stage (shown in Figure 3). The exact number depends on the requirements in a phase of pilot runs. Each image contains a copy of a seed database and all the necessary code for running the experiment chunks on the node. Executable code is tuned to the intended hardware. After that, image is migrated to the cloud service from where it can be easily deployed for creation of the computational nodes.

4.3. Optimizing the execution environment

It is hard to know exactly in advance what settings of the computational node are efficient. For that reason we need an optimization phase of an experiment with regards to the settings of the experiment run. This is a stochastic combinatorial optimization problem as well and it can be stated as a problem of minimizing the monetary cost of renting cloud instances under a constraint of

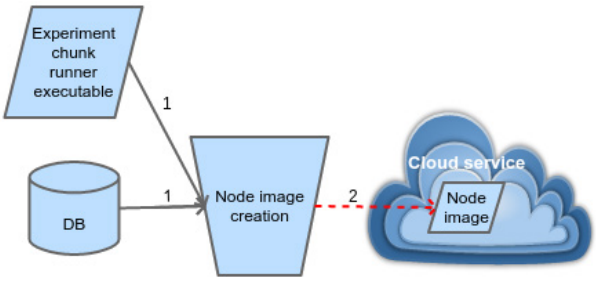


Figure 3. Creating the image for virtual machines in computational nodes

achieving desired probability of finishing before the selected deadline. The problem can be formulated and expressed as:

$$\min \mathbf{E} \sum_{x \in S} C_x(n_x) \quad (1)$$

with the constraints:

$$\sum_{x \in S} \delta_x = 1, \quad (2)$$

$$\mathbf{P}\{T_x(n_x) \leq D\} \geq \delta_x \cdot p, \ \forall x \in S, \quad (3)$$

$$\delta_x \in \{0,1\}, \ \forall x \in S, \quad (4)$$

$$n_x \in \mathbf{N}_0, \ \forall x \in S. \quad (5)$$

In the above formulation **E** is the expectation operator and **P** is a probability measure of the set. The above problem (1)-(5) has several parameters with values known prior to optimization: D is a desired deadline, S is a set of node-type options, and p is a desired probability of achieving the deadline. Also, C_x is a random cost of running the experiment on n_x instances of node-type x. The cost is random as it depends on the utilization durations of n_x nodes. T_x is a random experiment finish time when running it on n_x instances of node-type x. Decision variables are n_x and δ_x where n_x represents a number of instances of type x (5), and δ_x represents exclusive choice between the node-types (4). Randomness in this problem originates from the noise in execution due to hardware reasons and uncertainty in computational requirements of each experiment unit. That randomness is reflected in C_x and T_x. These functions can be created using statistical analysis or machine learning on the data from pilot runs. The objective function (1) is the expected value of total cost of the experiment. It sums the costs across all possible node-types but only one of those costs is going to be non-zero. The first constraint (2) ensures that only one node-type is selected. A set of chance-constraints (3) ensures that for selected node-type x a number n_x of nodes is selected so as to achieve a desired level of safety. All zero-valued δ_x make these constraints trivially satisfiable for all non-selected node-types x. For the selected node-type, a constraint enforces that a probability of finishing before the deadline must not be lower than the prescribed p.

We propose to make a parameter search over a set of node settings by running a simple and small experiment (using only a subset of experiment units) on each setting in S. A set of options should be small in order to reduce the cost of doing the pilot runs. That set can be composed by a careful pre-selection based on available data from previous general analyses such as [22] and based on analysis of the code for experiment execution as in [26].

The settings with the most impact are the selected cloud-service provider and cloud-instance type. The latter can be further customized across some set of subsettings, but usually the types are predefined less flexibly. The most important subsettings pertain to hardware components: CPU, disk size, RAM, etc. Pilot runs are identical to executing the experiment described in the following subsection. The difference from the latter is in the scale; in pilot runs just one experimental node per different setting is used , while many experimental nodes with identical settings are used during the experimental run. Collected data are analyzed by a procedure described in subsection 4.5. Based on the collected data and optimization model, economically most efficient node setting is chosen for the full experiment. Also, for that choice we get an estimate of the total experiment run time and cost by taking the summary statistics of T_x and C_x. Based on a desired due date, we infer the necessary number of computational instances N. N is the value of only non-zero n_x in optimal solution of the the optimization model.

4.4. Executing the experiment

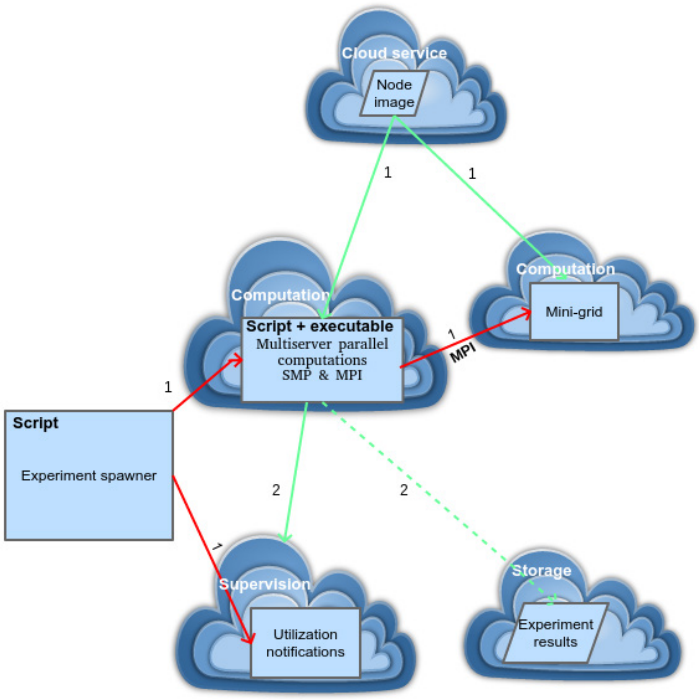


Figure 4. Execution of the experiment

The experiment is executed by distributing chunks across computational nodes (Figure 4). Experiment spawner is a script that balances the load between instantiated computational nodes. The experiment is partitioned into disjoint experimental chunks. Each chunk is a set of experiment units to be executed on a single computational instance.

1. Experiment spawner creates N cloud nodes and sends over network the parameters that define their workload chunks. These parameters constitute a small amount of information. Each experimental node runs a chunk runner which processes its assigned workload. Each runner, depending on the type of a problem, can instantiate additional computational nodes to form a mini-grid using Message Passing Interface (MPI). That is done if some of the algorithms require such execution architecture by design. SMP can be switched on by the parameters sent from the experiment spawner. Each node sets triggers for utilization alarms at the performance supervisor. Triggers improve the efficiency of a system by notifying the subscribers of different events. This information can be used for better, more prompt load balancing and it can reduce the renting costs.

2. At the end of processing a chunk, a computational node triggers supervisor's alarm when it can get another chunk to execute. When the node finishes with processing, it migrates its results to a permanent cloud storage and it gets terminated.

4.5. Result analysis

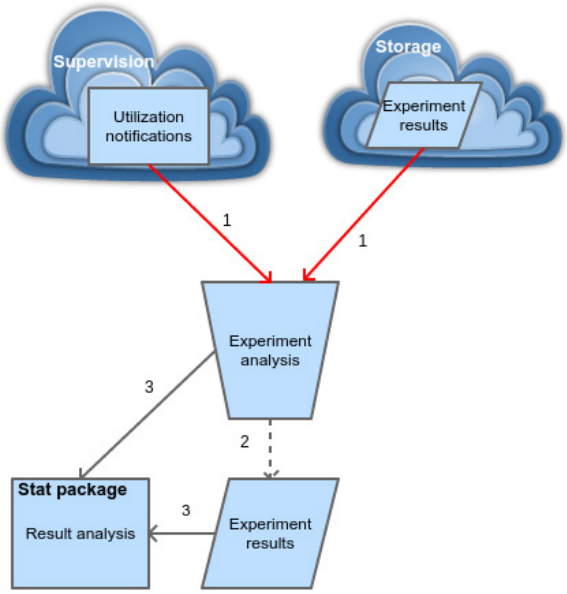


Figure 5. Analysis of experimental results

Users are notified by supervisor's alarm about the finalized jobs. A procedure is subscribed to notifications and reacts to them as shown in Figure 5. At first, the data is migrated from a computational node to some storage, locally or in the cloud. If the node has no more jobs to process, it is terminated to reduce costs. Analysis is done iteratively over the partial data as they pool to the storage. At the end of the process, with all the experimental results available, concluding results are created.

5. The database

The main objective of our system is to produce detailed experimental data as fast as possible. We have decided to store all the data into a database as it simplifies manipulations with overwhelming amounts of data.

It is assumed that running computational experiments is expensive both temporally and monetary. Storage, on the other hand, is much cheaper on both accounts. For that reason, as much data as possible should be stored for future analyses to avoid experiment re-runs.

Practicing scrutiny is important in science in order to prune mistakes and misconduct. Replication studies simply repeat experiments to check if the results match. In computer science, this

can be done by sampled replications that repeat small sampled set of experiment units. In each repetition, the same settings for pseudo-random number generators (PRNG), algorithms and problem instances should be used to create near-identical conditions as with the initial experiment. The results of replication study must match the stored values for everything except the execution time which includes incontrollable system noise. This is more efficient than replications of physical experiments where the exact conditions cannot be repeated and results should only statistically match. In the latter case, a greater number, if not all, of experimental units needs to be repeated in order to make conclusions. Also, executing experiment units in computational experiments is often cheaper than for physical experiments. The probability of non-matching results gets exponentially smaller with the size of randomly generated sample for replication.

5.1. Design

As described in section 2, communication overheads are serious performance detractors to scalability in the cloud in comparison to the in-house grids. Grids have tight interconnection and synchronization. For that reason, we have decided to minimize communication frequency between the cloud instances. All instances have their own database for storing results that springs from the initial seed database which is copied to all instances during creation of the node image as described in subsection 4.2. Seed database holds metadata and identification/replication data. The former defines all necessary structures to store the experiment data. The latter are the basic data necessary for identification of experiment units. Such data include the shared information for all instances of CO problems, used optimization algorithms, PRNG types and uncertainty scenarios in the case of stochastic or robust CO. Hence, all the data that define and describe experiment units (Figure 1) are present in a seed database.

Each node is created with a separate copy of a seed database. These copies make up a distributed database. The writeable relations, which record the experimental run data, are horizontally fragmented. Horizontal fragmentation keeps table schemas and distributes table rows across the nodes, as depicted in Figure 6.

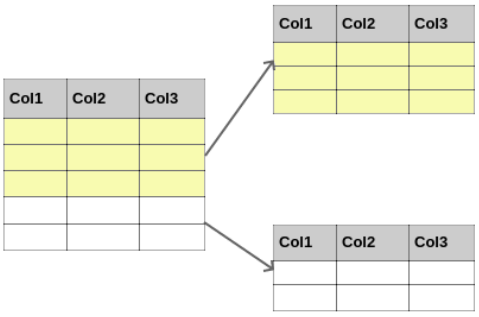


Figure 6. An illustrative example of horizontal fragmentation

During the experiment execution, each node saves two types of data: identification/replication data, and performance data. The former describe the settings of an executed experiment unit. This consists of the settings of the used optimization algorithm, PRNGs and other components that hold the key to replicating the execution of the experimental unit during replication studies. Performance data track decisions made during the execution and measurements of their effects. Quantitatively they make up the majority of generated and stored data, and the final conclusions are based on them.

Distributed data uses the identification scheme that combines together node-specific identification/replication information with shared information that originated from a seed database. That scheme makes data aggregation from all chunks unambiguous.

6. Application

The primary motivation for creating previously described architecture was a practical problem. As a result of different branch of research, we have worked with the new type of stochastic project scheduling problem, Cost-based Flexible Stochastic Resource Constrained Project Scheduling Problem (CBF-SRCPSP), defined in [27]. This problem deals with proactive-reactive project scheduling which makes the synchronization between project collaborators easier. It uses an agreed upon baseline schedule that stores the time-agreements for starting times of different project activities. Deviations of real start times from these agreements are penalized for inflexible, change-sensitive activities. The additional element in this model is that the changes to the baseline schedule can be profitable if they can be undertaken sufficiently far into the future. This extension to the features of baseline schedule makes the solution space much more complex due to the aforementioned proactive rescheduling operations. Searching the solution space becomes more time-consuming, though it pays off. We were lead empirically through the design and creation of optimization algorithms for CBF-SRCPSP.

The experimental set was created based on the standard Project Scheduling Problem Library (PSPLIB) set of instances for deterministic resource constrained project scheduling problems [28]. Cluster sampling was used to select sets of template instances from J30, J60, and J120 problem sets. The latter consist of projects that have 30, 60, and 120 project activities, respectively. The templates were expanded to fit CBF-SRCPSP model by modeling the stochasticity in activity durations with discretized beta distributions with a combination of selected and randomly generated distribution parameters. Additional parameters for each activity were randomly generated, according to the selected triangular distributions. After this procedure we ended up with 300 instances of each size which sprang up from template instances extended to fit the new model. The templates were instantiated according to the two experimental factors regarding to the project deadline: tightness of the deadline and bonus for early-finishing the project. For each instance, we generated 1000 activity duration scenarios from the discretized beta distributions. Activity durations were the only source of uncertainty. All of this data: project instances and uncertainty scenarios per each project were fed into a seed database. The used database was sqlite3 as it fit the needs of our experiment. Its simplicity trades-off well with its shortcomings in comparison to the more elaborate database management systems.

The seed database was also populated with the metadata about used PRNGs and CO algorithms. We have used two PRNGs: Mersenne twister with careful parameterization [29] and Threefry [30]. Several search algorithms were developed during the algorithm design. We had at our disposal implementations of optimization algorithms from the CO literature that were already

available in the used simulation library [26]. The final experiment included one benchmark CO algorithm from the literature and two selected newly-developed algorithms that were hypothesized to significantly outperform the benchmark. However, the analysis necessary for proving the hypothesis necessitated sufficient number of samples. The newly-developed algorithms were computationally expensive due to a search in more complex solution space than the benchmark.

The node images used Linux with gcc and necessary libraries. New algorithms were developed upon the C++ simulation library from [26]. This means that the most computationally intense parts, namely simulation-based experimentation and database logging were coded in C++ for better performance. The mini-grids were not used, since all the algorithms used only shared memory parallelization with two threads using OpenMP to speed up the search. This choice was based on the experiment in [26] regarding the parallel efficiency of the used simulation library. That also determined the number of computational cores per instance.

The total experiment was run on two occasions. The second experiment was done due to improvements to the developed algorithms and the results from the second study were used in the final experiment report. Here we shall cover both experiment runs as they utilized different experimental choices. In both cases, the raw results were downloaded locally using the python script. Statistical language R was used for results analysis. RSQLite R package was used to query the databases for the relevant data and to gather them together from all the sources.

6.1. The first experimental run

During the first experiment run we have selected Amazon Web Services (AWS) [31] as a service provider. Therefore, we did not use service provider as the experimental factor in pilot runs during the selection of the ideal instance type. We have used EC2 for computations, and have opted, after pilot runs, for c3.large instances with 2 dedicated physical cores of Intel Xeon E5-2680 v2 (Ivy Bridge) processors and 3.75 GiB of RAM. Simple Storage Service (S3) was used to store experimental results. The combination of CloudWatch (CW) and Simple Notification Service (SNS) was used for supervision and utilization notifications. Experiment spawner was coded in python, using the boto API [32] for accessing AWS and paramiko module [33] for controlling SSH2 connections.

Experiment consisted of 1.6 million experiment units. Experiment chunking was done across the problem instances. The chunk runner accepts parameters that describe chunk boundaries. These parameters are just several bytes in size and it is all the information needed for initiating the experiment execution on a computational instance. The total computational workload of the experiment was estimated based on running small sample of experiment units during the pilot run. It was estimated that the total workload is three and a half months of computational labor on the available hardware. The deadline was set to 7 days which we wanted to achieve with 90% probability. We calculated the necessary number of computational instances to satisfy this requirement. This resulted in using up to 26 cloud instances and reducing the total duration to 6 days. The database for each instance was migrated to S3 at the end of chunk execution and it awaited further analysis there. The cost of the first experiment was 445$.

6.2. The second experimental run

The second experiment was run on improved algorithms. The number of the most computationally expensive units was significantly increased, resulting in the total of 2.1 million experimental units. The optimization stage was used to select the node-type, also taking into account the service provider. We have used prior knowledge of the characteristics of our experiment runner - low working memory consumption (under 500MB) and high CPU utilization - to narrow down a range of instances. We tested the instances with 2 processors and as close as possible to 2GiB of memory. Market research was used to select the small set of service providers: Online Virtual Hosting (OVH), AWS (due to the use in the first experiment), and Linode. The selected instance-types are listed in Table 1.

Table 1. Members of alternative set *S* for optimization of execution environment

	OVH	**Amazon EC2**	**Linode**
instance-type	2vCores@ 2.4GHz 8GB RAM	c3.large 2vCores@ 2.8GHz 3.75GB RAM	2vCores@ 2.5GHz 4GB RAM
price	13.49$ (monthly)	0.105 $(hourly)	20$ (monthly)

The identical pilot run was used on all instance-types. Seven different types of experiment units were sampled into the pilot's unit set. The price of pilot runs in the optimization stage was 1.1$. The measurements were used to model execution durations for each combination of instance-type x and unit type u as Gaussian random variables $d_{x,u}$. Then, the total experiment duration for each instance-type is Gaussian random variable where w_u is the number of units of type u in our experiment. We approximated T_x by assuming that the total work modeled by d_x is simply equally divided among n_x computational instances.

It was estimated with the probability of 90% for the fastest option that the total workload is just above three years. The ranking and the necessary number of instances were calculated in order to satisfy the selected deadline of 30 days with the probability of 90%. The deadline was set to 30 days to take advantage of the monthly pricing. Figure 7 shows the optimal expected total costs for each option in Table 1. The results in Figure 7 were anonymized due to the legal concerns. Based on the available data, we have selected the C3 type instance as the most efficient with the estimated cost just below 500$. The experiment was run and it finished after 23 days of executing.

7. Future work

Possible future research ideas include improvement to the robustness of the total workload estimator. In our application, different simplifying assumptions were made and the error of their approximation effect should be investigated.

In order to reduce costs of the research community, general data about computational experiments (duration and prices) could be shared online with the public. This data can be useful for

creating promising and efficient alternative sets for optimizing the execution environment.

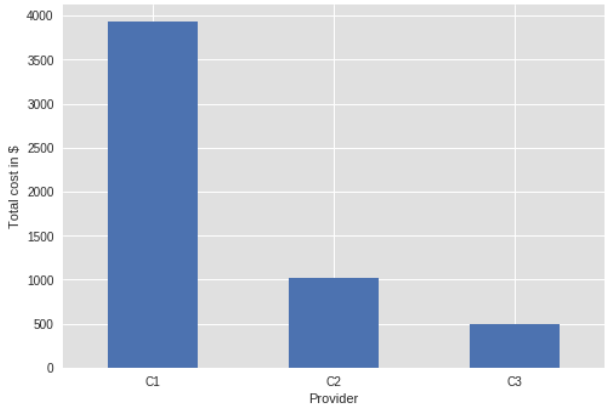


Figure 7 Estimates of total expected cost for anonymized permutation of options from Table 1

An automated tool for the composition of alternative sets could be utilized based on the experiment requirements. There is a potential in using pilot run on just one carefully selected instance type and then utilizing machine learning to find the best option and the number of necessary instances N for the desired target duration and the probability of achieving it. More advanced notions of risk and chance constraints can be used to account for the uncertainty in estimations and its economic effect. That can inform budget planning and the project management that undertakes the experiment.

Many parts of the architecture presented in section 4 are abstracted from the details of the particular experiments and can be reused in different settings. Component-based framework for general experimenting can be created. In that way, reuse of existing components can be increased and a code generator for repetitive parts can speed up the development, especially if the user does most of the manual process through an intuitive graphical user interface. Cloud costs can be reduced further by better node tracking and possibly using the cheaper spot instances for non-critical computations, especially during the algorithm design, prototyping and various pilot runs. The performance penalty in case of using such instances should be investigated in order to drive recommendations for configurations that utilize them.

8. Conclusion

Cloud computing is still not a simple and clear choice for high performance computing due to the issues pointed out in [11]: communication overhead, virtualization and system noise. The efficiency of cloud depends greatly on the specifics of the problem that we are trying to solve.

We have presented a system for distributing combinatorial optimization experiments over the cloud. Doing computational experiments for validation and guiding the design of CO algorithms has a specific property that it can be parallelized across experimental units that tend to be independent. This allows for low coupling between the simultaneous tasks and circumvents the issues related to the communication overhead.

Our system records rich data about the experimental runs in order to reduce the need for re-runs of experiments which may be computationally and monetarily expensive. In order to keep communication overhead to the minimum, distributed database with horizontal fragmentation was used. Each cloud node populates only the data related to its assigned disjoint experiment chunk. The unambiguity of the data across the system is, hence, kept without additional effort. When the tested algorithms use distributed computations in mini-grids, they should keep the grid size within the limits of recent studies, such as given in [11], to get the best performance benefits. It is expected that additional tuning and advances in cloud computing technology will increase the limits found by these studies.

We described two successful applications of our proposed system on the newly developed algorithms for complex stochastic combinatorial optimization problem, CBF-SRCPSP. Initial estimated sequential duration of several months to several years was reduced to under a month (the first experiment under a week) by distributing the execution. The optimization stage of our architecture finds the best settings for the execution environment. This enables the selection of the best instance-type across different cloud-service providers. In our case, that significantly reduced the cost of running the experiment.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. Brcic and N. Hlupic, "Cloud-distributed computational experiments for combinatorial optimization," in 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017, pp. 197–201., http://doi.org/ 10.23919/MIPRO.2017.7973417

[2] S. Westphal, "Scheduling the German Basketball League," Interfaces, vol. 44, no. 5, pp. 498–508, Oct. 2014.

[3] Y. Sheffi, "Combinatorial Auctions in the Procurement of Transportation Services," Interfaces, vol. 34, no. 4, pp. 245–252, Aug. 2004.

[4] G. Kant, M. Jacks, and C. Aantjes, "Coca-Cola Enterprises Optimizes Vehicle Routes for Efficient Product Delivery," Interfaces, vol. 38, no. 1, pp. 40–50, Feb. 2008.

[5] J. C. Goodson, J. W. Ohlmann, and B. W. Thomas, "Rollout Policies for Dynamic Solutions to the Multivehicle Routing Problem with Stochastic Demand and Duration Limits," Oper Res, vol. 61, no. 1, pp. 138–154, Jan. 2013., http://dx.doi.org/10.1287/opre.1120.1127

[6] P. Lamas and E. Demeulemeester, "A purely proactive scheduling procedure for the resource-constrained project scheduling problem with stochastic activity durations," J. Sched., vol. 19, no. 4, pp. 409–428, Aug. 2016., http://doi.org/ 10.1007/s10951-015-0423-3

[7] H. P. Simão, A. George, W. B. Powell, T. Gifford, J. Nienow, and J. Day, "Approximate Dynamic Programming Captures Fleet Operations for Schneider National," Interfaces, vol. 40, no. 5, pp. 342–352, Jul. 2010., http://doi.org/ 10.1287/inte.1100.0510

[8] C. Vecchiola, S. Pandey, and R. Buyya, "High-Performance Cloud Computing: A View of Scientific Applications," in 2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks, 2009, pp. 4–16., http://doi.org/ 10.1109/I-SPAN.2009.150

[9] C. Evangelinos and C. N. Hill, "Cloud Computing for parallel Scientific HPC Applications: Feasibility of Running Coupled Atmosphere-Ocean Climate Models on Amazon's EC2," in In The 1st Workshop on Cloud Computing and its Applications (CCA), 2008.

[10] D. Kondo, B. Javadi, P. Malecot, F. Cappello, and D. P. Anderson, "Cost-benefit analysis of Cloud Computing versus desktop grids," in 2009 IEEE International Symposium on Parallel Distributed Processing, 2009, pp. 1–12., http://doi.org/ 10.1109/IPDPS.2009.5160911

[11] A. Gupta et al., "The Who, What, Why, and How of High Performance Computing in the Cloud," in 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, 2013, vol. 1, pp. 306–314., http://doi.org/ 10.1109/CloudCom.2013.47

[12] D. Tomić, Z. Car, and D. Ogrizović, "Running HPC applications on many million cores Cloud," in 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017, pp. 209–214., http://doi.org/ 10.23919/MIPRO.2017.7973420

[13] A. Marathe et al., "A Comparative Study of High-performance Computing on the Cloud," in Proceedings of the 22Nd International Symposium on High-performance Parallel and Distributed Computing, New York, NY, USA, 2013, pp. 239–250., http://doi.acm.org/10.1145/2462902.2462919

[14] S. S. Shringarpure, A. Carroll, F. M. D. L. Vega, and C. D. Bustamante, "Inexpensive and Highly Reproducible Cloud-Based Variant Calling of 2,535 Human Genomes," PLOS ONE, vol. 10, no. 6, p. e0129277, Jun. 2015., http://doi.org/ 10.1371/journal.pone.0129277

[15] "[1703.03864] Evolution Strategies as a Scalable Alternative to Reinforcement Learning." [Online]. Available: https://arxiv.org/abs/1703.03864. [Accessed: 10-Sep-2017].

[16] S. Teerapittayanon, B. McDanel, and H. T. Kung, "Distributed Deep Neural Networks Over the Cloud, the Edge and End Devices," in 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), 2017, pp. 328–339., http://doi.org/10.1109/ICDCS.2017.226

[17] "Ray — Ray 0.2.0 documentation." [Online]. Available: http://ray.readthedocs.io/en/latest/. [Accessed: 11-Sep-2017].

[18] N. Hansen, A. Auger, O. Mersmann, T. Tusar, and D. Brockhoff, "COCO: A Platform for Comparing Continuous Optimizers in a Black-Box Setting," ArXiv160308785 Cs Stat, Mar. 2016.

[19] S. Cahon, N. Melab, and E.-G. Talbi, "ParadisEO: A Framework for the Reusable Design of Parallel and Distributed Metaheuristics," J. Heuristics, vol. 10, no. 3, pp. 357–380, May 2004., http://doi.org/ 10.1023/B:HEUR.0000026900.92269.ec

[20] D. R. White, "Software review: the ECJ toolkit," Genet. Program. Evolvable Mach., vol. 13, no. 1, pp. 65–67, Mar. 2012., http://doi.org/ 10.1007/s10710-011-9148-z

[21] "MOEA Framework, a Java library for multiobjective evolutionary algorithms." [Online]. Available: http://moeaframework.org/index.html. [Accessed: 05-Feb-2017].

[22] A. Li, X. Yang, S. Kandula, and M. Zhang, "CloudCmp: Comparing Public Cloud Providers," in Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, New York, NY, USA, 2010, pp. 1–14., http://doi.acm.org/10.1145/1879141.1879143

[23] A. S. Ganapathi, "Predicting and Optimizing System Utilization and Performance via Statistical Machine Learning," University of California at Berkeley, Berkeley, CA, USA, 2009.

[24] D. N. Hieu, T. T. Minh, T. V. Quang, B. X. Giang, and T. V. Hoai, "A Machine Learning-Based Approach for Predicting the Execution Time of CFD Applications on Cloud Computing Environment," in Future Data and Security Engineering, 2016, pp. 40–52., http://doi.org/ 10.1007/978-3-319-48057-2_3

[25] S. Venkataraman, Z. Yang, M. Franklin, B. Recht, and I. Stoica, "Ernest: Efficient Performance Prediction for Large-scale Advanced Analytics," in Proceedings of the 13th Usenix Conference on Networked Systems Design and Implementation, Berkeley, CA, USA, 2016, pp. 363–378.

[26] M. Brčić and N. Hlupić, "Simulation library for Resource Constrained Project Scheduling with uncertain activity durations," in 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014, pp. 1041–1046., http://doi.org/ 10.1109/MIPRO.2014.6859722

[27] M. Brčić, D. Kalpić, and M. Katić, "Proactive Reactive Scheduling in Resource Constrained Projects with Flexibility and Quality Robustness Requirements," Comb. Optim., pp. 112–124, Aug. 2014., http://doi.org/ 10.1007/978-3-319-09174-7_10

[28] R. Kolisch and A. Sprecher, "PSPLIB - A project scheduling problem library: OR Software - ORSEP Operations Research Software Exchange Program," Eur. J. Oper. Res., vol. 96, no. 1, pp. 205–216, Jan. 1997.

[29] M. Matsumoto and T. Nishimura, "Dynamic Creation of Pseudorandom Number Generators," presented at the Proceedings of the Third International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, 1998, pp. 56–69.

[30] J. K. Salmon, M. A. Moraes, R. O. Dror, and D. E. Shaw, "Parallel random numbers: As easy as 1, 2, 3," in High Performance Computing, Networking, Storage and Analysis (SC), 2011 International Conference for, 2011, pp. 1–12., http://doi.org/ 10.1145/2063384.2063405

[31] "Amazon Web Services (AWS) - Cloud Computing Services," Amazon Web Services, Inc. [Online]. Available: //aws.amazon.com/. [Accessed: 14-Dec-2014].

[32] "boto GitHub repository," GitHub. [Online]. Available: https://github.com/boto/boto. [Accessed: 29-Sep-2014].

[33] "Welcome to Paramiko! — Paramiko documentation." [Online]. Available: http://www.paramiko.org/. [Accessed: 25-Oct-2017].

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 138-145 (2018)
www.astesj.com
Special Issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Efficient Limited Feedback Technique for FDD MIMO Systems

Papis Ndiaye, Moussa Diallo, Idy Diop

Department of Computer Science, Polytechnic Institute (ESP), Université Cheikh Anta Diop de Dakar, Senegal

ARTICLEINFO

Article history:
Received: 15 November, 2017
Accepted: 05 February, 2018
Online: 17 March, 2018

Keywords:
MIMO OFDM
Beamforming
Feedback overhead

ABSTRACT

In this paper an efficient feedback quantization technic for beamforming in MIMO systems is presented. The proposed technic named time domain quantization TD-Q is based on the feedback of time domain parameters necessary for the reproduction of the beamforming matrix at the transmitter. This TD-Q presents the same performance than the conventional Givens rotation quantization GR-Q approach which is adopted in IEEE 802.11ac standardand. The performance and amount of feedback of the proposed TD-Q are studied and compared with the GR-Q in IEEE 802.11ac context.

1 Introduction

The combination of Multiple-Input Multiple-Output (MIMO) and Orthogonal Frequency Division Multiplexing (OFDM) technologies (MIMO-OFDM) is now adopted in several communication standards, including the 5^{th} generation of mobile communication network (5G) [1], the IEEE WLAN 802.11ac [2] and the IEEE 802.16 standards (WiMax)[3].

On the one hand, OFDM is a worthwhile tradeoff between bit-error rate performance and spectral efficiency. OFDM consumes part of the channel bandwidth, but it is robust to frequency selective fading environment. In addition, it enables the use of several advanced technics to further enhance the system throughput, as for instance the bit loading technic [4] and the subcarriers allocation in orthogonal frequency division multiple access (OFDMA) [5].

On the other hand, the MIMO system has the potential to improve the system capacity. In IEEE 802.11ac wireless local area network (WLAN) standard, there are five transmit and receive MIMO technics: Cyclic Shift Diversity (CSD), Space Time Block Coding (STBC), Spatial Division Multiplexing (SDM), Maximal Rotation Combining (MRC) and Transmit Beamforming (TxBF) [2]. Among these five technics, this is the Transmit Beamforming which can maximize the system capacity. Indeed, the Beamforming with precoding and postcoding eliminates the co-channel interferences (CCI) which are the fundamental problem faced by the practical MIMO system [6-8].

However, in beamforming technic the channel state information (CSI) must be available at the transmitter and the receiver. This CSI is estimated at the receiver and fed back to the transmitter. The CSI, which indicates amplitude and phase for each transmit antenna, receive antenna, and each OFDM subcarrier in the RF channel may reduce the overall throughput. To reduce the feedback amount, the receiver computes the beamforming matrix (precoding matrix) which is an unitary matrix and compresses it before sending back to the transmitter.

There are several proposals in the literature for the beamforming matrix compression. Many authors propose codebook based approach. In codebooks based technics, the channel distribution is taken into account during the codebook design. Instead of sending the precoding matrix, only the index of the selected precoding matrix (after channel estimation and SVD decomposition) is sent. Intrinsically, the codebook is restricted to have fixed cardinality. Thereby, the selected pre-coding matrix, which is the most similar is not necessarily the most optimal. Unfortunately, the codebook size has an impact on the system performance and requires high storage. For traditional MIMO systems, several codebooks have been proposed, such as Kerdock codebook [9], codebooks based on vector quantization [10], Grassmannian packing [11], discrete Fourier transform (DFT) [12] and quadrature amplitude modulation [13]. The coodbook principle is adopted in LTE [14]. In [15], the codebook was designed and optimized by selecting matrices having 8 phase-shift keying (PSK) entries and coping with a large range of propagation conditions. However the cost of this scheme is that the performance gain decreases dramatically due to the channel quantized er-

https://dx.doi.org/10.25046/aj030216

ror caused by the codebook feedback[16].

Another alternative is the Givens rotation (GR) approach. Here, Givens rotation is used to decompose the unitary beamforming matrix. After the decomposition, the receiver only feeds back the GR parameters necessary for the reproduction of the beamforming matrix by the transmitter [17]. The GR approach is adopted in IEEE 802.11ac standard for TxBF mode[2]. The compression ratio of the GR quantization is 75% for a MIMO 2×2 configuration [18]. However, this ratio decreases depending on the number of antennas. It then becomes necessary to find an alternative to the GR qunatization since the 802.11*ac* standard has adopted an 8×8 MIMO configuration which may reach $10Gbps$ and this number of antennas will undoubtedly increase due to the growing demand for bitrate.

This paper proposes a quantization approach named time domain quantization (TD-Q). Instead of the quantized precoder matrix as in GR, we propose the feedback of the quantized time domain channel coefficients. In addition, an optimization based on a metric updated during the channel estimation process and the MIMO channels impulse response is also proposed in order to further reduce the feedback overhead.

The rest of the paper is organized as follows. The section 2 presents the beamforming in MIMO OFDM systems. Section 3 is dedicated to the GR-Q which is adopted in the IEEE 802.11ac standard. Following that, the proposed TD-Q is presented in the section 4. Next, simulation results and comparison between GR-Q and TD-Q are shown in the section 5. Finally, the conclusion on the works presented in this paper is done in the section 6.

2 System Model

Consider a MIMO channel with N_t transmit antennas and N_r receiver antennas. In classical MIMO-OFDM context, for each subcarrier, the $N_r \times 1$ received signal y can be expressed as:

$$y = Hx + n \tag{1}$$

where H is the $N_r \times N_t$ frequency channel coefficients, x the $N_t \times 1$ transmitted signal and n the $N_r \times 1$ zero mean Gaussian noise vector.

By using singular value decomposition (SVD), the matrix H can be decomposed into:

$$H = UDV^H \tag{2}$$

where $U \in \mathbb{C}^{N_r \times R}$ and $V \in \mathbb{C}^{N_t \times R}$ are unitary matrix, $D \in \mathbb{C}^{R \times R}$ is diagonal, $(.)^H$ denotes the hermitian of $(.)$ and R the rank of the matrix H.

In MIMO-OFDM with beamforming context as in TxBF mode of IEEE 802.11ac, the received signal y is expressed this time by:

$$y = U^H HVx + U^H n = U^H UDV^H Vx + U^H n = Dx + U^H n \tag{3}$$

where V and U^H are used as a precoder matrix and a postcoder matrix respectively. The postcoder matrix U^H is an unitary matrix and thus the noise is neither colored nor enhanced.

Since D is exactly diagonal, there is no CCI and the MIMO channel matrix is reduced on R separate and independent SISO channels.

However, the precoder matrix V is needed at the transmitter. Two solutions are available in order to provide CSI at the transmitter. The first solution consists of exploiting the channel reciprocity in case of time division duplex (TDD) transmission modes. But, it is not straightforward due to the mismatch in the radio front end. The alternative, which is the most appropriate and that can be applied for both TDD and frequency division duplex (FDD) transmission modes, consists in using the reverse link to feed back the matrix V to the transmitter. Note that, for each subcarrier the unitary matrix V has to be quantized and fed back to the transmitter.

3 Givens Rotation Quantization and feedback compression ratio analysis

3.1 Givens Rotation Quantization principle

Givens Rotation Quantization GR-Q is proposed in [17]. The authors propose a reduction of the feedback overhead by exploiting the unitary property of the precoder matrix V. The idea is to represent the matrix V as a special form of Givens rotation with a complex diagonal matrix. Thereby, the $N_t \times R$ unitary matrix V can be factorized as follows.

$$V = \prod_{i=1}^{min(R,N_t-1)} [D_i \prod_{l=i+1}^{N_t} G_{li}^T(\psi_{li})] \times I_{N_t \times R} \tag{4}$$

where

$$D_i = \begin{bmatrix} I_{i-1} & 0 & 0 & \dots & 0 \\ 0 & e^{j\phi_{i,i}} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & e^{j\phi_{N_t-1,i}} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5}$$

$$G_{li} = \begin{bmatrix} I_{i-1} & 0 & 0 & \dots & 0 \\ 0 & cos(\psi_{li}) & 0 & sin(\psi_{li}) & 0 \\ 0 & 0 & I_{l-i-1} & 0 & 0 \\ 0 & -sin(\psi_{li}) & 0 & cos(\psi_{li}) & 0 \\ 0 & 0 & 0 & 0 & I_{N_t-l} \end{bmatrix} \tag{6}$$

The matrix D_i is an $N_t \times N_t$ diagonal matrix, $I_{N_t \times R}$ is an $N_t \times R$ identity matrix and G_{li} is an $N_t \times N_t$ Givens rotation matrix.

According to the equations (4-6), the parameters to be determined to identify the precoder matrix V are:

$$\begin{array}{llllll} \psi_{li} & for & i = 1,2,\dots,m & and & i < l \leq N_t \\ \phi_{j,i} & for & i = 1,2,\dots,m & and & i < j \leq N_t - 1 \end{array} \tag{7}$$

where $m = min(R, N_t - 1)$.

Take a 3×3 unitary matrix V as an example, it can be expressed as:

$$V = D_1 \times G_{21}^T(\psi_{21}) \times G_{31}^T(\psi_{31}) \times D_2 \times G_{32}^T(\psi_{32}) \times I_{3\times 3} \quad (8)$$

The precoder matrix V can then be reconstructed through the six parameters $\psi_{21}, \psi_{31}, \psi_{32}, \phi_{1,1}, \phi_{2,1}$ and $\phi_{2,2}$.

Therefore, instead of all the elements of the matrix V, it is sufficient to considere the parameters ψ_{li} and $\phi_{j,i}$. These parameters can vary from 0 to 2π for ϕ and from 0 to $\pi/2$ for ψ.

Now, ψ and ϕ can be quantized according to equations (9) and (10) where $\widehat{\psi}$ and $\widehat{\phi}$ represent the quantized angles.

$$\widehat{\psi} = \frac{k\pi}{2^{nb_\psi+1}} + \frac{\pi}{2^{nb_\psi+2}} \quad (9)$$

where $k = 1, 2, \ldots, 2^{nb_\psi} - 1$ and nb_ψ the number of bits per angle ψ

$$\widehat{\phi} = \frac{k\pi}{2^{nb_\phi-1}} + \frac{\pi}{2^{nb_\phi}} \quad (10)$$

where $k = 1, 2, \ldots, 2^{nb_\phi} - 1$ and nb_ϕ the number of bits per angle ϕ

Finally, the receiver feeds back the quantized parameters $\widehat{\psi}_{li}$ and $\widehat{\phi}_{j,i}$ to the transmitter which can recover the quantized precoder matrix $\widehat{V}$ by using:

$$\widehat{V} = \prod_{i=1}^{min(R,N_t-1)} [\widehat{D}_i \prod_{l=i+1}^{N_t} \widehat{G}_{li}^T(\widehat{\psi}_{li})] \times I_{N_t \times R} \quad (11)$$

where

$$\widehat{D}_i = \begin{bmatrix} I_{i-1} & 0 & 0 & \ldots & 0 \\ 0 & e^{j\widehat{\phi}_{i,i}} & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & e^{j\widehat{\phi}_{N_t-1,i}} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (12)$$

It is important to note that:

- The receiver has to feed back a precoder matrix $\widehat{V}$ for each subcarrier.
- The quantization error Σ with $\widehat{V} = V + \Sigma$ exists on each subcarrier.

Consider the equation (3) with the quantized error Σ and $\widehat{V}$ as a precoder matrix.

$$\begin{aligned} y &= U^H U D V^H \widehat{V} x + U^H n \\ &= U^H U D V^H (V + \Sigma) x + U^H n \\ &= Dx + \overline{D} V^H \Sigma x + U^H n \\ &= Dx + M_{CCI} x + U^H n \end{aligned} \quad (13)$$

The term $M_{CCI}x = (DV^H\widehat{V} - Dx$ is the corresponding CCI caused by the quantization error. However, in telecommunication standrd, one always chooses a number of quantization bits (nb_ψ and nb_ϕ) that allows the minimization of resulting CCI. This is the case of IEEE 802.11ac where $nb_\psi = nb_\phi = 10$.

Finally, the feedback overhead length (in bit) noted by N_{FGR} can be calculated using the formula $N_{FGR} = N_u * nb * N_t(N_t - 1)$ assuming that $N_r = N_t$, $nb_\psi = nb_\phi = nb$ and N_u the number of used subcarriers.

3.2 Feedback overhead length analysis in IEEE 802.11ac context

The bandwidth in IEEE802.11ac has increased from a maximum of 40 MHz with the old standard up to 80 or even 160 MHz. The number of subcarriers and used one according to the bandwidth are grouped in the Table 1.

Table 1: Number of used subcarriers according to the de bandwidth in IEEE 802.11ac

Bandwidths	N	N_u
20 MHz	64	52
40 MHz	128	108
80 MHz	256	234
160 MHz	512	486

Table 2 contains for any MIMO configurations and bandwidth:

- the feedback overhead length when no compression is performed,
- the feedback overhead length when GR quantization is performed,
- the compression ratio achieved by GR quantization.

As we can see, the GR quantization can reach a compression ratio up to 75% for a MIMO 2×2 configuration. This compression ratio decreases when the number of antennas becomes large. Because of the high demand in terms of transmission throughput, the number of antennas will increase in the evolution of IEEE 802.11 standard. The next WLAN generation will undoubtedly consider the massive MIMO concept which is more and more studied for WLAN [19-20]. In this context, it is necessary to find an alternative to the GR quantization because its compression ratio, as demonstrated by the equation (14), will be around 50%:

$$\lim_{N_t=N_r\to+\infty} \frac{N_t(N_t-1)}{N_t * N_t * 2} = \frac{1}{2} \quad (14)$$

4 Time domain quantization

4.1 Time domain quantization principle

Consider a MIMO channel with N_t transmit antennas and N_r receive antennas. The discrete time domain channel response between the transmitting antenna i

Table 2: Feedback overhead (in bit) with and without GR quantization for any configurations ($N_t \times N_r$ and bandwidth).

MIMO configuration	Number of GR parameters	feedback overhead N_{FGR} for GR				feedback overhead without quantization				Ratio
		20 MHz	40 MHz	80 MHz	160 MHz	20 MHz	40 MHz	80 MHz	160 MHz	
2x2	2	1040	2160	4680	9720	4160	8640	18720	38880	75%
3x3	6	3120	6480	14042	29160	9360	19365	42120	87480	66.67%
4x4	12	6240	12960	28080	58320	16640	34560	74880	155520	62.5%
6x6	30	15600	32400	70200	145800	37440	77760	169730	349920	58.33%
8x8	56	29120	60480	131040	271410	53248	110592	239616	622080	56.25%

and the receiving antenna j under the multipath fading environments can be expressed as:

$$h_{ij}(n) = \sum_{l=0}^{L-1} h_{ij,l}\delta(n - \tau_{ij,l}) \tag{15}$$

where L is the number of paths, $h_{ij,l}$ and $\tau_{ij,l}$ the complex time varying channel coefficient and delay of the l^{th} path.

Note that in practical OFDM system, in order to eliminate the intersymbol interference (ISI) between consecutive OFDM symbols, the maximum multipath delay is within the cyclic prefix, of length (CP) *ie* $L \leq CP$. Thereby, we propose, instead of the quantized precoder matrix V for each subcarrier, to feed back the quantized time domain channel coefficients for the entire system. The main goal of this proposition is to considerably reduce the feedback overhead because in this case, the feedback is composed at most by $N_t * N_r * CP$ complex coefficients. As can be seen, the number of OFDM subcarriers has no impact on the feedback overhead length. In addition, this proposed time domain quatization method can be performed jointly with the channel estimate.

4.2 Joint time domaine channel estimation and feedback quantization

In closed loop MIMO-OFDM systems, a preamble is inserted at the beginning of frames to facilitate the channel estimation. Comb-type pilot method can also be used on the rest of the frame (as in IEEE 802.11ac [2] or LTE [1]) to improve the accuracy of the channel estimation during the transmission. In this paper, we focus on time domain channel estimation (TD-CE). This technique is known to provide very good results by significantly reducing the noise on the estimated channel coefficients [21][22].

Firstly, without using any knowledge of the statistics on the channels, least square (LS) channel estimation can be performed at the receiver side using the demodulated pilots signal and the known pilots symbol in the frequency domain for each subcarrier.

Next assuming that $H_{ij,k}$ is the discrete LS estimated channel response on subcarriers k between the ith transmit antenna and the jth receive antenna, the CSI can be converted into the time domain by the inverse discrete fourier transform (IDFT) algorithm as:

$$\begin{aligned} h_{ij}(n) &= IDFT(H_{ij}) \\ &= \sqrt{\tfrac{1}{N_s}} \textstyle\sum_{k=0}^{N_s-1} H_{ij,k} e^{\frac{j2nk\pi}{N_s}} \end{aligned} \tag{16}$$

where N_s is the number of subcarriers and $n = 1, \ldots, N_s$.

Note that the LS method is widely used due to its simplicity and minimum requirements for the knowledge of channel statistics . However, the LS estimator does not consider the noise effect. That is why its performance is often degraded by the noise. By taking into account the equation (15) and the fact that $L \leq CP$, the time domain samples $h_{ij}(n)$ from ((16)) can be divided into two parts:

- the first CP samples in which are the paths of the channel,
- the other samples which are only composed by noise.

A smothing process which keeps only the first CP samples by eliminating the other one can considerably improve the accurate of the estimated channel [21]. However, it is possible to estimate a mean of the noise noted by $\overline{m}$ from the noise samples before its deletion.

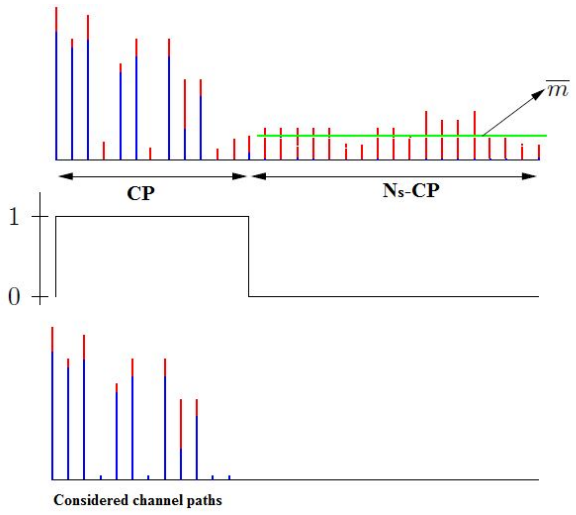


Figure 1: Mean Noise

This proposed process is illustrated in the figure 1. the metric $\overline{m}$ estimated from the noise samples allows to keep only the channel paths which are among the

first CP samples. As can be seen, all samples below the $\overline{m}$ were eliminated.

Now, the receiver have to do the two following actions:

- Perform a DFT of size N_s by considering only the remaining samples $h_{ij}(n)$. The resulting estimated channel frequency response is then more accurate than the LS one.

- Perform the feedback of the remaining samples $h_{ij}(n)$. To achieve this goal, each of the remaining samples $h_{ij}(n)$ whose number is lower than CP is expressed as:

$$h_{ij}(n) = \rho_{ij} e^{\phi_{ij,n}} = tan(\psi_{ij,n}) e^{\phi_{ij,n}} \qquad (17)$$

  where $n \leq CP$, $\psi_{ij,n}$ and $\phi_{ij,n}$ angles parameters can vary from 0 to $\pi/2$ and 0 to 2π respectively.

  After that, $\psi_{ij,n}$ and $\phi_{ij,n}$ is quantized and fed back to the transmitter according to equations (9) and (10) where $\widehat{\psi}_{ij,n}$ and $\widehat{\phi}_{ij,n}$ represent the quantized angles parameters.

Thus, the feedback overhead length (in bit) noted by N_{FTD} can be calculated using the formula $N_{FTD} = N_{rs} * 2 * nb * N_t^2$ assuming that $N_r = N_t$, $nb_\psi = nb_\phi = nb$ and N_{rs} the number of remaining samples.

Thanks to this feedback, the transmitter can reconstitute the accurate estimated channel. Thereby, the MIMO channel response is available both at the reciever and the transmitter side. Finally, transmitter and reciever perform SVD decomposition which allows it to respectively have the pre-coder and the post-coder.

5 Performance comparison in IEEE 802.11ac context

In this section, we compare the Givens rotation quantization GR-Q and the proposed time domain quantization TD-Q one in terms of feedback overhead and BER performance in IEEE 802.11ac context [2]. Note that the bandwidth in IEEE802.11ac has increased from a maximum of 40 MHz with the old standard up to 80 or even 160 MHz. For each bandwidth, the cyclic prefix length (CP) is equal to a quarter of the number of subcarriers (in Table 1). We consider in the similations MIMO 2×2, 3×3, 4×4, 6×6 and 8×8 with $4-QAM$ and $64-QAM$ modulations.

In the sequel, all simulations are performed using the ITU Channel Model for Outdoor to Indoor and Pedestrian Test Environment. Since the delay spread can vary significantly, the ITU recommendation specifies two different delay spreads for each test environment: low delay spread (channel model A), and medium delay spread (channel model B). The parameters of these channel's model are grouped in Table 3. In this paper we focus on the channel model B.

5.1 feedback overhead comparison

Figures 2 to 6 show the evolution of the total number of feedback complex coefficients of the proposed time domain quantization according to the signal to noise ratio (SNR) for MIMO 2×2, 3×3, 4×4, 6×6 and 8×8 configurations. For each MIMO configuration, simulations are performed for the 4 existing bandwidth of the IEEE 802.11 ac standard (20 MHz,40 MHz,80 MHz and 160 MHz). Note that the total number of feedback complex coefficients represents the sum of all the remaining samples and it can be calculated by $N_{rs} * N_t^2$ assuming that $N_r = N_t$.

As can be seen from these figures:

- whatever the MIMO configuration, the total number of feedback complex coefficients increases according to the bandwidth and the number of antennas.

- whatever the SNR, the plotted cumulative distribution function (CDF) shows that it is statistically possible for each MIMO configuration and bandwidth to know the maximum value of the total number of feedback complex coefficients. Following this study the total number of feedback complex coefficients for any configurations ($N_t \times N_r$ and bandwidth) in IEEE 802.11ac ontext are grouped in the table 4.

Therefore, according to equation (17) the feedback overhead length of the proposed time domaine quatization (in bit) noted by N_{FTD} can be calculated using $N_{FTD} = N_{rs} * 2 * nb * N_t^2$ and the number of remaining samples N_{rs} in the Table 4.

Except for MIMO 2×2 configuration and bandwidth $20MHz$, the feedback overhead of the proposed time domain quantization is much shorter than that of the Givens Rotation quantization. The compression ratio of the proposed time domain quantization compared to Givens Rotation quantization can be up to almost 70%.

The most important to note here is the fact that unlike the Givens Rotation quantization for which the compression ratio decreases when the number of antennas becomes large (see section 3.2), the proposed time domain compression becomes more efficient. This last remark makes the proposed time domain compression very interesting for the next WLAN generation which will undoubtedly consider the massive MIMO concept which is more and more studied for WLAN [19-20].

Table 3: ITU Channel Model for Outdoor to Indoor and Pedestrian Test Environment.

Tap	Channel A		Channel B	
	Relative delay(ns)	Average power (dB)	Relative delay(ns)	Average power (dB)
1	0	0	0	0
2	110	−9.7	220	−0.9
3	190	−19.2	800	−4.9
4	410	−22.8	1200	−8
5			2300	−7.8
6			3700	−23.9

Table 4: Total number of feedback complex coefficients of the proposed time domain quantization for any configurations ($N_t \times N_r$ and bandwidth) in IEEE 802.11ac ontext.

MIMO configuration	feedback overhead N_{FTR}			
	20 MHz	40 MHz	80 MHz	160 MHz
2x2	55	90	161	293
3x3	111	192	334	613
4x4	191	323	582	1069
6x6	415	703	1254	2337
8x8	718	1215	2205	4106

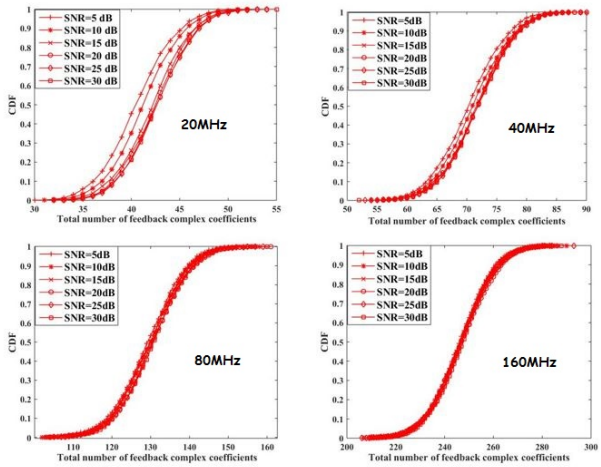


Figure 2: The Cumulative Distribution Function of the total number of feedback complex coefficients according to the Signal to Noise ratio and Bandwidth for MIMO 2×2 configutation

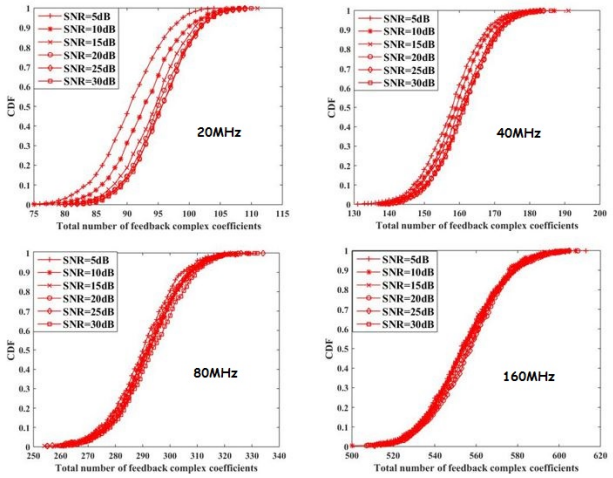


Figure 3: The Cumulative Distribution Function of the total number of feedback complex coefficients according to the Signal to Noise ratio and Bandwidth for MIMO 3×3 configutation

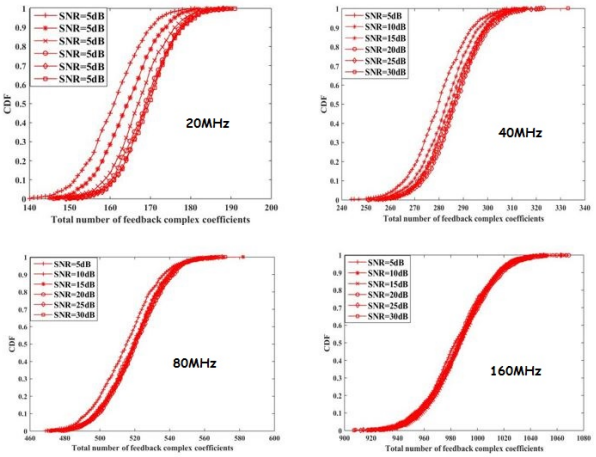


Figure 4: The Cumulative Distribution Function of the total number of feedback complex coefficients according to the Signal to Noise ratio and Bandwidth for MIMO 4×4 configutation

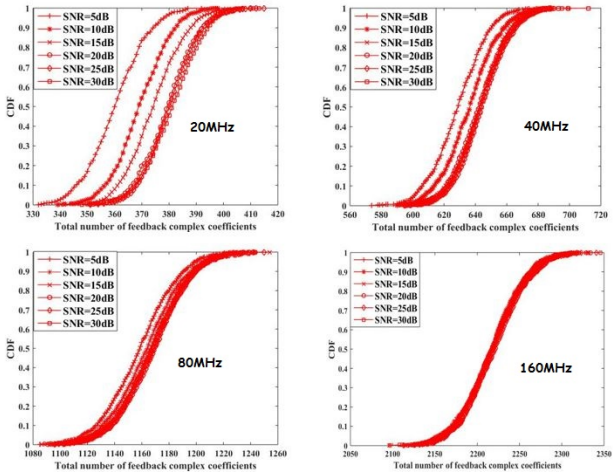


Figure 5: The Cumulative Distribution Function of the total number of feedback complex coefficients according to the Signal to Noise ratio and Bandwidth for MIMO 6×6 configutation

5.2 BER performance

In this subsection, we compare the beamforming with Givens rotation quantization (GR-Q) and that with the time domain quantization in terms of bit error rate (BER) performance. The considered simulation parameters are listed in Table 6.

Table 5: Feedback overhead (in bit) comparison between Givens Rotation quantization and the proposed time domain quatization for any configurations ($N_t \times N_r$ and bandwidth).

MIMO configuration	feedback overhead N_{FGR} for Givens Rotation quantization				feedback overhead N_{FTR} for the proposed time domain quatization				Ratio	
	20 MHz	40 MHz	80 MHz	160 MHz	20 MHz	40 MHz	80 MHz	160 MHz	min	max
2x2	1040	2160	4680	9720	1100	1800	3220	5860	−0.5%	39.7%
3x3	3120	6480	14042	29160	2220	3840	6680	12260	28.8%	58%
4x4	6240	12960	28080	58320	3820	6460	11640	21380	38.7%	63.3%
6x6	15600	32400	70200	145800	8300	14060	25080	46740	46.8%	68%
8x8	29120	60480	131040	271410	14360	24300	44100	82120	50.6%	69.7%

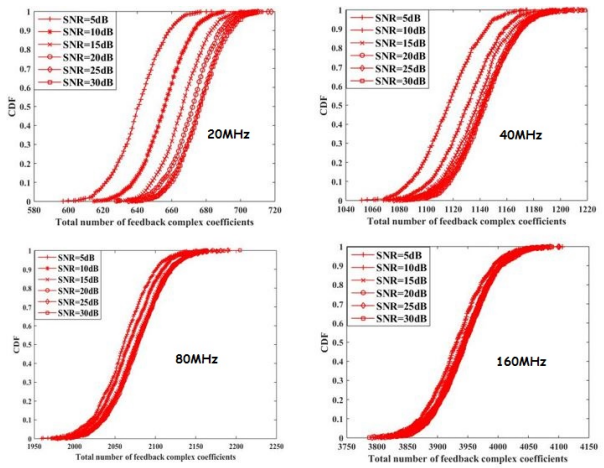


Figure 6: The Cumulative Distribution Function of the total number of feedback complex coefficients according to the Signal to Noise ratio and Bandwidth for MIMO 8×8 configutation

Table 6: Simulation Parameters

Channel Model	ITU Channel Model
Bandwidth	40MHz
Number of FFT points (N)	128
Number used subcarriers	108
Number of bits per angle	10
cyclic prefix	*Long ie* 32 samples
Number of $N_t \times N_r$ antennas	2×2 and 4×4
Modulation	QPSK and 64-QAM
FEC	Convolutional
Coding Rate	1/2

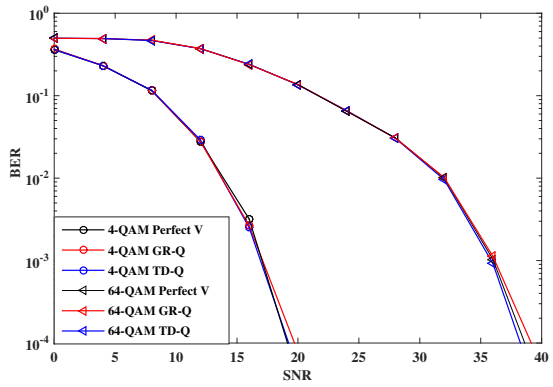


Figure 7: BER performance for perfect beamforming (Perfect V), beamforming with GR-Q and beamforming with the prposed time domain quantization (TD-Q) in MIMO 2×2 context. The ITU Channel Model is considered for simulations.

The BER performance is shown in Figures 7 and 8 for 4–QAM and 64–QAM in MIMO 2×2 and MIMO 4×4 context respectively. The number of bits per angle feedback is taken to 10. It can be seen that the proposed TD-Q presents the same performance, for all considered configurations in these simulations, than the GR-Q which is adopted in IEEE 802.11ac standard. As considered in the 802.11ac standard, the simulations confirm that 10 bits per angle feedback minimizes the quantization error and allows the quantized beamforming (GR-Q and TD-Q) to have the same performance than the perfect one.

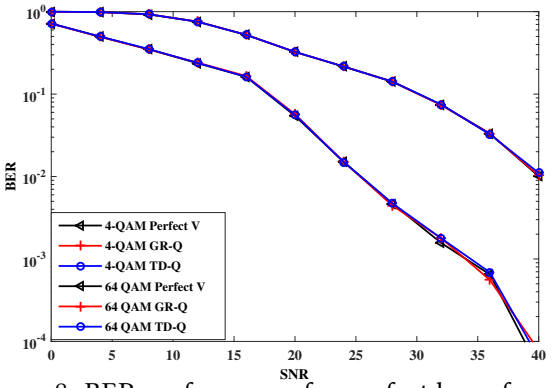


Figure 8: BER performance for perfect beamforming (Perfect V), beamforming with GR-Q and beamforming with the prposed time domain quantization (TD-Q) in MIMO 4×4 context. The ITU Channel Model is considered for simulations.

6 Conclusion

The proposed TD-Q requires less amount of feedback than the GR-Q one in IEEE 802.11ac TxBF mode. In addition, simulations in 802.11ac context with standard TGn channel model show that the TD-Q beamforming allows to have the same performance than the adopted GR-Q. The highlight of this proposal is that the proposed TD-Q can be considered as a credible alternative to the GR-Q for the IEEE 802.11ac evolution. The proposed quantization technic is not only for WLAN systems. It can be adapted for all SISO or MIMO system using closed loop beamforming.

References

1. Q.Cui, H. Wang, P. Hu, X. Tao, P.Zhang, J. Hamalainen and L. Xia, “Evolution of limited feedback CoMP systems from 4G

to 5G " IEEE vehicular technologie magazine, **9**(3), 94–103, 2014. DOI: 10.1109/MVT.2014.2334451

2. "IEEE P802.11ac/D0.2" IEEE standard Draft, 2011.
3. "Mobile WiMAX Part II: A Comparative Analysis" IWiMAX Forum, 2006
4. D. Wang, Y. Cao and L. Zheng, "Efficient Two-Stage Discrete Bit-Loading Algorithms for OFDM Systems" IEEE Trans. on Vehicular Technology, **59**(7), 3407–3416, 2010. DOI: 10.1109/TVT.2010.2052937
5. T. Ramji, B. Ramkumar and M. S. Manikandan "Resource and subcarriers allocation for OFDMA based wireless distributed computing system" IEEE Intern Adv Comp. Conf. 2014. DOI: 10.1109/IAdCC.2014.6779345
6. A.Goldsmith, S. A. Jafar, A. Jindal and S. Vishwanath "ACapacity limits of MIMO Channels" IEEE J. Select. Areas Commun **21**, 684-702, 2003.
7. H. Kim, Y. Jung, J. Park and J. Kim "Adaptative CSI Feedback Scheme to maximize the throughput in IEEE 802.11ac system," IEEE ISCE 1-2, Jeju, Korea, June 2014.
8. M. Mbaye, M. diallo and M. Mboup"LU based Beamforming schemes for MIMO systems" IEEE Trans. on Vehicular Technology, **66**(3), 1-9, 2016 DOI: 10.1109/TVT.2016.2573046.
9. T. Inoue and R. W. Heath Jr"Kerdock codes for limited feedback precoded MIMO systems" IEEE Transactions on Signal Processing, **57**(9), 37113716, 2009.
10. J. C. Roh and B. D. Rao"Design and analysis of MIMO spatial multiplexing systems with quantized feedback" IEEE Transactions on Signal Processing, **54**(8), 28742886, 2006.
11. D. J. Love and R. W. Heath Jr"Limited feedback unitary precoding for spatial multiplexing systems" IEEE Transactions on Information Theory, **51**(8), 29672976, 2005.
12. K. Schober, R. Wichman, and T. Koivisto"MIMO adaptive codebook for closely spaced antenna arrays" European Signal Processing Conference, 1-5, 2011.
13. D. J. Ryan, I. V. L. Clarkson, I. B. Collings, D. Guo, and M. L. Honig"QAM codebooks for low-complexity limited feedback MIMO beamforming" IEEE International Conference on Communications, 41624167, 2007.
14. LTE Physical layer procedures "Evolved Universal Terrestrial Radio Access (E-UTRA)" 3GPP TS 36.213 version 13.4.0 Release 13, 2017.
15. G. Berardinelli and T. B. Sorensen and P. Mogensen and K. Pajukoski "SVD-Based vs. Release 8 Codebooks for Single User MIMO LTE-A Uplink" IEEE 71st Vehicular Technology Conference, 1-5, 2010, DOI: 10.1109/VETECS.2010.5493709.
16. Zheng Jiang; Bin Han; Liang Lin; Peng Chen; Fengyi Yang; Qi Bi "On compressive CSI feedback beamforming scheme for FDD massive MIMO" IEEE International Conference on Communications in China, 1-5, 2016.
17. C. Yuen, S. Sun, M. Ho and Z. Zhang "Beamforming matrix quantization with variable feedback rate" EURASIP journal on wireless communications and networking, 2012.
18. J. Kim and C. Aldana"Efficient feedback of the channel information for closedloop beamforming in WLAN" IEEE VTC-spring, pp. 2227-2230, Melbourne, 2006.
19. Y. Morino, T. Hiraguri, H. Yoshino and K. Nishimori"Proposal of overhead-less access control scheme for multi-beam massive MIMO transmission in WLAN systems" Med-Hoc-Net, pp. 1-5, Budva, 2017.
20. J. Lee, K. J. Choi and K. S. Kim"Massive MIMO full-duplex for high-efficiency next generation WLAN systems" International Conference on Information and Communication Technology Convergence, pp. 1152-1154, Jeju, 2016.
21. Moussa Diallo; Laurent Boher; Rodrigue Rabineau; Laurent Cariou; Maryline Helard "Transform domain channel estimation with null subcarriers for MIMO-OFDM systems" IEEE International Symposium on Wireless Communication Systemse, pp. 209-213, Reykjavik, 2008.
22. Zhi Zheng, Caiyong Hao and Xuemin Yang"Least squares channel estimation with noise suppression for OFDM systems" IEEE ELECTRONICS LETTERS, Vol. 52 No. 1, pp. 3739, 2016.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 145-150 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Relational Database Model and Tools for Environmental Sound Recognition

Yuksel Arslan*, Abdussamet Tanış, Huseyin Canbolat

Electrical and Electronics Engineering, Ankara Yıldırım Beyazıt University, 06010, Turkey

ARTICLE INFO

Article history:
Received: 29 October, 2017
Accepted: 17 November,2017
Online: 10 December, 2017

Keywords:
Database Design
Environmental Sound Recognition
Machine Learning

ABSTRACT

Environmental sound recognition (ESR) has become a hot topic in recent years. ESR is mainly based on machine learning (ML) and ML algorithms require first a training database. This database must comprise the sounds to be recognized and other related sounds. An ESR system needs the database during training, testing and in the production stage. In this paper, we present the design and pilot establishment of a database which will assists all researchers who want to establish an ESR system. This database employs relational database model which is not used for this task before. We explain in this paper design and implementation details of the database, data collection and load process. Besides we explain the tools and developed graphical user interface for a desktop application and for the WEB.

1. Introduction

This paper is an extension of work presented in 25th Signal Processing and Communications Applications Conference (SIU), 2017 [1]. The database design and implementation described in that paper was mainly for impulsive sound detection and the database was for a hazardous sound recognition application. Here the extended database is for all kinds of environmental sounds and it is aimed for all kinds of ESR applications.

Historically non-speech sound recognition has not received as much attention as automatic speech recognition (ASR). ASR has well established algorithms and databases while research has begun much earlier. Automatic ESR (AESR) is getting attraction since last two decades. We can list the following applications of AESR: In military, forensic and law enforcement domain there are studies on gunshot detection systems. In [2], a gunshot detection system is proposed. In [3], the gunshot blast is used to identify the caliber of the gun. In [4] and [5] ESR is used for robot navigation. ESR can be used for home monitoring. It can be used to assist elderly people living in their home alone [6], [7]. In [8], it is used for home automation. In [9] and [10], ESR is used for recognition of animal sounds. In the surveillance area, it is used for surveillance of road [11], public transport [12], elevator [13] and office corridor [14].designations. ESR system design is started with the training phase. How a sound database is used in training phase of an ESR system is explained in Figure 1.

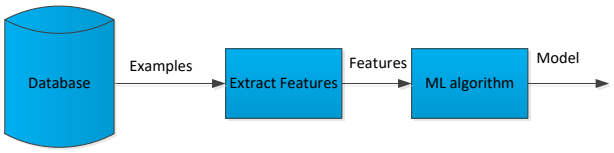


Figure 1: ESR system in training phase

During the development stage of an ESR system the desired ML algorithm must be trained with the sounds to be recognized. The database provides the sound clips to be recognized and other sound clips which are negative examples. After training, a model is developed and this model is used for testing. In Figure 1, the database provides positive and negative examples; features are inputs to the ML algorithm, ML algorithm using these features produce the model.

In Figure 2, testing phase of an ESR system is shown. In testing phase database provides the positive and negative examples, model produces the predictions about the examples provided and at last predictions are compared with the truth provided by the database and a performance rate is reached. According to this performance rate, ML algorithm or the feature set or other parameters of the ML algorithm may be needed to change. Then the whole training and testing phase start again. This process continues until an acceptable performance rate is reached.

After model creation and testing, this model is used in the production phase (Figure 3). In production phase, sounds come

*Yuksel Arslan, Ankara, Cankaya Ilkbahar Mah. Guneypark Kumeevleri A35/35 Turkey, Email: yuksel_arslan@yahoo.com

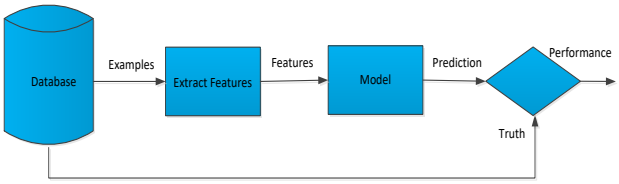


Figure 2: ESR system in testing phase

from the environment, model makes the predictions about the classes of sounds and we actually may not know the real truth. In production stage we may update sounds in database, update the model even we may change the ML algorithm. In each case we need the database.

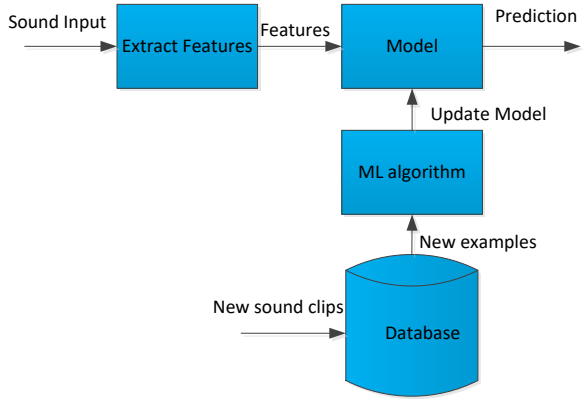


Figure 3: ESR system in production phase

This paper is organized as follows: In section II we will make a literature review of ESR databases. In section III a new relational database model for ESR systems will be explained. At the last, section IV, the contribution of this work and other planned activities will be explained.

2. Literature Review

In this section we will review the mostly referenced databases in ESR related papers. The structure of the databases will be explored; pros and cons of the structure will be argued if applicable.

One of the databases that have been mostly reverenced is Real World Computing Partnership's (RWCP) non-speech database [15]. Using a standard single microphone, acoustic signals of about 100 types of sound sources were measured in an anechoic room as the dry source. By using the impulse response of three kinds of rooms, convolution with the 48 kHz sampling signal was carried out to reconstruct the sound signal in respective rooms. The database structure (structure of directories) and navigation to sounds is explained and provided by an HTML page as shown in Figure 4 for dry sources. As an ESR developer we must examine the structure of HTML file and find the desired sound clips. Maintenance of the data and usage is not so easy with this kind of structure.

Detection and Classification of Acoustic Scenes and Events (DCASE) is an official IEEE Audio and Acousti Acoustic Signal Processing challenge. For challenge a database is prepared and it is also a resource for ESR researchers. DCASE 2016 challenge consists of four tasks [16]. The goal of Task-2 is to detect sound events for example "bird singing", "car passing by" that are present within an audio segment. To be prepared for the challenge two datasets, train and development are given. Train is used to create the model and development is to test the model.

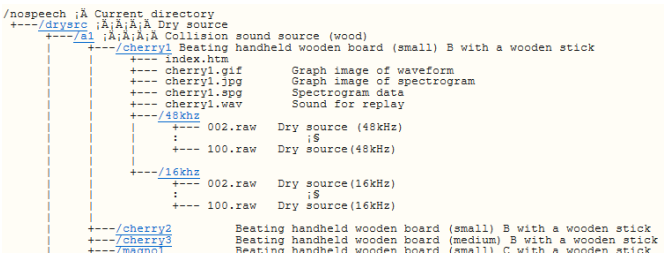


Figure 4: RWCP dry source sound clips structure

For training all sound files are in one directory and a readme file explains the details, such as sampling frequency, quantization bit depth, etc. For development dataset there is an annotation text file for each sound clip. Each .txt file contains information about the onset, offset, and event class for each event in the scene, separated by a tab.

DCASE 2017 Task-2 dataset contains ".yaml" files for annotations [17]. ".yaml" files (Figure 5) can be read by a Matlab command, so it is easy to work with these structured files.

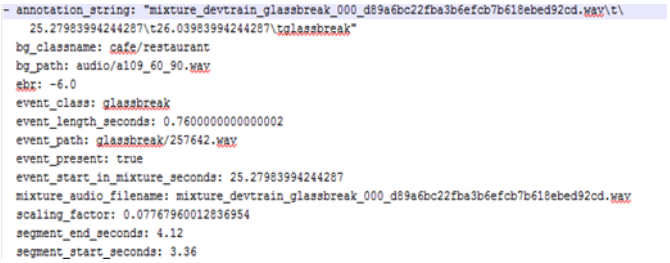


Figure 5 DCASE 2017 task-2 development dataset "yaml" file for glass break

Dataset for Environmental Sound Classification (ESC) contains two databases. There are 10 classes and each class has 40 clips in ESC-10 dataset. ESC-10 is subset of ESC-50 which contains 50 classes and each class has 40 clips. There is a readme file in which there is a line for each clip explaining the details of the clip for each dataset [18].

In [19], a dataset of annotated urban sound clips are recorded and taxonomy for urban sound sources is proposed. The dataset contains 10 sound classes with 18.5 hours of annotated sound event occurrences. The dataset contains a CSV file (Figure 6) explaining the details of each recording.

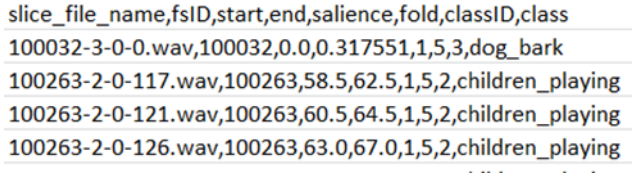


Figure 6: CSV file explaining Urbansound database

Another ESR database is http://www.desra.org. In [20], the aim of this database, details of the design and the sources used are

explained. It is designed as multi-purpose database. The database contains variety of sounds from different events with thorough background information for each sound file. It is accessible from the Internet (The database is not fully functional at the moment). The database was designed considering for admin tasks and general user level tasks. Web front end provides the functionality for user level tasks.

http://www.auditorylab.org is [21] another database. This database was constructed by Carnegie Mellon University to examine the human ability to use sounds to understand what events are happening in the environment. All the sounds in the database are recorded in a controlled way. The laboratory and the recording media used are technically detailed. In the construction of database sounds are grouped by the event that makes the sound. Sounds of events like impact, rolling can be downloaded from this database.

3. Database Design

The databases reviewed in Section 2 are prepared for just resources for the development of ESR applications. Many of them provide the data and the files for correctly handling the data. In this context they are valuable resources for all ESR researchers. Desra [20], provides extra tools such as a web graphical user interface for searching and testing.

In the development process of an ESR system, some main functions training, testing and production explained in Section I. During this development process, many sound clips are used; new sound clips may be added or deleted. We deal with many features extracted from these sound clips; we create models using different ML algorithms. Then we compare the models; try to find best features and best ML algorithms. This loop continues. Our first goal in this database design is to help researchers as much as possible to ease the burden of handling data. The second goal is to help maintenance of the data. Our data is sound clips, features, algorithms and models. Addition, deletion, searching, annotation, backup and restore can be thought as the maintenance task.

3.1 Taxonomy of Environmental Sounds

We need a taxonomy to be able to store, search and retrieve the data from the database. During literature review we see two kinds of taxonomy. In the first taxonomy, the sounds can be grouped by the event that makes the sound. In Figure 7, grouping of sounds defined in [21] is seen.

title	author	type	modified
LEGAL NOTICE	Auditory Lab User	Page	2009-06-24 21:16
Impact Events	Auditory Lab User	Folder	2009-06-25 10:15
Deformation Events	Auditory Lab User	Folder	2009-06-25 10:16
Rolling Events	Auditory Lab User	Folder	2009-06-25 10:17
Air Events	Auditory Lab User	Folder	2009-06-25 10:17
Liquid Events	Auditory Lab User	Folder	2009-06-25 10:18
Experiment Stimuli	Auditory Lab User	Folder	2009-06-25 10:19

Figure 7: Grouping of sounds based on sound producing events [21]

Another hierarchical grouping is seen in Figure 8. This is also classified in the first taxonomy. This is the grouping defined in [22] based on sound producing events and the listeners' description of the sound. Second taxonomy is based on the sound source [19][23]. The subset of the taxonomy defined in [23] is taken into consideration for urban sounds given in [19].

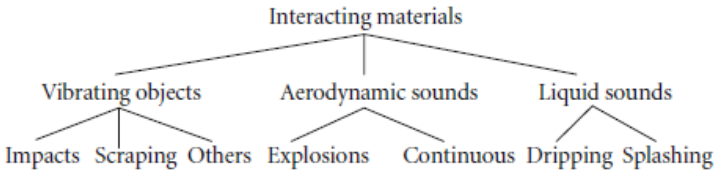


Figure 8: Grouping of sounds defined in [22]

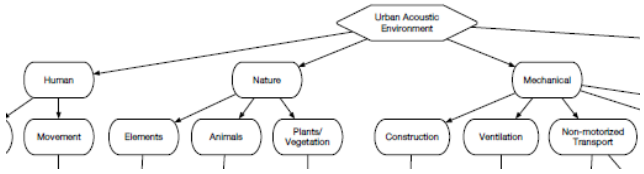


Figure 9 A part of the urban sound taxonomy defined in [19]

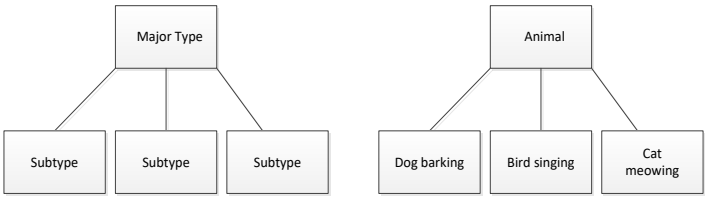


Figure 10: Environmental sounds grouping in the database

3.2 Non Functional Requirements of the Database

- There will be sound clips in the database. These sound clips will contain environmental sounds. Each sound clip can have more than one environmental sound. These environmental sound clips can belong to different major environmental sound types. Origin of these sound clips, such as own recording, from an internet site or from another database, should be entered to database. Recording details should be entered. Size, file type of sound clip and the path where the sound clips are recorded should be entered to the database.
- There should be major types, such as human, nature, animal, etc. There should be subtypes, such as dog bark, gunshot etc. and these must belong to major type. Start and end sample index or start/end time of these clips should be known.
- There will be environmental sound clips in the database. These clips should be extracted from sound clips defined in first paragraph. Each of these sound clips should have a subtype and it should be known from which sound clips it

is extracted. Extraction method should be entered. If it is extracted by an algorithm the algorithm name otherwise as "manual" should be entered to database.

- Recording details should be entered to the database such as, sampling rate, quantization bit, channel size, etc.
- Background clips should be entered to the database. Each background clip should have a type. Each background clip should also have file type, file size, recording detail and the path where it is recorded.

3.3 Functional Requirements

- There should be scripts which will take algorithms as arguments and extract the environmental sounds from sound clips. These extracted clips should be entered to the database as explained in 3.2.
- There should be scripts which will embed environmental sound into noise clips at desired Signal to Noise Ratio (SNR) level.
- There should be scripts which will extract features from environmental sound clips and store to the database.
- The scripts, their help files, paths should be stored in the database.

3.4 Graphical User Interface (GUI) Requirements

- Administrators can use the GUI for meta data entrance, deletion and update such as major types, subtypes, recording details.
- Administrators can use the GUI for data load manually or using the data loader script.
- Researchers can use the GUI for searching and downloading the desired environmental sounds.
- Researchers can load their features and models to the database.

3.5 Database Implementation

The implemented database will include some data which can be stored by way of data types found in the standard database software and also sound files. These sound files will not be stored to database instead they will be stored in the operating system file structure and the path to this file is just recorded in the database.

Microsoft SQL 2008 Database Server is used to create tables. The tables and the relationships between them are shown in Figure 11. The scripts providing functionality are coded in Matlab 2011a. The files are stored in "mat" type when required.

For functional requirements the following scripts are coded using Matlab.

- Environmental Sound Detector: This script is an interface between algorithms that extracts environmental sounds from the sound clips. Different algorithms can be used here for extraction of the environmental sounds. The algorithms must conform to this script interface definition. This script takes the algorithm name that will be started and the environmental sound types which will be searched are given as arguments.
- Environmental Sound Embedder: This script takes the type of the environmental sound, noise type, SNR level and at last the number of required record count to be created. The script merges the environmental sound clip with the noise clip at the desired SNR level and records it to the database table.
- Feature Extractor: This script acts as an interface between feature extractors. It takes the path of the feature extractor from the features table, feature name and the environmental sound type from command line. After extracting features, it is saved as a mat file.
- Data Loader: To load data from other databases this script is used to interface with data loading scripts.

The GUI is developed using Microsoft Visual Studio with C# language. GUI has admin utilities and end user tools. In Figure 12(a), it is seen the part of GUI providing admin operations. These admin operations are additions and deletions of major types, subtypes and sampling information. Besides the GUI provides the administrators load data one by one or using a script to load as a batch.

The GUI provides the users some tools. The tools are for searching the database and testing their algorithms on the sound clips given by the database. Users can search the database to see the sound clips with desired type.

End users can search the database with noise clip type, sound clip type and with SNR interval then select a clip and copy it to their own computer. After finding the start index of the embedded environment sound clip, by writing the start index of this environmental sound to the edit box on the GUI and by clicking the test button, they can test their algorithm correctness. The WEB GUI for end user tools is seen in Figure 12(b).

After database and scripts implementation we loaded the data from Urbansound [19] database. Now many operations on the database can be done either using GUI or an SQL command line.

4. Conclusion

The lack of common database for environmental sound recognition is an important obstacle in front of the researchers. The development of and ESR system is a tough process during which someone have to deal with lots of sound clips with different types, algorithms and models. In this paper we explained a relational database model which will make the data handling easier. The database developed is different from other counterparts which are just providing the data. The relational database model described here provides easy maintenance as well as easy usage.

Although our goal of designing this database is for ESR, other areas that deal with environmental sounds can use it.

By improving the database by adding data from general databases mostly used and by adding more functionality we aim it to be a common database for research activities of ESR.

Conflict of Interest

The authors declare no conflict of interest.

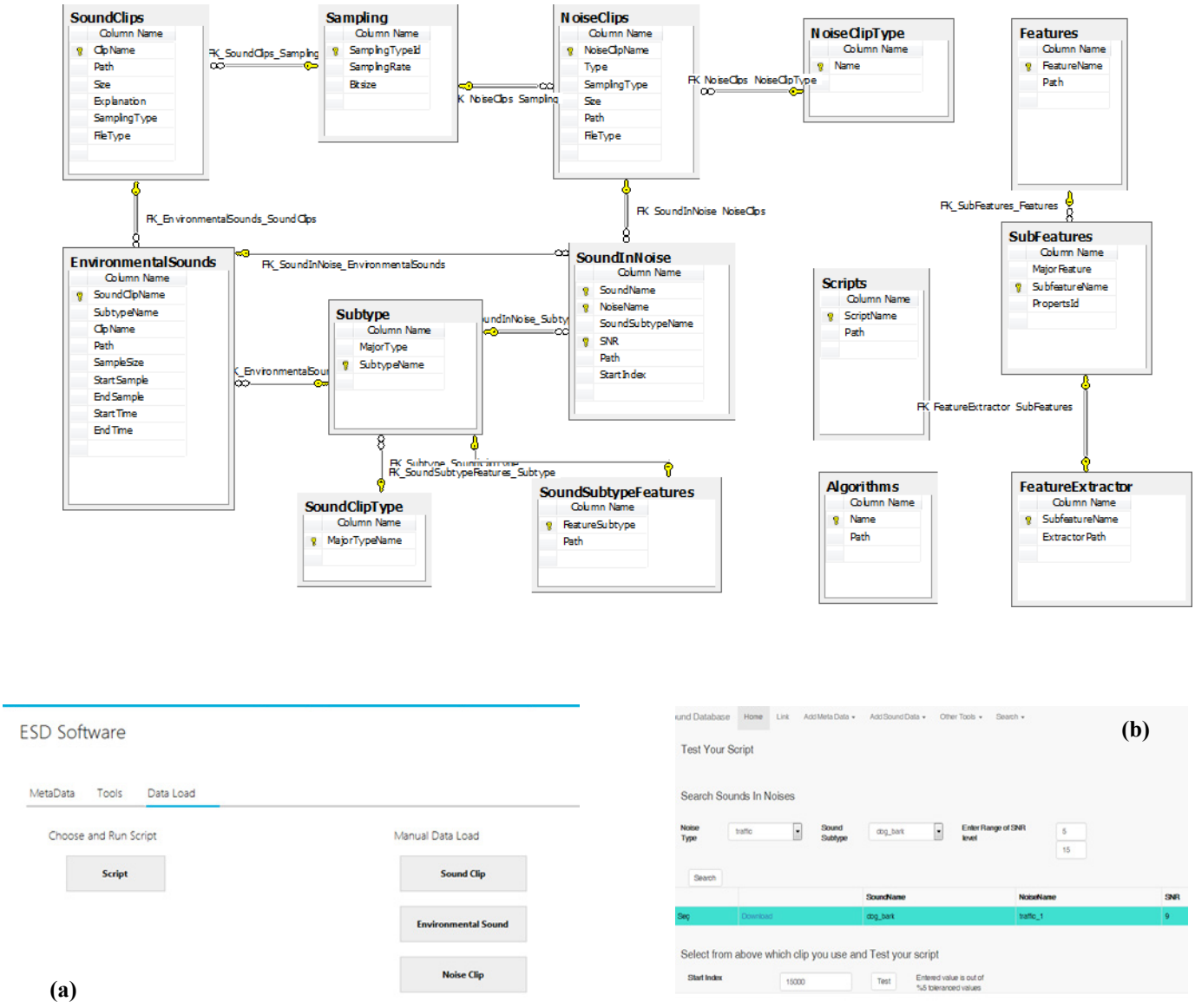


Figure 12 (a) Desktop application GUI for administrators to data load (b) WEB GUI for end user tools

References

[1] Y.Arslan and H. Canbolat, "A sound database development for environmental sound recognition", Signal Processing and Communications Applications Conference (SIU), 25th, 2017.

[2] T. Ahmed, M. Uppal and A. Muhammad, "Improving Efficiency and Realibility of Gunshot Detection Systems", IEEE, ICASSP 2013.

[3] P. Thumwarin, T. Matsuura and K. Yakoompai, "Audio forensics from gunshot for firearm identification", Proc. IEEE 4th Joint International Conference on Information and Communication Technology Electronic and Electrical Engineering Tailand, pp. 1-4, 2014.

[4] Chu, S.;Narayanan, S.; Kuo, C.-C.J.;Mataric, M.J.:Where am I? Scene recognition for mobile robots using audio features, in 2006 IEEE Int.Conf. on Multimedia and Expo. IEEE, 885–888, 2006.

[5] Yamakawa, N.; Takahashi, T.; Kitahara, T.; Ogata, T.; Okuno,H.G.: Environmental sound recognition for robot audition using Matching-Pursuit, in Modern Approaches in Applied Intelligence, in K.G. Mehrotra, C.K. Mohan, J.C. Oh, P.K. Varshney & M. Ali (Eds), Springer Berlin Heidelberg, 1–10, 2011.

[6] Chen, J.; Kam, A.H., Zhang, J.; Liu, N.; Shue, L.: Bathroom activity monitoring based on sound, in Pervasive Computing, in H.W. Gellersen, R.Want, & A. Schmidt (Eds), Springer Berlin Heidelberg, 47–61, 2005.

[7] Vacher, M.; Portet, F.; Fleury, A.;Noury, N.: Challenges in the processing of audio channels for ambient assisted living, in 2010 12th IEEE Int. Conf. on e-Health Networking Applications and Services (Healthcom), IEEE, 330–337, 2010.

[8] Wang, J.-C.; Lee, H.-P.; Wang, J.-F.; Lin, C.-B.: Robust environmental sound recognition for home automation. Automation Science and Engineering, IEEE Transactions on, 5 (1) (2008), 25–31.

[9] Bardeli, R.;Wolff,D.; Kurth, F.; Koch,M.; Tauchert, K.-H.; Frommolt, K.-H.: Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recognit. Lett., 31 (12) (2010), 1524–1534.

[10] Weninger, F.; Schuller, B.; Audio recognition in the wild: static and dynamic classification on a real-world database of animal vocalizations. in 2011 IEEE Int.Conf. on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2011, 337–340.

[11] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, and N. Petkov. Car crashes detection by audio analysis in crowded roads. In Advanced Video and Signal Based Surveillance e (AVSS), 2015 12th IEEE International Conference on, pages 1-6, Aug 2015.

[12] J. Rouas, J. Louradour, and S. Ambellouis, "Audio Events Detection in Public Transport Vehicle," Proc. of the 9th International IEEE Conference on Intelligent Transportation Systems, 2006.

[13] R. Radhakrishnan and A. Divakaran, "Systematic acquisition of audio classes for elevator surveillance," in Image and Video Communications and Processing 2005, vol. 5685 of Proceedings of SPIE, pp. 64–71, March 2005.

[14] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, "Audio based event detection for multimedia surveillance," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06), vol. 5, pp. 813–816, Toulouse, France, May 2006.

[15] Nakamura, S.; Hiyane, K.; Asano, F.; Nishiura, T.; Yamada, T.: Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, in LREC, 2000.

[16] DCASE 2016 WEB site available at: http://www.cs.tut.fi/sgn/arg/dcase2016/ (27.09.2017 last accessed)

[17] DCASE 2017 WEB site available at: http://www.cs.tut.fi/sgn/arg/dcase2017/ (27.09.2017 last accessed)

[18] K. J. Piczak, "ESC: Dataset for environmental sound classification", Proceedings of the ACM International Conference on Multimedia, 2015.

[19] Justin Salamon , Christopher Jacoby , Juan Pablo Bello, A Dataset and Taxonomy for Urban Sound Research, Proceedings of the 22nd ACM international conference on Multimedia, November 03-07, 2014.

[20] Brian Gygi and Valeriy Shafiro, "Development of the Database for Environmental Sound Research and Aplication (DESRA):Design, Fuctionality and Retrieval Considerations", EURASIP Journal on Audio, Speech, and Music Processing Volume 2010.

[21] http://www.auditorylab.org (28.09.2017 last accessed)

[22] W. W. Gaver, "What in the world do we hear? An ecological approach to auditory event perception," Ecological Psychology, vol. 5, no. 1, pp. 1–29, 1993.

[23] A. L. Brown, J. Kang, and T. Gjestland. Towards standardization in soundscape preference assessment. Applied Acoustics, 72(6):387-392, 2011

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 146-156 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Limitations of HVAC Offshore Cables in Large Scale Offshore Wind Farm Applications

Tiago Antunes*,1, Tiago Alexandre dos Reis Antunes[1], Paulo Jorge da Costa Santos[2], Armando José Pinheiro Marques Pires[3]

[1]*Siemens, S.A. & MIT Portugal Program, Sustainable Energy Systems, Universidade Técnica de Lisboa, 2744-016, Portugal*

[2]*EST Setúbal Instituto Politécnico de Setúbal / INESC Coimbra, Electrical Engineering, 2914-76, Portugal*

[3]*EST Setúbal Instituto Politécnico de Setúbal / UNINOVA, Electrical Engineering, 2914-76, Portugal*

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 24 January, 2018
Online: 17 March, 2018

Keywords:
Alternate Current
Transmission Line Simulation
Offshore Wind Farms

ABSTRACT

The energy marathon is becoming increasingly based on renewable sources, whereas the continuous decrease on the cost of energy production has supported in last decade, for instance, the development of large-scale offshore wind applications. The consistency and availability of the AC-working equipment portfolio is only limited by physical application boundaries, which are quite evident in this sort of accomplishments. The focus of this article is to present a tool developed under the MATLAB environment which allows for a quick and real-time analysis of HVAC links, assessing the impact of voltage, current, power factor or distance conditions. The conclusions are drawn directly by means of a stress-point and operational acceptance range.

1. Introduction

This paper is an extension of work originally presented in 2017, 11th IEEE International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG) [1]. Offshore wind applications are clearly under the spotlight. The wind conditions, the ability to increase the rated power of the turbines and thus the size of the plant, in addition to the operational and technical advantages of having the units dispersed over the sea, provide the arguments to position it as one of the viable alternatives for the energy demands of the future [2]. However, there are known limitations for the Alternate Current (AC) Transmission Links.

As wind turbines now compete with gas turbines – in rated power – and the cost of production is consistently approaching the one of traditional carbon-based plants [3], innovation pushes the offshore power flow to the MW-unit level. That, added to the long distance the submarine cables should meet, reduces, in a sense, the reliability of traditional AC solutions [4]. This is the starting-point of this investigation. Would that information provide the basics, the goal is clear – the model aims to provide the operational range and the restrictions for any given AC link (meaning a set of initial conditions). Moreover, that was proven based on the results obtained shown later.

AC-based plant-to-shore links do not provide a one-size fits-all option and a careful analysis still is mandatory to ensure the proper conversion of the bulk wind onto usable power preferably with limited losses.

Although it is common knowledge that current HVAC Overhead Lines (OHL) can easily be deployed in distances in upwards of 100km, the same is not true for insulated-cable based lines, especially in offshore applications, which are based in submarine cables. For both of these cable types, their fabrication characteristics – hence the electromagnetic characteristics seriously restrict the application of AC on the offshore [5]. Once more, the goal of this paper is to present a model that allows for a reliable estimation of the maximum operational distance in such regard without requiring a time-consuming case-by-case investigation and still providing detailed results.

This paper divides in seven sections. The first part provides a description of the wind farm (WF) model under appreciation, which constitutes the baseline for the operation of the model. The intention is to provide a set of results, for equivalent air and cable-based lines, and ultimately, also to allow for an extrapolation for other configurations. The transmission-line equivalent and the GUI-based analysis models are presented in sections 2 and 3, respectively.

Sections 4 to 6 present the experimental results of the analysis performed, for standard OHL submarine cables, as well as, for

*Tiago Antunes, Email: tiago.antunes@siemens.com

offshore compensated-cable solutions. The last chapter includes the conclusions.

2. Simplified Model of the Offshore Wind Farm

2.1. Characteristics of the Wind Farm

This section consists of two main parts; The first one aims to present the simplified model of the wind farm under consideration, whereas the second one introduces the line model, later on used on the MATLAB-based software (section 3). Based on actual available portfolio [5] and existing projects [6], a decision was made to assess the link of an 80MW offshore wind cluster, composed by ten 8MW turbines. The schematic of the model is shown on Figure 1.

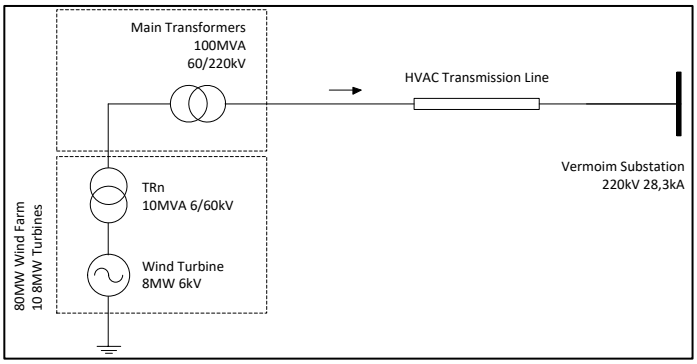


Figure 1. Layout of the Wind Farm

As shown on the figure above, the authors chose to use a 220kV link to the shore, given the rated power of the wind farm and the existing transmission infrastructure on that region. REN' Vermoim Substation located approximately 40km from the district capital Oporto was the onshore Point of Connection (PoC). The tables 1 and 2 present, respectively, the characteristics of the wind farm and the ratings at the given substation (at the 220kV busbar).

Table 1. Characteristics of the Wind Farm

Nr. Units	$S_{N\text{-}Turb}$ [MW]	$S_{N\text{-}Farm}$ [MW]	L_{Link} [km]
10	8	80	200

Table 2. Short-circuit ratings at Vermoim 220kV Busbar

$S_{CC\ MAX}$ [MW]	$S_{CC\ MIN}$ [MW]	$I_{CC\ MAX}$ [kA]	$I_{CC\ MIN}$ [kA]
10.7	8.3	28.3	21.7

2.2. Characteristics of the HVAC Link

The usage of insulated cables as the means to transmit power in HVAC is understandingly common. Nonetheless, offshore (submarine) applications have a different behavior that ultimately and most importantly, significantly reduce the acceptable distance of application. Limitations, such as high charging current and higher *Ferranti* effect [8], are widely known, the model must allow for the analysis of several scenarios, with real-time adjustment and performance assessment, still ensuring a less time-consuming computation of each one.

There are two main cable-sizing criteria, which are the rated voltage and admissible temperature [8] and the model is based on this simplification.

The nominal current is given directly by the rating of the wind farm whereas the fault current value is given at the HV OHL-bay at the onshore substation as it is the most significant contributor in case of short-circuit at the PoC. Both allow for an estimation of the minimum cable cross-section [7]:

$$I_N = \frac{P_N}{\sqrt{3}\,V_N \cos\varphi} = 262{,}43\ \text{A} \quad (1)$$

$$s_{min} = \frac{I_{CC}\sqrt{t}}{K_c} = 521{,}3\ \text{mm}^2 \quad (2)$$

Where *t* represents the fault-clearance time.

Having this as a reference, and given the cable standardization, a core section of 630mm^2 shall be selected. The OHL and cable will be selected having such as a reference. The upstream grid is not taken into consideration on the analysis (meaning the onshore PoC is assumed a PV-type node).

3. Software Implementation

3.1. Electrical Line-equivalent Model

When in comparison between lumped-sum or parabolic line-equivalent models, the differences are significant. In the case of OHL, the parabolic components of the equivalent are almost unitary, which allows for simplification and a lump-sum provides acceptable results [8].

In cable applications in general, the wave impedance is higher and as so the propagation constant decreases, resulting in non-unitary trigonometric components. These models compute the opposite-end results based on a linear and concentrated impedance section, the propagation speed is not taken into account (which drives from the simplification itself).

In order to proper translate the effects of submarine cable applications, the model has to take into account the influence of the propagation speed and the geometrical distance. An accurate model has to take effectively into account the distribution and variation of the parameters along the line, being therefore distance and transmission velocity- dependent (which lastly decreases for a higher cable length).

This knowledge is widely applied in very-fast transient analysis for atmospheric discharges or other high-velocity voltage and/or current variations on transmission lines. An *x* section of the line is presented in Figure 2.

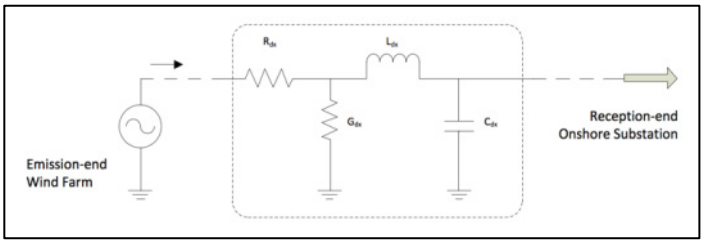


Figure 2. Infinitesimal section of the cable ([9], adapted)

As the results of that equivalent scheme are much more accurate, the respective equations (3) to (6) are used on the MATLAB model. These, on the first approach, provide a distance-based analysis of a given transmission link in regards to estimated voltage and current:

$$Z_L = (R + jX)\, l \text{ [ad]} \tag{3}$$

$$Y_T = (G + jB)\, l \text{ [ad]} \tag{4}$$

The voltage and current for *x* distance are given by:

$$V_x = V_1 \cosh(\gamma x) - Z_0\, I_1 \sinh(\gamma x) \text{ [kV]} \tag{5}$$

$$I_x = -V_1 \frac{1}{Z_0} \sinh(\gamma x) + I_1 \cosh(\gamma x) \text{ [A]} \tag{6}$$

All data shown on the GUI (hence plot options) are also based on a set of calculations detailed on the script. The calculations are performed for any given initial conditions, non-related to any previous tests – although a multi-set analysis can be provided, as shown on sections 4 and 5.

3.2. Model Design

The necessity to address a multitude of scenarios to draft earlier conclusions has led itself to the conclusion that a software, rather than casual calculation, had to be taken into practice. The development of a fast and custom-made tool to the analysis the design of HV transmission lines led the authors almost directly to the implementation of a GUI-based MATLAB model.

That allowed for the creation of a versatile and digital tool, which allows for a fast and straightforward verification, given a set of any initial conditions, if power transmission via a selected line (or power cable) is acceptable in technical terms. And, the model spans virtually to any range of, not only cables, but voltage, current, power factor or distance.

The possibility of checking a certain scenario in real-time and in on-hand is another plus, possible via the multitude of adjustments possible on the model itself, with instantaneous calculation and results. As the GUIDE is, basically, the user-friendly window of MATLAB C-based programming, also quite an user-friendly design was achieved.

In practical terms, the model operates by taking into account a set of transmission line conditions – stored on pre-loaded, yet adjustable .XLS sheet, of OHL and cables, then applies the equivalent model earlier addressed (section 3), and computes all the results that illustrate the steady-state operation of a given transmission "line". That includes voltage, current, no-load voltage & current, voltage drop, efficiency (losses evaluation), power factor, stability, surge impedance level (SIL) and model parameters (Z_L & Y_T).

3.3. Model Key Advantages

The segmentation and expandability of the model can be defined as two of the most important characteristics. All the information used and key features, aside from cables, are actually .XLS-based. The model starts and loads-up the majority of its functionalities with reference to the options and fields set on the configuration Excel sheets.

Features such as the stress parameters used for the calculation – further on shown – or the actual regulations, such as voltage or reactive power boundaries, can be adjusted with ease. The model also loads using acceptable and recommended voltage, current, power factor and distance values, further facilitating the usage. The plot section, on the other hand, allows for the simultaneous visualization of three different calculation windows, thus allowing the user do address a scenario in a much more complete manner.

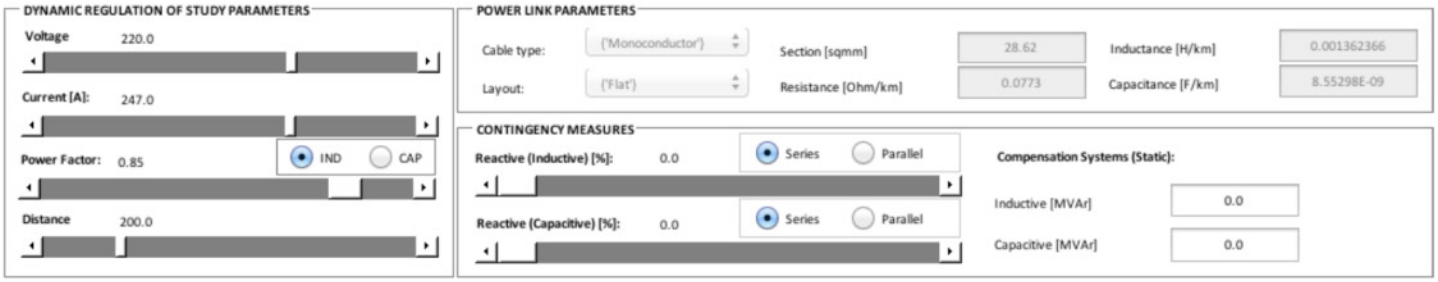
Finally, the information provided has a key focus. As said, the ability to implement the line-equivalent is wide-spread in literature. The difference is to provide such information and the data outcomes in simpler yet much more complete manner. That information is provided to the user using overlay layers on the results plotted, and may include operational boundaries, compensation expectations, stress-range distance pin-pointing, or even, simultaneous calculation results – for example, for difference one-end power factors. The summary of the available results is shown on Table 3.

Table 3. Available Results of the Model

Ref	Description
1	Powerline Voltage
2	Powerline Current
3	No-load Current
4	No-load Voltage
5	Voltage Angle
6	Current Angle
7	Voltage Phase Displacement
8	Current Phase Displacement
9	Active Power
10	Reactive Power
11	Apparent Power
12	Transmission Losses
13	Transmission Efficiency
14	Voltage Drop
15	Power Factor
16	Tan-Phi
17	Stability Limits (Pe-Phi)
18	Voltage Drop (DV-PF)
19	SIL (impedance)
20	SIL (apparent power)
21	Longitudional Impedance
22	Transverse Admittance

The critical factors that guide the operational ranges are shown in Table 5. Generically (and, in fact changeable), the critical factors are presented based on two conditions: "C" (meaning Critical, for 80% of target) and "F" (for Failure, or 100% of target value). Other conditions may apply as mention on Table 5. The numbers refer to the references and 80/100 thresholds can be adjusted as the intention of the user with minor core programming adaptation.

Once again, model design is as simple yet flexible and expandable as possible. The majority of the reference and configuration-data is XLS-based and new features are easy to implement. The same can be done for the (line) data, as the routine reads any given organized sheet and performs the calculation. After the automated analysis is complete, the user can make a search, distance or variable-dependent, for a value-at-point or given location of the "line".



As observed in Figure 3, the real-time interface is possible via the adjustable sliders – which adapt the basic parameters. Re-calculation is done fast and automatically after any change on those. The advantages of a tailor-made GUI model span for other areas. The plotting windows included have been adapted to allow for the *online* setting of the units to be used: International system (SI) units, per-unit (P.U.) or %-based results.

The adjustment of the ranges and precision of the axis is adjusted directly on each respective axis line. Exporting is also made directly from the working plotting window, allowing for quick integration on written reports. This is, once more, allowed by a strong background interface with .XLS files, and further on, with facilities such as additional interfaces, adaptation, parameter-variation mode and export facilities, are all intricate on the model.

Table 4. Critical Factor References for Model

Ref	Variable Checked	Slope	Critical Value	Failure Value
1-4	Line Voltage	>=	REN regulations	
5-8	Line Voltage	<=	REN regulations	
9	Line Current	>=	Thermal limit of cable	
10-13	No-load Voltage	>=	REN regulations	
14-17	No-load Voltage	<=	REN regulations	
18	No-load Current	>=	50%	100%
19	Reactive Power	>=	40%	50%
20	Losses	>=	40%	50%
21-22	Voltage Angle	>=	±30º	±40º
1-4	Line Voltage	>=	REN regulations	
5-8	Line Voltage	<=	REN regulations	
9	Line Current	>=	Thermal limit of cable	
10-13	No-load Voltage	>=	REN regulations	
14-17	No-load Voltage	<=	REN regulations	

What is also worth mentioning is the ability to automatically see the stress ranges on a color-coded plot-overlay for any given calculation. These ranges are recognized and the maximum operating distances are identified. Ranges are loaded based on the Portuguese regulations (e.g. REN) and those can also be adjusted.

For a comparison to be made on different transmission solutions a simulation is made on a 220kV air-insulated single-circuit single-pole transmission line. The rated voltage is selected taking once again into account the grid topology at the PoC.

4. Evaluation of Air-Insulated Connection

4.1. Parameter Calculation

The overall manufacturer-issued characteristics of such cable, later on introduced on the MATLAB model, are shown on Table 5. The maximum active power output is stationed at 80MW (having a cos φ of 0,85 lag). The matrix therefore can be calculated as presented in equations (7) and (8).

Table 5. Characteristics of the simulated Overhead Line

U_N [MW]	r [mΩ/km]	L [mH/km]	C [nF/km]
220.0	77.3	1.362	8.553

$$Z_L = (R + jX)\, l = 10.381 + \text{j}26.3894 \text{ [ad]} \quad (7)$$

$$Y_T = (G + jB)\, l = \text{j}10.1 \times 10^{-3} \text{ [ad]} \quad (8)$$

These values are automatically calculated on the model.

4.2. Experimental Results

Computation is done assuming the reference data at the reception-end (onshore PoC). The simulation of the voltage and current at the emission-end of the line are shown on Figures 4 and 5. Ultimately, the results confirm the reliability of the model with existing accurate OHL-analysis data [9]. The calculation method is than transposed to an offshore (submarine) framework, as presented on section 5.

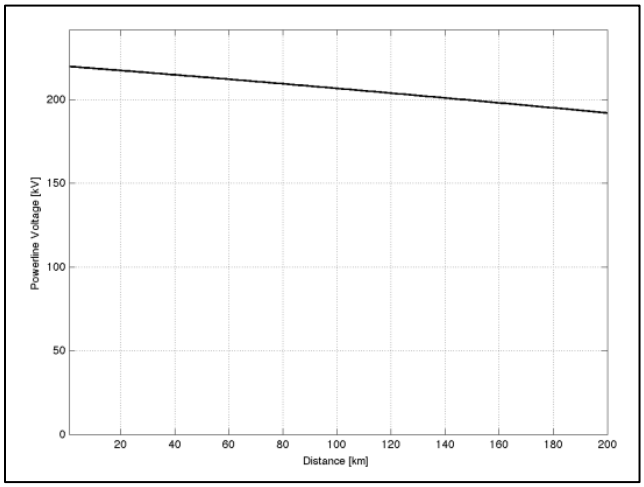


Figure 4. Powerline voltage simulation (220kV OHL)

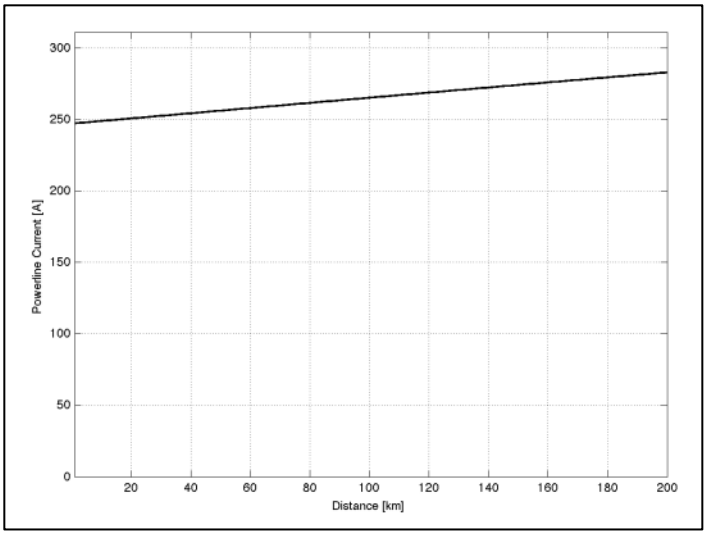


Figure 5. Powerline current simulation (220kV OHL)

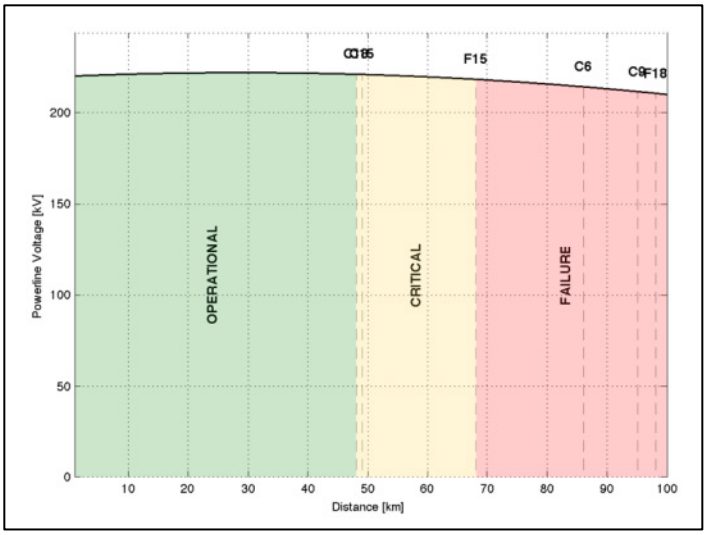


Figure 6. Powerline voltage simulation (220kV Submarine Cable)

The analysis of the submarine cable, being the focus, shall include much more significant data, however all being supported on the positive result of the voltage & current simulation shown on these two test plots.

5. Evaluation of Submarine Cable Connection

5.1. Parameter Calculation

The analysis of the submarine cable, hence parameter calculation, is done on the same basis as before. The cable data obtained from the manufacturer (ABB) is presented in Table 6. Cables are laid in a single pole linear and evenly distributed along the seabed.

Table 6. Characteristics of the simulated Submarine Cable

V_N [kV]	I_N [A]	Size [mm2]	R [Ω/km]	L [mH/km]	C [μF/km]
220	600	630	0.04935	1.4000	0.1600

$$Z_L = (R + jX)\text{l} = 9.87 + \text{j}87.9646\ [\text{ad}] \quad (9)$$

$$Y_T = (G + jB) = \text{j}10.1 \times 10^3 [\text{ad}] \quad (10)$$

Once again, these are automatically calculated on the model. Equations (5) and (6) apply for the calculation of all data plotted, including the calculation of the ABCD equivalent model parameters.

5.2. Voltage & Current Results

The operation, critical failure ranges can be seen on the figures shown. The reference for criticality/failure leads back to the REN regulations on the voltage boundaries. The same is shown on the current plot. A distance shorter than 100km is also sufficient to show the relevant information yet that range facilitates the condensation of data on the plot snapshot.

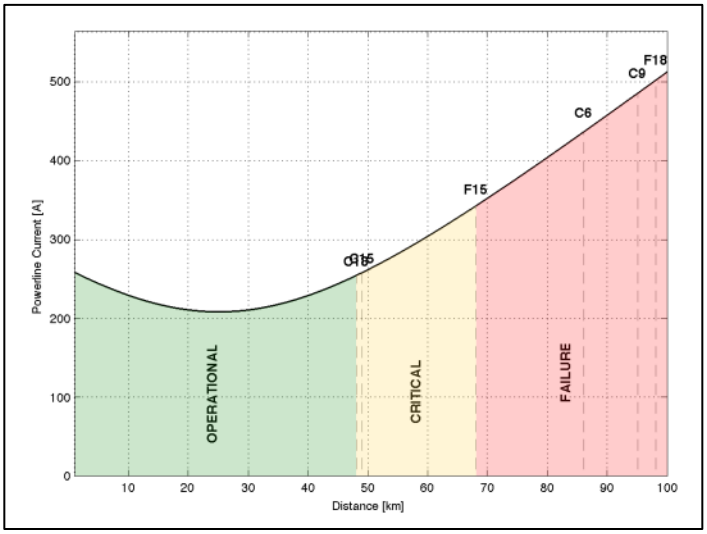


Figure 7. Powerline current simulation (220kV Submarine Cable)

The emission-end (WF) values are 210kV (Fig. 6) and 513A (Fig. 7). As it can be verified, the voltage drop reaches close to 5%. As for the current, it's final value is around two times the initial current value at the reception. This allows for an insight of one of the most important downsides of cables – the charging effect.

As far as the current values are concerned, the high capacitive characteristics of the submarine cable, lead to a significant increase of the line' current and emphasize the reactive power flow. This not only increases the power losses along the line but also confirms that cable solutions are not feasible after 30-50km [10] alongside theoretical information available.

5.3. Voltage Drop on Underground & Submarine Cables

As a basis of comparison, Figures 8 and 9 include the voltage drop along transmission cables using underground (8) and submarine (9) solutions. Evidently, as expected, the reduction of the voltage is much more significant on the second option.

The authors emphasize that, on both land and sea cables, the voltage drop encountered is actually a voltage rise along the line, due to the capacitive characteristics of the cables themselves. This has to be later on compensated on the WF-side, forcing the generators to operate under very low voltages.

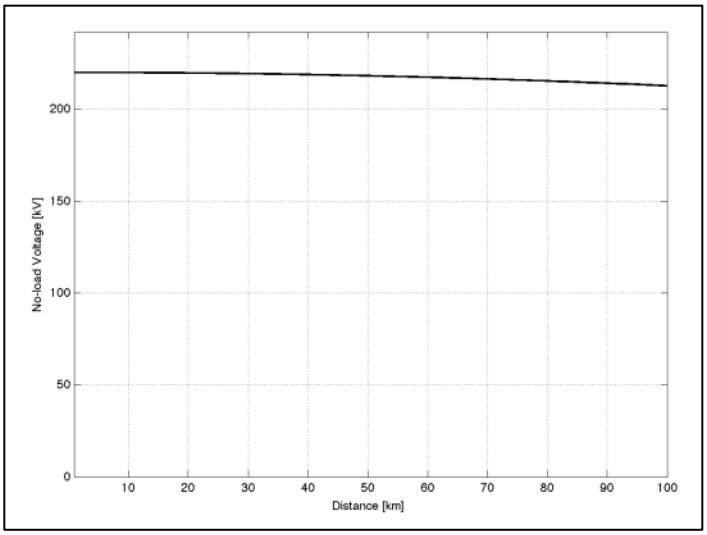


Figure 8. No-load voltage simulation (220kV Underground Cable)

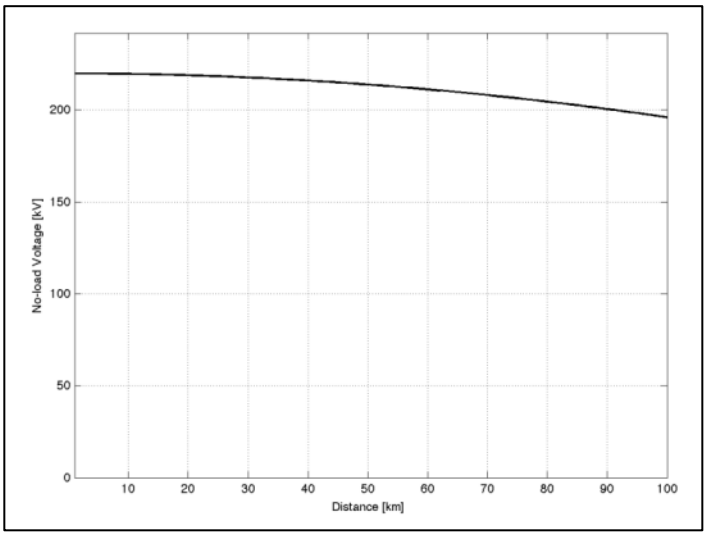


Figure 9. No-load voltage simulation (220kV Submarine Cable)

The inductance value has a tremendous influence on the voltage and current behaviour of the transmission line, but with more focus on the first one mentioned. Aside from the voltage decline – and as addressed earlier – with a higher impact, the charging and Ferranti effects are evaluated for this particular cable.

5.4. Charging Current & Ferranti Effect

The charging and Ferranti effects are the two most distinguishable plots which can be made from the behavior of insulated cables (both underground and submarine. The high capacitance of these cables leeds to a early increase on the current flow and therefore the transmittable power is severely restricted. No-load current, based on the F18 marker on the plot, is actually the failure cause of the cable under investigation.

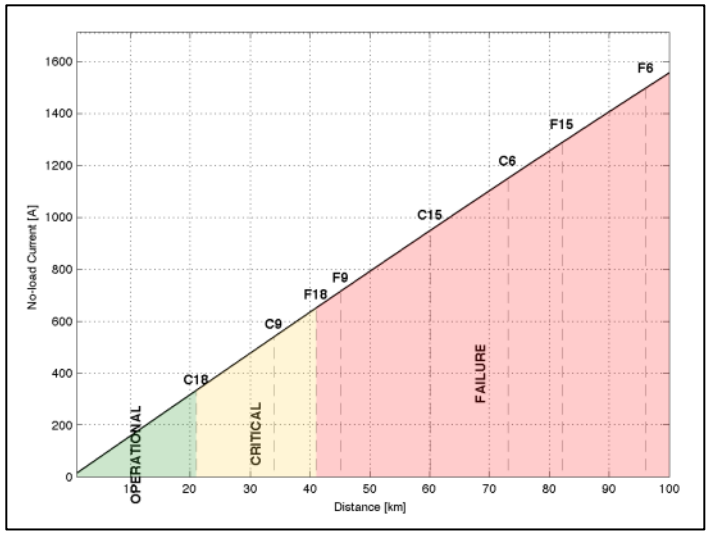


Figure 10. No-load current simulation (220kV Submarine Cable)

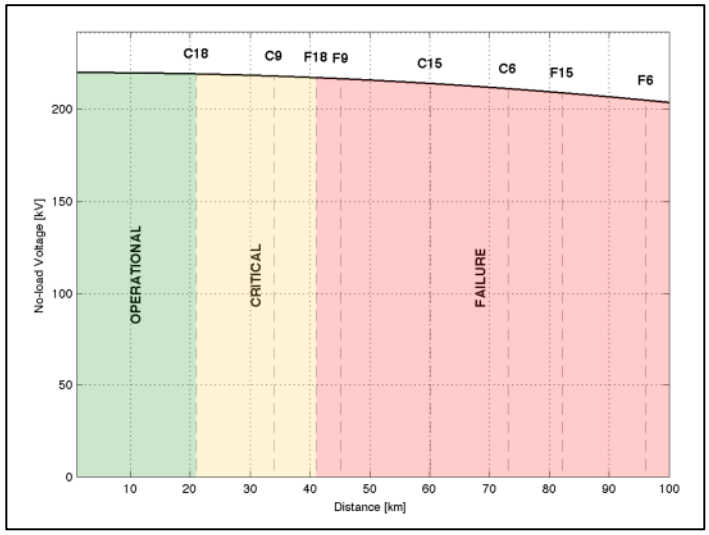


Figure 11. No-load voltage simulation (220kV Submarine Cable)

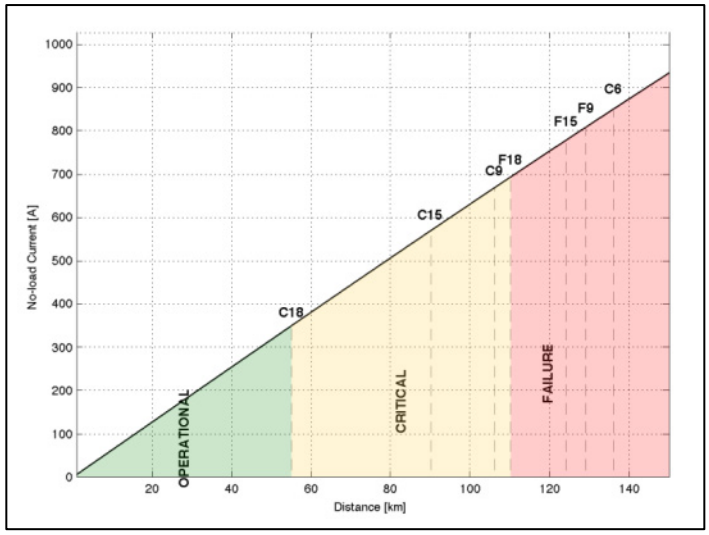


Figure 12. No-load voltage simulation (220kV Submarine Cable)

When compared to the equivalent underground cable included on the examples of the model, the differences are tremendous. The effect on the submarine option is almost 100% more. The operation of such "large capacitor" under no-load conditions also translates onto a voltage increase along the same cable – as shown previously on Figure 9.

5.5. Power Factor & Reactive Power Results

The behaviour of the power factor for a given section of the transmission cable is highly dependent on the initial conditions, especially the reception-end (PoC) power factor. For that reason, this section also includes a multi-scenario analysis, thus allowing for a better understanding of such impact.

A very high drop on the power factor is observed (close to 50% of the initial value), to a minimum of 0,44. Due to the increasing current, the active power still increases. However, it is the reactive power which increases the most, leading again to the failure of this setup for more than 70 kilometres (F15).

What is also verified is that, up to a certain distance, the capacitive characteristics of the line help to compensate for the inductive grade of the load (0,8 lag). Once the characteristics of the transmission line match the PoC, there is an inversion on the slope.

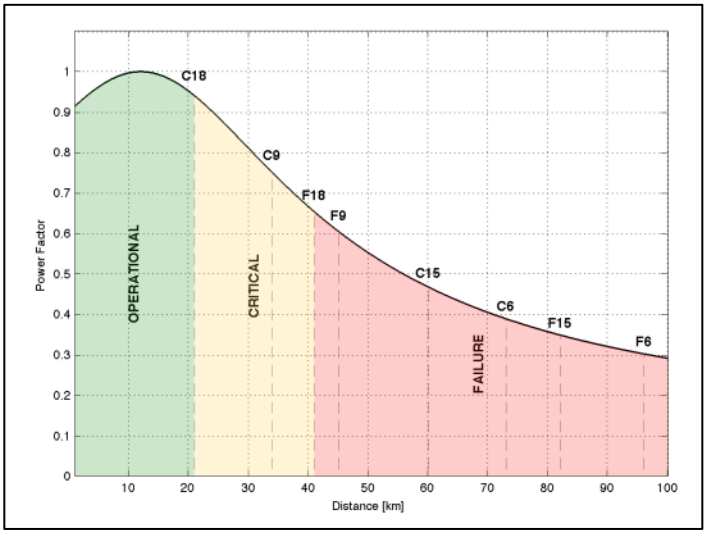


Figure 13. Power factor simulation (220kV Submarine Cable)

What is also verified is that, up to a certain distance, the capacitive characteristics of the line help to compensate for the inductive grade of the load (0,8 lag). Once the characteristics of the transmission line match the PoC, there is an inversion on the slope. After a certain distance (around 25km), the cable is actually acting as a reactive power provider. This fault is not triggered because, in percentage, it is still below the 50% threshold shown on Table 4.

What can also be understood from Fig. 8 is that the unitary power factor is obtained for a distance just under 15km. Despite failure only occurs after 40km (once more confirmed by literature), because the line has a significantly higher capacitance than the OHL, it acts much like a very large capacitor, consuming a highly capacitive current during operation. After 100km, the current required is about the same as the rating of the load (Figure. 7).

These results are quite conclusive on the balance that must be met. They also provide a further understanding on how much the compensation solutions must be tailor-made and real-time adjustible to proper ensure the safety of the link.

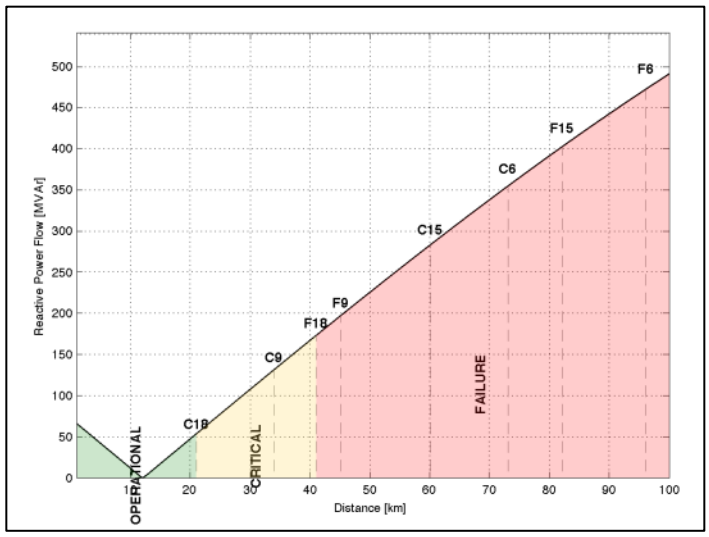


Figure 14. Reactive Power simulation (220kV Submarine Cable)

5.6. Losses & Efficiency

Losses are a significant part of the link evaluation. The model includes several paths to illustrate the losses. Two of the options are provided here: transmission efficiency and transmission (Joule) losses, plotted in Figures 15 and 16.

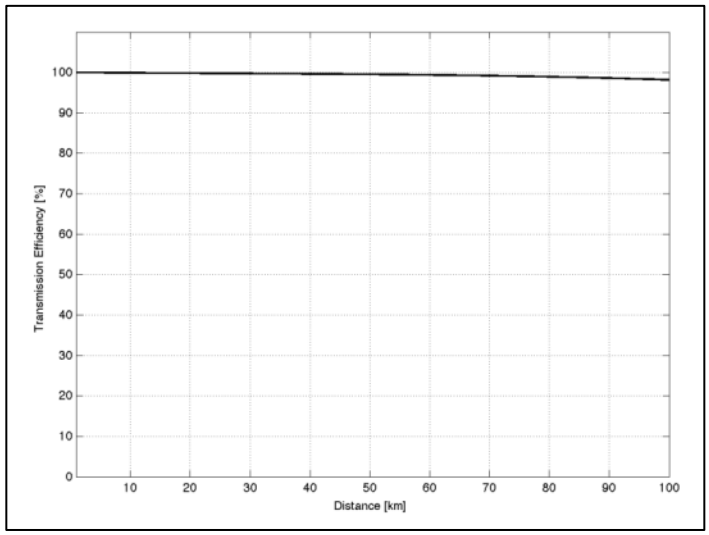


Figure 15. Transmission efficiency simulation (220kV Submarine Cable)

The efficiency remains under an acceptable range (above 90%). This easily confirms that the insulated cables have equivalent losses to the OHL, whereas the capacitive restrictions are the actual restrictions to the operation of these AC offshore links.

In a way, due to the higher inductance value, the capacitive current is not so high and Joule losses are slightly reduced. Nevertheless, the values commented are still very high and above acceptable boundaries. Losses are also similar to the equivalent land cable.

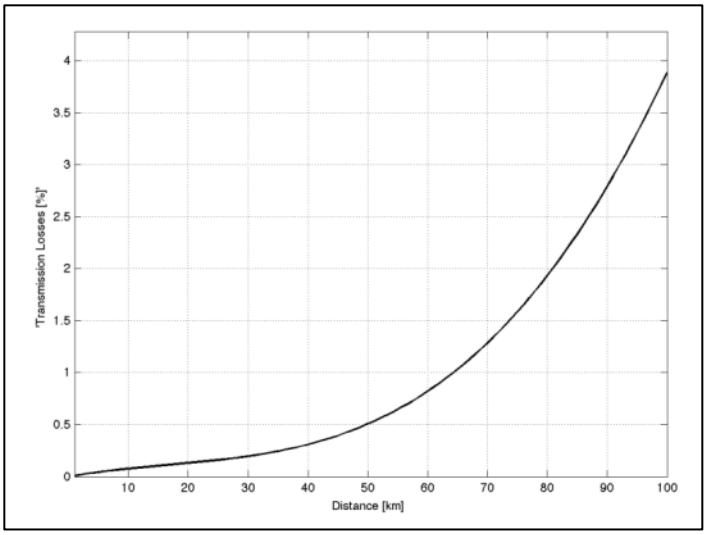


Figure 16. Transmission losses simulation (220kV Submarine Cable)

5.7. Voltage, Current & Displacement Angles

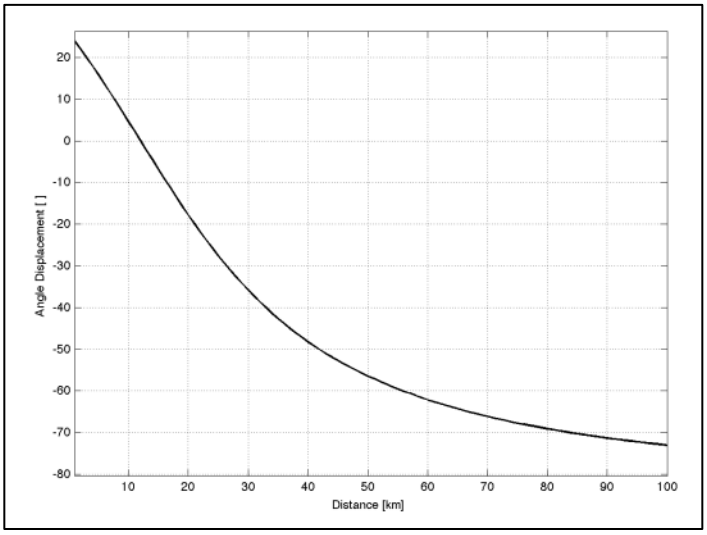


Figure 17. Angle displacement simulation (220kV Submarine Cable)

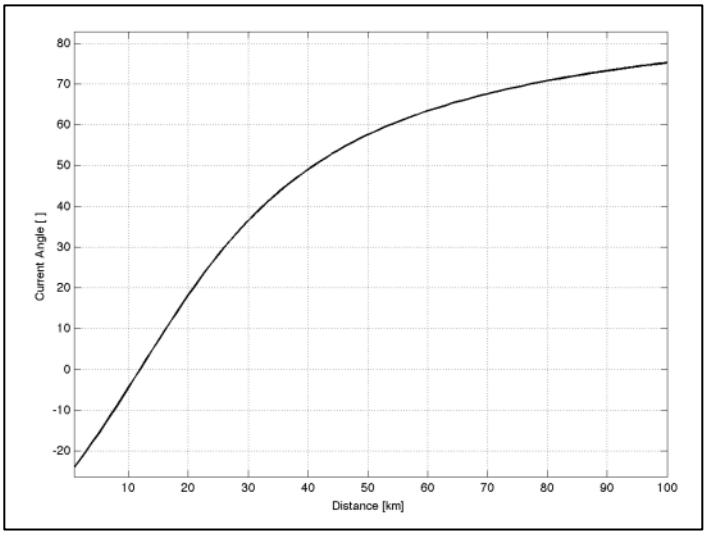


Figure 18. Current angle displacement simulation (220kV Submarine Cable)

The plot of the voltage and current angle displacements between the two ends of the cable is also a clear sign that, above the mentioned 40km, the cable is working under unacceptable operational conditions. The voltage is assuming a strongly capacitive characteristic. A comparison on the voltage angle variation for equivalent land and submarine cables is on Fig. 19.

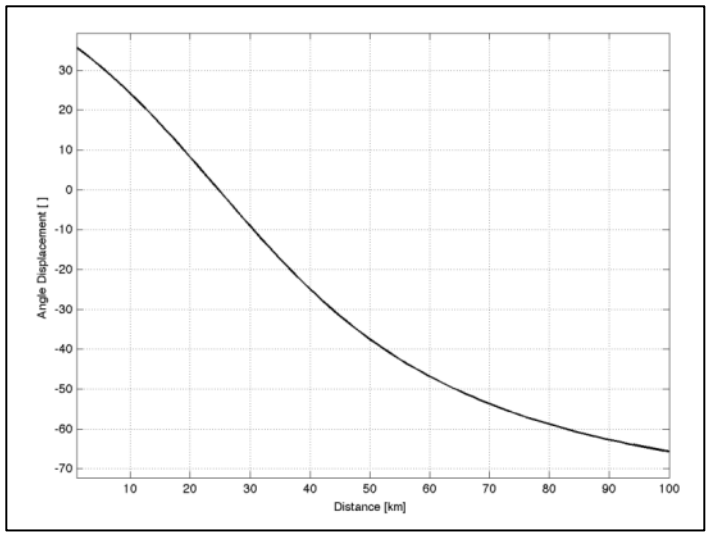


Figure 19. Angle displacement simulation (220kV Submarine Cable)

The angle and power factor values provide once more a good understanding path of the range of operation of the submarine link. The tendency of both curves is much alike the equivalent land cable. The power factor, here stated at 0,33 continues to be very low, inducing that the cable still behaves like a large capacitor. The analysis performed for the land cable, in order to assess the natural load and connectable distance is now demonstrated.

5.8. Variation of Cable Loading

With an inherit highlight of one of the additional major features of the model, this part presents the effects on the variation of the cable loading – current at the reception-end, once more, in regards to power factor and line current. The results are shown next. Hatches were disable to facilitate the visualization.

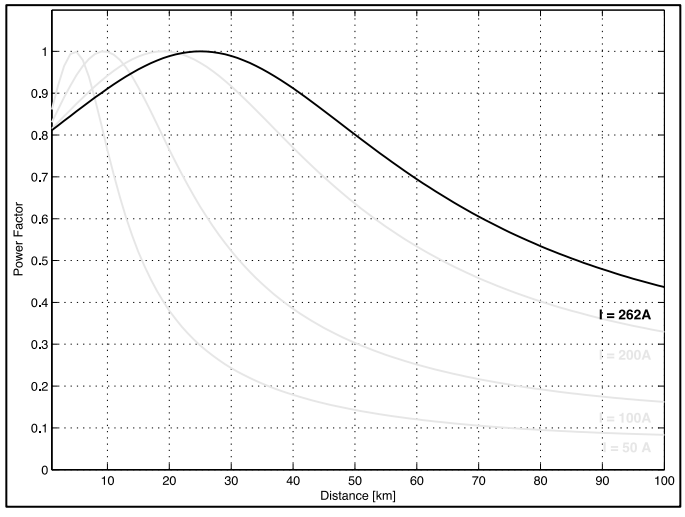


Figure 20. Power factor (load variation) simulation (220kV Submarine Cable)

If the impedance of the load is below the line's impedance, it means that the active power delivered is higher than the natural power for such voltage / cable. Land cables, which have a lower inductance and higher capacitance, are usually operated below their natural power [9], which is calculated, for this particular case, as follows:

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{87.9646}{10.1 \times 10^{-3}}} = 93.32\ \Omega \tag{11}$$

$$P_n = \frac{V^2}{Z_0} = \frac{200 \times 10^3}{93.32} = 518.65\ MW \tag{12}$$

However, subsea cables are characterized to a much higher inductance value, thus changing that approach. With a higher rating yet under the thermal capabilities of the cable, such cable can operate and obtain better operational values.

Addressing the effect on the cable loading, both in terms of rated power, but also, in terms of power factor is crucial. The higher the current provided by the wind farm, a negative shift of the peak power factor occurs, thus leading to conclude that the cable has a wider operation range if the loading (current) is close to its rated value.

On the other hand, for an equivalent unitary capacitance, the sea cables have a much higher inductance, which for certain ranges and loads, it can help to compensate the power factor variation on the line.

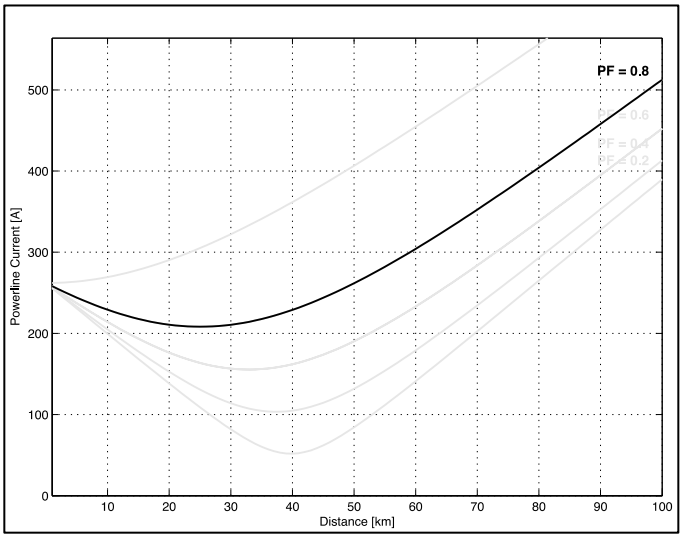


Figure 21. Current (load variation) simulation (220kV Submarine Cable)

As shown on Fig. 14, the lower the power factor of at the WF-side, the lower the current fed from the WF itself. On the other hand, the impact on the reference voltage is important. The voltage and current angles are also exposed to significant changes due to the variation of the initial power factor. The results are included on Figures 16 and 17.

This means that the lower the power factor (lag) is assumed, the voltage and current angles have smother slopes. This translates the mutual "compensation" between the power factor and the PoC and the capacitive characteristics of the long submarine cable.

For the current angle, the values at the end of scale (full distance) do not pose significant differences. But, if the center range of the angle curve is observed, it is clear that the acceptable operational boundary is improved with such power factor decrease.

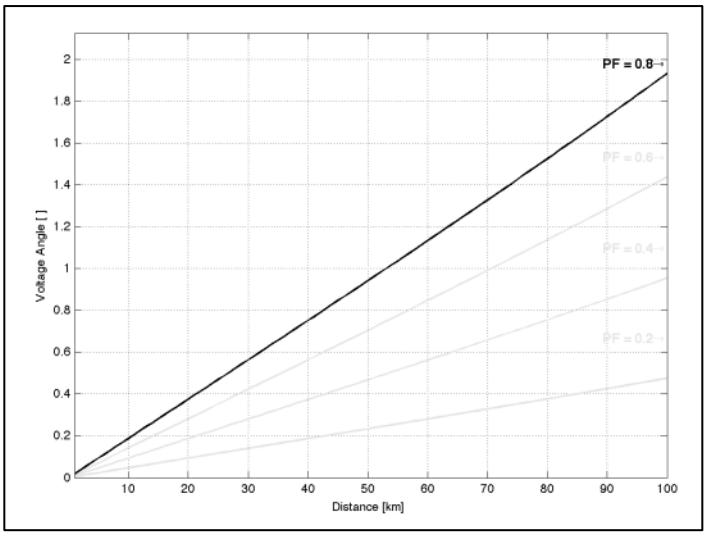


Figure 22. Voltage angle (load variation) simulation (220kV Submarine Cable)

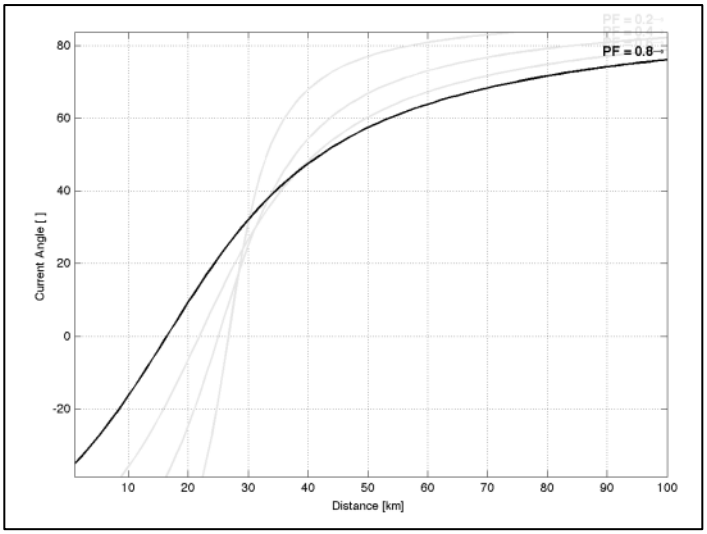


Figure 23. Current angle (load variation) simulation (220kV Submarine Cable)

Bearing in mind the purely academic approach for this – as cables used differ for distinct nominal voltages, what the model quickly provides is a framework on how an higher voltage can facilitate higher power flow. This is addressed via the SIL plot shown in Fig. 18.

The results on Fig. 16. further prove that model designed, as the conclusion, voltage increase allows for a higher SIL rating of the cable, which is aligned with the bibliography. The main result however is the easiness how this is selected on the GUI, by using the "Label" and "Freeze" capabilities. This allows the user to shown several identified scenarios {voltage, current, power factor, distance and compensation}.

Regardless, the what the authors conclude is that is now much easier, both in an academic or (risks measured) practical manner, to quickly assess the best conditions for which an AC offshore link should be designed. One route for the improvement of these results (links) is compensation.

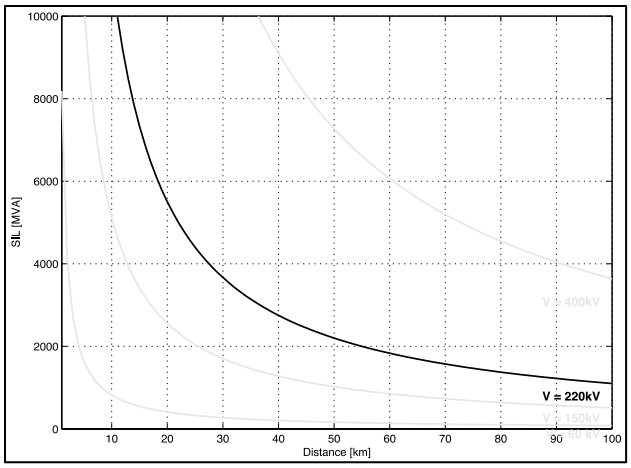


Figure 24. Current (load variation) simulation (220kV Submarine Cable)

6. MATLAB Compensated Results

The analysis of the compensated transmission cables is done still using the multi-scenario approach. As provided earlier, as the parameter changed in each scenario is automacally overlayd on the plot itself, concludes are easily drafted.

6.1. Estimation of Compensation

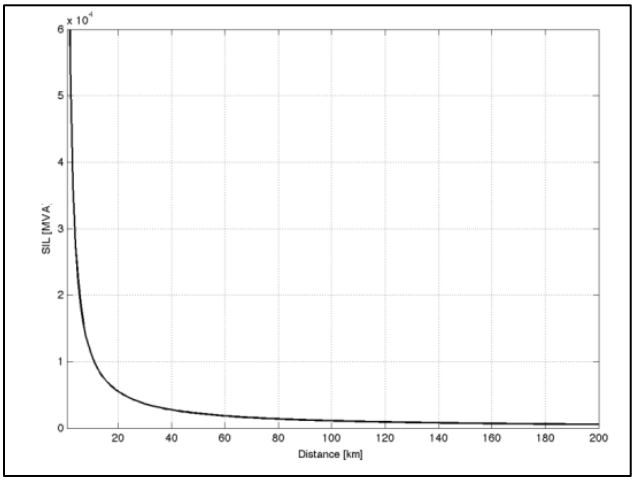


Figure 25. SIL simulation (220kV Submarine Cable)

Based on the results of the reactive power and considering a unitary power factor load, it is possible to estimate the rating of a shunt reactive compensation system – previously decided by the authors after execution a several scenario investigations using the same model. The reactive power is included in Figure 14 whereas the SIL (loading) evaluation is provided in Figure 25.

Given such and based on the experimental values shown on Fig. 16, 265MVAr was put into experiment. The effects on the voltage and power factor, two of the main images of the cable, are shown on the next section.

6.2. Voltage after Compensation

What is immediately observed is a shift of the stress (failure) point earlier identified, thus a clear improvement on the operational range. The reactive power at the emission, for nod distance, represents the full rating of the compensation.

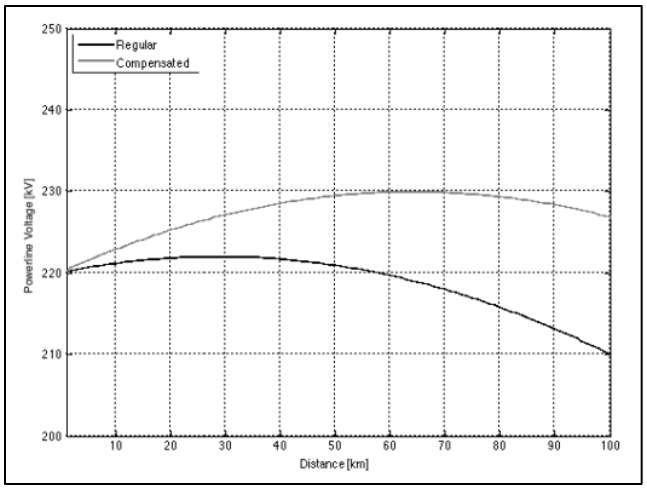


Figure 26. Voltage (after compensation) simulation (220kV Submarine Cable)

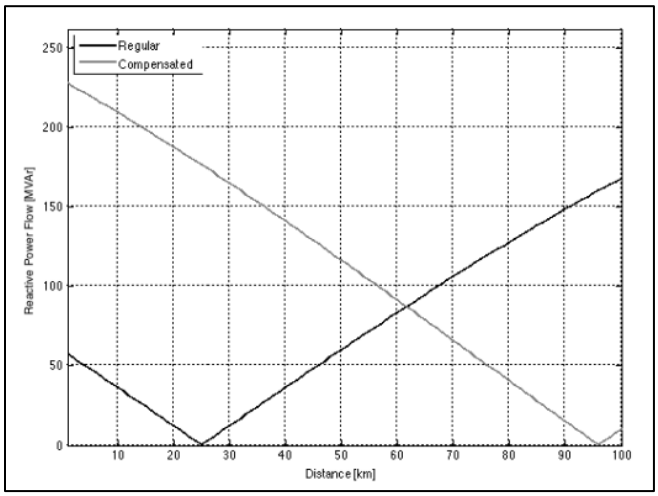


Figure 27. Current (after compensation) simulation (220kV Submarine Cable)

The GUI interface easily presents several scenarios, without going through a labour-intensive calculation routine. Taking once more the approach of adapting the cable loading (current at the reception-end), additional conclusions can be gathered (as provided on Figures 27 and 28).

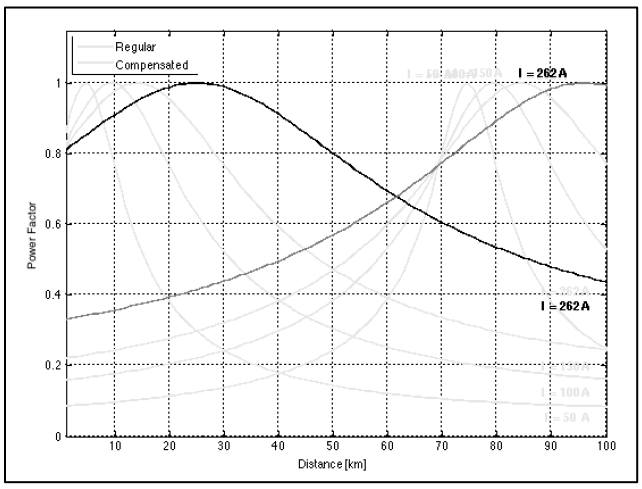


Figure 28. Power factor (load variation after compensation) (220kV Cable)

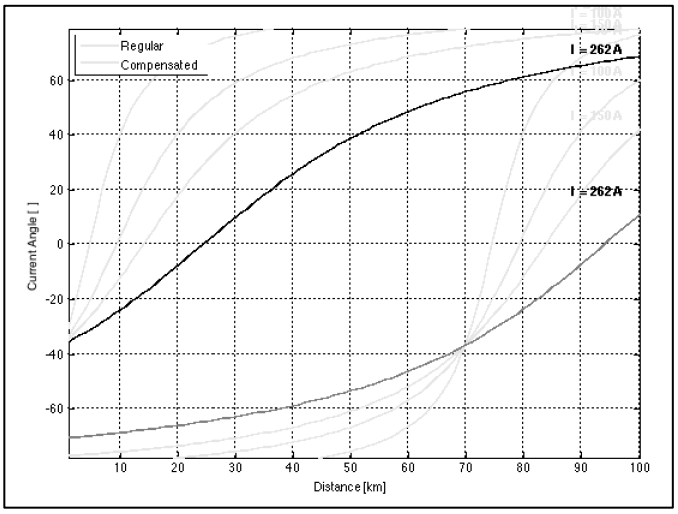


Figure 29. Current (load variation after compensation) (220kV Submarine Cable)

Voltage, reactive power and the power factor are now within perfectly acceptable boundaries, for the full 100km length. The power factor must be analysed depending on the point of connection of the compensation system, eventually using a split-off between the two ends.

Especially after the introduction of compensation, several loading scenarios have to be taken into account, once more, taking advantage of the GUI-model. For lower loading scenarios, the compensation system output has to be reduced or, eventually, taken out of service. Otherwise, what is identified in Figs. 19 and 20 is a clear overcompensation for a load drop (current). Nowadays, online FACTS systems effectively support the reactive power flow demanded by large-scale offshore wind farms. Although not addressed on this paper, once again, the model points to a quick rough estimation of a possible compensation solution.

7. Final Remarks & Conclusion

The authors firmly believe that this computational simulation tool is an important guide on the assessment of HVAC transmission links, especially for offshore applications such as wind farms.

As expected, the submarine cables presented the authors with the worst operational conditions from all the three options addressed on this document. The high inductance of such cables, mostly due to their type of insulation, introduces serious limitations, which translate in very short feasible distances. If this was the baseline for the investigation, the model as shown that the conclusions are obtained experimentally with ease.

The authors believe that the results shown before allow for a sufficient and accurate perception of the problematic associated with large HVAC transmission lines. The results were conclusive, with minor errors taking into consideration the objective proposed. So much more can be done in a further stage.

There is however still room for further improvement on the existing model, addressing topics such as different types of compensation of lines or introducing further characteristics of the line, of course, that can affect the line's behavior in some way.

The compensation system here studied also has some limitations, namely in terms of type (only fixed-step compensation is addressed) and in terms of distribution, eventually on several stages of the cable or transmission line. Future developments of the model should address this topic and allow for a more broad study of compensation effects.

The work can also extended and applied across other industries, such as Oil&Gas – for offshore energy supply to platforms and interplatform energy safety-grids. The goal of providing a safe and reliable tool, able to provide link estimations in a window of seconds to minutes was achieved.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors thank EDP Distribution (the Distribution System Operator in Portugal). This work also has been once more partially supported by the Portuguese Foundation for Science and Technology under projects grants PEst-OE/ EEI/UI308/2014 and UID/EEA/00066/2013.

References

[1] T. A. Antunes, P. J. Santos and A. J. Pires, "HVAC transmission restrictions in large scale offshore wind farm applications," in *2017 11th IEEE International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG)*, Cadiz, Spain, 2017.

[2] H. Holttinen, J. Kiviluoma, J. McCann, M. Clancy, M. Millgan, I. Pineda, P. B. Eriksen, A. Orths and O. Wolfgang, "Reduction of CO2 emissions due to wind energy - methods and issues in estimating operational emission reductions," in *Power & Energy Society General Meeting, 2015 IEEE*, Denver, USA, 2015.

[3] International Renewagle Energy Association, "The Power to Change: Solar and Wind Cost Reduction Potential to 2025," IRENA, 2016.

[4] B. Gustavsen and O. Mo, "Variable Transmission Voltage for Loss Minimization in Long Offshore Wind Farm AC Export Cables," in *IEEE Transactions on Power Delivery* , 2016.

[5] Siemens AG, *Wind Turbine SWT-8.0-154: Technical Specifications,* Hamburg: Wind Power and Renewables Division, 2016.

[6] Siemens AG, "Powered by partnership: Sustainable solutions for your offshore wind power project," 2016.

[7] ABB, "XLPE Cables up to 420kV," 2013.

[8] D. Moura, Técnicas de Alta Tensão - Curso Introdutório, Lisboa: Técnica - Revista de Engenharia, Lda., 1980.

[9] J. P. Sucena Paiva, Redes de Energia Eléctrica, Uma Análise Sistémica, 2nd Edition ed., vol. 1, I. Press, Ed., Lisbon: IST Press, 2007.

[10] Siemens, "High Voltage Direct Current Transmission: Proven Technology for Power Exchange," Siemens AG, Alemanha, 2011.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 329-341 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Mission Profile Analysis of a SiC Hybrid Module for Automotive Traction Inverters and its Experimental Power-loss Validation with Electrical and Calorimetric Methods

Ajay Poonjal Pai [*1], Tomas Reiter[1], Oleg Vodyakho[2], Martin Maerz[3]

[1]*Infineon Technologies AG, Neubiberg, 85579, Germany*

[2]*Infineon Technologies Americas Corp, El Segundo, 90245, USA*

[3]*Fraunhofer IISB, Erlangen, 91058, Germany*

ARTICLEINFO

Article history:
Received: 15 November, 2017
Accepted: 09 January, 2018
Online: 02 February, 2018

Keywords:
Silicon Carbide
Automotive Traction Inverter
Calorimetric Measurement
Mission Profile
Power Loss Calculations
Power Analyser

ABSTRACT

This paper investigates the efficiency benefits of replacing the Silicon diodes of a commercial IGBT module for the main inverter application of an electric vehicle with Silicon Carbide diodes, leaving the package, operating conditions and the system unchanged. This ensures that the comparison is directly between the chip technologies without any scope for discrepancies arising out of differences in the packaging, gate-driver circuit etc. A behavioral power loss calculation model is used to investigate the performance of the two modules for various drive cycles (Artemis, WLTP, NEDC). The behavioral power loss model is experimentally validated using two independent measurement methods, namely, power analyser based electrical input output method, and a calorimetric method which was developed especially for the low lossy light load condition. Furthermore, it is shown that the electrical method has close to 30% inaccuracy making it unsuitable for the main inverter applications, especially for comparing two different chip technologies, e.g., Silicon versus Silicon Carbide. The developed calorimetric method in contrast offers lower than 3% uncertainty.

1 Introduction

This paper investigates the efficiency benefits of replacing a Si IGBT based power module of an automotive traction inverter with a SiC Hybrid module, for public mission profiles such as NEDC, WLTP and Artemis. This paper is an extension of the work presented in [1] and [2], where the hybrid-Silicon Carbide (SiC) module was characterized and mission profile analysis was performed for public mission profiles. In this work, additionally, the inverter power loss model used for analysis will be experimentally validated using two independent methods, firstly with the input-output based electrical method, and secondly with the calorimetric method presented in [3], suitable for automotive main inverters which operate at light load conditions, i.e., less than a quarter of the nominal current more than 90% of the time.

2 Review of Literature and Motivation

A vast number of papers have been published in the last two decades investigating the advantages of SiC in automotive converters, especially by car makers like Toyota [4] and Ford [5] among others. While a vast majority of the publications have been on dc-dc applications with high switching frequency, one can find only a few papers addressing the traction inverter application which is generally a low switching frequency application (8 to 20 kHz). These papers that address the topic of SiC for automotive traction inverters often have one or more of the following drawbacks:

1. The considered switching frequencies are far higher than typical application requirements, e.g., [6] presents a full-SiC automotive inverter, but the switching frequency considered

[*]Ajay Poonjal Pai, Infineon Technologies AG, Am Campeon 1-12, Neubiberg, Germany 85579 & AjayPoonjal.Pai@Infineon.com

https://dx.doi.org/10.25046/aj030140

is 50 kHz.

2. The considered devices are rated for low currents, in the range of 4 to 20 A, e.g., [7, 8, 9]. The resulting inverters are quite far from the typical application requirements of the traction inverter (above 100 A).

3. An impractical number of smaller devices are considered connected in parallel to meet the power ratings of a high power Insulated Gate Bipolar Transistor (IGBT) module. For example, [10] provides a comparison of SiC Junction gate Field Effect Transistors (JFETs) against Silicon (Si) IGBTs in traction inverter application for typical mission profiles. But in order to compare the 5 A SiC JFETs against a 300 A Si IGBT module, it is assumed that 60 JFETs are connected in parallel, which makes it an impractical and unfair comparison!

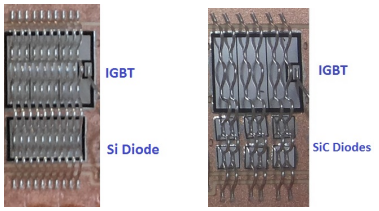
4. The devices and/or packages chosen for comparison are not suitable for mass production, but merely design studies, e.g., [11]. The constraints for a mass produced module can be quite different than for a design study module.

5. The compared Si and SiC chips are in completely different packages or application conditions, e.g., [12, 13, 14]. This makes it difficult to evaluate if the reported benefits of SiC are really coming from the advantages offered by the technology itself or simply due to the difference in package/operating conditions.

In short, a clear investigation of the efficiency benefits of using SiC as a direct replacement for a commercial Si- IGBT module at various boundary conditions, without giving any scope for discrepancies arising out of differences in the package, system, gate-driver circuit etc., is still missing in literature. This paper investigates the benefits of replacing a commercial Si IGBT module with a prototype Hybrid SiC module, as a first step, leaving the other components of the package and the system unchanged.

3 The Compared Modules

Infineon HybridPACK Drive [15] FS820R08A6P2B is chosen as the Si-IGBT module owing to its best-in-class low stray inductance (L_{sCE}=8nH) which makes it suitable for very fast switching applications. FS820R08A6P2 is an automotive qualified B6 bridge power module based on the new EDT2 Micro-pattern Trench-Field-Stop technology, with an implemented current rating of 820A per phase, and a blocking voltage rating of 750V. It has three IGBTs of 100 mm^2 each in parallel per switch (in total, A_I = 300 mm^2 per switch) and three anti-parallel diodes of 50 mm^2 each in parallel per diode (in total, A_D = 150 mm^2 per switch). Figure 1(a) shows an IGBT-diode pair of one switch. It comes with a Pin-Fin baseplate which is suitable for direct water cooling, and can operate up to T_j=175°C. It has one NTC per-phase integrated directly on the DCB which can be used for temperature sensing. For brevity, this module shall be referred to simply as "HPD" in this paper.



(a) With Si diode (b) With SiC diodes

Figure 1: An IGBT-diode pair of HybridPACK Drive Module

For the purpose of evaluating the benefits of SiC, a prototype SiC Hybrid module has been produced by replacing the Si diodes of HPD with Generation-5 650V SiC Schottky diodes from Infineon [16]. The SiC diodes have a die size of 7.12 mm^2 and a nominal current rating of 40A. Each of the 50 mm^2 Si diodes are replaced by a parallel connection of six SiC diodes as seen from figure 1(b), leaving the remaining construction of the module as it is, for a direct evaluation of the SiC diodes vis-a-vis the Si diodes. This module shall be referred to as "HPD-Hyb-SiC" in this paper.

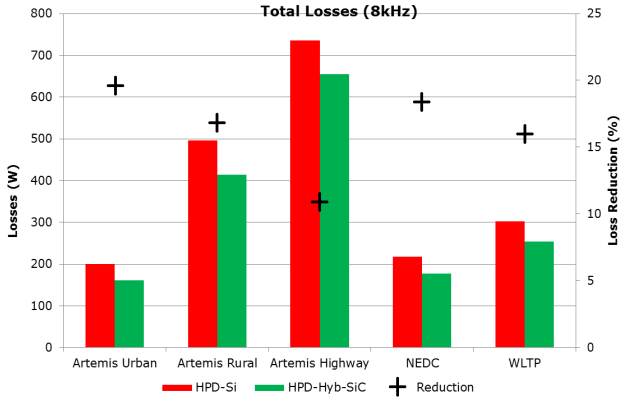
3.1 Inverter Power Loss Calculation Model

For a good comparison of the power loss performance of different chip technologies, it is imperative to have an accurate inverter power loss model. This model should offer an uncertainty significantly lower than the performance difference between the technologies compared. For example, it is meaningless to compare two different technologies which differ in loss performance by 20% with a model which suffers from 15% uncertainty, because it is not possible to ascertain if the apparent difference between the technologies is indeed because the technology is better than the other, or if it is merely due to the high uncertainty of the model used. Since this paper would compare Si and SiC technologies which usually differ by 10 to 20 percent, a model with an uncertainty of < ±5% presented in [17, 18] is used and the parameters A_{11}- E_{43} are determined for both the modules. The parameters for HPD can be found in [17, 18], and those for HPD-Hyb-SiC in [2].

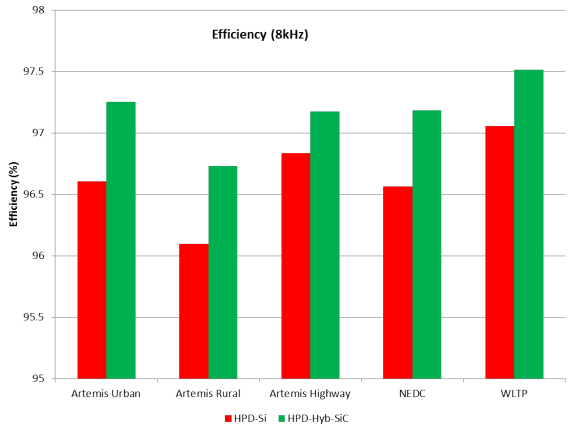
4 Mission Profile Analysis

The model discussed in the previous section is used to compare the performance of the two modules for 5 different mission profiles, viz., Artemis-Urban, Artemis-Highway, Artemis-Rural, WLTP and NEDC.

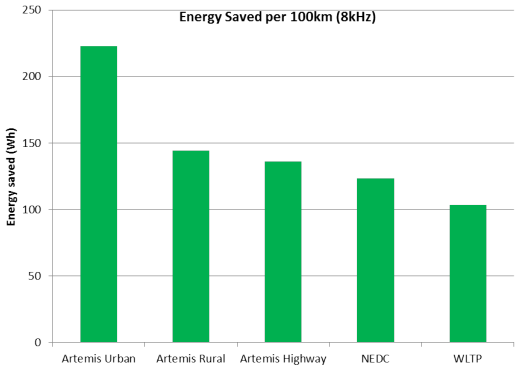
For the mission profile investigation, a typical mid size sedan similar to Volkswagen Golf presented in [2] is chosen as the reference vehicle. The results at 8kHz are given in figure 2.



(a)



(b)



(c)

Figure 2: Results of Mission Profile Analysis for the modules

As a result of the higher average driving speed, the overall losses are the highest for Artemis Highway driving Cycle. Artemis Urban on the contrary, has the least average losses due to its low average speed. HPD-Hyb-SiC has about 10-20% lower losses compared to HPD-Si. The reduction in the losses is the highest for the Artemis Urban cycle, where the inverter operates in the low current regime. This is because, at low currents the switching losses (which are significantly lesser in the SiC diode compared to the Si diode) dominate over the conduction losses (which are higher in the SiC diodes than the Si diodes). Next, the study is repeated at different switching frequencies, 8-15kHz. The improvement in efficiency of the SiC module over the Si module is summarized in figure 3. Again, the highest benefit of increasing f_{sw} is for the Artemis Urban drive cycle, which sees more than 1.2% improvement in efficiency at 15kHz compared to about 0.7% at 8kHz. Artemis Rural and Artemis Highway cycles too are not far behind. This is because in all the artemis cycles, the switching losses dominate over the conduction losses, and the scope for improvement with SiC diodes is high. The NEDC and WLTP cycles are mostly dominated by conduction losses, and as a result the benefit of using SiC diodes is not high.

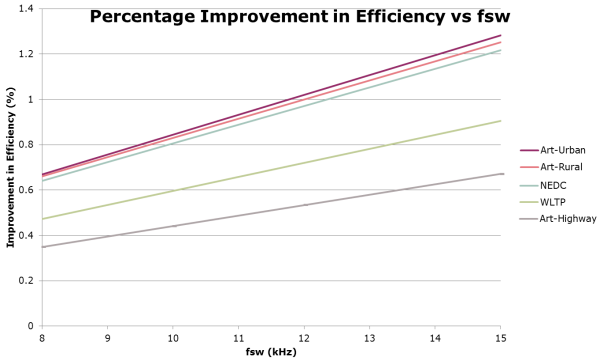


Figure 3: Mission Profile analysis at different f_{sw}

5 Experimental Validation of the Behavioral Power Loss Calculation Model

To have a good confidence level in the behavioral power loss calculation model used for mission profile analysis and to verify it experimentally, it is important to measure the inverter power losses at various operating points, preferably with two independent methods. Firstly, power loss measurements are performed at various inverter operating points with the power-analyser based input-output electrical method, which, being relatively easier to perform, is the most commonly used method for such applications. The sources of uncertainty are thoroughly analysed and it will be shown that this method has high uncertainty due to the switched nature of the output voltage waveform. A common solution to this problem is to add a sine output filter. However, the results include also the losses in the filter leading to wrong results, as will be shown in this paper. Thus, it is necessary to resort to the calorimetric method proposed in [3], which is nearly as easy to perform as the electrical method and is particularly suitable for low-lossy conditions.

This method does not require the use of an expensive calorimeter.

6 The Electrical Input-Output-based Method

6.1 Test Setup

A test platform is built as per the schematic in figure 4. The inverter (described in section 7) is connected to a 3-phase inductive load. The input dc current I_{dc} and output rms currents I_{rms} are sensed using high accuracy closed loop current transducers IT 200-S ULTRASTAB [19] and LF 510-S [20] respectively, which are then fed into a state-of-the-art precision Power Analyser WT1800 [21]. WT1800 has a high sampling frequency of 2 MHz with 16-bit resolution and a bandwidth of 5 MHz. It is also equipped with a digital line filter which can be set for frequencies from 100 Hz to 100 kHz. The impact of using this line filter on the accuracy will be also discussed later in this paper. The inverter is run in open-loop mode, and the desired output voltage is requested by setting the modulation index accordingly. The power analyser measures the input dc power P_{dc} and the output ac power P_{ac}, and the difference is equal to the power loss P_{loss}, from the basic definition, as per equation 1.

$$P_{loss} = P_{dc} - P_{ac} \tag{1}$$

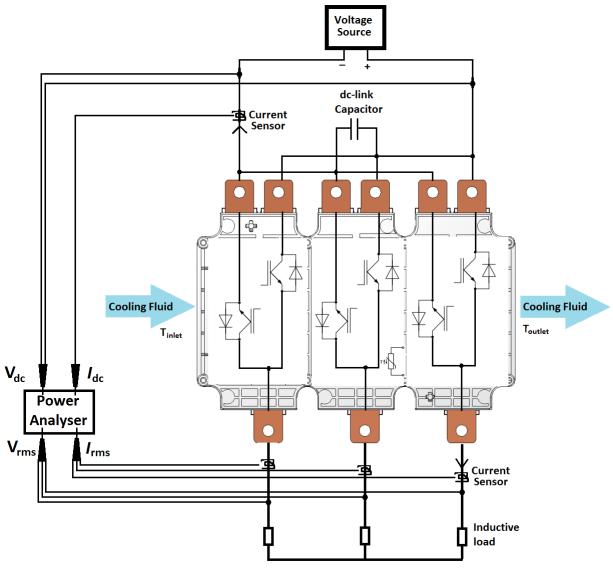


Figure 4: Schematic for measurement of power losses with the electrical method

6.2 Sources of Uncertainty

Uncertainties occur in this method, on account of the delays introduced by the probes and unintended phase-shifts between the different measured signals. In this section, the uncertainty involved in this method will be systematically derived. Out of the scope of this work are the errors introduced on account of the Radio Frequency Interference (RFI) and Electro Magnetic Interference (EMI) emanating from the high di/dt and dv/dt prevailing in hard switched power converters.

The uncertainty in the measurement of the power loss ΔP_{loss} from equation 1 can be derived using the Gaussian law of error propagation.

$$\frac{\Delta P_{loss}}{P_{loss}} = \sqrt{\left(\frac{\Delta P_{dc}}{P_{dc}}\right)^2 + \left(\frac{\Delta P_{ac}}{P_{ac}}\right)^2} \tag{2}$$

Equation 2 gives the uncertainty with a confidence value of about 68%. This will simply be referred to as *uncertainty* in the rest of this work. The upper bound for the uncertainty $\Delta P_{loss,max}$ can be simply calculated as

$$\frac{\Delta P_{loss,max}}{P_{loss}} = \frac{\Delta P_{dc}}{P_{dc}} + \frac{\Delta P_{ac}}{P_{ac}} \tag{3}$$

This will hence be referred to as the *maximum uncertainty*.

6.2.1 Uncertainty in P_{dc}

To calculate ΔP_{dc}, we have to go back to the fundamental equation of dc power, i.e.,

$$P_{dc} = V_{dc} \cdot I_{dc} \tag{4}$$

and the uncertainty can again be calculated by the Gaussian law described above as:

$$\frac{\Delta P_{dc}}{P_{dc}} = \sqrt{\left(\frac{\Delta V_{dc}}{V_{dc}}\right)^2 + \left(\frac{\Delta I_{dc}}{I_{dc}}\right)^2} \tag{5}$$

and the maximum uncertainty as:

$$\frac{\Delta P_{dc,max}}{P_{dc}} = \frac{\Delta V_{dc}}{V_{dc}} + \frac{\Delta I_{dc}}{I_{dc}} \tag{6}$$

From the reference manual of the power analyser [21], the uncertainties for the dc voltage measurement are given as a function of the set range $V_{dc,range}$ and the reading itself as follows

$$\Delta V_{dc} = 0.0005 \cdot V_{dc} + 0.001 \cdot V_{dc,range} \quad [V] \tag{7}$$

Additionally, it must be remembered that there is also a ripple in V_{dc} consisting mainly of a second harmonic of the fundamental frequency, and small magnitudes of harmonics of the switching frequency. These harmonics also contribute to the uncertainty, as will be covered in section 6.2.2. But, as the magnitude of this ripple is quite small, their contribution to the uncertainty can be neglected.

The uncertainties for the measurement of I_{dc} are similarly given as

$$\Delta I_{dc,analyser} = 0.0005 \cdot I_{dc} + 0.001 \cdot I_{dc,range} \quad [A] \tag{8}$$

$\Delta I_{dc,analyser}$, however, is the uncertainty of just the power analyser and since we are using an external LEM current transducer, we also have to take its uncertainty into account. From [19], the uncertainty

for the LEM transducer on the secondary side can be given as:

$$\Delta I_{\mathrm{dc,LEM,max,sec}} = I_{\mathrm{OE}} + \epsilon_{\mathrm{L}} \cdot I_{\mathrm{pry,range}} \cdot K_{\mathrm{N}} \quad [A] \quad (9)$$

and

$$\Delta I_{\mathrm{dc,LEM,sec}} = \sqrt{I_{\mathrm{OE}}^2 + (\epsilon_{\mathrm{L}} \cdot I_{\mathrm{pry,range}} \cdot K_{\mathrm{N}})^2} \quad (10)$$

where, $I_{\mathrm{OE}} = 80 \cdot 10^{-6}$A is the electrical offset current, $\epsilon_{\mathrm{L}} = 3 \cdot 10^{-6}$A is the linearity error, $I_{\mathrm{pry,range}}$=200A is range on the primary side and K_{N}=0.001 is the turns ratio. It is to be noted that, the effects due to self heating in the current transducer are neglected and it is assumed that the transducer is at a constant temperature of 25 °C. Since we are interested in the primary current for calculating P_{dc}, the absolute uncertainties get divided by the turns ratio as follows:

$$\Delta I_{\mathrm{dc,LEM}} = \frac{\Delta I_{\mathrm{dc,LEM,sec}}}{K_{\mathrm{N}}} \quad (11)$$

$$\Delta I_{\mathrm{dc,LEM,max}} = \frac{\Delta I_{\mathrm{dc,LEM,max}}}{K_{\mathrm{N}}} \quad (12)$$

Clubbing equations 8, 9 and 11/12, we obtain the effective uncertainty for I_{dc} as:

$$\frac{\Delta I_{\mathrm{dc}}}{I_{\mathrm{dc}}} = \sqrt{\left(\frac{\Delta I_{\mathrm{dc,analyser}}}{I_{\mathrm{dc,analyser}}}\right)^2 + \left(\frac{\Delta I_{\mathrm{dc,LEM}}}{I_{\mathrm{dc,LEM}}}\right)^2} \quad (13)$$

$$\frac{\Delta I_{\mathrm{dc,max}}}{I_{\mathrm{dc}}} = \frac{\Delta I_{\mathrm{dc,analyser}}}{I_{\mathrm{dc,analyser}}} + \frac{\Delta I_{\mathrm{dc,LEM,max}}}{I_{\mathrm{dc,LEM,max}}} \quad (14)$$

6.2.2 Uncertainty in P_{ac}

From the fundamental equation for the ac power for a line-line voltage of V_{rms}, line current I_{rms} and a power factor angle ϕ (in radians),

$$P_{\mathrm{ac}} = \sqrt{3} \cdot V_{\mathrm{rms}} \cdot I_{\mathrm{rms}} \cdot \cos\phi \quad (15)$$

and the uncertainty can again be calculated by the Gaussian law described above as:

$$\frac{\Delta P_{\mathrm{ac}}}{P_{\mathrm{ac}}} = \sqrt{\left(\frac{\Delta V_{\mathrm{rms}}}{V_{\mathrm{rms}}}\right)^2 + \left(\frac{\Delta I_{\mathrm{rms}}}{I_{\mathrm{rms}}}\right)^2 + (\Delta\phi \cdot \tan\phi)^2} \quad (16)$$

When ϕ is expressed in degrees, we have:

$$\frac{\Delta P_{\mathrm{ac}}}{P_{\mathrm{ac}}} = \sqrt{\left(\frac{\Delta V_{\mathrm{rms}}}{V_{\mathrm{rms}}}\right)^2 + \left(\frac{\Delta I_{\mathrm{rms}}}{I_{\mathrm{rms}}}\right)^2 + \left(\frac{\pi}{180} \cdot \Delta\phi \cdot \tan\phi\right)^2} \quad (17)$$

The calculation of uncertainty in the measurement of P_{ac} is less straightforward than that for P_{dc}. This is because, the inverter output voltages and currents used for determining P_{ac} are not pure sinusoids and contain a spectrum of various frequencies, and the uncertainties of the power analyser are differently defined for the spectral components of different frequencies. Moreover, the reference manuals for the power analyser define the uncertainties only for pure sine waves and do not explicitly describe how these spectral components have to be treated. One simplified approach, commonly adopted in literature, is to assume that the output voltage spectrum would be dominated by the fundamental output frequency f_{o} and then calculate the uncertainty as if the output voltage were a pure sine wave with f_{o}. However, as can be expected and as shall be shown later in this paper, such an approach results in a significant discrepancy in the estimation of uncertainty. Therefore, in this work, we propose the following spectrum-based approach.

Spectrum-based Approach to Calculating Uncertainty of Non-sinusoidal Signals

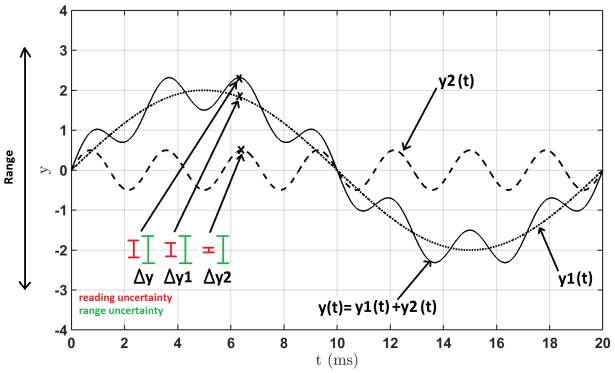


Figure 5: Reading- and range-uncertainties for a summation of signals

To understand how reading- and range- uncertainties should be calculated for a signal that is a combination of several signals, consider figure 5 which shows three signals, viz. $y_1(t)$ at 50 Hz, $y_2(t)$ at 350 Hz and $y(t)$= $y_1(t)$+$y_2(t)$. Lets assume that all the signals are measured with the same range r. Lets consider the points (t,y_1), (t,y_2) and (t,y) on the three signals respectively. The uncertainties for the three signals are also shown. The reading uncertainties $\Delta y_{1/2,\mathrm{reading}}$ can be calculated as follows:

$$\Delta y_{1,\mathrm{reading}} = \epsilon_{\mathrm{reading}}(f)\Big|_{f=50Hz} \cdot y_1 \quad (18)$$

$$\Delta y_{2,\mathrm{reading}} = \epsilon_{\mathrm{reading}}(f)\Big|_{f=350Hz} \cdot y_2 \quad (19)$$

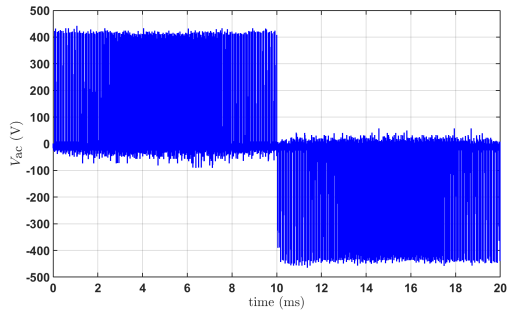
where, $\epsilon_{\mathrm{reading}}(f)$ is the frequency-dependent reading uncertainty coefficient normally specified in the reference manual. The range uncertainties $\Delta y_{1/2,\mathrm{range}}$ can be calculated as follows:

$$\Delta y_{1,\mathrm{range}} = \Delta y_{2,\mathrm{range}} = \epsilon_{\mathrm{range}} \cdot r \quad (20)$$

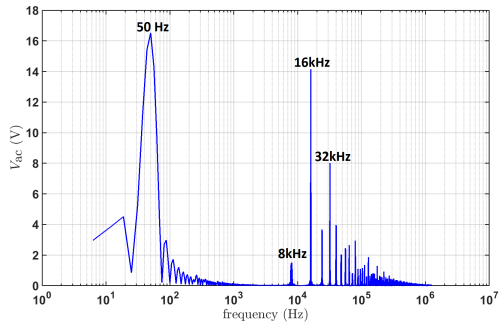
where, $\epsilon_{\mathrm{range}}$ is the uncertainty coefficient for the range, also specified in the reference manual. As the dependency of the range uncertainty on the spectral

components is small, it is assumed that ϵ_{range} is independent of f. The reading uncertainty $\Delta y_{\text{reading}}$ for the summed signal can be calculated simply by the sum of the reading uncertainties for the two signals as follows: The uncertainties for the three signals are also shown. The range uncertainties Δy_{range} can be calculated as follows:

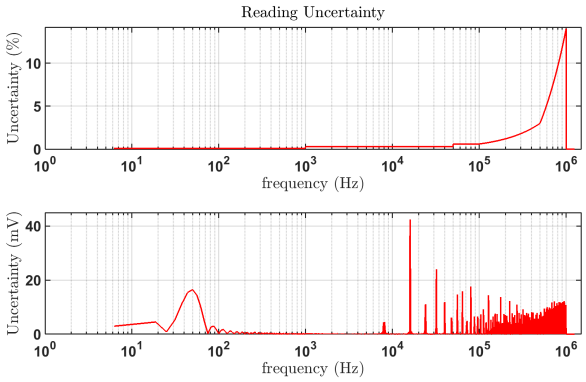
$$\Delta y_{\text{reading}} = \Delta y_{1,\text{reading}} + \Delta y_{2,\text{reading}} \tag{21}$$



(a) A fundamental period of V_{ac} in time domain



(b) Spectral components



(c) Reading uncertainty for the spectral components

Figure 6: Measured output line-line voltage V_{ac} of a hard-switched inverter

The range uncertainty Δy_{range}, however, could be expected to be the same for Δy as for $\Delta y_{1/2}$ as we are using the same range. This means that, unlike in the case of the reading uncertainty, the range uncertainty for the sum of two signals is not equal to the sum of the range uncertainties for the two signals. Therefore, Δy_{range} is equal to that calculated by equation 20.

Let us now consider a practical example. Figure 6(a) shows one fundamental period of the measured waveform of a typical hard-switched inverter output line-line voltage V_{ac}. First, the measured voltage waveform is subject to a fourier transformation to decompose it into its spectral components, as shown in figure 6(b). It can be seen that apart from the fundamental frequency 50Hz, there is significant contribution at the switching frequency 8kHz and multiples thereof. The reading uncertainties for the different components are calculated using the respective equations in table 1, as defined by the reference manual. Figure 6(c) shows the calculated reading uncertainty, both absolute and percentage, as a function of the spectral frequency. The original spectrum taking into account the uncertainties is now transformed back into the time domain. The range uncertainty is calculated as described previously and is added to this time-domain signal to yield the required $V_{\text{ac}} + \Delta V_{\text{ac}}$.

Table 1: Definition of Uncertainty for different Spectral Components

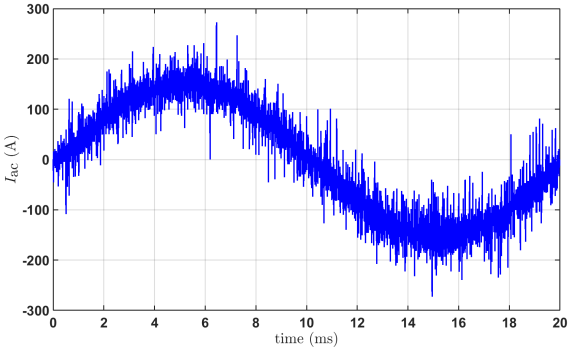
DC	$\Delta V_{\text{rms}} = 0.0005 \cdot V_{\text{rms}} + 0.001 \cdot V_{\text{rms,range}}$
0.1-10 Hz	$\Delta V_{\text{rms}} = 0.001 \cdot V_{\text{rms}} + 0.002 \cdot V_{\text{rms,range}}$
10-45 Hz	$\Delta V_{\text{rms}} = 0.001 \cdot V_{\text{rms}} + 0.001 \cdot V_{\text{rms,range}}$
45-66 Hz	$\Delta V_{\text{rms}} = 0.001 \cdot V_{\text{rms}} + 0.0005 \cdot V_{\text{rms,range}}$
66-1000 Hz	$\Delta V_{\text{rms}} = 0.001 \cdot V_{\text{rms}} + 0.001 \cdot V_{\text{rms,range}}$
1-50 kHz	$\Delta V_{\text{rms}} = 0.003 \cdot V_{\text{rms}} + 0.001 \cdot V_{\text{rms,range}}$
50-100 kHz	$\Delta V_{\text{rms}} = 0.006 \cdot V_{\text{rms}} + 0.002 \cdot V_{\text{rms,range}}$
100-500 kHz	$\Delta V_{\text{rms}} = (0.00006 * f) \cdot V_{\text{rms}} + 0.005 \cdot V_{\text{rms,range}}$
500-1000 kHz	$\Delta V_{\text{rms}} = (0.00022 * f - 0.08) \cdot V_{\text{rms}} + 0.01 \cdot V_{\text{rms,range}}$

This spectrum-based approach is also used to calculate the uncertainty for the current I_{ac}. The measured waveform is shown in figure 7(a). It can be seen that, unlike V_{ac}, I_{ac} is nearly sinusoidal which is on account of the inductance of the load. This can also be verified from the frequency spectrum shown in figure 7(b) where it can be seen that most of the energy is concentrated at the fundamental frequency. The contribution of the harmonics of the fundamental and the switching frequency is significantly lesser than in the case of V_{ac}. The calculated uncertainty as a function of the frequency spectrum is shown in figure 7(c).

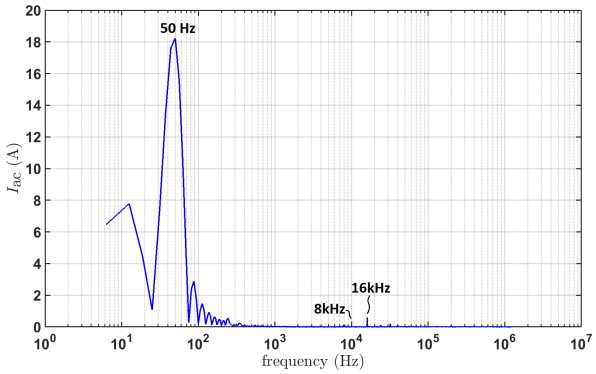
The total uncertainties thus calculated for the rms values of V_{ac} and I_{ac} are tabulated in table 2[1]. The approximate approach underestimates the uncertainty in V_{ac} by a factor of three compared to the spectrum-based approach, due to the abundance of high-frequency content in the waveforms. For I_{ac}, on the other hand, both approaches result in nearly the same value, owing to the current waveform being nearly sinusoidal. Therefore, it can be summarized that the approximated approach is sufficient for calculating uncertainty in I_{ac}, but it is necessary to go to

[1] As with I_{dc}, we are using an external LEM current transducer for measurement and therefore, we must separately calculate the uncertainties $\Delta I_{\text{rms,analyser}}$ and $I_{\text{rms,LEM}}$ for the power analyser and the LEM transducer [20] respectively.

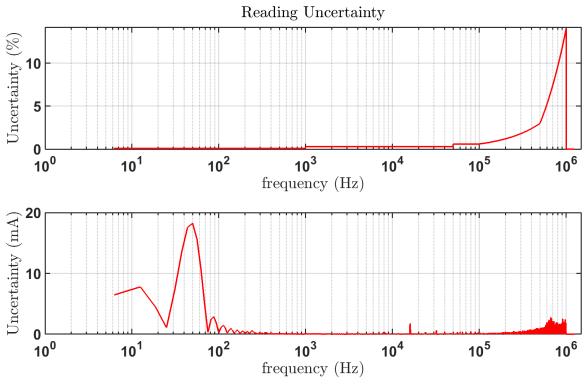
the more exact spectrum-based approach for calculating uncertainty in V_{ac}. Moreover, in applications with a higher switching frequency or a higher operating dc-link voltage V_{dc}, which is typical for a Silicon-Carbide-based application, the uncertainty in V_{ac} is higher respectively due to a wider distribution in the spectrum and a higher range that has to be chosen. In such applications, a higher deviation can be expected between the approximate and the spectrum-based approach which makes it more meaningful to use the spectrum-based approach.



(a) A fundamental period of I_{ac} in the time-domain



(b) Frequency spectrum



(c) Reading uncertainty for the spectral components

Figure 7: Measured line current I_{ac}

The uncertainty contribution due to the current transducer $\Delta I_{rms,LEM}$ can be calculated as follows:

$$\Delta I_{rms,LEM} = 0.005 \cdot I_{rms,nom} \tag{22}$$

where, the nominal current $I_{rms,nom}$=500A for the device used. The uncertainty ΔI_{rms} can be now written as

Table 2: Calculated measurement uncertainty for V_{rms} and I_{rms}

V_{rms}	ΔV_{rms}	
	Approximate	Spectrum-based
185.39 V	0.39 V (0.2 %)	1.11 V (0.6 %)
I_{rms}	$\Delta I_{rms,analyser}$	
	Approximate	Spectrum-based
106.158 A	0.181 A (0.17%)	0.203 A (0.19%)

$$\frac{\Delta I_{rms}}{I_{rms}} = \sqrt{\left(\frac{\Delta I_{rms,analyser}}{I_{rms,analyser}}\right)^2 + \left(\frac{\Delta I_{rms,LEM}}{I_{rms,LEM}}\right)^2} \tag{23}$$

and the maximum uncertainty as

$$\frac{\Delta I_{rms,max}}{I_{rms}} = \frac{\Delta I_{rms,analyser}}{I_{rms,analyser}} + \frac{\Delta I_{rms,LEM}}{I_{rms,LEM}} \tag{24}$$

For WT-1800, $\Delta\phi$ is given as

$$\Delta\phi = \left(\left|\phi - \cos^{-1}\left(\frac{\lambda}{1.002}\right)\right|\right) + \sin^{-1}\left(0.001 \cdot V_{rms} \cdot I_{rms}\right) \tag{25}$$

where λ is the power factor. Lastly, it has to be remembered that the uncertainty of the power analyser depreciates over the passage of time from its recent calibration. For WT-1800, the uncertainty at one year is 1.5 times that at 6 months.

6.2.3 A Common Measurement Mistake while applying Line Filters in Power Analysers

Most state-of-the-art power analysers come equipped with digital line filters which can be used to attenuate spectral components in the measured signals with a frequency higher than a certain cut-off frequency (generally programmable individually for each input channels). These filters are meant to be used on measurement signals where there is high frequency noise due to the limitation of the measuring equipment. Let us suppose that such a filter is used on the I_{ac} signal with a cut-off fequency of, say, 1kHz. The high frequency components in I_{ac} are predominantly due to measurement noise as can be seen from figure 7(b) and table 2, and using such a filter would help in attenuating this noise, thereby making the measurements more meaningful. However, suppose we use the same filter, either intentionally or accidentally, on V_{ac} which inherently has a high-frequency content (see figure 6(b) and table 2), other than measurement noise. In such a case, the filter would attenuate not only the noise, but also these high frequency components. This would, in turn, result in a lower-than-real measured value for P_{ac} and therefore, a higher P_{loss}. Therefore, the line filter should be used only on signals which do not inherently have high frequency components, like V_{dc}, I_{dc} and I_{ac}, but not on signals

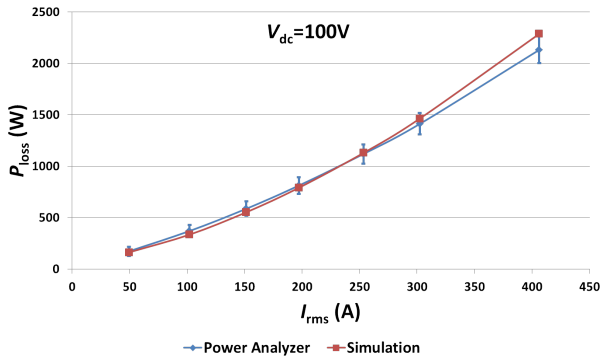
like V_{ac} which have a high frequency content. This is a common mistake while performing measurements incorporating such line filters, and will be demonstrated in the next section 6.3.

6.3 Measurement Results

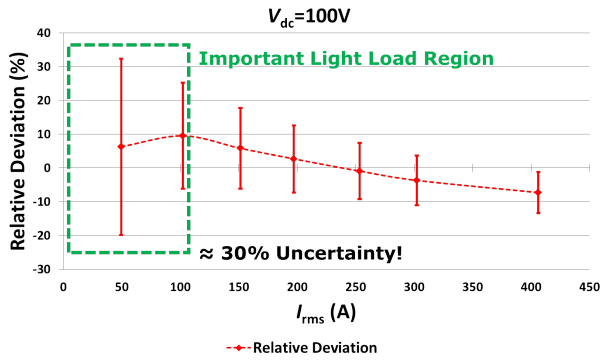
Figures 8-12 show the power losses measured with the electrical method for different application conditions. The uncertainties calculated for each of the measurements, using the spectrum-based approach described in the previous section, are shown as bars around the measurement points. Also shown are the simulated results based on the behavioral model discussed previously, and the relative deviation of the measured values from the simulation.

Figure 8(a) and 9(a) show P_{loss} measured for different output rms currents, I_{rms}, at V_{dc}=100V and 300V respectively. Across the entire range of the measured current, it can be seen that the simulations are within the tolerance of this state-of-the-art electrical input-output measurement approach, thereby validating the behavioral power loss calculation model.

In figures 10(a) and 11(a), the results are shown for different values of the gate resistances $R_{g,on}$ and switching frequencies f_{sw}. It can again be seen that the simulations are within the tolerance of this state-of-the-art input-output measurement approach.

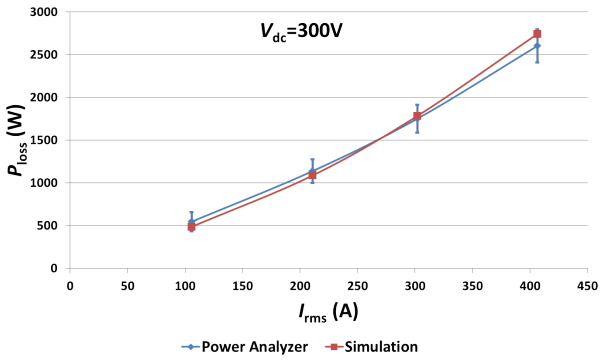


(a) P_{loss}

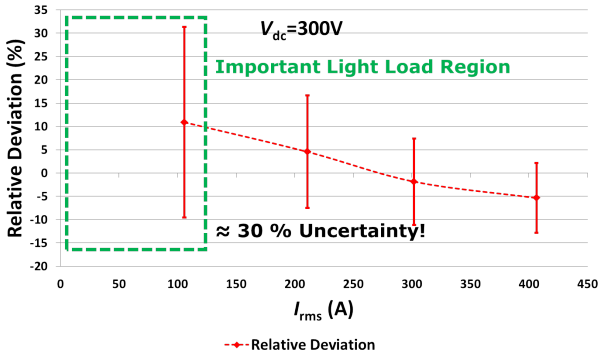


(b) Relative Deviation (%)

Figure 8: Comparison of the electrical method with simulations: P_{loss} vs. I_{rms} at V_{dc}=100V

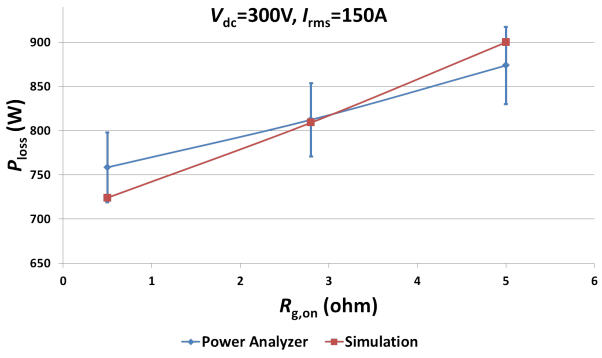


(a) P_{loss}

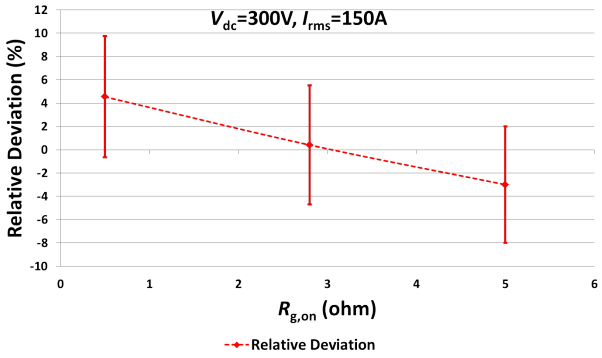


(b) Relative Deviation (%)

Figure 9: Comparison of the electrical method with simulations: P_{loss} vs. I_{rms} at V_{dc}=300V



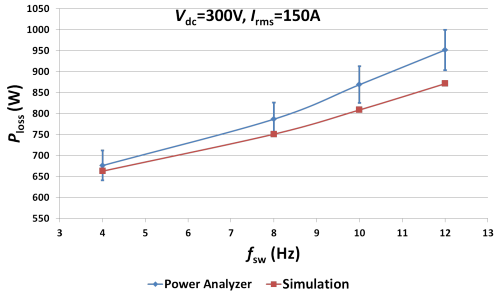
(a) P_{loss}



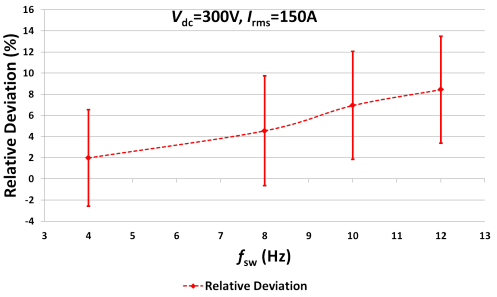
(b) Relative Deviation (%)

Figure 10: Comparison of the electrical method with simulations: P_{loss} vs. $R_{g,on}$

Lastly, figure 12 shows the measurements performed with the line filter (cut-off frequency=1kHz) enabled for the output ac voltage V_{ac}, and compares them with the measurements without the filter. As explained in section 6.2.3, it can be seen that the losses measured are higher without the line filter because the measured P_{ac} is lower-than-real. This shows that enabling the line filter on V_{ac} can lead to wrong measurements. Overall, it must be observed that the electrical input-output based method has nearly 30% uncertainty in the light-load condition, making this method unsuitable for the main inverter application, especially when comparing different chip generations.



(a) P_{loss}



(b) Relative Deviation (%)

Figure 11: Comparison of the electrical method with simulations: P_{loss} vs. f_{sw}

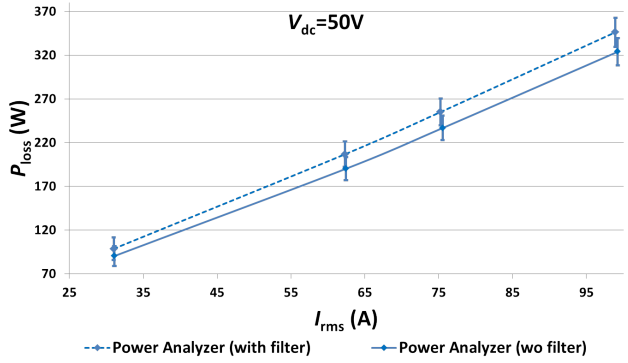


Figure 12: Comparison of the electrical method with simulations: Effect of Line Filter

7 Calorimetric Measurement of Power Losses

A more accurate approach, compared to the electrical method, is to measure power losses with a calorimeter [22]. For applications such as the automotive main inverter which operate at less than a quarter of the inverter nominal current more than 90% of the time[17], the light-load low-lossy condition is of interest. With traditional calorimetric methods, however, a sufficient rise in the fluid temperature is hard to obtain at these conditions. To overcome this problem, without compromising on the accuracy, the inverter is subjected to a calorimetric method particularly suitable for low-lossy conditions. This method does not require the use of an expensive calorimeter and is presented in detail in [3]. A requirement for the test is that all the energy losses in the module must ideally go into heating up the baseplate, with no convection. Therefore, the module baseplate is thermally insulated with a layer of polystyrene as can be seen in figure 13. EVAL-6ED100HPDRIVE-AS, a 6-channel gate-driver board based on the EiceDriver 1EDI2001AS from Infineon, is connected on the top of the module. The gate-driver board is controlled by a micro-controller logic board connected on top of it. The logic board is connected to a computer through a CAN bus, and the parameters such as f_{sw}, m, f_{out} can also be controlled by software, in open loop or closed loop modes. Moreover, it is also possible to read the temperatures sensed by the NTCs and log them. The complete inverter system is shown in figure 14. This method comprises the following two stages.

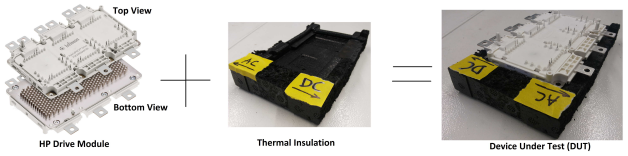


Figure 13: The DUT with the Thermal Package for the Calorimetric Test Bench

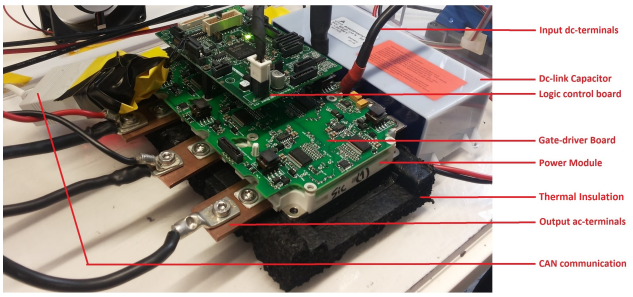
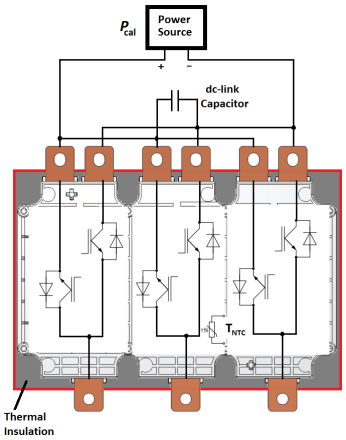


Figure 14: The Complete Inverter System in the Calorimetric Test Bench with the Thermally Isolated Baseplate

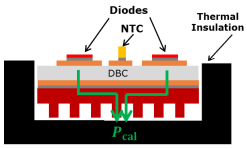
7.1 Calibration Stage

In this stage, the inverter system is connected to a dc source (with opposite polarity) with voltage sense pins capable of serving as a constant power source. The output ac terminals are disconnected as shown

in the schematic in figure 15(a). A known amount of constant power P_{cal} is injected into the module through the diodes. As there is no output power, all the injected power is dissipated in the module as heat, which is trapped in the heatsink on account of the thermal insulation, as seen in figure 15(b). As the temperature of the diode increases, its voltage drop changes and the dc source must be capable of suitably adjusting the current to maintain the power constant. Due to the thermal insulation, and the absence of convection, almost all the heat is trapped in the capacitance of the baseplate, and goes on to increase its temperature exponentially as shown in figure 16(a).



(a) Schematic Diagram



(b) Thermal Stack

Figure 15: Calibration Setup

There is a small amount of heat that is radiated into the ambient, or leaves the modules through any surface other than the heatsink. However, this effect will be cancelled out and have no impact on the accuracy of this method if the test setup during the calibration and measurement stages is identical. The temperature sensed by the NTCs is recorded until it rises from $T_1 = 50°C$ to $T_2 = 90°C$[2], after which the dc source is switched off and the system is let to cool down to the ambient temperature. In order to filter out measurement noise, an exponential curve is fitted to the measurement. From this fitted curve, the time taken for the temperature to reach T_1 from T_2 is taken as the *rise time* t_r. The temperature *slew rate* s_r is calculated as

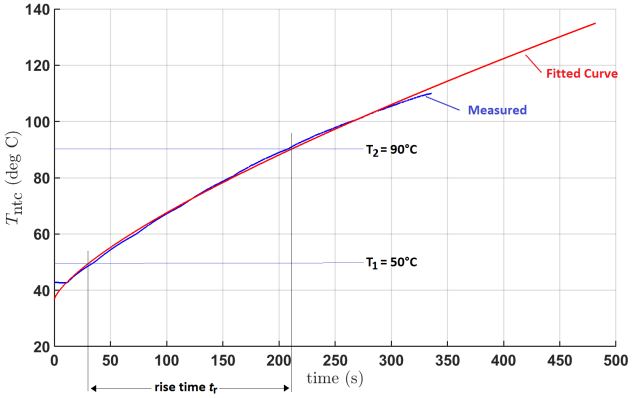
$$s_r = \frac{T_2 - T_1}{t_r} \tag{26}$$

This experiment is repeated at different values of injected power and P_{cal} is plotted against s_r as shown in figure 16(b) for both the modules and first order curves are now fitted to the two curves respectively, and the equations of the fitted curves are given below:

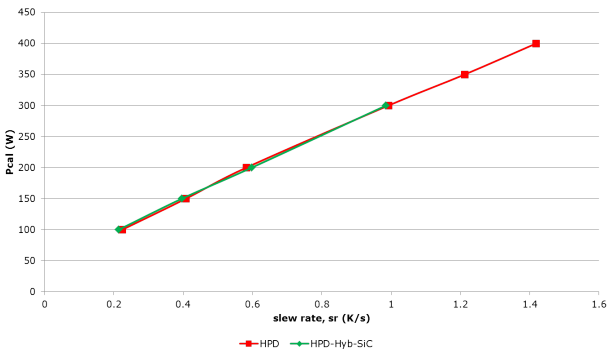
$$P_{loss} = 249.68 \cdot s_r + 48.655 \quad \text{[HPD Module]} \tag{27}$$

$$P_{loss} = 258.15 \cdot s_r + 45.905 \quad \text{[HPD-Hyb-SiC Module]} \tag{28}$$

As seen in figure 16(b), the calibration curves for the two modules match very closely, owing to the similar construction of the module. This is in line with the objective of this work to have minimum discrepancies arising out of differences in the packaging, for a fair comparison of Si and SiC. The slight difference seen between the curves can be attributed to the difference in the NTCs and the temperature measurement tolerances, which are cancelled out due to this method.



(a) NTC temperature at P_{cal} =100W



(b) Calibration Curve

Figure 16: Calibration

7.2 Measurement Stage

In this stage, the inverter is connected to a dc voltage-source (with normal polarity). Care must be taken to see that the setup, particularly the thermal insulation, is not disturbed between this stage and the calibration stage. The output ac terminals are connected to a three-phase star-connected passive load as shown in

[2]The choice of $T_1 = 50°C$ and $T_2 = 90°C$ is based on the observation that between these temperatures, the exponential curve is nearly linear. However, a different set of values may be chosen, provided that the temperature rise curve is nearly linear in this interval. But it must be ensured that the temperature limits chosen during calibration and measurement stages are the same.

figure 17 and the inverter is run in open-loop mode. A suitable rms current I_{rms} is established in the load, by adjusting m appropriately. As with the calibration stage, s_{r} is determined.

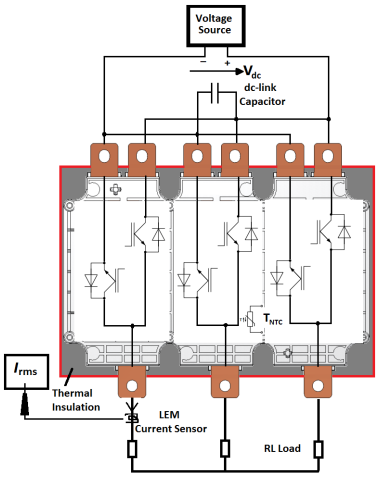
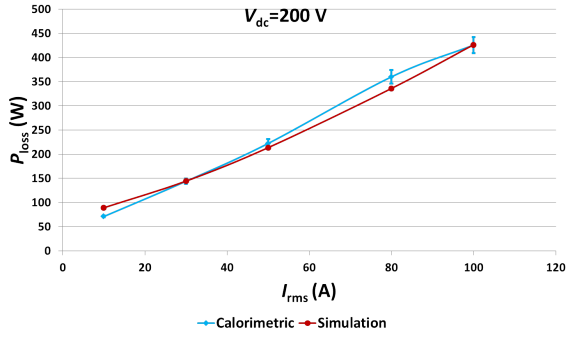
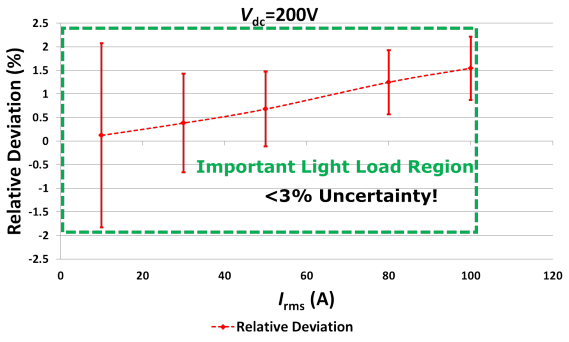


Figure 17: Measurement Stage



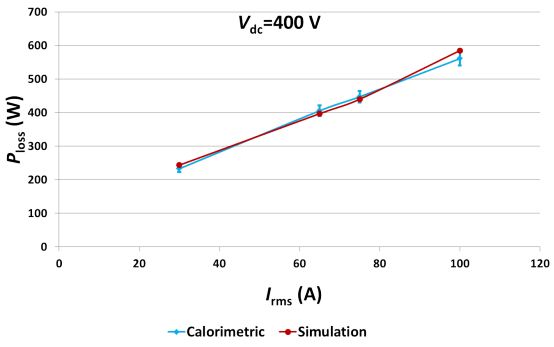
(a) P_{loss}



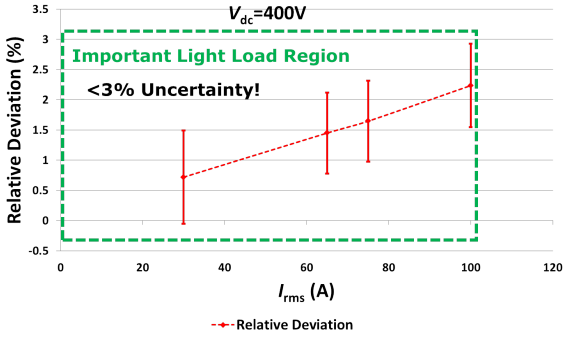
(b) Relative Deviation (%)

Figure 18: Comparison of the calorimetric method with simulations: P_{loss} vs. I_{rms} at V_{dc} = 200V

Now, the inverter losses can be obtained by substituting the measured s_{r} in equation 27 and 28 obtained from the calibration stage. The inverter power losses for different I_{rms} at two different working voltages 200 V and 400 V measured for HPD with the proposed calorimetric method are shown in figures 18(a) and 19(a). These are compared against simulations at the respective points. Also shown are the calculated uncertainties (this topic is described in [3]) for each of the points as bars around the measurement points. The relative deviation between the measurements and the simulations, expressed as percentage, is shown in figures 18(b) and 19(b). Across the entire range of measurement, it can be seen that the simulations are within the tolerance of the calorimetric approach, again validating the behavioral simulation model. Furthermore, in each of these cases, it can be seen that the uncertainty is well below 5%, especially at partial-load which makes the calorimetric method suitable for automotive main inverter applications, particularly for comparing chip technologies whose total power losses may differ in the range of 10-20%.



(a) P_{loss}



(b) Relative Deviation (%)

Figure 19: Comparison of the calorimetric method with simulations: P_{loss} vs. I_{rms} at V_{dc} = 400V

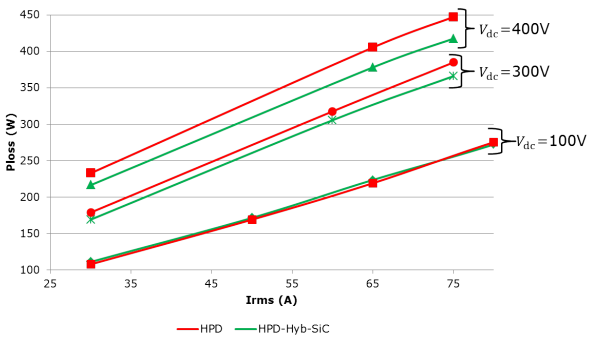


Figure 20: Measured Inverter Losses vs I_{rms} at various V_{dc} for HPD and HPD-Hyb-SiC

7.2.1 Comparison of Measured Inverter Losses for HPD versus HPD-Hyb-SiC

The tests are repeated for different operating points of V_{dc}, I_{rms} and f_{sw} for both the modules. Figure 20 shows a summary of the measured losses as a function of I_{rms} at different dc-link voltages. It can be seen that at V_{dc}=100V, the SiC module has almost the same losses as the Si module, offering no benefit. This is because at low V_{dc}, the switching losses do not contribute much to the total losses, and the conduction losses are dominant. As the conduction losses are higher in the SiC diodes as can be seen from the static curves presented in [2], HPD offers better overall performance than HPD-Hyb-SiC. At V_{dc}=300V, the benefits of SiC become prominent. At V_{dc}=300V, I_{rms}=75 A, there is about 5% reduction in the total losses. This gap widens as we increase V_{dc} as the switching losses become more dominant, and at V_{dc}=400V, 75A, we can see a reduction of around 7%.

8 Conclusions

In this paper, the benefits of replacing the Si diodes of a commercial automotive IGBT module with SiC diodes have been investigated for the main inverter application, maintaining the operating conditions, package and the rest of the system the same, to ensure a fair comparison of the devices without any external influence. A behavioral power loss model, suitable for mission profile analysis, is used to compare the performance of the two modules over several mission profiles. The highest benefit of using SiC diodes is seen for the Artemis Urban drive cycle, where the SiC diodes help reduce the overall losses by 20% at f_{sw}=8kHz. This translates to a saving of around 200Wh of battery energy per 100km. The behavioral model is experimentally verified by comparing it against two independent measurement methods, namely, electrical input output method and a calorimetric method. In each case, the simulation results are found to be within the tolerance of the measurements, thereby validating the simulation model used. At V_{dc} =400V, I_{rms} =75A, the inverter losses were found to be reduced by over 5% with the SiC module, due to the absence of reverse recovery in the unipolar SiC schottky diodes. This reduction is even better at higher dc-link voltages, due to the switching losses becoming more prominent.

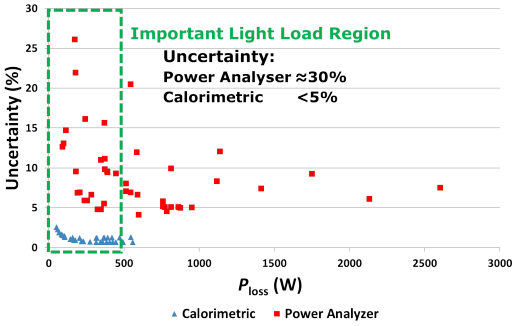


Figure 21: Scatter Plot of the uncertainties (%) at various measurement points for the two methods vs P_{loss}

Further, as summarized in figure 21, it can be concluded that the commonly used power analyser based electrical method has an uncertainty of nearly 30% in the light load condition, mainly due to delays and phase-shifts in the probes. It is to be noted that automotive traction inverters operate most of the time in the light-load condition, which means that the electrical method is not suitable for such applications. The developed calorimetric method outperforms the standard electrical input-output based method and achieves, especially in the important light-load region, a measurement uncertainty of lower than 5%. Furthermore, as the expensive calorimeter is not required for this method, it is nearly as easy to perform as the power-analyser based electrical method. This makes it ideal for comparing device technologies such as Si versus SiC in automotive main inverter applications.

9 Future Work

This work considered the advantages of replacing only the diodes with SiC. However, for higher benefits, it is desirable to replace the IGBTs with SiC MOSFETs, and it would be interesting to investigate the benefits they bring in terms of higher efficiency for different mission profiles. This will considered in a future publication.

References

[1] Ajay Poonjal Pai, Tomas Reiter, and Martin Maerz. Mission profile analysis and calorimetric loss measurement of a sic hybrid module for main inverter application of electric vehicles. In *Integrated Power Packaging (IWIPP), 2017 IEEE International Workshop On*, pages 1–5. IEEE, 2017.

[2] Ajay Poonjal Pai, Tomas Reiter, and Martin Maerz. Characterization and mission profile analysis of a SiC hybrid module for main inverter application of electric vehicles. In *EEHE 2017 Bamberg; Proceedings of*. Haus Der Technik, 2017.

[3] Ajay Poonjal Pai, Tomas Reiter, Oleg Vodyakho, Inpil Yoo, and Martin Maerz. A calorimetric method for measuring power losses in power semiconductor modules. In *EPE 2017 Warsaw; Proceedings of*. ECCE Europe, 2017.

[4] Kimimori Hamada. Great potential of SiC devices for environmentally friendly vehicles. In *APE Automotive Power Electronics*. IEEE, 2015.

[5] Ming Su, Chingchi Chen, Shrivatsal Sharma, and Jun Kikuchi. Performance and cost considerations for SiC–based HEV traction inverter systems. In *Wide Bandgap Power Devices and Applications (WiPDA), 2015 IEEE 3rd Workshop on*, pages 347–350. IEEE, 2015.

[6] Benjamin Wrzecionko, Dominik Bortis, and Johann W Kolar. A 120 C ambient temperature forced air–cooled normally–off SiC JFET automotive inverter system. *IEEE Transactions on Power Electronics*, 29(5):2345–2358, 2014.

[7] SK Singh, F Guédon, PJ Garsed, and RA McMahon. Half-bridge SiC inverter for hybrid electric vehicles: Design, development and testing at higher operating temperature. In *Power Electronics, Machines and Drives (PEMD 2012), 6th IET International Conference on*, pages 1–6. IET, 2012.

[8] Fei Shang, Alejandro Pozo Arribas, and Mahesh Krishnamurthy. A comprehensive evaluation of SiC devices in traction applications. In *Transportation Electrification Conference and Expo (ITEC), 2014 IEEE*, pages 1–5. IEEE, 2014.

[9] Justin K Reed, James McFarland, Jagadeesh Tangudu, Emmanuel Vinot, Rochdi Trigui, Giri Venkataramanan, Shiv Gupta, and Thomas Jahns. Modeling power semiconductor losses in HEV powertrains using Si and SiC devices. In *2010 IEEE Vehicle Power and Propulsion Conference*, pages 1–6. IEEE, 2010.

[10] Hui Zhang, Leon M Tolbert, and Burak Ozpineci. Impact of SiC devices on hybrid electric and plug-in hybrid electric vehicles. *IEEE transactions on industry applications*, 47(2):912–921, 2011.

[11] Burak Ozpineci, Madhu Sudhan Chinthavali, Leon M Tolbert, Avinash S Kashyap, and H Alan Mantooth. A 55-kW three-phase inverter with Si IGBTs and SiC schottky diodes. *IEEE Transactions on industry applications*, 45(1):278–285, 2009.

[12] M Chinthavali, Leon M Tolbert, Hui Zhang, Jung H Han, F Barlow, and Burak Ozpineci. High power SiC modules for HEVs and PHEVs. In *Power Electronics Conference (IPEC), 2010 International*, pages 1842–1848. IEEE, 2010.

[13] Timothy Junghee Han, Jim Nagashima, Sung Joon Kim, Srikanth Kulkarni, and Fred Barlow. Implementation of a fully integrated 50 kW inverter using a SiC JFET based six–pack power module. In *2011 IEEE Energy Conversion Congress and Exposition*, pages 3144–3150. IEEE, 2011.

[14] Florian Hilpert, Klas Brinkfeldt, and Stefan Arenz. Modular integration of a 1200 V SiC inverter in a commercial vehicle wheel–hub drivetrain. In *Electric Drives Production Conference (EDPC), 2014 4th International*, pages 1–8. IEEE, 2014.

[15] Product brief HybridPACK drive. *Infineon AG*, 2014.

[16] Datasheet 5th generation thinQ 650V SiC schottky diode IDW40G65C5, author=, journal=Infineon AG, year=2013.

[17] Ajay Poonjal Pai, Tomas Reiter, and Martin Maerz. A new behavioral model for accurate loss calculations in power semiconductors. In *PCIM Europe 2016; Proceedings of*. VDE, 2016.

[18] Ajay Poonjal Pai, Tomas Reiter, and Martin Maerz. An improved behavioral model for loss calculations in automotive inverters. In *EEHE 2016 Wiesloch; Proceedings of*, pages 412–427. Haus Der Technik, 2016.

[19] Lem current transducer IT 200-S ULTRASTAB, author=, journal=LEM, year=2014.

[20] Lem current transducer LF 510-S, author=, journal=LEM, year=2015.

[21] Yokogawa WT1800 getting started guide, author=, journal=Yokogawa, year=2015.

[22] C. Xiao, G. Chen, and W. G. H. Odendaal. Overview of power loss measurement techniques in power electronics systems. *IEEE Transactions on Industry Applications*, 43(3):657–664, May 2007.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 148-154 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Development of Indicators for Technical Condition Indexing of Power Transformers

Gints Poiss*, Sandra Vitolina, Janis Marks

Riga Technical University, Department of Electrical Machines and Devices, Latvia

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 08 January, 2018
Online: 30 January, 2018

Keywords :
Power transformer
Condition indexing
Fuzzy logic
Transformer windings
Oil insulation

ABSTRACT

Reliable operation of a power transformer with a certain load depends on the technical condition of individual construction parts and the ability to prevent defects that can cause a failure. During the lifecycle of a transformer, valuable data is constantly accumulated, which forms the basis for technical or risk assessment of the equipment. Therefore it also serves as a ground for the decisions on further operation, or repairs, or replacement. To achieve this goal, data need to be systematized. Since technical condition indexing allows combining various types of data including results of diagnostic tests is used within the framework of this research.

As part of a larger risk assessment methodology, algorithms for two indicators are proposed in this paper, and they are based on results of electrical measurements and analysis of oil parameters, respectively. The novelty of the algorithms for indicators introduced in this paper is based on analysis of features specific to the power system in Latvia such as large proportion of aged transformers, low loading level, significant variation in oil volumes, and statistics on typical faults. Proposed limits of parameters are verified with data from operation history. Taking into account the differences in the measurement periodicity, the indicator that is based on electrical measurements assesses the individual constructive parts of the transformer (windings and core, bushing and on-load tap changer) separately, whereas the other, indicator combines the results of oil parameters into a single assessment. These indicators were verified by using 30 transformers from the Latvian power system and the obtained results coincide well with the operation history.

1. Introduction

This paper is an extension of the results disseminated in the international conference paper relating to the development of risk indicator for power transformers based on electrical measurements [1]. This risk indicator is part of a condition indexing algorithm that allows categorizing a large volume of technical data obtained during operation and maintenance of a power transformer into discrete risk categories. It provides information in a fast and systematic way, so that transformers can be easily compared and critical ones can be selected for further investigation.

The main goal of this extended paper is to present an algorithm for determining the numerical value of the indicator based on results of electrical measurements (EM indicator), as well as for establishing the numerical value of the indicator based on results of oil parameters analysiss. The outcome for both indicators is the risk category derived on a scale from 1 (lowest risk) to 5 (highest risk, outage expected or immediate action necessary). The proposed indicators correspond to diagnostic tests regularly performed on power transformers in Latvian transmission system. However, significant differences in the periodicity of oil checks and electrical measurements determined the disparate structure of both algorithms and the mathematical methods selected.

The proposed indicators have been verified with 30 case studies. Results of diagnostic tests both from transformers that required repairing, and transformers in a normal operation condition were used. The authors express their gratitude to the Latvian Transmission System Operator (Augstsprieguma tikls, JSC) for providing data for the case studies.

Different approaches are applied to develop appropriate and efficient algorithms for technical condition index which includes such common power transformer parameters as oil characteristics, dissolved gas analysis and electrical measurements [2].

Artificial neural networks (ANN) and adaptive neuro-fuzzy inference system (ANFIS) models, for instance, are no longer limited just for diagnosing faults and aging of power transformers.

*Corresponding Author: Gints Poiss, Email: gints.poiss@edu.rtu.lv

These models can also be used to quantify the general condition of a power transformer [3], also as synthetic risk assessment methods [4] and fuzzy logic [5].

The different diagnostic tools and methods, periodicity and power system charchteristics are actually driving the need to develop a unique, expert-made assessment models for a certain power system. The novelty of the algorithms for indicators introduced in this paper is based on analysis of features specific to the power system in Latvia such as large proportion of aged transformers, low loading level, significant variation in oil volumes, and statistics on typical faults. Proposed limits of parameters are verified with data from operation history. The algorithms provide evaluation of the main constructive parts of a power transformer, and also the structure of indicator algorithms is planned as adjustable, if the amount of applied diagnostics tests is changed.

2. Technical Condition Index as a Part of Risk Matrix

Configurations of the risk matrix proposed for Latvian power system within the framework of this research is shown in Figure 1, and can be effectively used to evaluate each power transformer in the system, as well as to plan the maintenance or replacement of aged units. Risk matrix is divided into 3 parts, where green indicates a low-risk with no concern, blue indicates a moderate risk region or a transformer in a normal operation and technical condition, whereas the red region indicates a need for immediate action. The greatest attention will be paid to 4 cases depicted in Figure 1, that fall in the high-risk region.

In Figure 1, the ordinate axis of the risk matrix, shows operation characteristics and it is based on such important parameters as transformer age, load, maintenance history, the existence of monitoring system, importance in the system (as additional option). Three indicators are used to determine the technical condition index. Together with scoring and weighting factors it is depicted as a value on abscise axis of the risk matrix. Parameters used to determine the technical condition index are given in Table 1 and they correspond to diagnostic tests regularly performed on power transformers in Latvian transmission system.

The algorithm of the indicator based on dissolved gas analysis (DGA) is provided in [6]. It is based on assessment of 7 key gases shown in Table 1 and includes analysis of features specific to the power system in Latvia where power transformers are aged and variations in oil volumes are significant. Transformer loading, operation of the on-load tap changer (OLTC), and oil treatment are also taken into account in order to quantify this indicator.

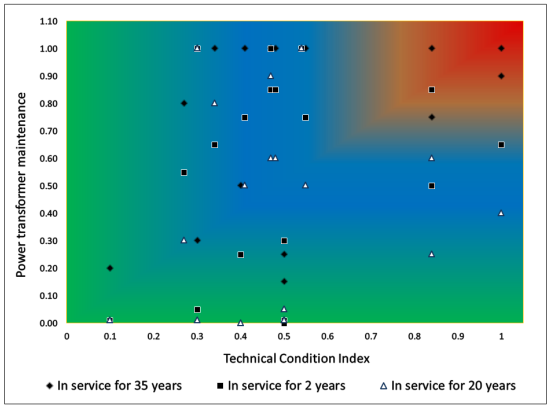


Figure 1. Proposed configuration of risk matrix

The indicator based on electrical measurements for technical condition indexing of a power transformer is developed by evaluating separately 3 main parts of a power transformer: windings, bushings and OLTC. A failure of any of these parts may be critical for a power transformer, the surrounding equipment, environment and even service personnel. The indicator based on analysis of oil parameters such as flash point, dissipation factor, acidity and moisture content is developed as a combined approach of binary and fuzzy logic which allows evaluating the risk on basis of several oil parameters as a single entity.

Both indicators are described in detail in the following chapters of this article.

Table 1. Indicators for technical condition indexing

Electrical measurements		
Windings and core	***Bushings***	***OLTC***
Power factor, PF	Power factor, PF	Static resistance, SR
Insulation resistance, R_m	Insulation resistance, R_m	Transition time & current ripple from dynamic resistance measurement (DRM)
Short-circuit impedance, Z_k	Capacitance, C1	
No-load losses, P_0		

Transformer oil	
Oil analysis	***DGA***
Dissipation factor	H_2, CH_2
Flash point	C_2H_4 C_2H_6
Moisture content	C_2H_2
Acidity	CO, CO_2

3. Development of Electrical Measurement Indicator

3.1. Winding and core

Power factor, insulation resistance and short-circuit impedance allow verifying the technical condition of transformer windings, whereas variations in no-load losses reflect the condition of a transformer core.

A flowchart of a multi-step algorithm is shown in Figure 2. where scoring system for windings and core as a part of the EM indicator is derived on a scale from T=1 (lowest risk) to T=5 (highest risk, outage expected).

The first step is input of measured parameter values from test reports: insulation resistance (R_m), power factor (PF), short-circuit impedance (Z_k), no-load losses (P_0), and winding temperature readings from a thermometer on a transformer.

Next step is data processing. As a reference point, it requires data from the factory test report for a particular power transformer, as well as background information such as age, rated voltage, etc. This information is stored in a specific technical file that is prepared for each transformer in the system. This file is also used for other condition indexing indicators. Since the technical file contains limits for the evaluation of parameters shown in Table 2, modifications are restricted.

During data processing the input data are modified into a suitable form for further analysis. PF values from the test report are calculated at base temperature of 20℃. Thus the limits proposed in Table 2 and stored in the technical file can be applied for risk evaluation (E1 indicates low risk, and E5 – high risk). PF limits are based on recommendations provided in the standard IEEE C57.152-2013. On basis of maintenance experience, additionally value for level E5 were specified. PF=1.5% at 20℃ is an effective parameter for detecting high risk since moisture and contamination of windings in this case in practice proved to be too high for reliable operation. Another parameter used for high risk detection is the D_t coefficient. It is a ratio between the measured resistance and rated voltage of high voltage winding (R_m/U_n). Coefficient D_t=1MΩ/kV at base temperature of 20℃ is set as the minimum insulation resistance limit [7]. For further analysis, values from the factory test report R_{fact} adjusted to base temperature are used as reference to evaluate if the decrease of measured R_m is acceptable.

Limits for evaluation of short-circuit impedance Z_k are based on [8] and given in Table 2. Two values of Z_k between phases should be within E4 of each other. Lager deviation can indicate on mechanical deformation of windings. A single-phase connection no-load losses P_0 measurement is a test that has been historically used for evaluating the magnetic system of power transformers in Latvia. A specific feature is that different test voltage is used (either 220V or 380V). It causes difficulties if evaluation is based on traditional comparative analysis to a factory test. In order to use results of a single measurement in this algorithm, limits based on statistical data and recommendations of [9] and performed study of measurement history of no-load losses from 100 transformers [1].

3.2. Bushings

Electrical measurement results can provide information about degradation of bushing insulation and help to avoid potential failures, such as bushing explosion and transformer winding deformation. Flowchart for bushing assessment is given in Figure 3, and evaluation process is similar to the windings assessment described before. Although a 3-level scoring system is proposed as more suitable for bushings as more polarizing. It starts with evaluation of results of PF measurements for bushings by applying respective limits from Table 2. for data from test reports are used for input. The input is followed by the next step – data processing where all data are modified in a suitable form just like in the algorithm for windings and the core part.

Table 2. Limits of parameters included in EM indicator

Parameter	Level			
	E1	E3	E4	E5
Windings and core				
PF, %, at 20℃	<0.5	0.5	1	≥1.5
$P_{0\ phase\text{-}phase}$, %		$P_{0\ A\text{-}B}$>40 $P_{0\ A\text{-}B}$<20 $P_{0\ A\text{-}C}$>4		
$Z_{k\ phase\text{-}phase}$, %			≥3	
Bushings				
PF, %, at 20℃	<0.5			≥1
On Load Tap Changer				
$SR_{ph\text{-}ph}$, %	<2		>2 (at least in 3 taps)	≥5
Transition time, t_t, ms	<100			≥200
$I_{N1,2}$, %			60	

If the measured PF value is above the limit E5 as it shown in Table 2 or the rated capacitance C1 of a bushing differs from factory measurement by more than 10 %, to reduce the risk of failure, it is advisable to remove the bushing from service [10].

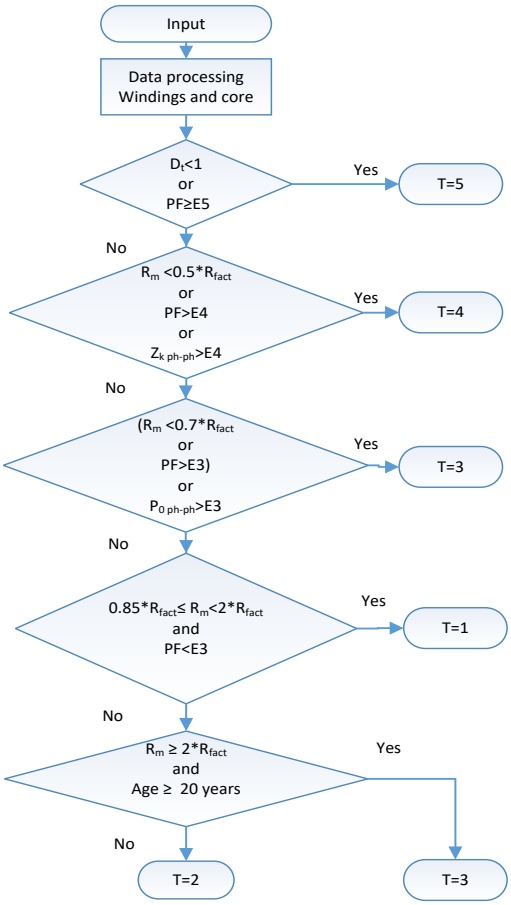


Figure 2. The algorithm of EM indicator for windings and core

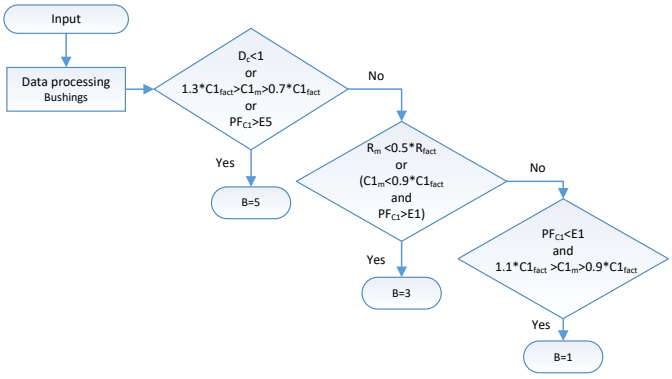


Figure 3. The algorithm of EM indicator for bushing

3.3. OLTC

The static resistance measurement (SRM) is an important tool in this industry and it is used to check for loose connections, broken strands, as well as high contact resistance in tap changer. For decades, SRM is performed for each tap in the tap changer for all power transformers in Latvia and the experience is significant. Therefore SRM is included in the OLTC algorithm as the exclusion rule which can only indicate if the risk is low or high.

For input in Figure 4, SRMs of all phases and all taps are assessed without the factory test report results; therefore the technical file is not needed.

In accordance with [11], the agreement with limit E5 indicates on high risk and it is an important signal to consider the need to repair tap changer immediately. If the SRM difference between phases is within the limit E1, it indicates on low risk. If resistance in 3 taps exceeds the limit E1, it should be investigated and TC=4 is assigned.

Dynamic resistance measurement (DRM) allows detecting defects caused by contact coking, contact wear, oil-film layer deposition, deviating transition times, contact timing problems, maintenance errors and damaged transition resistors. Therefore as additional criterion for the OLTC assessment – the dynamic resistance measurement (DRM) is proposed [1] evaluating two parameters form test report: transition time, ms, and current ripple change $I_{N1,2,}$ %. The proposed limit values E1 and E5 shown in Table 2 are assigned on basis of [12] research and were verified by authors with analysis of 76 test reports of DRM.

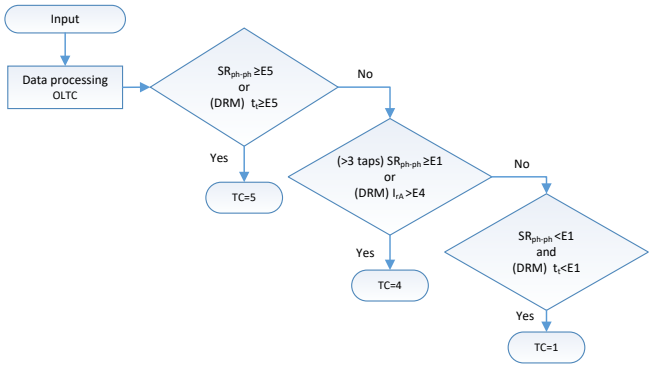


Figure 4. The algorithm of EM indicator for OLTC

4. Development of Oil Indicator

4.1. General description of oil indicator algorithm

Generally mineral transformer oil is used in power transformers in power system of Latvia. Regularly measured oil parameter are dielectric strength, moisture content, acidity, dielectric dissipation factor, interfacial tension, viscosity, flash point, and sludge content. Method for assessment of the measurement results described in the national standard involves comparison of data obtained in laboratory with the given assessments limits, and conclusions are obtained for each oil parameter separately rather than as a single entity. The standard does not provide guidance on how to obtain one conclusion based on multiple contradicting input parameters.

Fuzzy logic is widely applied for dissolved gas analysis (DGA) of power transformers [13, 14] since DGA contains relatively many input parameters. If binary logic is used a contradiction can occur, where some of the input parameters correspond with a good condition but others with a bad condition. Fuzzy logic can provide more detailed assessment of technical condition. For example, fuzzy logic based algorithm with acidity and interfacial tension of oil as input parameters is used to estimate remnant life of a power transformer [15]. Similar approach of data processing is described in [16] in which remaining service life of transformer in years is predicted based values of viscosity, resistivity, particle count, acidity and moisture content of oil. Another research [17] shows the possibility to combine measured values of moisture and acidity of oil, and the power factor of the winding to determine the type of transformer defect based on the physical and chemical condition of insulation in cases with contradicting input data values.

Therefore within the framework of this research fuzzy logic as data processing method is used to develop a technical condition indicator for a power transformer based on measurement results of oil parameters. Four independent oil parameters (flash point, dissipation factor, acidity and moisture content) and additional parameter (changes in flash point value in two subsequent measurements) are used as input data as shown in Figure 5. Limits shown in Table 3 are based on typical values observed for transformers installed in transmission network in Latvia. As a result, algorithm calculates the value of transformer oil indicator within limits from K=1 – (low risk) to K=5 (high risk, immediate actions required).

Flash point (FP) is the lowest temperature at which oil in certain conditions releases such amount of vapour that, together with air, creates a flammable mixture leading to fire hazard. Therefore, this parameter is proposed for indicating high risk. Based on practical experience in decision-making for transformer repairs dissipation factor (tanδ) is proposed as another parameter indicating high risk since it is sensitive to oil aging, moisture, as well as changes in contamination levels.

Table 3. Proposed Limits for Oil Parameters

Flash point, °C	Upper limit	Acidity, mgKOH/g	Moisture content, ppm	Tanδ, %
125 or ≥Δ5 in 2 subseqent measurements	Level 1	<0.05	<5	<1
	Level 2	≤0.075	≤10	≤2.5
	Level 3	>0.1	>15	>4

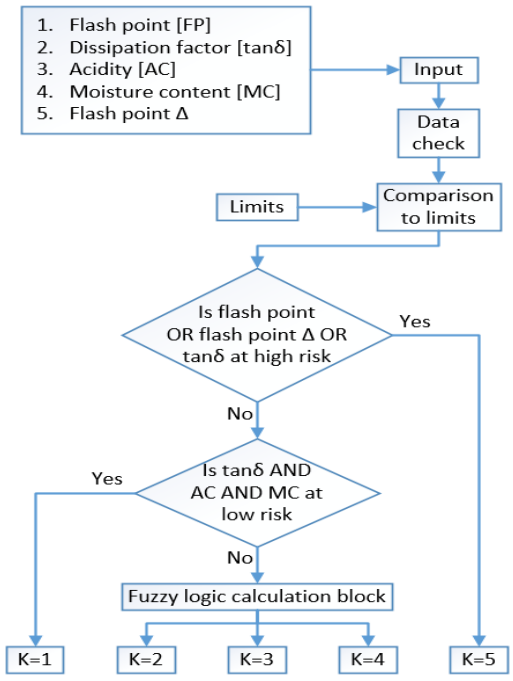


Figure 5. Overall flowchart of oil indicator algorithm

Dissipation factor tanδ along with acidity and moisture content are proposed as criteria for medium risk evaluation. In order to reach a unified conclusion regarding the correlation between these independent oil parameters, the use of fuzzy logic is suggested.

The operation of the algorithm begins with reading the input parameters, namely results from the most recent oil measurement which then are stored in technical file that is prepared for each transformer in the system for possible repetitive usage. If any of the necessary oil parameters have not been measured in this particular testing, the algorithm assumes the value of that parameter as in Level 2 (see in Table 3).

If the value of the parameter is repeatedly unknown, the approximation is performed in a logarithmic scale as shown in Figure 6 where approximated value aims for the Level 3, yet cannot reach it, since the difference from the maximum boundary is halved per iteration.



Figure 6. Assumed values for unknown parameter value

Next step of the proposed algorithm is the evaluation of exclusion cases indicating high risk (either flash point or changes in value of flashpoint or dissipation factor exceeds the limits of Level 3) or low risk (values of dissipation factor and moisture and acidity are all below limits of Level 1). Fuzzy logic calculation block activates if neither of the exception cases checks included in the algorithm is positive.

4.2. Fuzzification

Developed shape and slope for membership functions used within fuzzy logic calculation block of proposed algorithm is shown in Figure 7 originally, the selection of membership function shape and slope was carried out for each of the three oil parameters. Illustration of membership function for tanδ is shown in Figure 8. The triangle and trapeze forms were tested for M (medium risk) membership function, whereas slope was altered for L (low risk) and H (high risk) membership functions. All variations were tested by using the results of oil tests for 10 transformers with given maintenance history and different technical condition.

Results show that output values are not affected by the shape of M membership function, however, in some cases value changes were observed by alteration of the slope of L and H membership functions. After evaluation it was concluded that the combination of increased H membership function effect has the best correlation with technical condition based on maintenance history.

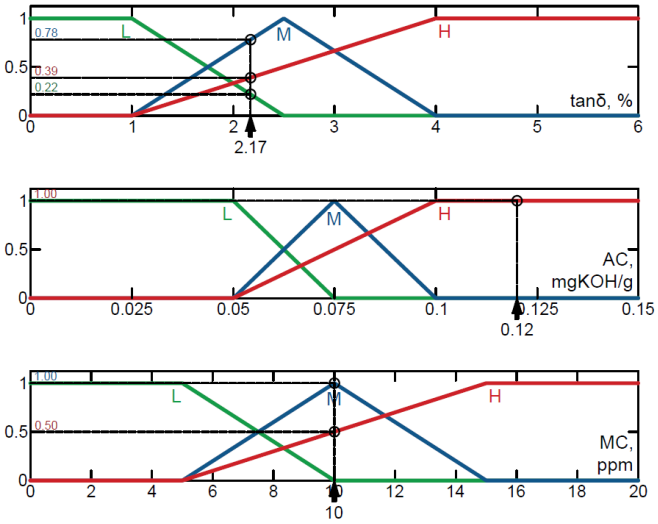


Figure 7. Membership functions of oil indicator

For obtaining corresponding output values, a rulebox is created as shown in Table 4. It contains all possible combinations. Altogether, there are 27 combinations, since for each of the 3 input parameters used in the fuzzy logic block 3 membership functions are assigned. The specific weight value is the minimum of all three membership values for a given combination. The specific weights are summed together for each result value and thus a fuzzified output is obtained. Afterwards, the gravitation centre of this output function is calculated. The obtained result is more precise, as more points are used for the calculation.

4.3. Fuzzy inference and defuzzification

An example is illustrated in Figure 7 with following values of oil parameters as input data from regular oil sample test of a particular 110 kV transformer: flash point, 139°C; changes in flash point value between two subsequent measurements, 2°C; dissipation factor tanδ, 2.17%; acidity, 0.12mgKOH/g; moisture content, 10 ppm. Since exception cases included in the algorithm are negative in this case fuzzy logic calculation block activates.

Table 4. Rulebox for fuzzy logic calculation block

#	Input Variables			Corresponding output	#	Input Variables			Corresponding output	#	Input Variables			Corresponding output
	tan δ	*AC*	*MC*			*tanδ*	*AC*	*MC*			*tanδ*	*AC*	*MC*	
1	L	L	L	2	10	H	M	L	3	19	L	M	H	3
2	M	L	L	2	11	H	L	M	3	20	M	M	H	3
3	L	M	L	2	12	H	M	M	3	21	H	H	L	4
4	L	L	M	2	13	L	H	L	3	22	H	H	M	4
5	M	M	L	2	14	M	H	L	3	23	H	L	H	4
6	M	L	M	2	15	L	H	M	3	24	H	M	H	4
7	L	M	M	2	16	M	H	M	3	25	L	H	H	4
8	M	M	M	2	17	L	L	H	3	26	M	H	H	4
9	H	L	L	3	18	M	L	H	3	27	H	H	H	4

The membership of dissipation factor is 0.78, 1 for acidity and 0.5 for moisture content. The specific weight is calculated as 0.5. This process is repeated for each matching combination. The gravitation centre of this output function is calculated, as well as the position of gravity centre on x axis is estimated as shown in Figure 9. The third of the position of gravity centre is the result and final output, in this case K=4 for oil indicator is obtained, which indicates a rather high operation risk.

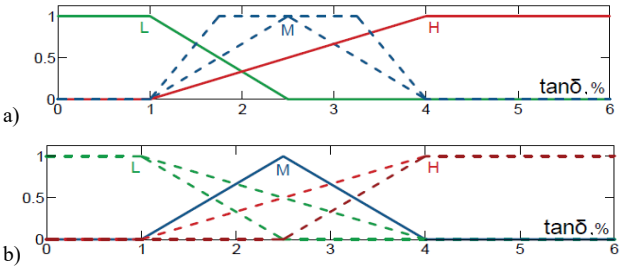


Figure 8. Selection of the shape and slope for membership functions for tanδ: a) shape for M membership function; b) slope for L and H membership function

	Membership function	Membership value
tanδ	L	0.22
	M	0.78
	H	0.39
AC	L	0
	M	0
	H	1
MC	L	0
	M	1
	H	0.5

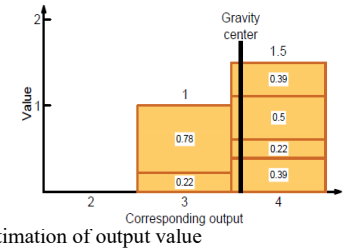


Figure 9. Estimation of output value

5. Verification of EM and Oil Indicator

Results of measurements within a time period of 8 years for 30 power transformers are used to verify both indicators and see how they complement each other. It has to be noted that the power transformer park in Latvia is aged and power transformers installed between 1967 and 2000 were used for verification with different technical condition, repairing plans and failure rate. Figure 10 shows numerical values of EM and oil indicator obtained for the latest available measurement set of each transformer. Since polarized score is used omitting the values 2 and 4 for bushings and 2 and 3 for OLTC the combined score 3&4 is allocated in a separate column. Mainly a numerical value of 1 was assigned to parameters OLTC and bushings which reflects timely scheduled repairs. In those two cases where score value of 5 is assigned for bushings immediate decisions to replace them have been made by system operator. Similarly OLTC is repaired immediately if there1 is a variation from normal operation or a problem is detected.

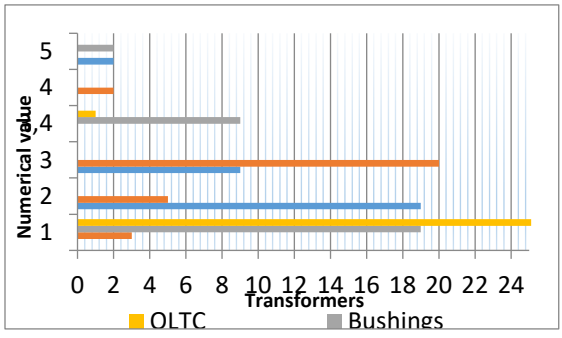


Figure 10. Numerical values of indicator parameters of 30 power transformers

Values 2 and 3 are obtained most frequently for oil analysis most showed result 2 and 3, which indicates either normal operation or marks out necessity for small maintenance works such as the change of silicagel in a thermosyphon. In two cases when score of 4 for winding and core is assigned result dramatic decrease of winding isolation resistance and increase of power factor therefore obtained results concise well with operation history.

For further analysis 4 cases from 30 were selected to illustrate how indicators together reveal the technical condition of a power transformer. Cases 1 to 3, plotted in Figure 11, depicts the latest measurements available for three different transformers. Case 2 reflects a transformer in an almost perfect technical condition. But in case 1, problems with OLTC can be noticed and analysis of measurement history revealed increased static resistance in 3 taps. Case 3 reveals bushing defect, caused by decreased insulation resistance in scheme C2 and increased power factor in scheme C1.

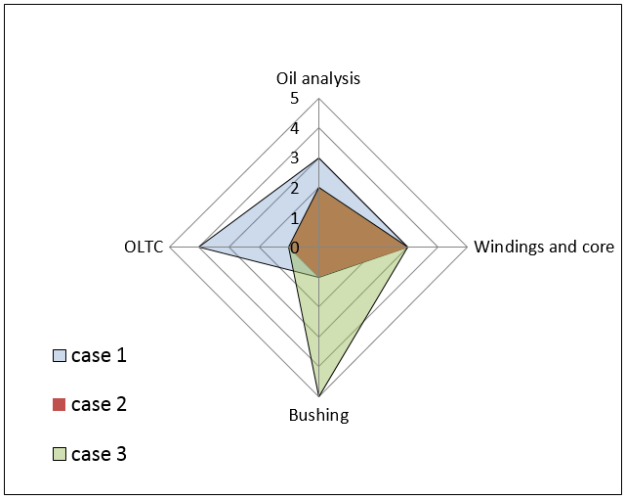


Figure 11. Case studies results

Since electrical measurements are performed on average once in four years EM indicator is more static. Case 4, plotted in Figure 12, illustrates the role of the oil indicator as the first implication of a possible fault in a transformer since its results are more dynamic. It can be observed that technical condition of this particular transformer has decreased due to increase in moisture level over the years.

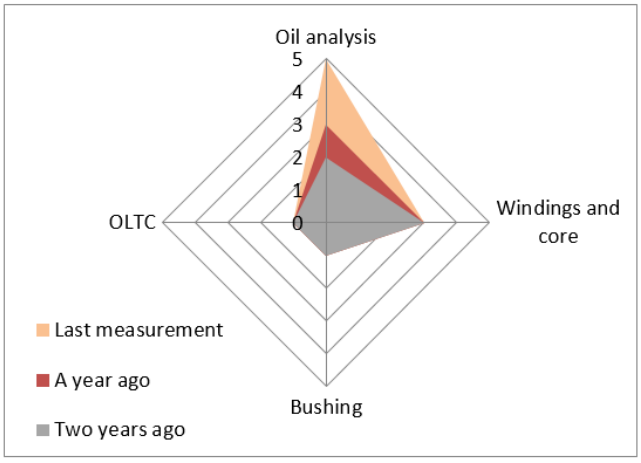


Figure 12. Oil analysis results

6. Conclusions

Literature review leads to a conclusion that methods for risk assessment of power transformers are mainly based on the results of performed diagnostic measurements and operation history. Since maintenance strategy may vary over the time and may depend on decisions by a particular transmission system operator the development of a modified risk matrix is a common global practice.

Configuration of risk matrix for power transformers in Latvian transmission network proposed within the framework of this research is composed of three regions indicating low risk, moderate risk and high risk the last one requiring immediate action. The place of a transformer within the risk matrix depends on its technical condition index (x-axis) and such parameters of operational history as transformer age, load, maintenance history, importance in the system (y-axis). Technical condition index is established by calculating the numerical values of the three indicators and applying a scoring and weighting factor.

To determine numerical values of the indicator based on oil analysis combined approach of binary and fuzzy logic proved to be successful since it allows evaluating risk based on several oil parameters as a single entity. Results from a verification confirm that higher output values are shown by transformers that have already attracted attention in practice. As oil analysis is done quite often, changes in values of this indicator are first indication of a possible defect in a transformer and they substantiate the necessity for further testing. However, an indicator based on results of electrical measurements more effectively serves to detect faulty constructive parts more effectively and can be used as a basis for decisions on replacement of a bushing or repairs of an OLTC.

Acknowledgment

This paper has been partly supported by the State Research Program „LATENERGI".

Conflict of Interest

The authors declare no conflict of interest.

References

[1] G. Poiss, S. Vitolina, "Development and Implementation of Risk Indicator for Power Transformers Based on Electrical Measurements", 18th International Scientific Conference on Electric Power Engineering (EPE), Czech Republic, Kouty nad Desnou, 17-19 May, 2017, pp.425-428.

[2] J. Haema, R. Phadungthin, "Development of Condition Evaluation for Power Transformer Maintenance", 4th International Conference on Power Engineering, Energy and Electrical Drives, Istanbul, Turkey, 13-17 May 2013.

[3] Hamed Zeinoddini-Meymand, Behrooz Vahidi, "Health index calculation for power transformers using technical and economical parameters ", IET Science, Measurement & Technology, Volume: 10, Issue: 7, 10 2016.

[4] Dun Lin, Yao-Yu Xu, Yu Liang, Yuan Li, Ning Liu, Guan-Jun Zhang, "A Risk Assessment Method of Transformer Considering the Economy and Reliability of Power Network", 1st International Conference on Electrical Materials and Power Equipment (ICEMPE), Xi'an, China, 2017.

[5] Juan. P. Lata , Diego. P. Chacón-Troya , R. D. Medina, "Improved tool for power transformer health index analysis", IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON), 2017.

[6] G.Poiss, "Development of DGA Indicator for Estimating Risk Level of Power Transformers," 17th International Scientific Conference on Electric Power Engineering (EPE 2016), Prague, 16-18 May, 2016.

[7] A Stitch in Time: The Complete Guide to Electrical Insulation Testing, Dallas: Megger, 2006.

[8] Scope and norms of testing electric equipment (in Russian), RAO "UES of Russia", RD 34.45-51.300-97, 1997.

[9] Kaganovich E. A., Reichman I. M., Power transformer tests up to 6300 kVA and voltages up to 35 kV, (in Russian), Energia, 1980.

[10] Electric Power Transformer Engineering, Third Edition, James H. Harlow CRC Press, May 16, 2012 by CRC Press 693 pages.

[11] "IEEE Guide for Diagnostic Field Testing of Fluid-Filled Power Transformers Regulators and Reactors," IEEE Std. C57.104, 2013.

[12] J. J. Erbrink, E. Gulski, J. J. Smit, R. Leich, B. Quak, and R. A. Malewski, "On-load tap changer diagnosis-an off-line method for detecting degradation and defects: Part 2," IEEE Electr. Insul. Mag., vol. 27, no. 6, pp. 27-36, 2011.

[13] A. Abu-Siada, S. Hmood, and S. Islam, „A New Fuzzy Logic Approach for Consistent Interpretation of Dissolved Gas-in-Oil Analysis," IEEE Transactions on Dielectrics and Electrical Insulation, vol. 20, pp. 2343 - 2349, 2013.

[14] B. Nemeth, S. Laboncz, and I. Kiss, „Condition Monitoring of Power Transformers using DGA and Fuzzy Logic" on Proc. 2009 IEEE Electrical Insulation Conference, Montreal, Canada, May 31-June 3, 2009, pp. 373-376

[15] S Forouhari, and A Abu-Siada, „Remnant Life Estimation of Power Transformer Based on IFT and Acidity Number of Transformer Oil" on Proc IEEE 11th International Conference on the Properties and Applications of Dielectric Materials (ICPADM), Sydney, Australia, July 19-22, 2015, pp. 552-555.

[16] A. K. Kori, A. K. „Sharma, and A. K. Singh Bhadoriya, Intelligent Diagnostic Method for Ageing Analysis of Transformer," Energy and Power Engineering, Vol. 4 No. 2, 2012, pp. 53-58

[17] W. C. Flores, E. E. Mombello, J. A. Jardini, G. Ratta, and A. M. Corvo, „Expert system for the assessment of power transformer insulation condition based on type-2 fuzzy logic systems," Expert Systems with Applications, vol. 38, pp. 8119-8127, 2011.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 151-156 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Analysis of Economic Load Dispatch with a lot of Constraints Using Vortex Search Algorithm

Mustafa Saka[*,1], Ibrahim Eke[2], Suleyman Sungur Tezcan[1], Muslum Cengiz Taplamacioglu[1]

[1]*Gazi University, Electrical and Electronics Engineering Department, 06500, Turkey*

[2]*Kirikkale University, Electrical and Electronics Engineering Department, 71450, Turkey*

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 17 November, 2017
Online: 11 December, 2017

Keywords:
Economic load dispatch
Valve point effect
Ramp rate limits
Prohibited zone
Transmission losses
Optimization
Vortex search algorithm

ABSTRACT

In modern powers systems, one of the most considerable topics is economic load dispatch (ELD). ELD is non-linear problem and it became non-convex and non-smooth problem with some constraints such as valve point loading effect. ELD is very crucial for energy generation and distribution in power systems. For solving ELD problem, a lot of methods were developed and used at different power systems. Vortex search algorithm (VSA) is proposed and applied for solving ELD problem in this paper. VSA method was developed in the form of stirring liquids. Transmission line losses, valve point loading effect, ramp rate limits and prohibited zones constraints were used to make the results of ELD problem the closest to the truth. The results which are obtained from VSA compared with PSO, CPSO, WIPSO, MFO, GA and MRPSO techniques. It can be clearly seen that VSA gave minimum cost values with optimum generator powers so it is very effective and useful method and it gave the best solutions for ELD.

1. Introduction

This paper is an extension of work originally presented in 4th International Conference on Electrical and Electronics Engineering [1]. The purpose of this work is solving economic load dispatch problem with a lot of constraints through with new optimization technique Vortex Search Algorithm.

Economic load dispatch of a power system is very important in terms of control and planning of that power system. Main goal of ELD is distributed total demand power among the committed thermal generation units with minimum production cost by satisfying set of equality and inequality constraints. If ELD problem is not solved for thermal power plants, demand power may be generated very costly. ELD problem can be basically modeled second order (quadratic) function [2]. However, this function may became more complex, non-smooth and non-convex with some constraints such as valve point loading effect, ramp rate limits, transmission line losses and prohibited zones.

Economic load dispatch plays very big role for operated power plants. For this reason, a lot of researcher studied this issue. A number of optimization techniques developed and applied to ELD problem. Quadratic Programming [3], Linear Programming [4], Non-Linear Programming [5], Lambda Iteration Method [6] etc. techniques were used for solving ELD problem. These traditionally techniques gives good results for basic ELD problem but these techniques may poor results when constraints and complexity are increased.

Together with the advances in the computer sciences, a lot of random search optimization techniques developed and used [7]. Different Evolution [8], Particle Swarm Optimization [9], Genetic Algorithm [10], Artificial Bee Colony [11], Harmony Search [12], Bacterial Foraging Optimization [13], Firefly Algorithm [14], Ant Colony Optimization [15] etc. are some of these techniques. User defined parameters are necessary for these optimization techniques. If the parameters are not chosen properly, the results obtained from these techniques may not be good results.

The organization of this paper as follows: Economic load dispatch, main objective of ELD, constraints of ELD and mathematical express of ELD are described in Section 2. Vortex Search Algorithm and its mathematical model are described in Section 3. Using test system, its parameters, obtained results and

[*]Corresponding Author: Mustafa SAKA, Gazi University, Graduate School of Naturel and Applied Sciences, Ankara, TURKEY, msaka@gazi.edu.tr

figures are described in Section 4. Finally evaluation of this paper is briefed in Section 5.

2. Economic Load Dispatch

There are a lot of operating cost for thermal power plants such as fuel cost, personal fees etc.. In these costs the biggest share is fuel cost of thermal generation units. For this reason solving economic load dispatch problem for thermal power plants is necessity. Main objective of economic load dispatch is keep the total fuel cost as minimum level while meet the total demand power. Basically defined cost function of ELD as a quadratic function as follows:

$$F_i = a_i + b_i \times P_i + c_i \times P_i^2 \quad (1)$$

2.1. Valve Point Loading Effect

Due to opening stream valves at the power systems losses are increased. This effect is called valve point loading effect. Due to system losses are increased wit valve point loading effect, total cost value is increased. Above equation has sinusoidal terms due to valve point loading effect. This situation can be seen in Figure 1. This transformed equation is expressed as follows:

$$F_i = a_i + b_i \times P_i + c_i \times P_i^2 + |e_i \times \sin(f_i \times (P_i^{min} - P_i)) \quad (2)$$

F_i represents resulting fuel cost value, P_i represent power of thermal generator, a_i, b_i, c_i are cost coefficients and e_i, f_i are valve point loading coefficients of thermal generator unit i.

Total cost value of system is obtained by summed cost values of every thermal generation units.

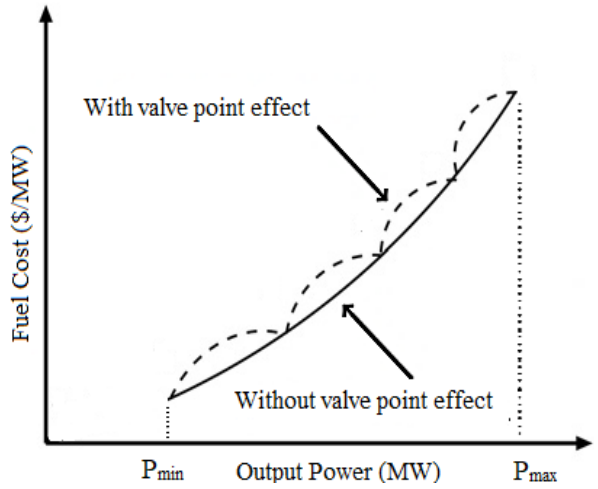


Figure 1 With and without valve point effect

2.2. Generators Limits

Thermal generators units must operate maximum and minimum power range. This power range can be different for different units:

$$P_{i,max} \geq P_i \geq P_{i,min} \quad (3)$$

Pi,min and Pi,max are represent minimum generator limit and maximum generator limit of unit i.

2.3. Power Balance

Total generated power at the thermal power plants meet the demand power by consumers. For this reason transmission line losses must be considered. The total generated power obtained sum of total demand power and total transmission line losses.

Transmission line losses is calculated as follows:

$$P_{loss} = \sum_{i=1}^{N} \sum_{j=1}^{N} P_i B_{ij} P_j + \sum_{i=1}^{N} B_{0i} P_i + B_{00} \quad (4)$$

Total generated power is calculated as follows:

$$\sum_{i=1}^{N} P_i = P_d + P_{loss} \quad (5)$$

N represent total thermal power plant. Pd and Ploss represent total demand power by consumer and total transmission line losses respectively. Bij, B0i and B00 are transmission loss coefficients.

2.4. Ramp Rate Limits

While a thermal generator unit is operating at a certain point, the operating point can only be increased to a certain level determined by the up ramp rate limit or decreased to a certain level determined by the down ramp rate limit.

This situation is shown as follows:

$$\max(P_i^{min}, P_i^0 - DR_i) \leq Pi \leq \min(P_i^{max}, P_i^0 + UR_i) \quad (6)$$

URi represents the up-ramp rate limit, DRi represent down ramp rate limit, is previous generated power and Pi present generated power of unit i.

2.5. Prohibited Zone

In some cases, thermal generation units do not worked and do not want to be worked some reasons such as mechanical corruption some particular power range (prohibited zone). These conditions can be expressed as follows:

$$P_i^{min} \leq Pi \leq P_{i,1}^{lowerbound} \quad (7)$$

$$P_{i,j-1}^{upperbound} \leq P_i \leq P_j^{lowerbound} \quad (8)$$

$$P_{i,ni}^{upperbound} \leq Pi \leq P_i^{max} \quad (9)$$

j represents number of prohibited zones of unit i and j=2, 3, 4 …ni. represents lower limit and represents upper limit of jth prohibited zone. Prohibited zone effect is shown in Figure 2.

3. Vortex Search Algorithm

Vortex search algorithm is a new optimization technique and inspired by stirring liquid materials.[17,18]. VSA is very influential and handy technique for solving economic load dispatch problem. problem. As a result of using an extensible step size modification arrangement a good balancing explorative and exploitative behavior of the search are obtained [18]. Vortex patterns can be represented in two dimensional space as a lot of nested circuits. The biggest and outer circuit is starting circuit of search space. The center of this circuit is calculated as follows:

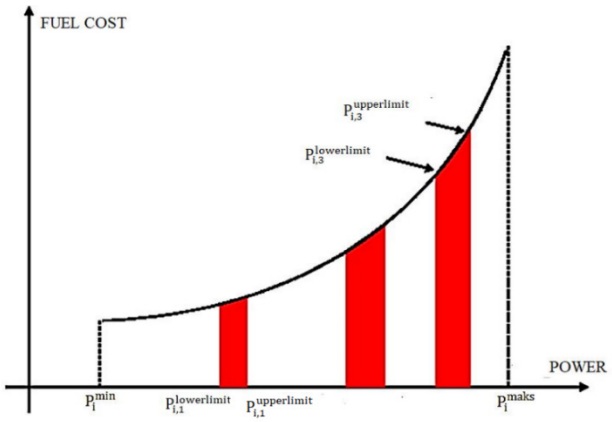


Figure 2 Shown of prohibited zones [16]

Upperlimit and lowerlimit are maximum and minimum constraints of the problem. A lot of candidate solutions are constituted into the outer circuit. The initial standard deviation is accepted as radius of this outer circuit and calculated as follows:

$$\mu_0 = (\text{upperlimit} + \text{lowerlimit}) / 2 \quad (10)$$

$$\sigma_0 = (\max(\text{upperlimit}) - \min(\text{lowerlimit})) / 2 \quad (11)$$

Generated candidate solutions must be controlled for ensure these solutions within the maximum and minimum limits of the problem. If these candidate solutions are not into the search space, they must be shifted into the search space. Some candidate solutions, which are not into the search space, shifted into the search space using equation as follows:

$$s_i^k = \text{lowerlimit}^i + (\text{upperlimit}^i - \text{lowerlimit}^i) \times \text{rand} \quad (12)$$

After this step a candidate solution, which is the best solution into the search space, is selected. This best solution is represented s'. This solution is memorized and current center μ0 is shifted to s'. New candidate solutions are constituted around this new center. All of candidate solutions are compared with previous best solution. If there is a better solution than s', it is selected as a new best solution [19]. This situation continues each iteration step. Radius of the circuits must be decreased every iteration. For this reason inverse gamma function (gammaincinv) is used and new radius is calculated every iteration as follows:

$$rt = \sigma 0 \times (1 / x) \times \text{gammaincinv}(x, a_t) \quad (13)$$

a_t is changeable parameter and it dependent a_0, t and MaxItr. This function shown as follows:

$$a_t = a_0 - (t / \text{MaxItr}) \quad (14)$$

Where t represents iteration number, MaxItr represents maximum iteration number. Due to cover the all search space a_0 is chosen 1. Outline flowchart of VSA method is shown in Figure 3:

4. Test System and Results

Vortex search algorithm was proposed and used for solving economic load dispatch problem. Six generation unit power system was selected. VSA algorithm was applied to this system considering various constraints. Each cases the number of iteration is limited to 1000 for obtained good solutions.

Case1 constraints are ramp rate limits and transmission losses;

Case2 constraints are ramp rate limits, prohibited zones and transmission losses;

Case3 constraints are transmission losses, valve point loading effect and ramp rate limits;

For every case transmission line losses coefficients are same and given as follows:

$B=10^{-4} \times$

0.17	0.12	0.07	-0.01	-0.05	-0.02
0.12	0.14	0.09	0.01	-0.06	-0.01
0.07	0.09	0.31	0.01	-0.10	-0.06
-0.01	0.01	0.00	0.24	-0.06	-0.08
-0.05	-0.06	-0.10	-0.06	1.29	-0.02
-0.02	-0.01	-0.06	-0.08	-0.02	1.5

$B_0=10^{-4} \times$

-3.9080	-1.2970	7.0470	0.5910	2.1610	-6.6350

$B_{00}=0.056$

4.1. Case 1: ELD with Ramp Rate

For this case two different constraints, which are ramp rate limits and transmission line losses, applied to system. Cost coefficients, ramp rate limits data and maximum – minimum limits of generators are shown in Table 1. Results of VSA was given Table 2 with different optimization techniques. Convergence behavior of VSA for this case was shown Figure 4. For this case selected power is 1263MW.

Table 1 Data of Case 1

Unit	a	b	c	P^0	UR	DR	Pmin	Pmax
1	240	7	0.0070	440	80	120	100	500
2	200	10	0.0095	170	50	90	50	200
3	220	8.5	0.0090	200	65	100	80	300
4	200	11	0.0090	150	50	90	50	150
5	220	10.5	0.0080	190	50	90	50	200
6	190	12	0.0075	150	50	90	50	120

Table 2 Results for Case 1

P(MW)	VSA	PSO	CPSO	WIPSO
P1	457.0630	493.24	471.66	454.39
P2	172.3751	114.63	140.03	164.279
P3	264.3900	263.41	240.06	264.223
P4	141.4373	139.71	149.97	123.21
P5	164.0545	179.65	173.78	167.22
P6	76.1690	84.83	99.97	120.00
P_{loss}	12.4889	12.22	12.38	12.24
Cost($)	15448	15489	15481.87	15453.13

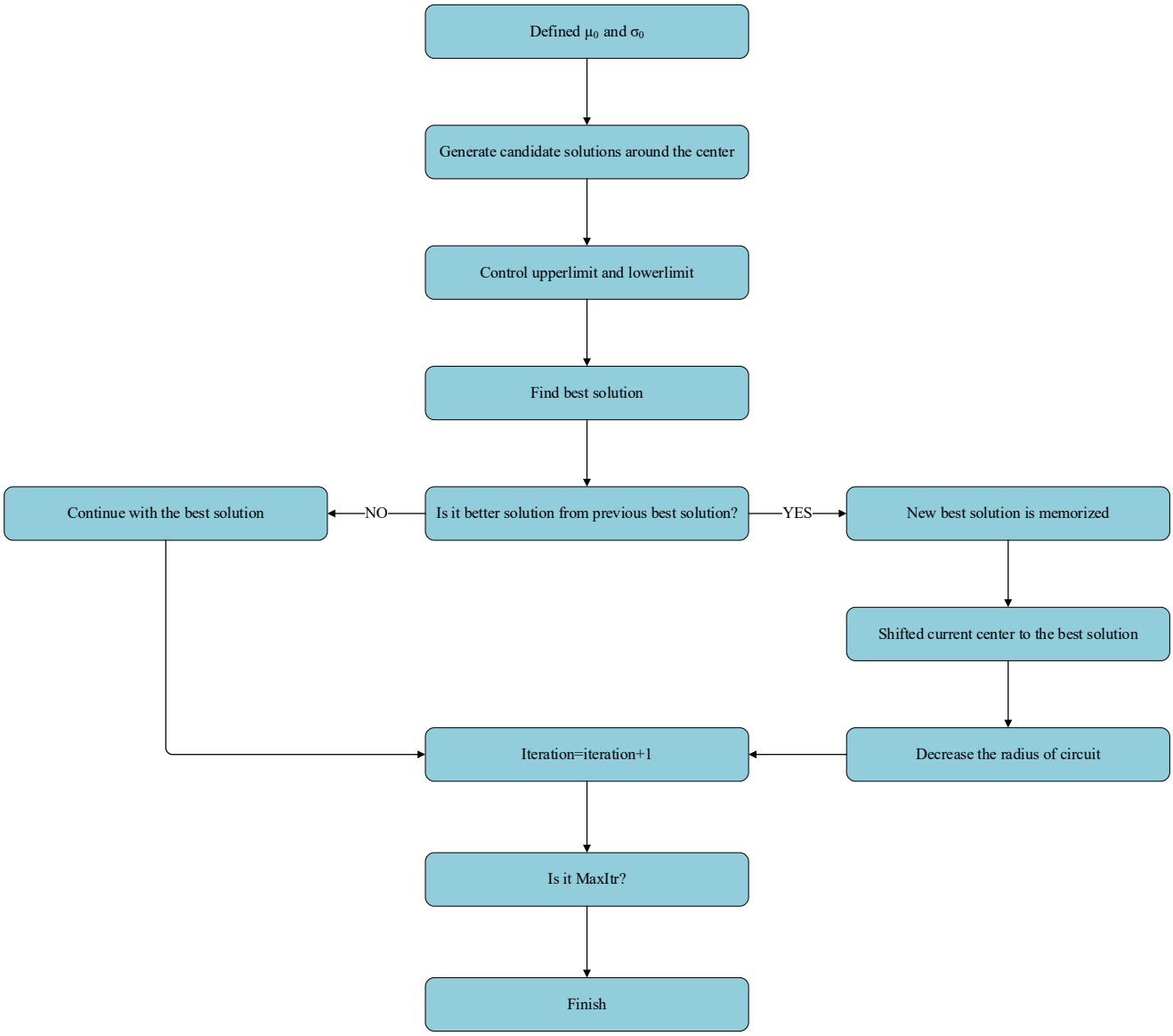


Figure 3 Outline flowchart of VSA

Results of vortex search algorithm compared with PSO, CPSO and WIPSO techniques from [20]. It is clear that from the Table 2 proposed VSA method has capable of the finding best solutions and minimum cost value. Figure 4 is shown convergence behavior of VSA for this case.

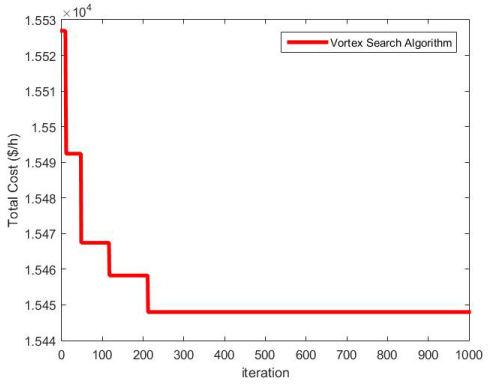


Figure 4 Convergence behavior of VSA for Case1

For this case three different constraints, which are ramp rate limits, prohibited zones and transmission line losses, applied to system. Cost coefficients was given in Table 3. Prohibited zone ranges and maximum – minimum limits of generators were given in Table 4. Ramp rate limits was given in Table 5. Convergence behavior of VSA for this case was shown Figure 5. 1263 MW power was selected for demand power. Results of VSA and other techniques results are shown in Table 6.

Table 3 Cost Coefficients

Unit	a	b	c
1	240	7	0.0070
2	200	10	0.0095
3	220	8.5	0.0090
4	200	11	0.0090
5	220	10.5	0.0080
6	190	12	0.0075

Table 4 Limits and Prohibited Zones of Generators

Unit	Pmax	Pmin	Prohibited Zone
1	500	100	210-240, 350-380
2	200	50	90-110, 140-160
3	300	80	150-170, 210-240
4	150	50	80-90, 110-120
5	200	50	90-110, 140-150
6	120	50	75-85, 100-105

Table 5 Ramp Rate Limits

Unit	P^0	UR	DR
1	440	80	120
2	170	50	90
3	200	65	100
4	150	50	90
5	190	50	90
6	150	50	90

Table 6 Results for Case 2

P(MW)	VSA	MFO	PSO	GA
P1	446.03	426.08	447.49	474.80
P2	181.09	199.80	173.32	178.63
P3	263.45	247.49	263.47	262.20
P4	133.96	136.94	139.05	134.28
P5	176.65	166.24	165.47	151.90
P6	74.53	98.93	87.12	74.18
P_{loss}	12.73	12.51	12.95	13.02
Cost($)	15447	15448.7	15450	15459

According to given results from Table 6, proposed VSA method has better cost value when compared with the MFO, PSO, GA techniques from [21].

Figure 5 is shown convergence behavior of VSA for this case.

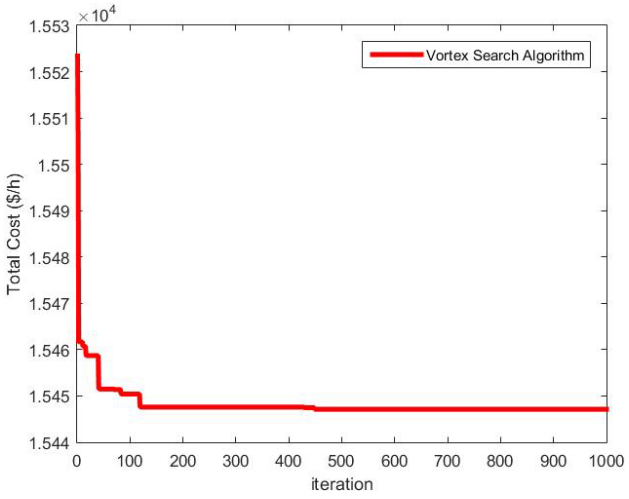


Figure 5 Convergence behavior of VSA for Case 2

4.3. Case 3: ELD with Valve Point Loading Effect

For this case three different constraints, which are transmission line losses, valve point loading effect and ramp rate limits, applied to system. Cost coefficients and valve point loading effect coefficients are given in Table 7. Obtained results were given in Table8 and compared with different methods. Convergence behavior of VSA for this case was shown Figure 6. 1263 MW power was selected for demand power.

Results of VSA and other techniques results was shown in Table 8. Maximum and minimum limits of generators are same with Table 4 and ramp rate limits are same with Table 5. Demand power is selected 1263 MW for this case.

Table 7 Valve point effect and cost coefficients

Unit	a	b	c	e	f
1	240	7	0.0070	300	0.031
2	200	10	0.0095	150	0.063
3	220	8.5	0.0090	200	0.042
4	200	11	0.0090	100	0.08
5	220	10.5	0.0080	150	0.063
6	190	12	0.0075	100	0.084

Table 8 Results for Case 3

P(MW)	VSA	PSO	CPSO	WIPSO	MRPSO
P1	495.29	443.03	467.55	437.82	442.07
P2	195.90	169.03	163.05	173.28	167.23
P3	235.79	262.02	253.41	271.97	267.09
P4	65.62	134.78	115.07	138.70	132.81
P5	197.82	147.47	169.45	146.98	155.02
P6	87.27	125.35	113.24	103.63	107.02
P_{loss}	12.92	18.68	18.70	18.08	18.03
Cost($)	15746	16372.9	16329.2	16327	16310.76

It is obviously seen that from the Table 8 VSA, method gave best cost value when compared with from PSO, CPSO, WIPSO and MRPSO techniques from [22].

Figure 6 is shown convergence behavior of VSA for this case

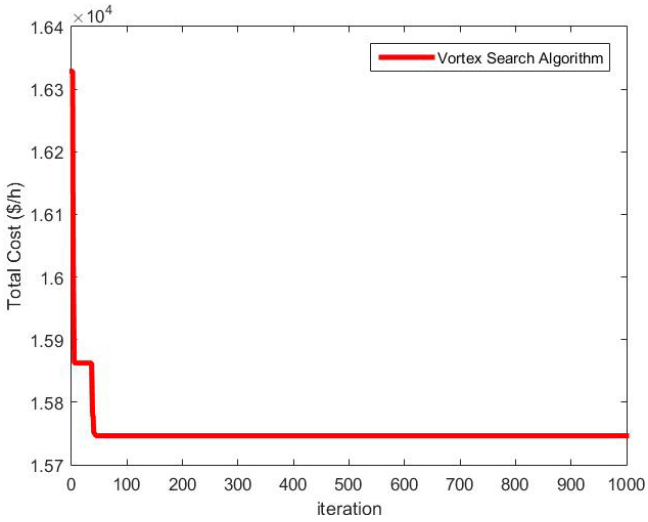


Figure 6 Convergence behavior of VSA for Case 3

5. Conclusion

In this paper vortex search algorithm was applied to six generator power system for solving economic load dispatch problem. ELD problem was became more complex and more difficult problem with valve point loading effect, ramp rate limits, transmission line losses and prohibited zones constraints. Three different situation were analyzed and for first case minimum cost value was found 15448 $, for second case minimum cost value was found 15447 $, for third case minimum cost value was found 15746 $. Obtained results from VSA compared with another techniques from the literature. These results clearly show that VSA is very capable,

feasible and effective method for solving non smooth and very complex economic load dispatch problem.

References

[1] M. Saka, I. EKE, S. S. Tezcan, M. C. Taplamacioglu., "Economic Load Dispatch Using Vortex Search Algorithm" in 4th International Conference on Electrical and Electronics Engineering, Ankara Turkey, 2017.

[2] T. Sen. H. D. Mathur, "A New Approach to Solve Economic Dispatch Problem Using A Hybrid ACO–ABC–HS Optimization Algorithm", Electrical Power and Energy Systems, 78, 735-744, 2016.

[3] L. S. Coelho, V. C. Mariani, "Combining of Chaotic Differential Evolution and Quadratic Programming for Economic Dispatch Optimization with Valve-Point Effect". IEEE Trans. Power Syst., 21, 989-996, 2006.

[4] A. Farag, S. Al-Baiyat, T. C. Cheng, "Economic Load Dispatch Multi Objective Optimization Procedures Using Linear Programming Techniques" IEEE Trans. Power Syst., 10, 731-738, 1995.

[5] L. Papageorgiou, E. Fraga, "A Mixed Integer Quadratic Programming Formulation for the Economic Dispatch of Generators with Prohibited Operating Zones" Electr. Power Syst. Res., 77, 1292-1296, 2007.

[6] C. L. Chen, S. C. Wang, "Branch-and-Bound Scheduling for Thermal Generating Units", IEEE Transactions on Energy Conversion, 8(2), 184–189, 1993.

[7] P. Bhui, N. Senroy, "A Unified Method for Economic Dispatch with Valve Point Effects" in IEEE Int. Con. Power Systems, 1-5, 2016.

[8] R. Balamurugan, S. Subramanian, "Differential Evolution-Based Dynamic Economic Dispatch of Generating Units with Valve-Point Effects," Electric Power Components and Systems, 36, 828-843, 2008.

[9] P. Sriyanyong, "An Enhanced Particle Swarm Optimization for Dynamic Economic Dispatch Problem Considering Valve-Point Loading," in Proc. the Forth IASTED International Conference on Power and Energy Systems (AsiaPES 2008), 167-172, 2008.

[10] C. L. Chiang, "Genetic-Based Algorithm for Power Economic Load Dispatch" IET Gener. Trans. Distrib., 1, 261-269, 2007.

[11] S. Hemamalini, S. P. Simon, "Artificial Bee Colony Algorithm for Economic Load Dispatch Problem with Non-Smooth Cost Functions" Electr. Power Compo. Syst., 38, 786-803, 2010.

[12] Y. C. Liang, J. R. C. Juarez, "A Normalization Method for Solving the Combined Economic and Emission Dispatch Problem with Meta-Heuristic Algorithms", Int. J. Electr. Power Energy Syst., 54, 163-186, 2014.

[13] R. Vijay, "Intelligent Bacterial Foraging Optimization Technique to Economic Load Dispatch Problem", Int. J. Soft Comput. Eng., 2, 55-59, 2012.

[14] X. S. Yang, S. S. S. Hosseini, A. H. Gandomi, "Firefly Algorithm for Solving Non-Convex Economic Dispatch Problems with Valve Loading Effect", Appl. Soft Comput., 12, 1180-1186, 2012.

[15] S. Pothiya, I. Ngamroo, W. Kongprawechnon, "Ant Colony Optimisation for Economic Dispatch Problem with Non-Smooth Cost Functions", Int. J. Electr. Power Energy Syst., 32, 478-487, 2010.

[16] M. Saka, "Economic Load and Emission Dispatch Analysis at Power Systems", M. Sc. Thesis, Gazi University, Ankara, 2017.

[17] B. Dogan, T. Olmez, "A New Metaheuristic for Numerical Function Optimization: Vortex Search Algorithm", Information Sciences, 293, 125-145, 2015.

[18] B. Dogan B., T. Olmez, "Vortex Search Algorithm for the Analog Active Filter Component Selection Problem", International Journal of Electronics and Comminications, 69, 1243-1253, 2015.

[19] M. Saka, I. Eke, S. S. Tezcan, M. C. Taplamacioglu, "Application of Vortex Search Algorithm to Economic Load Dispatch Including Transmission Losses with Valve-Point Loading Effect" in 1st International Mediterranean Science and Engineering Congress, Adana Turkey, 2016.

[20] N. Singh, Y. Kumar, "Economic Load Dispatch Using MRPSO with Generator Ramp Rate Limits Constraint" in Int. Con. Comp. Int. and Com. Networks, 632-637, 2012.

[21] N. Trivedi, A. H. Ranpariya, A. Kumar, P. Jangir, "Economic Load Dispatch Problem with Ramp Rate Limits and Prohibited Operating Zones Solve using Levy Flight Moth-Flame Optimizer" in Int. Conf. Energy Efficient Technologies for Sustainability, 442-447, 2016.

[22] N. Singh, Y. Kumar, "Economic Load Dispatch with Valve Point Loading Effect and Generator Ramp Rate Limits Constraint Using MRPSO", International Journal of Advanced Research in Computer Engineering & Technology, 2, 1472-1477, 2013.

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 155-159 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Influence of supply frequency on dissipation factor measurement and stator insulation diagnosis

Cyrille Caironi*, Bernhard Fruth[1], Detlef Hummes[2], Rudolf Blank[3]

[1]*PDSS, Zurich, SWIZERLAND*

[2]*American University of Kuwait, Safat, KUWAIT*

[3]*B2Electronic, Klaus, AAUSTRIA,*

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 08 January, 2018
Online: 30 January, 2018

Keywords:
Dielectric
Tangent Delta
Frequency measurement

ABSTRACT

This paper is an extension of work originally presented in EIC-2017. It deals with influence of the supply frequency for dissipation factor measurements, mainly for tests under power frequency and low frequency. After a theoretical reminder, we present some experiments on single coils and stators of high voltage motors. Finally, we discuss the results and the desirability of choosing one or the other of these methods.

1. Introduction

Dissipation factor, or tan δ, is currently used for the insulation diagnostics. While the tan δ measurements are mainly performed at power frequency, some methods based on the frequency steps [1] and very low frequency have been developed. The methods using very low frequency have some advantages in term of portability, power requirement, etc. but it is interesting to see the impact of the frequency on the quality of measurement and, subsequently, of the diagnostic.

This paper presents the influence of the supply frequency on tangent delta measurement. The paragraph II reminds the theoretical bases. The paragraph III presents some experiments, firstly on single coil and secondly on a stator of high voltage motor. Test are performed at 0.1 Hz and 50 Hz. The analysis of the result presents some difference, especially on the sensitivity at relaxation phenomena. These differences are discussed, and we can define some interests on insulation diagnostic.

2. Theory reminder

Dielectric losses are currently measured and used for electrical machines diagnostics. To perform the dielectric loss measurement, the stator is considered as a capacitor and the theory supporting these measurements is based on the complex permittivity. Of course, this capacitor is imperfect.

It is schematized by the capacitance with a resistance in parallel, both in series with a resistance and an inductance, in our case, we can neglect Ls (fig. 1).

The real part of impedance is due to the series and parallel resistances Rs and Rp. Rs becomes important at high frequency when Rp has a dominating effect essentially at low frequency.

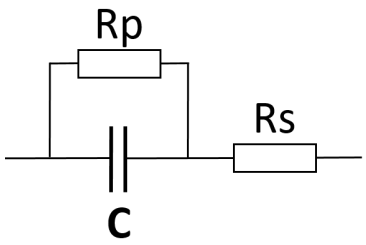


Fig. 1: Real capacitor model *(without Ls)*

*Cyrille CAIRONI, CAIRCY, 22, rue de Ceintrey 54160 Pulligny France Email : cyrille.caironi@caircy.org

Rp is the picture of the losses due to the dielectric polarization, the dielectric resistance, leakage current and also charge redistribution phenomena. All these phenomena are involved at low frequency especially at industrial frequencies and lower. De facto we currently use the model within resistance in parallel which excludes Rs (1).

$$\underline{I} = j\omega\underline{\varepsilon}C\underline{U} = \underline{I_{Rp}} + j\underline{I_C} \quad (1)$$

Due to the phenomena of relaxation and losses, the insulating materials permittivity is complex (2)

$$\underline{\varepsilon} = \varepsilon_0 \left(\varepsilon_r' - j\,\varepsilon_r''\right) \quad (2)$$

In this arrangement, ε_r' is the component which defines the capacitive current (3) and $\varepsilon_r''$is the component which defines the current I_{Rp} corresponding at the active losses in phase with the voltage (4).

$$\varepsilon_r' = \frac{\underline{I_C}}{j\omega C\underline{U}} \quad (3)$$

$$\varepsilon_r'' = \frac{\underline{I_{Rp}}}{\omega C\underline{U}} \quad (4)$$

The tangent delta, or dissipation factor is defined as (5):

$$tan\delta = \frac{\underline{I_{Rp}}}{\underline{I_C}} = \frac{\varepsilon_r''}{\varepsilon_r'} = \frac{1}{\omega CRp} \quad (5)$$

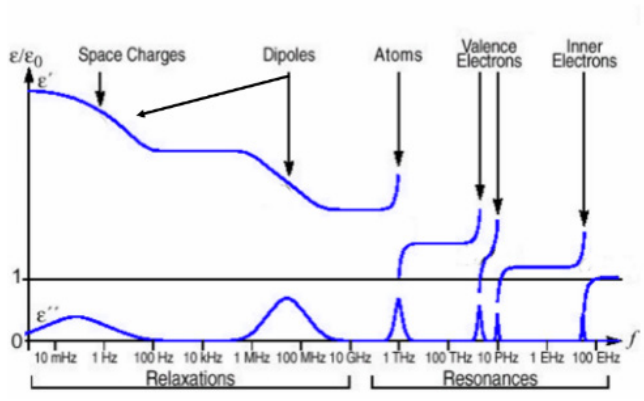


Fig. 2: effect of resonance and relaxation on permittivity *(source [2])*

Of course, we can see that the tangent delta is inversely proportional to the frequency. But the main interest of the low frequency is the physics of the material. The polarization and relaxation phenomena involved in dissipation of the insulation have different origins and each of them approximately can be linked to frequency (fig. 2) as presented in [2].

The electronic and ionic polarization have a high frequency level (around 10^{15} Hz for electronic polarization) this type of polarization is not interesting for the machine diagnostic in first approach. On the other hand, the relaxation phenomena linked to dipoles, space charges and interface charges are very interesting.

These relaxation phenomena enable us to check precisely details like pollution and polymerization, which are visible at lower frequencies (less than 1 Hz) by tan δ analysis.

Furthermore, the temperature influence is a particularly important factor as presented on figure 3 and figure 4 [3] and explained in [4].

It could be an advantage but also a risk because a variation on the temperature has a bigger impact at low frequency (less than 1 Hz) than at industrial frequency.

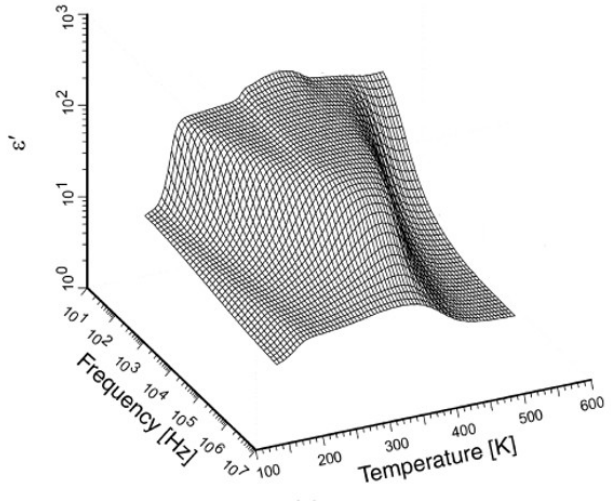


Fig. 3: Typical real part of the permittivity of porous glass versus frequency and temperature *(source [3])*

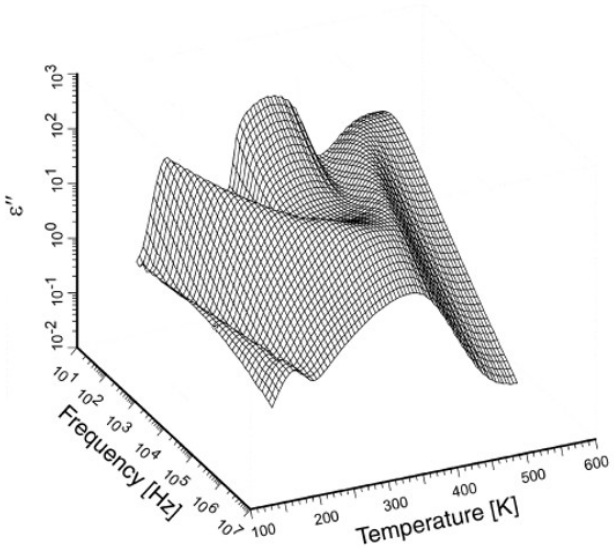


Fig. 4: Typical imaginary part of the permittivity of porous glass versus frequency and temperature *(source [3])*

3. Experimental studies

IEEE 433 recommends performing very low frequency tests at 0.1 Hz [5] and the monitoring present on the market uses this frequency. Therefore, the tests are performed at frequency of

0.1 Hz and 50 Hz, in accordance to IEEE 433 [5] and to IEC 60034-27-3 [6].

3.1. Experiments on coils

Tan δ measurements were performed on sample coils (7 kV rated) in steps of 1 kV from 1 kV to 9 kV. Measurements were performed successively at 0.1 Hz and 50 Hz.

These tests show higher tan δ values with a 0.1 Hz frequency supply than with a 50 Hz supply (fig. 5). Therefore, these measurements confirm some explanations of the theoretical reminder and they are also in accordance to the related literature and different diagnostic methods proposed in [1].

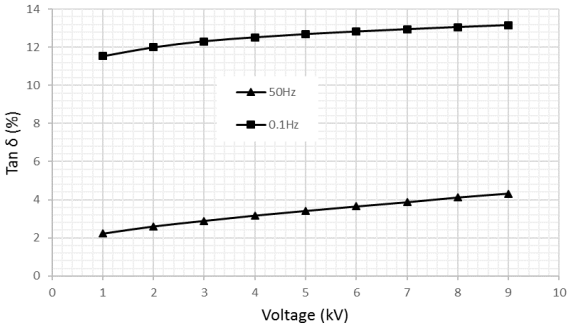


Fig. 5: tan δ versus voltage on single Coil

The difference of levels can be explained by a dielectric relaxation, which means a lowering of the dielectric losses with frequency, therefore 0.1 Hz is more sensitive than 50 Hz.

The underlying physical mechanism is usually related to the polymerization degree of the impregnating resin and interfacial polarization [4].

The voltage dependence of the losses of the sample is very similar for both frequencies. At this step, we could consider that the frequency just influences the level of tan δ. In fact, the influence is also on the behavior of the tan δ versus voltage.

3.2. Experiments on Stator

Measurements were performed on a 1.5 MW, 11 kV motor in accordance to IEC 60034-27-3 [6] and customer requirements (fig. 06).

Capacitance measurements at 50 Hz and 0.1 Hz have a similar behavior (fig. 08), on the other hand, the tan δ has a different shape at 50 Hz and 0.1 Hz (fig. 07)

As for a simple coil, the tan δ is higher at 0.1 Hz than at 50 Hz but of different shape (fig. 07). The measurement at 0.1 Hz shows a voltage dependence which is not present at 50 Hz.

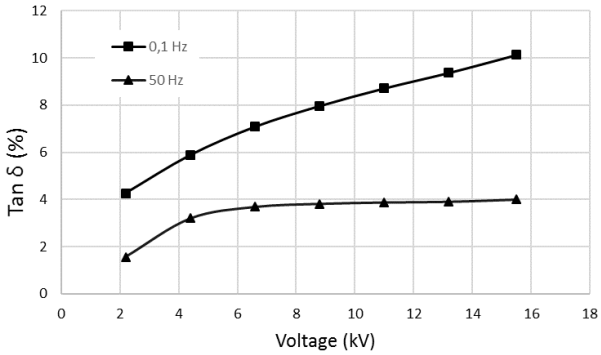
This nonlinear voltage dependences of rotating machine insulation properties typically stem from partial discharges and currents in endwinding corona protection.

Evidence of the structural influence on the dielectric properties is shown in [4] as mechanical losses and dielectric losses largely coincide. This indicates that the 0.1 HZ test method can provide information about insulation quality, e.g. brittleness, shrinkage and risk of resin cracking caused by post-polymerization

Partial discharges analysis, and especially PRPD, can complete the tan δ information. Of course, the tan δ and PRPD must be performed at the same frequency.



Fig. 6: VLF supply and measurement instrumentation on 11 kV Stator



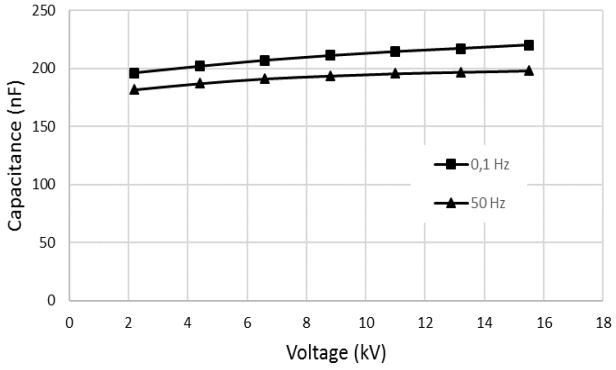
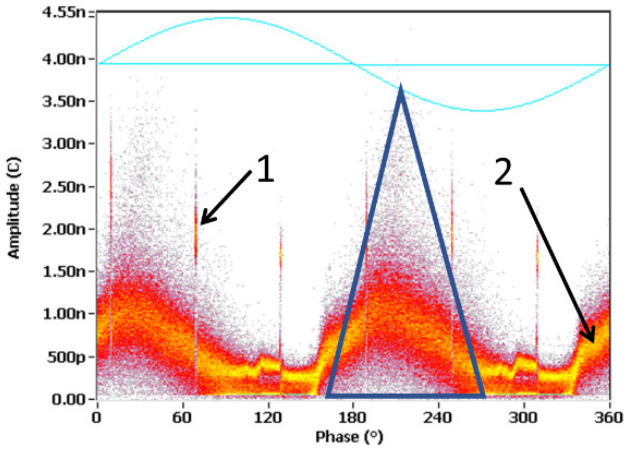


Fig. 8: capacitance versus voltage on 11 kV Stator

The patterns show solely the ionization phenomena and must be analyzed in accordance with the tan δ measurements. Furthermore, the noise phenomena have shown more important

using power frequency supplies. Phenomena linked to the noise can be solved by numerical tools presented in [7] and [8] especially in case of on-line PD monitoring.

Fig. 09 shows PRPD pattern at the network power supply and Fig. 10 shows PRPD pattern at 0.1 Hz. Despite differences, both patterns give similar information, but the 0.1 Hz can be easily compared to the tan δ.



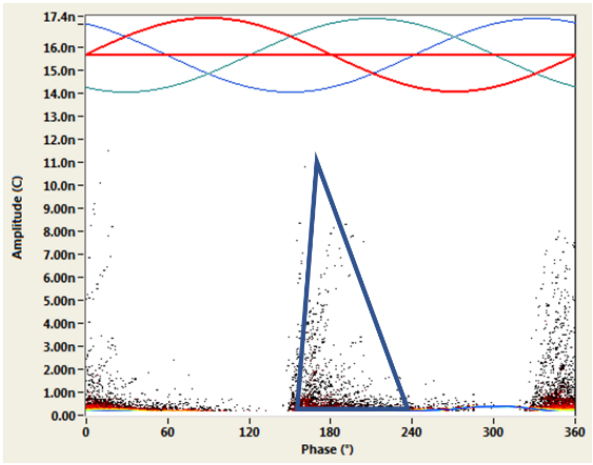


Fig. 10: PRPD at 0.1 Hz on 11 kV Stator

4. Extension to the Stator Insulation Diagnostic

The paragraph II confirms the theoretical reminder presented in paragraph I. Especially on complete stators, tan δ measurement allow to see some more phenomena at 0.1 Hz, being undetectable at 50 Hz.

In fact, the measurements at 0.1 Hz should be most efficient to the diagnostic phenomena linked to the complete insulation system [9]. Furthermore, the low frequency usage is not new. For example, measurements for CIGRE [10] was performed at 1Hz in 1990 in order to:

- do not have distortion of sinus wave (which is not perfect when it is provided by transformer at supply frequency)
- reduce the pulse pile-up
- have a cleaner PD pattern
- the higher sensitivity to the structural properties of the composite material allows to more sensitively follow the resin post-polymerization
- etc.

However, the low frequency usage was costly in the beginning of the 90th, which is not the case anymore and this off-line monitoring procedure can be easily implemented and could be interesting to complement on-line monitoring which is de facto happening at power frequency.

5. Conclusion

tan δ measurement. It reminds the theoretical bases before present two different cases: a coil and a stator. Measurement were performed on two frequencies: 50 Hz (power frequency) and 0.1 Hz.

If similar diagnostic results together with partial discharge measurements, different defects can be detected or identified at lower power frequency.

The low frequency measurements present some advantages. Especially the dielectric losses react more sensitive to structural changes.

However, while diagnostics can be improved using low frequencies, criteria must still be developed if the low frequency method should be used for acceptance testing. Indeed, most of the tan δ criteria are established for power frequency tests and we have to perform further research in order to define reliable acceptance criteria.

On the practical side, the instruments used by the authors have proven easily transportable, insensitive to noise and have very low power requirements in comparison to power frequency devices. So, that we feel, not only the enhanced diagnostic potential but also cost consideration, which justify more research efforts in that field.

References

[1] Christof Sumereder, Michael Muhr, Michael Marketz, Christian Rupp, Michael Kruger "Unconventionnal Diagnostic Methods for Testing Generator Stator Windings" IEEE Electrical Insulation Magazine , Vol. 25 No. 5, sept. 2009.

[2] A.K. Jonscher "Dielectric relaxation in solids" Chelsea Dielectric Press, London, 1983.

[3] Y. Feldmann, Y.A. Gusev, M.A. Vasilyeva 'Dielectric phenomena in complex systems" Tutorial Kazan Federal University, 2012.

[4] B. Fruth, G. Liptak, "Dielectric Properties of Mica Epoxy Composites Subjected to Thermal and Thermoelectrical Aging", Proc. 6th Int. Symp. High Voltage Engineering, New Orleans, 1989, paper 21.02

[5] IEEE 433-2009 : "IEEE Recommended Practice for Insulation Testing of AC Electric Machinery with High Voltage at Very Low Frequency".

[6] IEC 60034-27-3 "Rotating electrical machines – Part 27-3: Dielectric dissipation factor measurement on stator winding insulation of rotating electrical machine" 2015.

[7] C. Caironi "Contribution to the on-line forecast maintenance of electric machines by the analysis of signals related to physical phenomena relating thereto." PhD thesis, Université de Lorraine, France, 2002.

[8] C. Caironi, L. Durantay, D. Brie, A. Rezzoug "Interest & utility of Time-frequency and Time scale Transforms in the Partial Discharges Analysis" in proc. ISEI'2002, Boston USA.

[9] Nathaniel Taylor "Dielectric response and partial discharge measurements on stator insulation at varied low frequency", PhD thesis Stokholm, Sweden, 2010.

[10] B. Fruth, J. Fuhr “Partial discharges recognition – A Tool for diagnosis and Monitoring Ageing” In proc CIGRE’1990 Paris, France.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 157-161 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Numerical evaluation of ABS parts fabricated by fused deposition modeling and vapor smoothing

Sung-Uk Zhang*

Electro Ceramic Center, Dept. of Automotive engineering, Dong-Eui University, Busan, 47340, South Korea

ARTICLE INFO

Article history:
Received: 26 October, 2017
Accepted: 01 December, 2017
Online: 10 December, 2017

Keywords:
Additive manufacturing
Fused deposition modeling
ABS
Vapor smoothing
Finite element analysis
Response surface analysis

ABSTRACT

The automotive industry has focused to use polymer materials in order to increase energy efficiency. So, the industry pays attention to use 3D printing technologies using several polymers. Among several 3D printer technologies, fused deposition modeling (FDM) is one of the popular 3D printing technologies due to an inexpensive extrusion machine and multi-material printing. FDM could use thermoplastics such as ABS, PLA, ULTEM so on. However, it has a problem related to the post-processing because FDM has relatively poor layer resolution. In this study, the mechanical properties of ABS parts fabricated by FDM were measured. The ABS parts were divided into one with vapor smoothing process and the other without the vapor smoothing process which is one of the post-processing methods. Using dynamic mechanical analysis (DMA) and dilatometer, temperature-dependent storage modulus and CTE for ABS specimens were measured. Based on the measured thermo-mechanical properties of ABS parts, finite element analysis was performed for an automotive bumper made of ABS. Moreover, response surface methodology was applied to study relationships among design parameters of thickness of the bumper, ambient temperature, and application of the vapor smoothing process. In result, a design guideline for a ABS product could be provided without time-consuming experiments

1. Introduction

The automotive industry has tried to increase energy efficiency for vehicles. In order to achieve the goal, one of ways is to reduce vehicle weight. If weight of the vehicle increases by 10%, fuel economy also increases by about 7% [1]. Therefore, researchers in the automotive industry have focused to use materials lighter than steel. One of the substitute materials is polymer. Currently, about 15% to 20% of total weight of a vehicle is in charge of polymer materials. The polymer materials are widely used for automobile parts including front ends, bumpers, fascia systems, grills, fenders, intake manifolds, engine mounts, door structures, panels, door impact beams, bonnet panels, under hood parts, pedal boxes, body structures, steering columns and so on [1]. Among several polymers, acrylonitrile butadiene styrene (ABS) is applied for bumpers, seats, dashboard, interior trim, exterior trim, and lighting and so on. This study focuses on ABS because fused deposition modeling (FDM), which is one of additive manufacturing techniques, could use ABS to make 3D printed structures.

Recently, additive manufacturing also known as 3D printing is considered as the key driver of the fourth industrial revolution. Compared to the conventional subtractive manufacturing such as turning, cutting, milling and grinding, 3D printing has unique advantages in design flexibility, personal fabrication, cost of geometry complexity, dimensional accuracy, time and cost efficiency in production run and so on [2]. Based on these merits, rapid development has been made in the several industrial areas such as aviation parts, manufacturing, and medical science [3]. However, the 3D printing has demerits of, scalability, trade-off between building and layer resolution, cost for mass manufacturing, material heterogeneity, need of anchor and support material, requirement for post-processing and so on [2]. So many researchers have solved the problems for the next generation of the 3D printing technique [2].

This study focuses on the post processing technique for the additive manufacturing. Additive manufacturing techniques have been classified into seven categories by ASTM international, which are shown in Figure.1.

Among them, FDM belongs to a category of material extrusion, which is an inexpensive and very popular 3D printing technique [2]. But FDM sometimes struggles with poor surface finish so that the post processing technique is necessary. The poor surface finish occurred by stair stepping effect could be treated by

*Corresponding Author: Sung-Uk Zhang, 82-51-890-1645, zsunguk@deu.ac.kr

polishing process such as sanding and vapor smoothing. The sanding may not be possible for tiny or complex geometries. On the other hand, the vapor smoothing could be effective for the complex shapes fabricated by 3D printing as shown in Figure. 2.

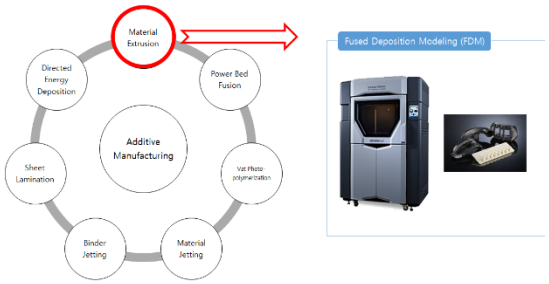


Figure.1. Seven categories for AM technologies

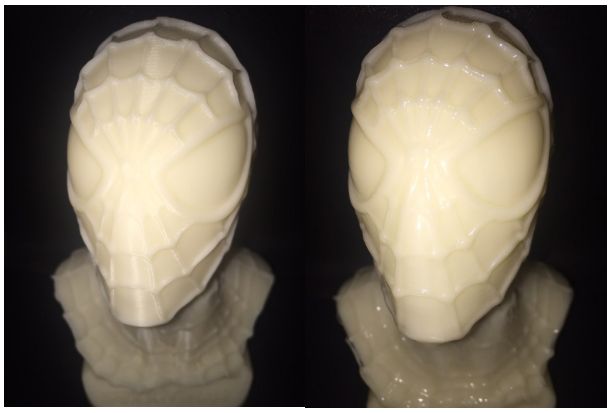


Figure.2. Vapor smoothing before and after

Many researchers have studied the vapor smoothing process and its effect in the mechanical properties. Kuo and Mao [4] developed an acetone-vapor smoothing system for ABS parts fabricated by FDM. They showed that surface roughness of the ABS parts could be significantly improved by using their system. Garg et al. [5] observed mechanical properties and surface roughness according to building parameters and vapor smoothing process for ABS specimens fabricated by FDM. In order to obtain the mechanical properties, they used ultimate testing machine (UTM). Singh et al. [6] studied vapor smoothing process for FDM parts. They had tried to improve the vapor smoothing process in term of surface roughness and dimensional accuracy by using DOE technique. Singh et al. [7] optimized building parameters of FDM and control parameters of vapor smoothing processes for hip joint as a biomedical application. They used Taguchi's method and response surface methodology to obtain the optimum level of process parameters. Cantrell et al. [8] measured mechanical properties of 3D printed ABS and PC parts with varying raster and building orientation of additive manufacturing. Gao et al. [9] investigates the impact of the vapor smoothing for 3D printed ABS parts in terms of the strength and elongation. They measured anisotropic mechanical properties of the ABS parts and showed that the material properties are affected by the length of process time for the vapor smoothing. Belter and Dollar [10] suggested a technique to strengthen 3D printed parts fabricated by FDM. They tried to remove voids in the printed parts by filling them with high-strength resins. They performed three-point bending test to obtain mechanical properties. Melenka et al. [11] evaluated mechanical properties and dimensional accuracy for 3D printed PLA parts fabricated by FDM. They used DOE technique to establish a relationship between building parameters and mechanical properties of the 3D printed parts.

In this study, temperature-dependent mechanical properties of 3D printed ABS parts fabricated by FDM and the acetone-vapor smoothing are measured by using dynamic mechanical analyzer (DMA) and dilatometer. Based on the measured data, finite element analysis and response surface methodology are performed in order to generate a design guideline for an automotive bumper made of ABS.

2. Methodology

The second section has four subsections. The first subsection includes geometry information of specimen for DMA and dilatometry. The second briefly explains about the vapor smoothing process. The third subsection describes about information measured by DMA and dilatometry. The last subsection introduces about finite element analysis and response surface methodology for an automotive bumper made of ABS.

2.1. Specimen fabrication for DMA and Dilatometry

Specimens for DMA and dilatometry were fabricated by using Fortus 3D printer which is based on Stratasys FDM technology. In modeler setup of Stratasys's software, building parameters were controlled, which are part interior style, visible surface style, and support style be set to sparse-low density, being enhanced, and being smart. The specimen for DMA has 27.8 mm length with 6.5 mm width and 2.0mm thickness and the size of specimen for dilatometry is 6 mm by 6 mm by 12 mm as shown in Figure. 3.

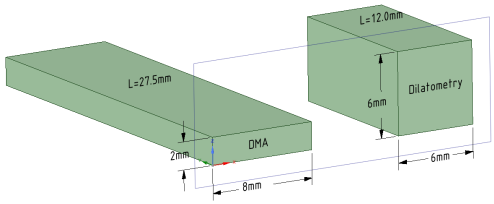


Figure.3. Geometry of the specimen for DMA and Dilatometer

2.2. Vapor smoothing process

Vapor smoothing process lets ABS plastic dissolve in acetone. So the secondary bonds between the ABS polymer chains is broken so that external layers of ABS 3D printed parts are softening. In result, this technique makes the surface finish of 3D printed parts be improved. A simple process for the vapor smoothing is shown in Figure. 4.

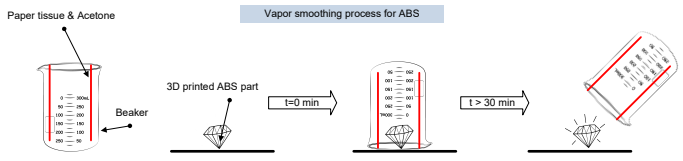


Figure.4. Vapor smoothing process

First of all, dust and debris from 3D printed ABS parts are removed. Small amount of acetone is poured on paper tissues. The tissues are attached inside of a beaker. Then, the 3D printed parts are covered with the beaker. Then, allow parts to be inside of the beaker for about 30 minutes. In this paper, amount of acetone was controlled from 5ml to 15ml in order to observe a relationship between the amount of acetone and mechanical properties of the 3D printed parts with the vapor smoothing.

2.3. Dynamic mechanical analyzer and Dilatometry

Dynamic Mechanical Analyzer (DMA), which known as DMA, is a powerful tool to characterize mechanical properties of a material as a function of temperature, time, frequency, stress, and so on. Q800 DMA was used, which is manufactured by TA instruments. Temperature-dependent storage modulus of ABS 3D printed parts could be observed by using DMA. As an applied mechanical load of DMA is a sinusoidal or oscillating force, complex modulus of E* could be obtained for polymer materials such as ABS, PLA, PC, and ULTEM, which consists of the storage modulus E' (real part) and the loss modulus E" (imaginary part) as follow

$$\mathbf{E}^* = E' + \mathbf{i}E'' \qquad (1)$$

$$\tan\delta = E''/E' \qquad (2)$$

where, tan δ is the ratio of loss modulus to storage modulus, which is an indicator of the energy lost. Figure.5 graphically shows temperature dependent storage modulus and tan δ measured by using DMA. In the graph of tan δ , the peak of tan δ indicates the glass transition temperature which means a midpoint between the glassy and rubbery states of a polymer.

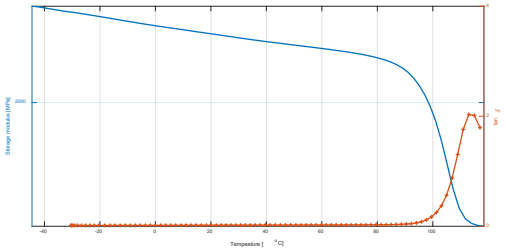


Figure.5.Temperature dependent storage modulus and tan δ obtained by using DMA

Moreover, a dilatometer is used to measure temperature dependent CTE of the polymer materials, which is an instrument to measure volume changes occurred by a physical or chemical event. In this study, the physical event is temperature change. The thermal expansion is expressed as

$$\alpha = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right) \qquad (3)$$

where, α is a coefficient of thermal expansion (CTE), V is a volume, and T is a temperature. We used DIL-402C manufactured by NETZSCH. Using the dilatometry, we could obtain temperature dependent CTEs for ABS 3D printed parts.

2.4. Finite element analysis and response surface methodology for a 3D printed bumper made of ABS

A portion of a bumper is modeled and structurally evaluated by using finite element analysis as shown in Figure. 6 and Figure. 7. Fig 6 shows the geometry of the automotive bumper. Fig 7 graphically shows fixed boundary conditions, load boundary conditions, and body temperature condition on the geometry. In Figure. 7, the symbols of A, B and C mean boundary conditions of fixed support, pressures, thermal conditions.

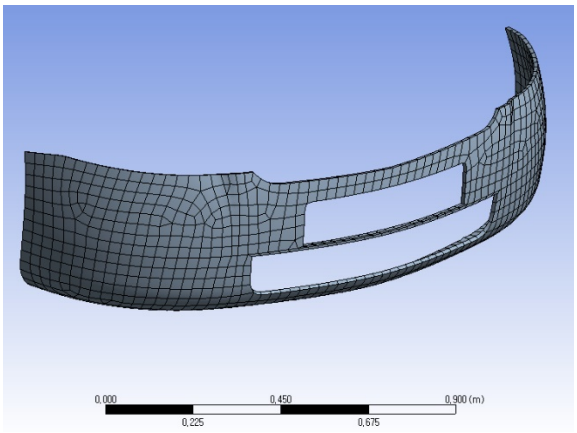


Figure.6. Finite element model for an automotive bumper

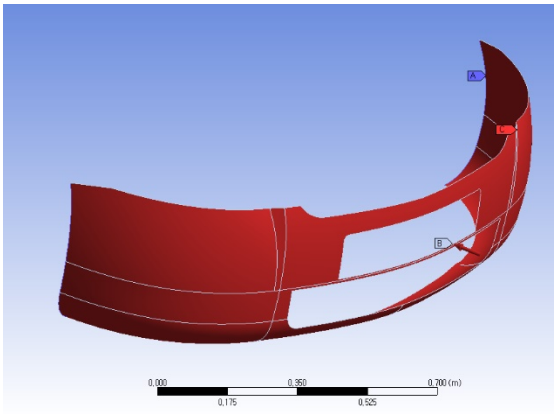


Figure.7. Boundary conditions for the finite element model

Input parameters for the response surface methodology were temperature, thickness of bumper, element size, and application of the vapor smoothing. The range of each parameter is shown in Table 1.

Table 1. The factor ranges for input parameters

Input parameter	[min, max]
Temperature [°C]	[-40, 70]
Thickness of Bumper [m]	[0.01,0.03]
Element Size [m]	[0.03,0.04]
Vapor smoothing	w/o, w/

Box-Behnken design was used for the response surface methodology. Design points were generated by using Minitab software as shown in Figure. 8. And the output parameter was a maximum of total deformation of the bumper.

3. Results and Discussion

The first describes thermo-mechanical properties of ABS specimens measured by DMA and dilatometry. The second subsection depicts numerical simulation for an automotive bumper made of ABS.

3.1. Thermo-mechanical properties of ABS specimens

Four different splits were measured, which were exposed to different amount of acetone vapor; 0ml, 5ml, 10ml and 15ml. Three specimens per each split were fabricated and measured in the same condition. Averaged thermo-mechanical properties of the

ABS specimens were obtained, which are the storage modulus and CTE as shown in Figure.9 and Figure.10.

	C1	C2	C3	C4	C5	C6	C7	C8-T
	StdOrder	RunOrder	PtType	Blocks	Temperature[C]	Thickness of Bumper [m]	Element size [m]	Vapor smoothing
1	1	1	2	1	-40	0.01	0.04	w/o
2	2	2	2	1	70	0.01	0.04	w/o
3	3	3	2	1	-40	0.03	0.04	w/o
4	4	4	2	1	70	0.03	0.04	w/o
5	5	5	2	1	-40	0.02	0.03	w/o
6	6	6	2	1	70	0.02	0.03	w/o
7	7	7	2	1	-40	0.02	0.05	w/o
8	8	8	2	1	70	0.02	0.05	w/o
9	9	9	2	1	15	0.01	0.03	w/o
10	10	10	2	1	15	0.03	0.03	w/o
11	11	11	2	1	15	0.01	0.05	w/o
12	12	12	2	1	15	0.03	0.05	w/o
13	13	13	0	1	15	0.02	0.04	w/o
14	14	14	0	1	15	0.02	0.04	w/o
15	15	15	0	1	15	0.02	0.04	w/o
16	16	16	2	1	-40	0.01	0.04	w/
17	17	17	2	1	70	0.01	0.04	w/
18	18	18	2	1	-40	0.03	0.04	w/
19	19	19	2	1	70	0.03	0.04	w/
20	20	20	2	1	-40	0.02	0.03	w/
21	21	21	2	1	70	0.02	0.03	w/
22	22	22	2	1	-40	0.02	0.05	w/
23	23	23	2	1	70	0.02	0.05	w/
24	24	24	2	1	15	0.01	0.03	w/
25	25	25	2	1	15	0.03	0.03	w/
26	26	26	2	1	15	0.01	0.05	w/
27	27	27	2	1	15	0.03	0.05	w/
28	28	28	0	1	15	0.02	0.04	w/
29	29	29	0	1	15	0.02	0.04	w/
30	30	30	0	1	15	0.02	0.04	w/

Figure.8. Design points generated by Minitab software

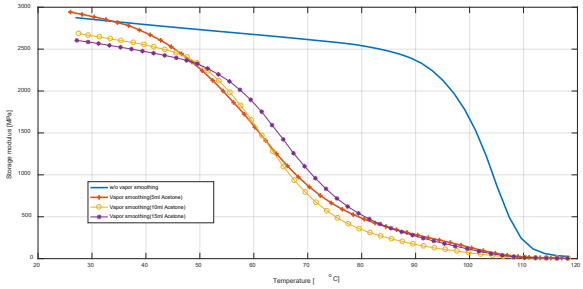


Figure.9. Average of temperature dependent storage moduli for four splits

Figure.9 graphically shows the difference between specimens without vapor smoothing process and ones with vapor smoothing process. The glass transition temperature of the specimens with the vapor smoothing process is much lower than one without the vapor smoothing process.

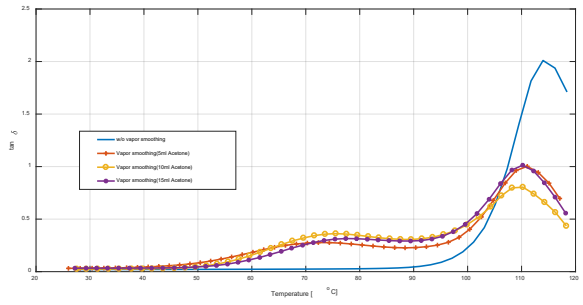


Figure.10. Average of tan δ for four splits

Figure.10 graphically shows the difference between specimens without vapor smoothing process and ones with vapor smoothing process in terms of tan δ. In the graph of first split, peak point of tan δ occurs at about 114ºC. However, the first peaks of the other specimens with the vapor smoothing are shown at 67ºC, 71ºC, and 74ºC. Figure.11 shows that CTEs vary as the temperature increases. CTE of the specimen without vapor smoothing is relatively constant.

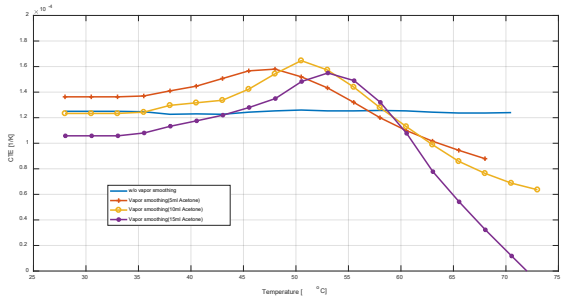


Figure.11. Average of the temperature dependent CTEs for four splits

However, CTEs of the others with vapor smoothing have large variations and have one peak between 45ºC and 60ºC. As the temperature increases more than 60ºC, CTEs of the samples with vapor smoothing decreases.

Figure.9, Figure.10 and Figure. 11 obviously shows which split does not have the vapor smoothing process. Figures represent that the vapor smoothing process weakens thermal stability of ABS 3D printed parts. However, above figures could not show any correlation between CTE and the amount of acetone vapor in the range of 5ml to 15ml.

3.2. Numerical evaluation for an automotive bumper made of ABS

The output parameter was the maximum value of total deformation of the automotive bumper as shown in Figure. 12. The maximum deformation occurs at the center of the bumper under the given boundary conditions.

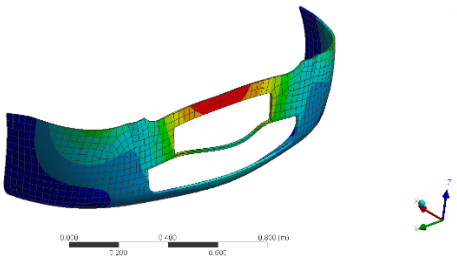


Figure.12. Total deformation of the automotive bumper

Using the response surface methodology and the finite element analysis, two regression equations with varying application of the vapor smoothing process were obtained as follow

$$D[m] = 0.1538 - 0.001126 \times T[C] - 5.97 \times Th[m] - 3.80 \times Elem[m] + 0.000006 \times T[C] \times T[C] + 108.4 \times Th[m] \times Th[m] + 51.7 \times Elem[m] \times Elem[m] - 0.01157 \times T[C] \times Th[m] + 0.02709 \times T[C] \times Elem[m] \quad (4)$$

$$D[m] = 0.182 - 0.000793 \times T[C] - 7.10 \times Th[m] - 3.81 \times Elem[m] + 0.000006 \times T[C] \times T[C] + 108.4 \times Th[m] \times Th[m] + 51.7 \times Elem[m] \times Elem[m] - 0.0116 \times T[C] \times Th[m] + 0.0271 \times T[C] \times Elem[m] + 0.2 \times Th[m] \times Elem[m] \quad (5)$$

where, D is the maximum deformation of the automotive bumper, T is the temperature, Th is the thickness of the bumper and Elem is the size of finite elements. When the vapor smoothing process is not applied, Equation (4) could be used. Otherwise, equation (5) could be used to estimate the maximum deformation of the bumper. R-square value of the response surface model was 81%. Two equations represent that the maximum deformation is

sensitive to the thickness of bumper and the temperature under given ranges. Based on two equations, contour plots were computed by using Minitab software as shown in Figure.12. According to the vapor smoothing process, two design guidelines could be obtained by using the response surface methodology, which might be useful information for bumper designers.

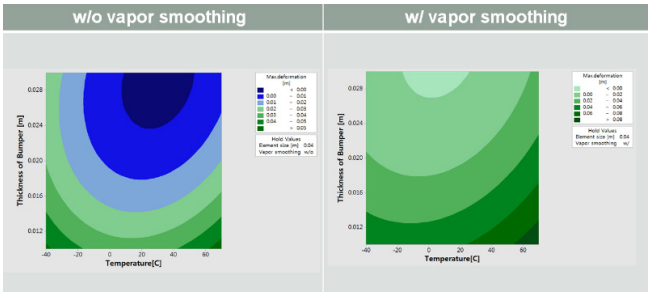


Figure.13. Contour plots according to the vapor smoothing process

4. Conclusion

In this study, the temperature-dependent mechanical properties of ABS parts were measured by used dynamic mechanical analyzer and dilatometry, which were fabricated by FDM and the vapor smoothing technique. The vapor smoothing process weakens thermal stability of ABS 3D printed parts. According to the temperature-dependent storage modulus of 3D printed ABS specimen, the glass transition temperature dramatically decreases by the process. The vapor smoothing process made relatively large variation in CTE of ABS 3D printed parts. Based on the measured thermos-mechanical properties of ABS specimen, finite element analysis and response surface methodology for an automotive bumper made of ABS were performed so that two design guidelines could be provided according to the vapor smoothing process.

Acknowledgment

This work was supported by Dong-Eui University Grant (201702720001) and the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. NRF-2017R1C1B5074219).

References

[1] Lyu MY, Choi TG. Research trends in polymer materials for use in lightweight vehicles 2015;16:213–20. doi:10.1007/s12541-015-0029-x.

[2] Gao W, Zhang Y, Ramanujan D, Ramani K, Chen Y, Williams CB, et al. The status, challenges, and future of additive manufacturing in engineering. Comput Des 2015;69:65–89. doi:10.1016/j.cad.2015.04.001.

[3] Chen L, He Y, Yang Y, Niu S, Ren H. The research status and development trend of additive manufacturing technology. Int J Adv Manuf Technol 2017;89:3651–60. doi:10.1007/s00170-016-9335-4.

[4] Kuo C-C, Mao R-C. Development of a Precision Surface Polishing System for Parts Fabricated by Fused Deposition Modeling. Mater Manuf Process 2016;31:1113–8. doi:10.1080/10426914.2015.1090594.

[5] Garg A, Bhattacharya A, Batish A. Chemical vapor treatment of ABS parts built by FDM: Analysis of surface finish and mechanical strength. Int J Adv Manuf Technol 2017;89:2175–91. doi:10.1007/s00170-016-9257-1.

[6] Singh R, Singh S, Singh IP, Fabbrocino F, Fraternali F. Investigation for surface finish improvement of FDM parts by vapor smoothing process. Compos Part B Eng 2017;111:228–34. doi:10.1016/j.compositesb.2016.11.062.

[7] Singh J, Singh R, Singh H. Repeatability of linear and radial dimension of ABS replicas fabricated by fused deposition modelling and chemical vapor smoothing process: A case study. Measurement 2016;94:5–11. doi:10.1016/j.measurement.2016.07.064.

[8] Cantrell JT, Rohde S, Damiani D, Gurnani R, DiSandro L, Anton J, et al. Experimental characterization of the mechanical properties of 3D-printed ABS and polycarbonate parts. Rapid Prototyp J 2017;23:811–24. doi:10.1108/RPJ-03-2016-0042.

[9] Gao H, Kaweesa D V., Moore J, Meisel NA. Investigating the Impact of Acetone Vapor Smoothing on the Strength and Elongation of Printed ABS Parts. Jom 2017;69:580–5. doi:10.1007/s11837-016-2214-5.

[10] Belter JT, Dollar AM. Strengthening of 3D printed fused deposition manufactured parts using the fill compositing technique. PLoS One 2015;10:1–19. doi:10.1371/journal.pone.0122915.

[11] Melenka GW, Schofield JS, Dawson MR, Carey JP. Evaluation of dimensional accuracy and material properties of the MakerBot 3D desktop printer. Rapid Prototyp J 2015;21:618–27. doi:10.1108/RPJ-09-2013-0093.

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 160-167 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Analysis of Outdoor and Indoor Propagation at 15 GHz and Millimeter Wave Frequencies in Microcellular Environment

Muhammad Usman Sheikh*, Jukka Lempiainen

Tampere University of Technology, Department of Electronics and Communications Engineering, Finland.

ARTICLE INFO

Article history:
Received: 26 November, 2017
Accepted: 07 January, 2018
Online: 30 January, 2018

Keywords:
Multipath propagation
Microcellular
3D ray tracing
System performance
5G
Millimeter wave frequencies

ABSTRACT

The main target of this article is to perform the multidimensional analysis of multipath propagation in an indoor and outdoor environment at higher frequencies i.e. 15 GHz, 28 GHz and 60 GHz, using "sAGA" a 3D ray tracing tool. A real world outdoor Line of Sight (LOS) microcellular environment from the Yokusuka city of Japan is considered for the analysis. The simulation data acquired from the 3D ray tracing tool includes the received signal strength, power angular spectrum and the power delay profile. The different propagation mechanisms were closely analyzed. The simulation results show the difference of propagation in indoor and outdoor environment at higher frequencies and draw a special attention on the impact of diffuse scattering at 28 GHz and 60 GHz. In a simple outdoor microcellular environment with a valid LOS link between the transmitter and a receiver, the mean received signal at 28 GHz and 60 GHz was found around 5.7 dB and 13 dB inferior in comparison with signal level at 15 GHz. Whereas the difference in received signal levels at higher frequencies were further extended in an indoor environment due to higher building penetration loss. However, the propagation and penetration loss at higher frequency can be compensated by using the antenna with narrow beamwidth and larger gain.

1. Introduction

This article is an extension of research work originally presented at International Wireless Communication and Mobile Computing (IWCMC'17) conference [1]. In reference [1], studies were made at 15 GHz only, and only outdoor propagation was studied. However, in this article the radio propagation at millimeter wave frequencies i.e. 60 GHz is additionally studied. Furthermore, the research work of this article also includes indoor propagation analysis, which was not done earlier in [1].

Nowadays, the Fifth Generation (5G) of the mobile communication system is being actively discussed in both industry and academia [2-4]. Currently, various advanced wireless access technologies including High Speed Packet Access (HSPA), Long Term Evolution (LTE), and LTE-Advanced (LTE-A) are in operation. However, the ultimate solution for the ultra high capacity requirement of the future system is expected to be provided by the 5G. It is strongly believed that the higher frequency bands will play a vital role in meeting the capacity targets of the next generation of the cellular networks. The frequency bands between 700 MHz to 4 GHz are currently used by the mobile operators for radio transmission, and are already over loaded with current mobile technologies. Higher frequency bands offer wider spectrum. On the other hand, the higher path loss is also attributed with the higher frequency of operation [5]. The R&D organizations are extensively putting their effort for investigating the utilization of higher frequency bands for mobile communications [6]. The utilization of advanced antenna technologies such as Massive MIMO (MMIMO) and beamforming help in extending the coverage at higher frequencies.

It is important to understand the radio propagation characteristics in order to properly design and to efficiently optimize the system. Classical coverage prediction models do- not provide the insight information about the channel (environment). However, the deterministic ray tracing models provide the multidimensional characteristics of the radio propagation environment [7]. Multipath propagation is a complex phenomenon and involves several propagation mechanisms e.g. specular

*Corresponding Author : Muhammad Usman Sheikh,
Email : muhammad.sheikh@tut.fi

reflection, diffraction from the corners, transmission, diffraction from the rooftops, and diffused scattering [8]. Ray tracing is a promising technique for finding the possible paths between the transmitter and receiver.

At reference [9], the NTT DOCOMO provides the field measurement results of 5G radio system operating at the frequency of 15 GHz in an outdoor and indoor microcellular environment. One of the core targets of this article is to provide the simulation result for the scenario considered in [9], so that a comparison between the measured and simulated data can be made. For this purpose, the sAGA ray tracing tool based on Image Theory (IT) algorithm is utilized for the simulations. The radio propagation properties such as received signal strength, Power Angular Spectrum (PAS), and Power Delay Profile (PDP) at higher frequencies are also provided in this article. The simulation results presented in this article highlight the difference of propagation at different considered frequencies i.e. 15 GHz, 28 GHz and 60 GHz, and also highlight the difference of propagation in an outdoor and indoor environment.

2. System Model and Environment

2.1. System Configuration

The Fifth Generation (5G) of the mobile communication system based on Time Division Duplex (TDD) operating at 15 GHz frequency is considered at reference [9]. The 5G communication system consists of four contiguous Component Carriers (CCs), and each CC is assumed to have 100 MHz bandwidth. Carrier Aggregation (CA) is employed to combine four component carriers. The transmission power per component carrier is set to 27.3 dBm, which makes a total transmission power of the transmitter equals 33.32 dBm for 400 MHz bandwidth. The base station is installed with a directional antenna at the height of 8 m. The transmitter is located on the wall of the building. The base station antenna has a horizontal Half Power Beamwidth (HPBW) of 90° and a vertical HPBW of 10.5° and has a maximum antenna gain of about 14.5 dBi. A mechanical downtilt of 13.5° is used to restrict the propagation in small cell environment. Whereas at the receiver end the MS is assumed to have an omni directional antenna with 0 dBi gain. As, the target of this paper was to provide the simulation results for the system and environment considered in [9], therefore the same set of parameters and environment are assumed in this paper.

2.2. Simulation Platform

The "sAGA" a MATLAB based three-dimensional ray tracing tool is used for the simulations. The sAGA tool is indigenously developed by the authors of this paper. Unlike other quasi 3D ray tracing tool, sAGA performs full three dimensional ray tracing. Multipath propagation involves several mechanisms of interactions e.g. reflection, diffraction, transmission, and scattering. The sAGA ray tracing tool uses Image Theory (IT) to find all the possible reflected, diffracted, ground reflected, and rooftop diffracted paths with the given number of reflections and diffractions [10-12]. In case of reflected paths, the loss in energy due to specular reflection is given by the reflection co-efficient and depends upon the incident and the reflected angle of path and material permittivity. Whereas for the diffracted paths the diffraction loss is given by Berg's recursive model [13]. The energy is scattered in the wide range of directions and the impact of scattering becomes significant at higher frequencies. Therefore, a concentric circle approach given at [14] is used to generate the scattering points on the walls of the building. Ray tracing requires detailed information about the simulation environment and provides accurate results. Three dimensional ray tracing requires 3D map of the simulation environment. The computational complexity of the ray tracing algorithm increases with the increase in the number of supported reflections and diffractions.

A. Simulation Environment

For simulations, a small area from the Yokusuka city of Japan is selected. The Google map view of the targeted area is shown in Figure 1. For outdoor microcellular environment simulation, the parking area of NTT DOCOMO R&D center in Japan is considered, and for indoor simulation the office building is considered in this article.



Figure 1. Google map view of targeted area.

Figure 2 illustrates the two dimensional simplified map of the considered area. The simplified map model consists of just five buildings. For outdoor coverage simulations, it is assumed that the MS is mounted on a vehicle at a height of 3.1 m. The MS is driven at the speed of 10 km/h along eight different routes (A1-A8) as shown in Figure 2. For indoor coverage simulations, two paths A9 and A10 are considered. The path A9 is close to the exterior wall of the building, and path A10 is 7 m away from the exterior wall of the building. The location of the transmitter is marked with the blue spot. The transmitter antenna has the azimuth angle of 90° (facing towards North). All the outdoor and indoor simulation routes have a clear LOS with the transmitter. Two outdoor location points are marked as Pt1 and Pt2, and one indoor location point is marked as Pt3 in Figure 2. These static points are selected for Power Angular Spectrum (PAS) analysis.

For PAS simulations, a directive antenna with 14° HPBW in the horizontal domain and 10.5° HPBW in the vertical domain with 20 dBi maximum gain is used. It is assumed that a directive antenna on a receiver side at a height of 1.65 m is rotated by 360° in the azimuth plane with a step size of 4°. For ray tracing, the reflected path with the maximum three reflections, and diffracted path with single diffraction were found. The scatterers are spread over the surface of the walls of the building. The general system simulation parameters are summarized in Table I.

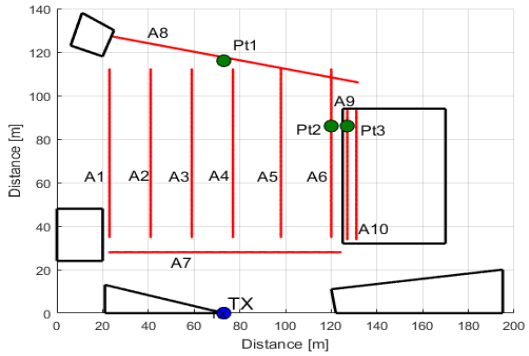


Figure 2. Illustration of simulation area and simulation routes.

Table I. General system simulation parameters for route simulations.

Parameters	Unit	Value
Frequency	GHz	15/28/60
Single carrier component (CC) bandwidth	MHz	100
System bandwidth (4 CCs)	MHz	400
Transmission power per CC	dBm	27.3
Total transmission power	dBm	33.32
Transmitter height	m	8
Antenna downtilt	°	13.5
MS height	m	3.1
Reflections		3
Diffractions		1
Diffuse scattering		Enabled

2.3. Building Penetration Loss (BPL)

The signal experiences a penetration loss while penetrating from the outdoor environment to the indoor environment. Outdoor to indoor penetration loss is generally termed as Building Penetration Loss (BPL). The building penetration loss is the function of frequency and depends heavily on the material characteristics of the building; therefore the BPL can be significantly different for different material types at different frequencies. Generally, the old houses are composed of plane standard glass windows and concrete wall, while the Infrared Reflective (IRR) glass windows are commonly used in the new modern energy saving houses. In reference [15], the old buildings are assumed to have 30 % of the standard glass windows and 70 % of the concrete wall. Similarly, the assumption for new modern building type corresponds to the 70 % of the IRR glass windows and 30 % of the concrete wall.

A simple model structure has been proposed in [15] to model a single material frequency dependent penetration loss. The penetration loss for different material types is provided at [16-20].

$$L_{\text{Single glass,dB}} = 0.1 * \text{Frequency}_{\text{GHz}} + 1, \quad (1)$$

$$L_{\text{Double glass,dB}} = 0.2 * \text{Frequency}_{\text{GHz}} + 2, \quad (2)$$

$$L_{\text{IRR glass,dB}} = 0.3 * \text{Frequency}_{\text{GHz}} + 23, \quad (3)$$

The penetration loss for the concrete wall as a function of frequency is modeled as

$$L_{\text{Concrete,dB}} = 4 * \text{Frequency}_{\text{GHz}} + 5, \quad (4)$$

As the buildings are composite of windows and concrete wall, the building penetration loss for old buildings and new buildings is modeled as shown in (5) and (6), respectively [21].

$$L_{\text{Old building,dB}} = -10\text{Log}_{10}\left[0.3 * 10^{\frac{-L_{\text{Double glass,dB}}}{10}} + 0.7 * 10^{\frac{-L_{\text{Concrete,dB}}}{10}}\right], \quad (5)$$

$$L_{\text{New building,dB}} = -10\text{Log}_{10}\left[0.7 * 10^{\frac{-L_{\text{IRR glass,dB}}}{10}} + 0.3 * 10^{\frac{-L_{\text{Concrete,dB}}}{10}}\right], \quad (6)$$

The building penetration loss as a function of frequency for different building types is shown in Figure 3.

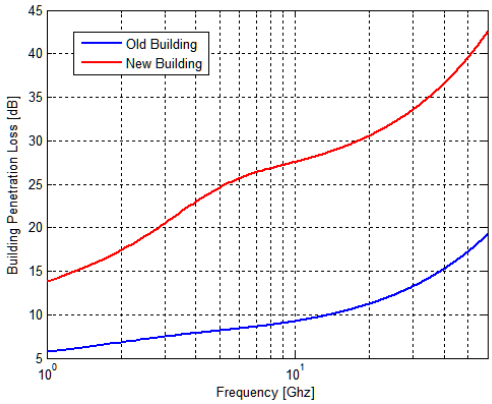


Figure 3. Building penetration loss as a function of frequency.

2.4. Indoor Propagation Loss

In an indoor environment, generally the indoor walls are made up of standard glass alternatively plaster. In [15], two different indoor wall loss models are presented as a function of the frequency assuming an average wall distance of 4 m. The Indoor Loss Model 1 assumes an indoor wall of standard glass, whereas Indoor Loss Model 2 is based on the measurements performed in [16]. Two indoor wall loss models are modeled as shown in (7) and (8). Indoor loss as a function of frequency for two different indoor wall loss models, expressed as dB/m is shown in Figure 4.

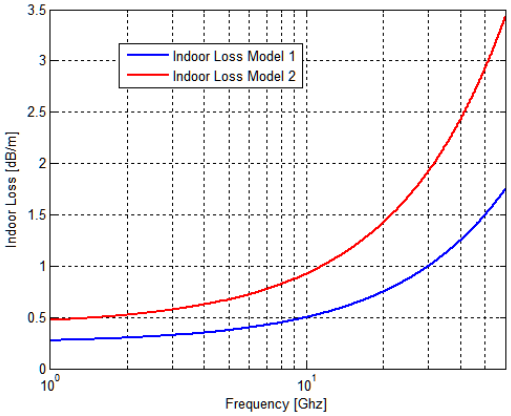


Figure 4. Indoor loss models as a function of frequency.

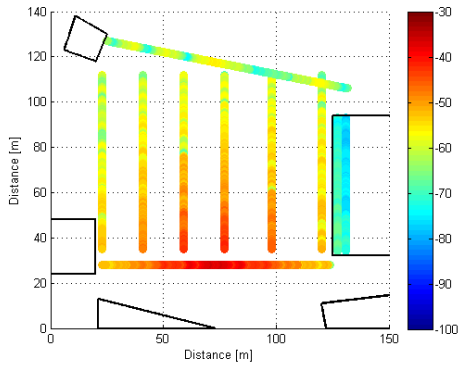
$$L^{(1)}_{\text{Wall loss,dB/m}} = L_{\text{Single glass,dB}}, \quad (7)$$

$$L^{(2)}_{\text{Wall loss,dB/m}} = 0.2 * \text{Frequency}_{\text{GHz}} + 1.7, \quad (8)$$

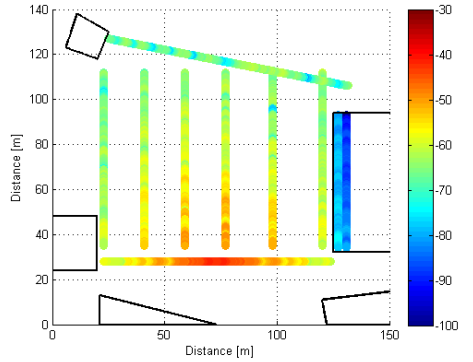
3. Simulation Results and Discussions

Figure 5 shows the received signal strength in dBm for eight outdoor and 2 indoor simulation routes. Figure 5(a), Figure 5(b) and Figure 5(c) shows the received signal strength along the paths at 15 GHz, 28 GHz and 60 GHz, respectively. It can be seen in Figure 5(a) that the maximum received signal strength of around -38 dBm is found in front of the BS antenna in the direction of the main lobe. However, the signal strength starts to degrade as the receiver starts to move away from the transmitter, and deviate from the direction of the main lobe of an antenna. The considered area for simulation was an open environment; therefore the LOS path always existed between the transmitter and receiver. Whereas there are signal fades along the routes due to the constructive and destructive addition of multipaths. The route A1 and A6, and path A2 and A5 are symmetrical and are at almost equal distance from the transmitter. It is interesting to see that the signal strength along the route A6 is better compared with A1 due to more reflected and scattered multipaths from the wall of the nearby building. As soon as the MS crossed the building in route A6, the signal degradation due to the absence of a reflected path is witnessed. The mean received signal level of path A6 is -57.01 dBm at 15 GHz. For indoor paths, old building type is considered. Path A9 is adjacent to path A6. However, due to building penetration loss the mean received signal level for path A9 is -69.78 dBm assuming old building type. The mean received signal level is further drops to -75.12 dBm in case of new building type which consists of IRR glass. Path A10 is located more deep inside the building, and due to additional indoor wall loss the signal strength degrades more rapidly in an indoor environment. The difference between the received signal level at 15 GHz and two other higher frequencies is clearly evident. However, the path loss at higher frequency can be compensated by using the antenna with higher gain.

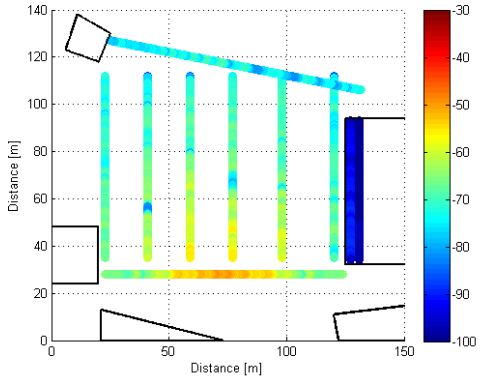
Table II presents the mean received signal level for the different simulation routes at 15 GHz, 28 GHz and 60 GHz. Considering the outdoor simulation routes, the mean path loss difference of around 5.7 dB and 13 dB was found between 15 GHz and 28 GHz transmission, and between 15 GHz and 60 GHz transmission, respectively. Similarly, for indoor simulation paths the mean path loss difference is around 10.57 dB and 27.82 dB between 15 GHz and 28 GHz transmission, and 15 GHz and 60 GHz transmission, respectively. As stated earlier, that the simulation routes A1 and A6 are symmetrical, however due to the presence of nearby wall for route A6 the mean received signal strength is around 1.25 dB and 1.43 dB better than A1 at 15 GHz and 28 GHz, respectively. Building penetration loss is the function of frequency and therefore the signal experiences higher penetration loss at higher frequencies. A significant difference was found between the received signal level of indoor user located in an old and new building type.



(a)



(b)



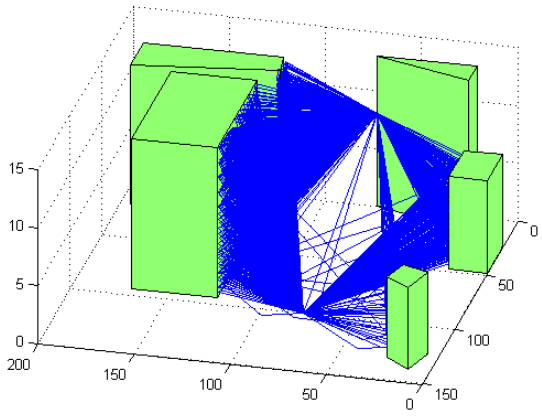
(c)

Figure 5. Received signal strength along the simulation route at, (a) 15 GHz, (b) 28 GHz, and (c) 60 GHz.

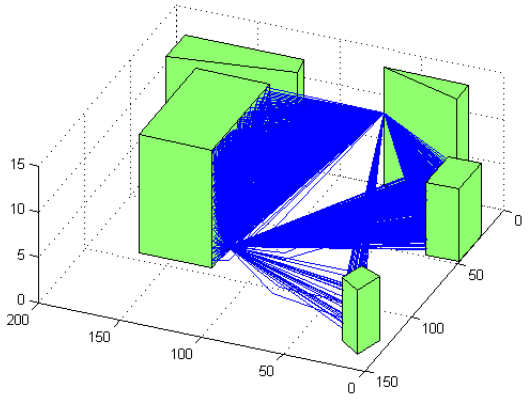
Table II. Mean RX level.

Simulation route	Mean RX level (dBm)		
	15 GHz	28 GHz	60 GHz
A1	-58.26	-64.16	-70.15
A2	-54.96	-60.61	-68.57
A3	-53.23	-58.46	-65.45
A4	-52.37	-58.32	-66.45
A5	-53.92	-60.88	-66.79
A6	-57.01	-62.73	-70.98
A7	-45.54	-51.49	-59.51
A8	-63.09	-67.17	-74.90
A9 (Old building)	-69.78	-78.54	-91.65
A10 (Old building)	-75.12	-86.20	-104.4
A9 (New building)	-88.61	-98.66	-114.97
A10 (New building)	-93.94	-106.33	-127.73

Figure 6(a) and Figure 6(b) show the cluster of rays at 15 GHz reaching the outdoor location Pt1 and Pt2, respectively.



(a)

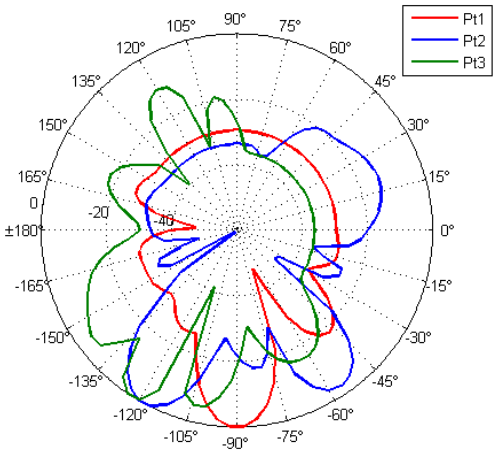


(b)

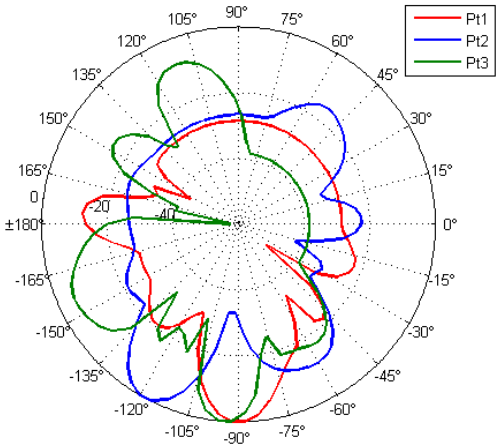
Figure 6. Multipath rays at 15 GHz from TX to (a) Pt1, and (b) Pt2.

The cluster of rays shown in Figure 6 includes wall reflected, ground reflected, diffracted and scattered paths. The impact of diffused scattering is clearly evident, and a considerable amount of energy at receiver points is coming through a large number of scattered paths. In Figure 6(a), the Pt1 is located in the middle of the simulation environment; therefore the walls of the building have clear visibility to the Pt1. Whereas, Pt2 is located closed to the wall of the nearby building, and the building on the right side of the TX does not have visibility to Pt2. Therefore, it can be seen in Figure 6(b) that there is no scattered path from the building on the right side of the transmitter.

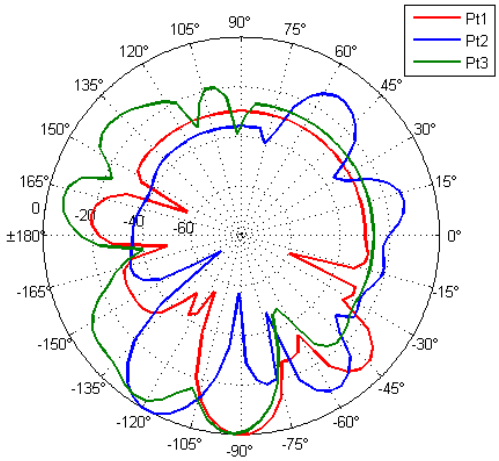
Figure 7(a), Figure 7(b) and Figure 7(c) shows the normalized power angular spectrum for location Pt1, Pt2, and Pt3 at 15 GHz, 28 GHz, and 60 GHz, respectively. The PAS shown in Figure 7 confirms that in the LOS condition the major contribution to the total received power comes through the LOS path, and through the specular reflected paths. However, at 28 GHz and 60 GHz the impact of diffused scattering is more significant than 15 GHz. An indoor environment has more scattering components due to nearby walls. It can be seen in Figure 7 that phenomenon of scattering has more significance at higher frequencies specifically in an indoor environment. The Pt1 is located at an azimuth angle of 90°, and therefore at Pt1 the strongest LOS path has a Direction of Arrival (DoA) of -90° as shown in Figure 7. The next strongest path is the reflected path with -44° of DoA.



(a)



(b)



(c)

Figure 7. Power angular spectrum for location Pt1, Pt2 and Pt3, (a) at 15 GHz, and (b) at 28 GHz.

The LOS path has a 3D path length of 116 m; whereas the reflected path has a path length of 156 m i.e. 40 m longer than the LOS path. Therefore, the strength of the reflected path is almost 17 dB and 22 dB less compared with the LOS path at 15 GHz and

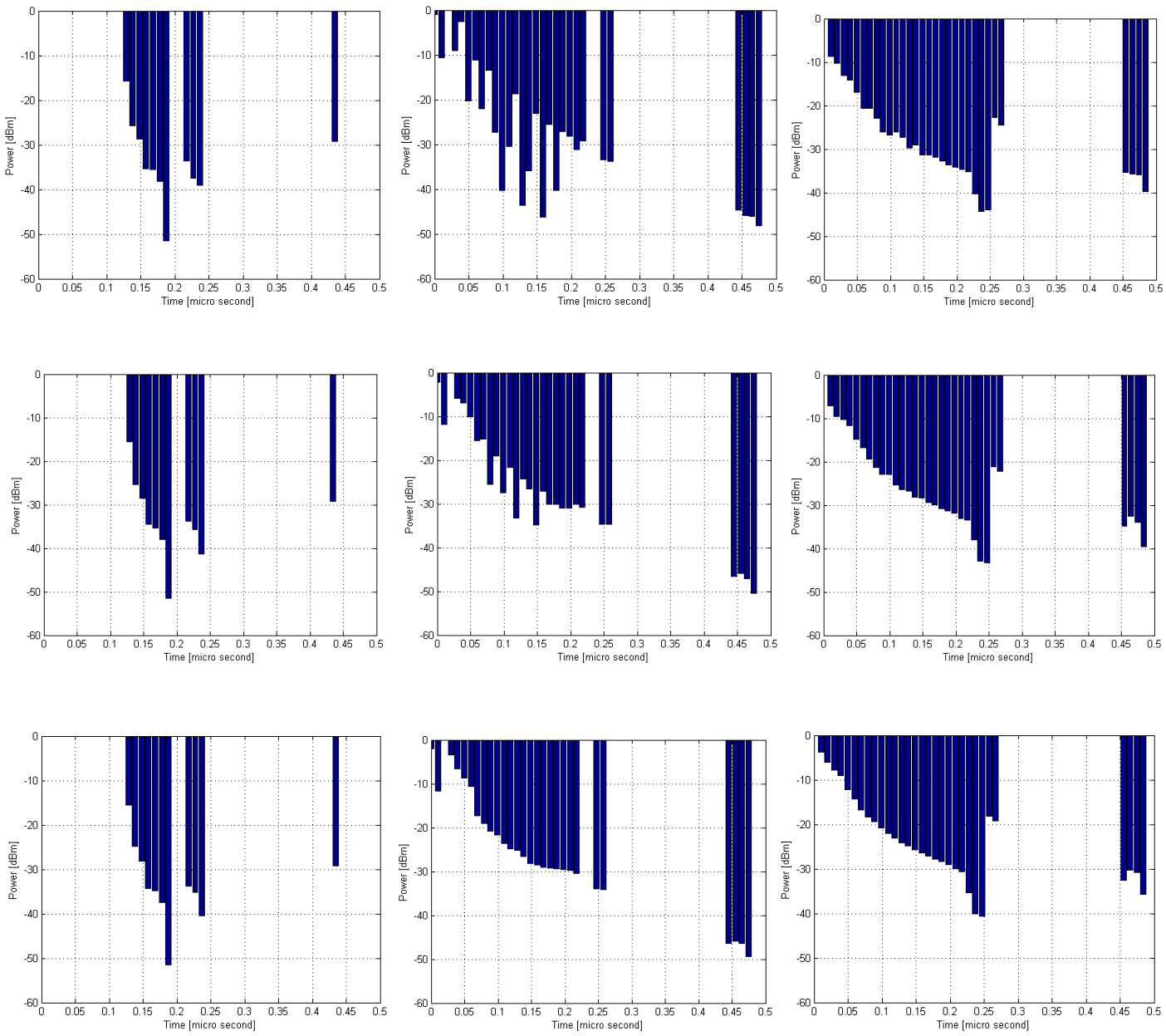


Figure 8. Power delay profile, (a) Pt1 (15GHz), (b) Pt2 (15GHz), (c) Pt3 (15GHz), (d) Pt1 (28GHz), (e) Pt2 (28GHz), (f) Pt3 (28GHz), (g) Pt1 (60GHz), (h) Pt2 (60GHz), and (i) Pt3 (60GHz).

28 GHz, respectively. However, at 60 GHz due to the presence of a large number of scattering components the composite strength of received signal components at -44° of DoA is almost 10 dB less compared with LOS direction. The Pt2 is located on the right side of the transmitter and has a valid LOS link with the transmitter

In Figure 7, the strongest path (LOS path) has around -120° of DoA and the other reflected path is reaching the receiver at -60°. At Pt2, the direct path has a path length of 98 m, and the reflected path has a path length of 104 m i.e. 6 m longer than the LOS path. Therefore, the strength of the reflected path is around 4 dB and 11 dB less compared with the LOS path at 15 GHz and 28 GHz, respectively. Also, the impact of diffused scattering from the wall at an angle of -60 ° at 28 GHz and 60 GHz is visible. The considered microcellular case is a simple environment, and not a multipath rich outdoor environment. Therefore, the most of the energy at the receiver points is coming from a narrow direction only. However, in an indoor environment, due to a rich indoor scattering environment the PAS of incoming multipath is wide spread, especially at higher frequencies i.e. 28 GHz and 60 GHz.

Figure 8 shows the Power Delay Profile (PDP) acquired through simulations of location Pt1, Pt2 and Pt3 at 15 GHz, 28 GHz and 60 GHz. For Pt1, the reflected path has 40 m longer path length therefore the reflected path reaches the receiver point with a delay of around 0.125 μs relative to LOS path, and the signal strength of the reflected path is around 16 dB less compared with LOS path. For Pt1, the power delay profile is almost identical at

15 GHz, 28 GHz, and 60 GHz. The Pt2 is located close to the wall of the building, and therefore the difference of path length between the reflected path and the LOS path is short. By comparing Figure 8(b), Figure 8(e), and Figure 8(h), it can be noted that the reflected path with slightly less signal strength than LOS path reaches the receiver point within a very short delay of 0.01 μs. Other diffracted and scattered paths from the nearby wall also reach the receiver Pt2 with a short delay period. The power delay profile of Pt2 got smoother at higher frequencies due to a large number of scattering paths at higher frequencies.

Indoor location Pt3 is located nearby the outdoor location Pt2. However, an indoor environment is a scattering rich environment and a larger number of scattering paths reach the receiver point with small delay. Therefore, the PDP of Pt3 is smoother compared with PDP of Pt2. It is interesting to find that the phenomenon of scattering becomes more prominent in higher frequencies, and the power of multipath components with short delay improves at 28 GHz and 60 GHz. The mean delay spread in nanoseconds for static location points are given in Table III.

Table III. Mean delay spread.

Location	Mean Delay Spread (ns)		
	15 GHz	28 GHz	60 GHz
Pt1	150	150	150
Pt2	109	107	106
Pt3	124	119	114

4. Conclusion

In this article, a multidimensional analysis of the multipath propagation in an outdoor and indoor microcellular environment has been carried out at higher frequencies i.e. 15 GHz, 28 GHz and 60 GHz, by means of a sophisticated ray tracing tool. The simulations are performed using a simplified 3D map of a small area from the Yokusuka city of Japan. The considered outdoor simulation routes and outdoor static points represent an open environment scenario with valid LOS link between the transmitter and the receiver in a microcellular environment. Indoor simulation routes and an indoor static point were also considered for studying indoor propagation at higher frequencies. The simulation results presented for different simulation routes at 15 GHz are in a close relationship with a measurement results presented in [9]. Similarly at 15 GHz for the selected static points the power angular spectrum (PAS) at receiver end acquired through simulations are in a satisfactory agreement with the measured PAS given at [9]. The power angular spectrum (PAS) at receiver shows the direction of arrival of LOS, specular reflected, diffracted and diffused scattered paths. It was found that mean received signal strength in an outdoor environment for the selected simulation routes at 15 GHz is almost 5.7 dB, and at 60 GHz the mean RX level is almost 13 dB inferior in comparison with the propagation at 15 GHz due to higher propagation loss at higher frequencies. The difference in a mean received signal in an indoor environment between 15 GHz and 28 GHz is extended to nearly 10. 5 dB, and the difference between the mean received signal in an indoor environment between 15 GHz and 60 GHz is further stretched to 28.5 dB due to extra building penetration loss at higher frequencies. However, the higher path loss at 28 GHz and 60 GHz can be compensated by using an antenna with higher gain. The power delay profile acquired through simulations was found helpful in distinguishing the multipath richness of the environment. The indoor was found more scattering rich environment compared with outdoor environment. The absolute values of performance results presented in this article may vary depending upon the modeling impairment, antenna radiation pattern, simulation parameters and simulation environment.

Acknowledgment

Authors would like to thank European Communications Engineering (ECE) Ltd for supporting this research work.

References

[1] M. U. Sheikh and J. Lempiäinen, "Analysis of multipath propagation for 5G system at higher frequencies in microcellular environment," 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, 2017, pp. 1660-1664.

[2] NTT DoCoMo, Inc, "5G Radio Access: Requirements, Concept and Technologies," White Paper, 2014.

[3] Huawei Technologies, "5G: A Technology Vision", White Paper, 2013.

[4] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, Y. Selén, "5G radio access," Ericsson Review, vol. 91, no. 6, pp. 42–48, 2014.

[5] T. S. Rappaport , J. N. Murdock and F. Gutierrez , "State of the art in 60 GHz integrated circuits & systems for wireless communications" , Proc. IEEE , vol. 99 , pp.1390 -1436 , 2011.

[6] S. G. Larew, T. A. Thomas, M. Cudak, A. Ghosh, "Air Interface Design and Ray Tracing Study for 5G Millimeter Wave Communications" in Proc. IEEE Globecom 2013, Atlanta, USA, 9-13 Dec, 2013.

[7] F. Fuschini, H. El-Sallabi, V. Degli-Esposti, L. Vuokko, D. Guiducci and P. Vainikainen, "Analysis of Multipath Propagation in Urban Environment Through Multidimensional Measurements and Advanced Ray Tracing Simulation," in IEEE Transactions on Antennas and Propagation, vol. 56, no. 3, pp. 848-857, March 2008.

[8] T. S. Rappaport, "Wireless Communications: Principles and Practice", 2nd Edition, 2001.

[9] K. Tateishi et al., "Field experiments on 5G radio access using 15-GHz band in outdoor small cell environment," Personal, Indoor, and Mobile Radio Communications (PIMRC), 2015 IEEE 26th Annual International Symposium on, Hong Kong, 2015, pp. 851-855.

[10] H. W. Son, and N. H. Myung, "A Deterministic Ray Tube Method for Microcellular Wave Propagation Prediction Model," in IEEE Transactions on Antennas and Propagation, vol. 47, no. 8, pp. 1344-1350, August 1999.

[11] S. Soni, and A. Bhattacharya, "An Efficient Two-dimensional Ray-tracing Algorithm for Modeling of Urban Microcellular Environment", in International Journal of Electronics and Communications, vol. 66, no. 6, pp. 439-447, June 2012.

[12] D. N. Schettino, F. J. S. Moreira and C. G. Rego, "Efficient Ray Tracing for Radio Channel Characterization of Urban Scenarios," in 12th Biennial IEEE Conference on Electromagnetic Field Computation, Miami, FL, 2006, pp. 267-267.

[13] J.-E. Berg, "A Recursive Method for Street Microcell Pathloss Calculations", Sixth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC'95, Wireless: Merging onto the Information Superhighway. Vol. 1, 1995.

[14] X. Li, "Efficient Ray Tracing Simulation", Master's thesis at LUND University, 2014.

[15] E. Semaan, F. Harrysson, A. Furuskär and H. Asplund, "Outdoor-to-indoor coverage in high frequency bands", 2014 IEEE Globecom Workshops (GC Wkshps), Austin, TX, 2014, pp. 393−398.

[16] C. Larsson, F. Harrysson, B.-E. Olsson, and J. E. Berg, "An outdoor-to-indoor propagation scenario at 28 GHz," in 8th European Conference on Antennas and Propagation (EuCAP 2014), The Hague,The Netherlands, April 2014, pp. 3301−3304.

[17] W. C. Stone, “Electromagnetic signal attenuation in construction materials, ”NIST Building and Fire Research Laboratory, Gaithersburg, Maryland, NISTIR 6055 Report No. 3 6055, Oct. 1997.

[18] L. M. Frazier, “Radar surveillance through solid materials,” in Proceedings of the SPIE - The International Society for Optical Engineering, vol. 2938, Hughes Missile Syst. Co., Rancho Cucamonga, CA, USA, 1997, pp. 139−146.

[19] R. Wilson, “Propagation losses through common building materials 2.4 GHz vs 5 GHz,” University of Southern California, CA, Tech. Rep. E10589, Aug. 2002.

[20] C. A. Remley, G. H. Koepke, C. L. Holloway, C. A. Grosvenor, D. G. Camell, J. M. Ladbury, R. Johnk, D. R. Novotny, W. F. Young, G. Hough, M. McKinley, Y. Becquet, and J. Korsnes, “Measurements to support modulated-signal radio transmissions for the public-safety sector,” NIST, Boulder, CO, Tech. Rep. Tech. Note 1546, Apr. 2008.

[21] White paper on “5G Channel Model for bands upto 100 GHz”.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 6, 162-174 (2017)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

Mealy-to-Moore Transformation - A state stable design of automata

Mustafa Oezguel*, Florian Deeg, Sebastian M. Sattler

Chair of Reliable Circuits and Systems, Friedrich-Alexander-Universität Erlangenen-Nürnberg, Paul-Gordan-Str. 5, 91052 Erlangen, Germany

ARTICLEINFO

Article history:
Received: 27 October, 2017
Accepted: 02 December, 2017
Online: 10 December, 2017

Keywords :
Asynchronous Feedback
Functional Safety
Stabilization
Hazard-free
Race-free
Parallel de-composition
Dual-rail

ABSTRACT

The paper shows a method of transforming an asynchronously feedbacked Mealy machine into a Moore machine. The transformation is done in dual-rail logic under the use of the RS-buffer. The transformed machine stabilizes itself and is safe to use. The transformation is visualized via KV-diagrams and calculated with formulas. We will present three use-cases for a better understanding. To underpin the stated transformation a simulation is also presented.

1 Introduction

A Synchronous circuits entail several benefits. Firstly performance advantages, e.g. faster system perfomance, no clock synchronization, fewer power consumption and fewer electromagnetic emission. Secondly asynchronous circuits also have safety advantages, e.q. no clock skew. Asynchronous circuits can also lead to simplicity, by connecting different parts modularly [1, 2]. On the other hand asynchronous circuits provide risks, such as faults like hazards, glitches and so on [3]. Therefore it is very important to design asynchronous circuits correctly, especially in safety critical applications such as in automotive, where human lives are at stake. To prevent faults like hazards, a mealy-to-moore transformation will be presented, which creates a function stable asynchronous machine. The Mealy machine with the state transfer function δ and the output function λ, see figure 1 be given. Each Mealy machine can be transformed into an equivalent Moore machine [3] for example by increasing the arity of δ by $|x|$ and coding it with x, $z' = (z,x)$ with $\delta'(z',x) \mapsto (\delta(z,x),x)$, and the output function is only dependent on the new state variable z' with $\mu : (z') \mapsto \mu(z')$.
By comparing the two machines the pros and cons can be shown. Therefore the idea of transforming the machines into each other to profit from the benefits of both is obvious. Mealy machines have the advantage of requiring less states since one state can produce a number of different outputs in combination with the input. A Moore machine's state on the other hand only produces one output. A Mealy machine is also faster by reacting directly to the input. This feature however is not always wanted, since it can lead to undesired outputs (e.g. hazards, glitches) when the input is variable. A Moore machine is more stable in this regard, since it only indirectly reacts to input changes. The output only changes when transferring into the next state. Transforming a Mealy machine into a Moore machine is therefore useful in case a direct dependence on the input is to be avoided [2].
In the paper the transformation of a Mealy machine into a Moore machine is presented. The machine will be designed under the use of dual-rail logic and the RS-buffer, and will stabilize itself. Firstly the transformation is made by breaking the feedback up and nesting it in the RS-buffer, secondly by encoding the inputs as states, so that the output function is only dependent on pseudo states. With the method presented in this article, function stable sequential circuit parts can be seen as combinational logic and moved over other blocks at will. With this method, individual machines can be strung together to realize complex circuits for safety relevant applications.

2 Theory

This paper describes the underlying theory of the transformation and provides an illustrative example.

Figure 1 shows a fully asynchronous Mealy machine. The branches entering a node in the graph should end reflexively, so that only transient states are allowed which are conscious and triggered from the outside [2]. To consider all branches of the Mealy machine as locally reflexively

*Corresponding Author: Mustafa Oezguel, mustafa.oezguel@fau.de

https://dx.doi.org/10.25046/aj020621

concluded, the feedback should be moved over λ, making the output y the feedback. In order to set the correct state z for the transfer function δ, a function λ^{-1} is realized. This function generates the reduced state $\underline{z}$ from the feedback y. This equivalent transformation is outlined in figure 2.

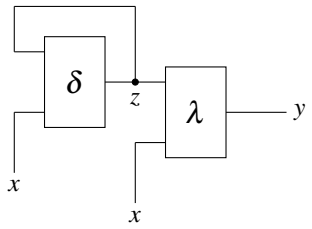


Figure 1: Fully asynchronous Mealy machine

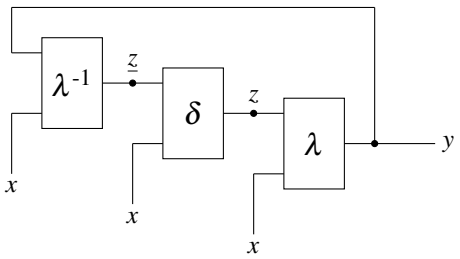


Figure 2: Equivalent transformation

The following applies:

$$\begin{aligned} \delta(\underline{z},x) &= \underline{z} \\ \delta\left(\lambda^{-1}(y,x),x\right) &= \lambda^{-1}(y,x) \\ \lambda\left(\delta\left(\lambda^{-1}(y,x),x\right),x\right) &= y \\ \text{mit } \lambda\left(\lambda^{-1}(y,x),x\right) &= y \end{aligned}$$

The next stage of transformation shows at first sight no feedback. Only the RS-buffer, which is in $\underline{\delta}$, has a nested feedback. This transformation is outlined in figure 3. Transient states can be triggered via the input, while state stable states are being held via the RS-buffer. In order to set the correct state $\underline{z}$ to μ, the function $\underline{\delta}$ must be well designed, which will be shown in the following sections.

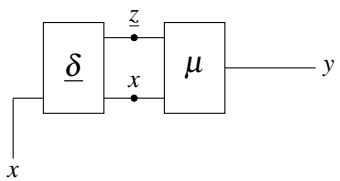


Figure 3: Moore transformation

The following applies:

$$\begin{aligned} \underline{\delta}(\underline{z},x) &= \underline{z} \\ \mu(\underline{z},x) &= y \end{aligned}$$

2.1 Dual-rail logic

For the implementation in dual-rail logic [4], the fully asynchronous circuit from figure 1 will be divided in the 1- and 0-share. This is done by partitioning the state transfer function and the output function, see figure 4.

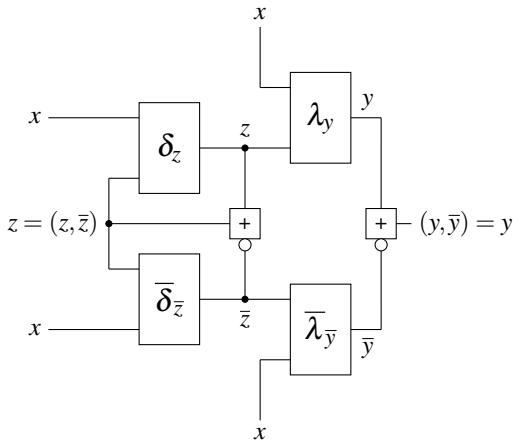


Figure 4: Mealy realized in dual-rail logic

The functions δ and λ are each w.l.o.g. realized in two blocks $\delta = (\delta_z, \overline{\delta}_{\overline{z}})$ and $\lambda = (\lambda_z, \overline{\lambda}_{\overline{z}})$. In order to guarantee this secure dual rail structure, RS-buffers are used. The mode of operation corresponds in a certain manner to the C-element [4, 5]. The function $\underline{\delta}$ of figure 3 will be designed by appropriate coding according to the RS-buffer, showing a general structure as outlined in figure 5.

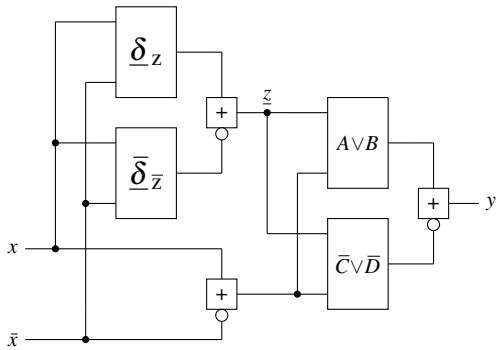


Figure 5: Transformed Moore

The stabilized 1-states, $\underline{\Lambda}_z \circ\!\!-\!\!\bullet \underline{z}$, are transformed into 1-outputs, $\Lambda_{\underline{y}} \circ\!\!-\!\!\bullet \underline{y}$. The corresponding Venn diagram of the stabilized 1-states in 1-outputs can be seen in figure 6. 1-states that appear as 0-outputs are declared as $\overline{\Lambda}_{\underline{z}}$, 0-states which appear as 1-outputs are declared as $\Lambda_{\overline{z}}$. The 1-partition is composed of:

$$\Lambda_{\underline{y}} \circ\!\!-\!\!\bullet \lambda_y(\underline{\delta}_z) + \lambda_y(\underline{\overline{\delta}}_{\overline{z}}) + \overline{\lambda}_{\overline{y}}(\underline{\delta}_z)$$

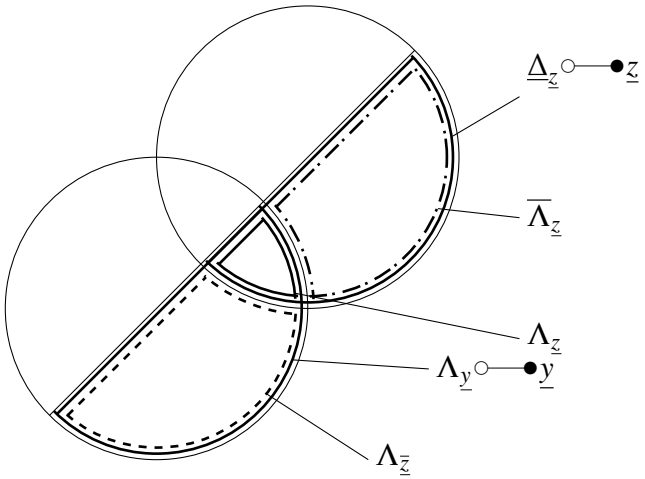


Figure 6: Venn diagram of 1-states in 1-outputs

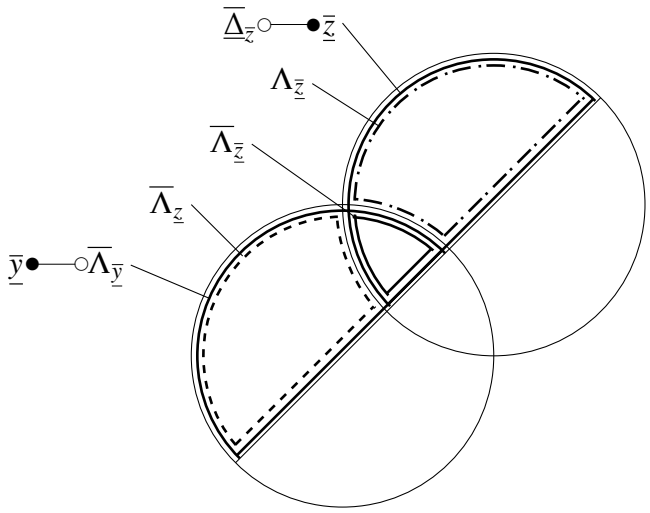


Figure 7: Venn diagram of 0-states in 0-outputs

The stabilized 0-states, $\overline{\underline{\Delta}}_{\overline{\underline{z}}} \circ\!\!-\!\!\bullet \overline{\underline{z}}$, are transformed into 0-outputs, $\overline{\Lambda}_{\overline{\underline{y}}} \circ\!\!-\!\!\bullet \overline{\underline{y}}$. The corresponding Venn diagram of the 0-states in 0-outputs is shown in figure 7. 0-states which appear as 1-outputs are declared as $\Lambda_{\overline{\underline{z}}}$, 1-states which appear as 0-outputs are declared as $\overline{\Lambda}_{\underline{z}}$. The 0-partition is composed of:

$$\overline{\Lambda}_{\overline{\underline{y}}} \circ\!\!-\!\!\bullet \overline{\lambda}_{\overline{y}}(\overline{\underline{\delta}}_{\overline{z}}) + \overline{\lambda}_{\overline{y}}(\underline{\delta}_{\underline{z}}) + \lambda_{y}(\overline{\underline{\delta}}_{\overline{z}})$$

To understand the coding, first the RS-buffer has to be understood.

2.2 RS-Buffer

The RS-Buffer used and its circuit symbol are represented in figure 8 and 9 and

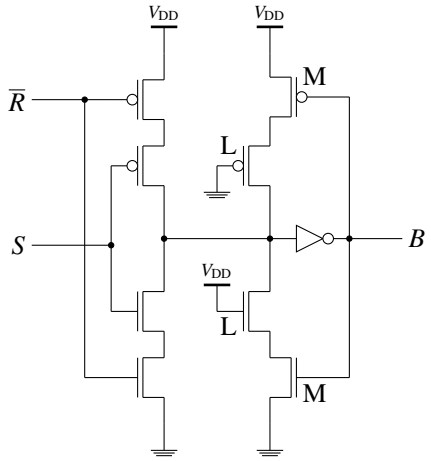


Figure 8: RS-Buffer (schematic)

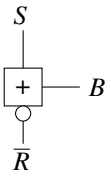


Figure 9: Circuit symbol of the used RS-Buffer

its truth table is specified in table 1.

R	0	0	1	1
S	0	1	0	1
B	B	1	0	B

Table 1: Truth table of the RS-Buffer

The RS-buffer is used to synchronize a dual-rail signal by only letting a state tranfer happen, if both incoming signals are disjoint (how it is desired) to each other. This means that there is no overlaying of signal, which could lead to fatal faults. The RS-buffer consists of the tri-state driver, which is the first state and the second stage the so called babysitter. The driving structure now provides 3 functions. Firstly the setting signal, which means the output is in the logical 1-state. Secondly the resetting signal, which means the output is in logical 0-state and least the hold state which holds the last state at the output by keeping the signal in the babysitter and having a high impedance state at the tri state. The babysitter consists of two complementary loops. The formula for the output B is therefore $B := \overline{R}(B \vee S) \vee S(B \vee \overline{R}) = \overline{R}S \vee \overline{R}B \vee SB$.

2.3 Design of the reduced state transfer function $\underline{\delta}$

As you can see the RS-buffer is coded via a 3-partitioning. There is the setting signal S=[1], the resetting signal R=[1] and the hold signal RS=[00]$\vee$[11]. The state transfer function of an automaton can in turn be split in 4 parts. Firstly

the transient 1-share, that means independent of the state the next state will be at logical 1, secondly the transient 0-share, that means output at logical 0 independent of the state, thirdly the state stable share of the state transfer function, that means the share, which is at constant input dependent on the value of the last state and the state instable state, which means a permanent change of the state at constant input, see figure 10. The resulting 3 parts, which are left, can now be easily compared to the 3-partitition of the RS-buffer. The transient 1-share $\underline{\delta}_z$ is realized by the setting signal of the RS-buffer, similar the transient 0-share $\overline{\underline{\delta}}_{\overline{z}}$ is realized by the resetting signal and the state stable share is realized by the hold signal of the RS-buffer.

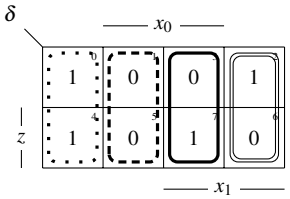


Figure 10: Parts of a state transfer function

To understand the transformation, first a few compositions must be defined:

$$
\begin{aligned}
{}_{z}\delta(z,x) &= \delta(z=1,x)\\
{}_{\overline{z}}\delta(z,x) &= \delta(z=0,x)\\
{}_{z}\overline{\delta}(z,x) &= \overline{\delta}(z=1,x)\\
{}_{\overline{z}}\overline{\delta}(z,x) &= \overline{\delta}(z=0,x)
\end{aligned}
$$

This work only concentrates on function stable machines, so that state instable parts are not allowed. State instable parts can be deduced via

$${}_{z}\overline{\delta} \wedge {}_{\overline{z}}\delta$$

Functions must not exhibit these parts. The other parts can be deduced by the following expressions

$$
\begin{aligned}
&\text{transient } 1: \underline{\delta}_z = {}_{z}\delta \wedge {}_{\overline{z}}\delta\\
&\text{transient } 0: \overline{\underline{\delta}}_{\overline{z}} = {}_{z}\overline{\delta} \wedge {}_{\overline{z}}\overline{\delta}\\
&\text{state stable}: {}_{z}\delta \wedge {}_{\overline{z}}\overline{\delta}
\end{aligned}
$$

This leads to the genereal formula for $\underline{\delta}_i$ for n states:

$$
\begin{aligned}
\underline{\delta}_{z_i} &= \underline{\delta}_{z_i} + \neg\overline{\underline{\delta}}_{\overline{z}_i}\\
&= ({}_{z_i}\delta_{z_i} \wedge {}_{\overline{z}_i}\delta_{z_i}) + \neg({}_{z_i}\overline{\delta}_{\overline{z}_i} \wedge {}_{\overline{z}_i}\overline{\delta}_{\overline{z}_i})\\
&= (\delta_{z_i}(z_i=1,x) \wedge \delta_{z_i}(z_i=0,x))\\
&+ \neg(\overline{\delta}_{\overline{z}_i}(z_i=1,x) \wedge \overline{\delta}_{\overline{z}_i}(z_i=0,x))
\end{aligned}
$$

with $i = 0...n-1$

2.4 Design of the output function

The detailed description of the Venn-diagrams given in figure 11 and 12 is reported in [6].

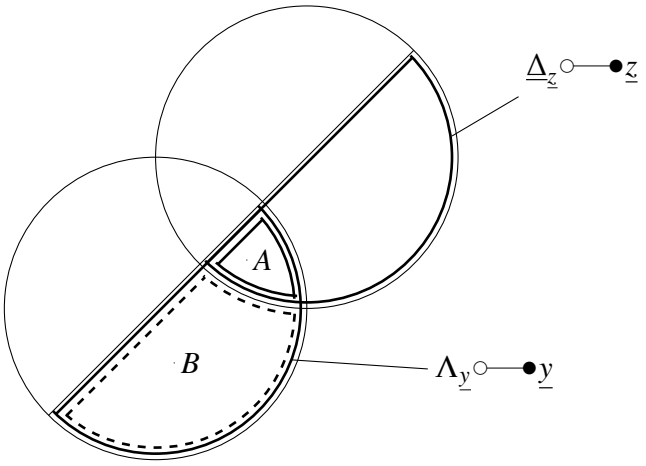


Figure 11: Venn-diagram of 1-states in 1-outputs

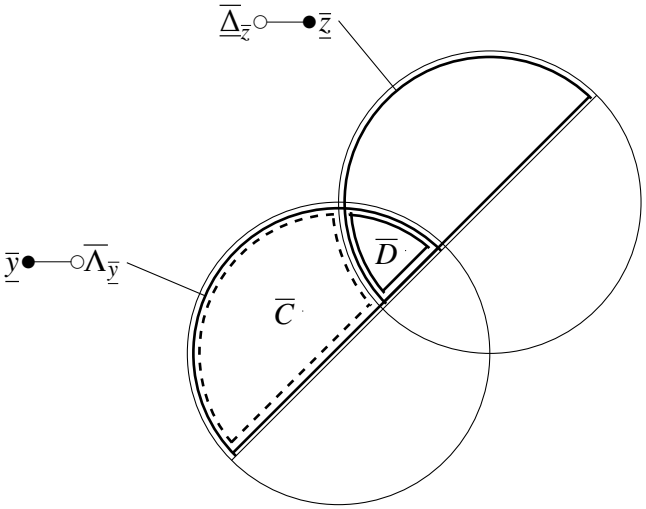


Figure 12: Venn-diagram of 0-states in 0-outputs

The coding of the parts A, B, $\overline{C}$ and $\overline{D}$ of the Venn-diagrams, see figure 11 and figure 12, is as follows:

$$
\begin{aligned}
&(z,y)\\
&(11) = A = \delta_z \wedge \lambda_y\\
&(01) = B = \overline{\delta}_{\overline{z}} \wedge \lambda_y\\
&(10) = \overline{C} = \delta_z \wedge \overline{\lambda}_{\overline{y}}\\
&(00) = \overline{D} = \overline{\delta}_{\overline{z}} \wedge \overline{\lambda}_{\overline{y}}
\end{aligned}
$$

The dual-rails y and $\overline{y}$ for y are now:

$$
\begin{aligned}
y &= A \vee B\\
\overline{y} &= \overline{C} \vee \overline{D}\\
y &= y + \neg\overline{y}
\end{aligned}
$$

The seperate parts of the machine have now been deduced and can simply be strung together to get the form of figure 5.

3 Use-Case

To get a better insight, three examples will be presented.

3.1 1-dimensional example

For $x = (x_1, x_0)$, $y = (y)$ and $z = (z)$ the transformation of the Mealy machine $(X, Y, Z, \delta, \lambda)$ with the state transfer function and output function

$$\delta(z,x) = x_1x_0 \vee z\overline{x}_1x_0 \qquad \lambda(z,x) = \overline{x}_1x_0 \vee z\overline{x}_1$$

will be executed. The automaton before the transformation can be seen in figure 13.

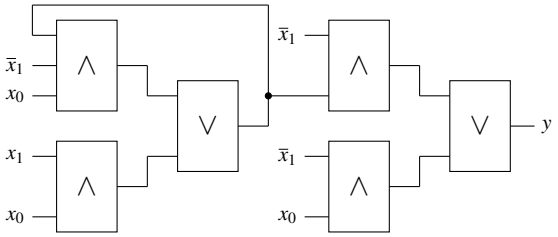


Figure 13: Mealy machine before the transformation

3.1.1 Design of $\underline{\delta}$

The KV-diagram of the state transfer function can be seen in figure 14.

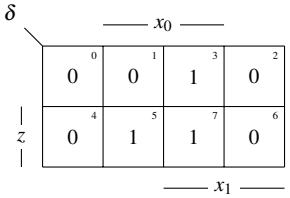


Figure 14: The state transfer function

To code the RS-buffer correctly, firstly the KV-diagram will be compacted, see figure 15.

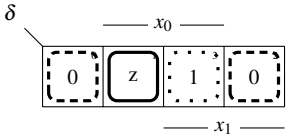


Figure 15: Compacted KV-diagram of δ

The parts of the state transfer function can easily be seen and the KV-diagrams for $\underline{\delta}_z$ and $\overline{\underline{\delta}}_{\overline{z}}$ can be deployed, see figure 16. For $\overline{\underline{\delta}}_{\overline{z}}$ 0 is coded as 1 and 1 is coded as 0 respectively, because of the dual-rail structure.

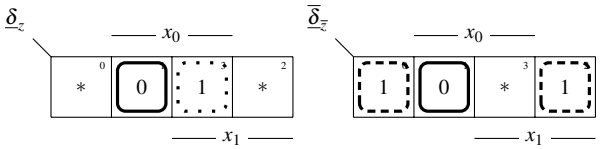


Figure 16: $\underline{\delta}_z$ and $\overline{\underline{\delta}}_{\overline{z}}$

For the next steps of the transformation, first the complement of δ has to be calculated:

$$\overline{\delta}(z,x) = \overline{z}\overline{x}_1 \vee \overline{x}_0$$

The parts of the state transfer function are:

$$\begin{aligned} {}_z\delta &= \delta(z=1,x) = x_0 \\ {}_{\overline{z}}\delta &= \delta(z=0,x) = x_1x_0 \\ {}_z\overline{\delta} &= \overline{\delta}(z=1,x) = \overline{x}_0 \\ {}_{\overline{z}}\overline{\delta} &= \overline{\delta}(z=0,x) = \overline{x}_1 \vee \overline{x}_0 \end{aligned}$$

It has to be checked, if there are any instable parts:

$${}_{\overline{z}}\delta_z \wedge {}_z\overline{\underline{\delta}}_{\overline{z}} = x_1x_0 \wedge \overline{x_0} = 0$$

${}_{\overline{z}}\delta \wedge {}_z\overline{\delta}$ is 0 means the function has no instable parts. The functions $\underline{\delta}_z$ and $\overline{\underline{\delta}}_{\overline{z}}$ can now be calculated via

$$\begin{aligned} \underline{\delta}_z &= {}_z\delta_z \wedge {}_{\overline{z}}\delta_z = x_1x_0 \\ \overline{\underline{\delta}}_{\overline{z}} &= {}_z\overline{\delta}_{\overline{z}} \wedge {}_{\overline{z}}\overline{\delta}_{\overline{z}} = \overline{x}_0 \end{aligned}$$

The coding for $\underline{\delta}$ of the RS-buffer is as follows:

$$\begin{aligned} &(r\ s) \\ \text{Set:} &(*1) = x_1x_0\ (\underline{\text{transient 1-share}}) = \underline{\delta}_z \\ \text{Reset:} &(1*) = \overline{x}_0\ (\underline{\text{transient 0-share}}) = \overline{\underline{\delta}}_{\overline{z}} \\ \text{Hold:} &(**) = \overline{x}_1x_0\ (\underline{\text{state stable share}}) \end{aligned}$$

Now $\underline{\delta}$ can be derived:

$$\begin{aligned} \underline{\delta} &= \underline{\delta}_z + \neg\overline{\underline{\delta}}_{\overline{z}} \\ &= x_1x_0 + \neg(\overline{x}_0) \end{aligned}$$

3.1.2 Design of λ

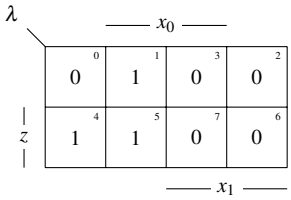


Figure 17: Output function λ

The stabilized state transfer function has now been designed, now the output function must be designed. Therefore the output function has to be layed over the state transfer function and the parts have to be specified:

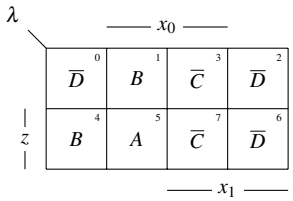


Figure 18: Parts of the output function

The parts are:

$$A(z,x) = z\overline{x}_1x_0$$
$$B(z,x) = \overline{z}\overline{x}_1x_0 \vee z\overline{x}_1\overline{x}_0$$
$$\overline{C}(z,x) = x_1x_0$$
$$\overline{D}(z,x) = x_1\overline{x}_0 \vee \overline{z}\overline{x}_0$$

This leads to the output function:

$$y = A \vee B = \overline{x}_1x_0 \vee z\overline{x}_1$$
$$\overline{y} = \overline{C} \vee \overline{D} = x_1 \vee \overline{z}\overline{x}_0$$
$$y = y + \neg\overline{y} = \overline{x}_1x_0 \vee z\overline{x}_1 + \neg(x_1 \vee \overline{z}\overline{x}_0)$$

The resulting automaton has the form of figure 19 with the determined functions $y = A \vee B$ and $\overline{y} = \overline{C} \vee \overline{D}$.

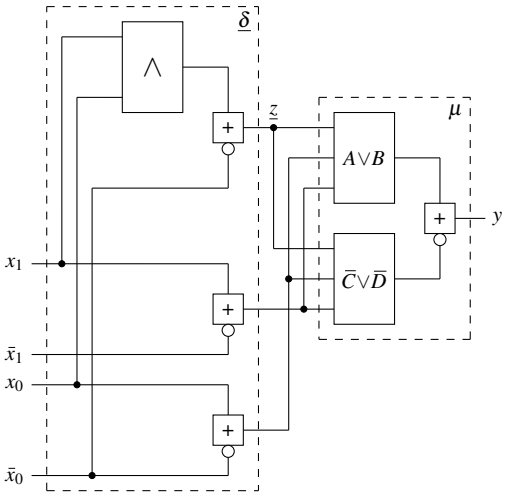


Figure 19: Transformed Moore

The truth table of the automaton:

#	z	x_1	x_0	δ	$\underline{\delta}_z$	$\overline{\underline{\delta}}_{\overline{z}}$	$\underline{\delta}$	λ	A	B	$\overline{C}$	$\overline{D}$	$A \vee B$	$\overline{C} \vee \overline{D}$	y
0	0	0	0	0	*	1	0	0	*	*	*	1	*	1	0
1	0	0	1	0	*	*	0	1	*	1	*	*	1	*	1
2	0	1	0	0	*	1	0	0	*	*	*	1	*	1	0
3	0	1	1	1	1	*	1	0	*	*	1	*	*	1	0
4	1	0	0	0	*	1	0	1	*	1	*	*	1	*	1
5	1	0	1	1	*	*	1	1	1	*	*	*	1	*	1
6	1	1	0	0	*	1	0	0	*	*	*	1	*	1	0
7	1	1	1	1	1	*	1	0	*	*	1	*	*	1	0

Table 2: Partial truth table of the resulting automaton

3.2 2-dimensional example

For $x = (x_1, x_0)$, $y = (y)$ and $z = (z_1, z_0)$ the transformation of the Mealy machine $(X, Y, Z, \delta, \lambda)$ with the state transfer functions and output function

$$\delta_{z_0}(z,x) = \overline{x}_0 \vee z_1x_1 \qquad \delta_{z_1}(z,x) = z_0x_1x_0 \vee z_1x_1\overline{x}_0$$
$$\lambda(z,x) = z_1x_1 \vee z_0\overline{x}_1\overline{x}_0$$

will be executed. The machine before the transformation is outlined in figure 20.

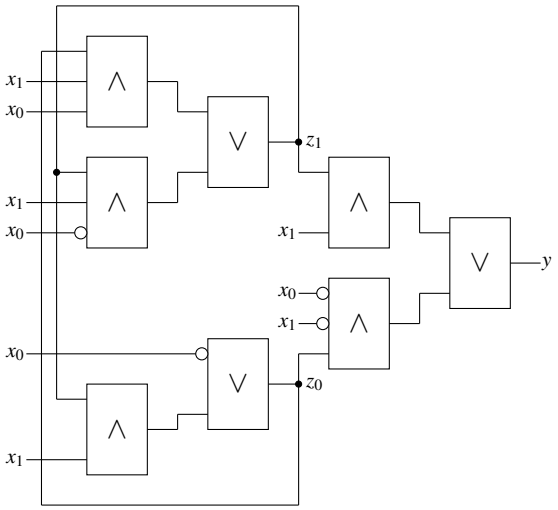


Figure 20: Automaton before the transformation

3.2.1 Design of $\underline{\delta}_{z_0}$

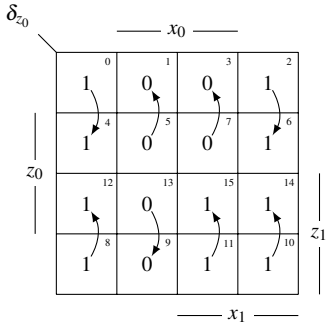


Figure 21: KV-diagram of δ_{z_0}

The KV-diagram of figure 21 will be lossless compacted given by the pointers in the diagram to the diagram of figure 22.

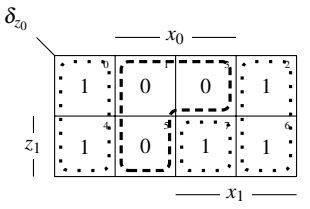


Figure 22: compacted KV-diagram of δ_{z_0}

With the compacted KV-diagram the RS-buffer for δ_{z_0} can be coded by specifying the functions $\underline{\delta}_{z_0}$ and $\overline{\underline{\delta}}_{\overline{z}_0}$, see

figure 23. The partial function $\underline{\delta}_{z_0}$ delivers the 1's of δ_{z_0} as 1's, while $\overline{\underline{\delta}}_{\bar{z}_{z_0}}$ delivers the 0's of δ_{z_0} as 1's. In this state transfer function, there is no state stable part only transient parts. Making the stable $\underline{\delta}_{z_0}$ a normal dual-rail structure.

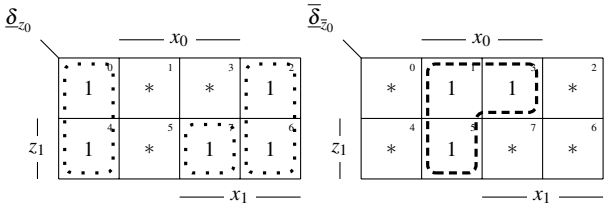


Figure 23: $\underline{\delta}_{z_0}$ and $\overline{\underline{\delta}}_{\bar{z}_0}$

State instable parts are not allowed in the state transfer function:

$$\bar{z}_0\delta_{z_0} \wedge_{z_0} \overline{\delta}_{\bar{z}_0} = (\bar{x}_0 \vee z_1x_1) \wedge (\bar{z}_1x_0 \vee \bar{x}_1x_0) = 0$$

Now the resulting $\underline{z}_0$ can be calculated:

$$\begin{aligned}
\underline{\delta}_{z_0}(z,x) &= {}_{z_0}\delta_{z_0} \wedge_{\bar{z}_0} \delta_{z_0} = \bar{x}_0 \vee z_1x_1 \\
\overline{\underline{\delta}}_{\bar{z}_0}(z,x) &= {}_{z_0}\overline{\delta}_{\bar{z}_0} \wedge_{\bar{z}_0} \overline{\delta}_{\bar{z}_0} = \bar{z}_1x_0 \vee \bar{x}_1x_0 \\
\underline{z}_0 &:= \underline{\delta}_{z_0} + \neg(\overline{\underline{\delta}}_{\bar{z}_0}) \\
\underline{z}_0 &:= \bar{x}_0 \vee z_1x_1 + \neg(\bar{x}_1x_0 \vee \bar{z}_1x_1)
\end{aligned}$$

3.2.2 Design of $\underline{\delta}_{z_1}$

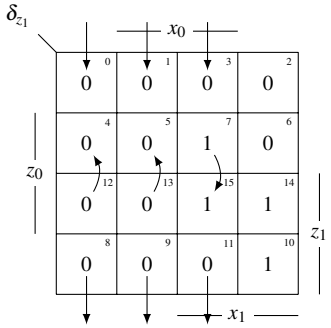


Figure 24: KV-diagram of δ_{z_1}

For $\underline{\delta}_{z_1}$ the same approach applies. First the KV-diagram, see figure 24, will be compacted:

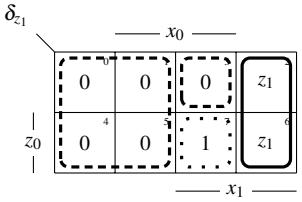


Figure 25: compacted KV-diagram of δ_{z_1}

Now the functions $\underline{\delta}_{z_1}$ and $\overline{\underline{\delta}}_{\bar{z}_1}$ can be deduced:

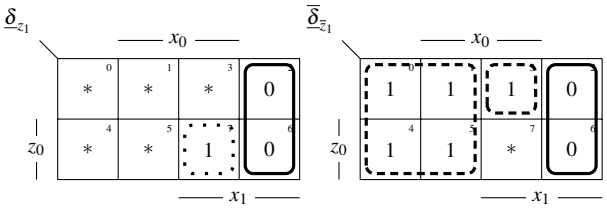


Figure 26: KV-diagrams for $\underline{\delta}_{z_1}$ and $\overline{\underline{\delta}}_{\bar{z}_1}$

First δ_{z_1} is checked for instable parts:

$$\bar{z}_1\delta_{z_1} \wedge_{z_1} \overline{\delta}_{\bar{z}_1} = z_0x_1x_0 \wedge (\bar{z}_0x_0 \vee \bar{x}_1) = 0$$

No instable parts, so the RS-buffer can be coded for δ_{z_1} by determining the functions $\underline{\delta}_{z_1}$ and $\overline{\underline{\delta}}_{\bar{z}_1}$:

$$\begin{aligned}
\underline{\delta}_{z_1}(z,x) &= {}_{z_1}\delta_{z_1} \wedge_{\bar{z}_1} \delta_{z_1} = z_0x_1x_0 \\
\overline{\underline{\delta}}_{\bar{z}_1}(z,x) &= {}_{z_1}\overline{\delta}_{\bar{z}_1} \wedge_{\bar{z}_1} \overline{\delta}_{\bar{z}_1} = \bar{x}_1 \vee \bar{z}_0x_0 \\
\underline{z}_1 &:= \underline{\delta}_{z_1} + \neg(\overline{\underline{\delta}}_{\bar{z}_1}) \\
\underline{z}_1 &:= z_0x_1x_0 + \neg(\bar{x}_1 \vee \bar{z}_0x_0)
\end{aligned}$$

3.2.3 Design of λ

The KV-diagram of λ can be seen in 27. Because of the two-dimensional state, λ has to be divided into eight parts. They exist as follows:

(z_1, z_0, y)	$(z_1, z_0, \bar{y})$
$(001) = A$	$(001) = \overline{E}$
$(011) = B$	$(011) = \overline{F}$
$(101) = C$	$(101) = \overline{G}$
$(111) = D$	$(111) = \overline{H}$

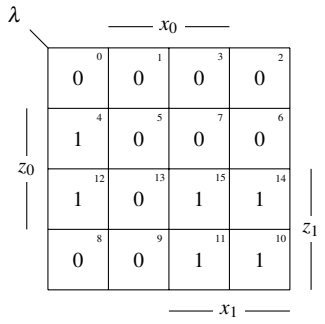


Figure 27: KV-diagram of λ

The parts of λ are set up in the KV-diagram, see figure 28.

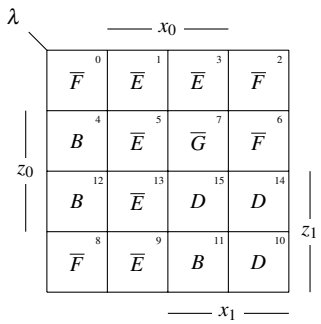


Figure 28: Parts of λ

The output functions y and $\bar{y}$ can be inferred from this:

$$
\begin{aligned}
y &= A \vee B \vee C \vee D \\
&= z_0\bar{x}_1\bar{x}_0 \vee z_1\bar{z}_0x_1x_0 \vee z_1x_1\bar{x}_0 \vee z_1z_0x_1 \\
&= z_1x_1 \vee z_0\bar{x}_1\bar{x}_0 \\
\bar{y} &= \overline{E} \vee \overline{F} \vee \overline{G} \vee \overline{H} \\
&= \bar{x}_1x_0 \vee \bar{z}_1\bar{z}_0x_0 \vee \bar{z}_0\bar{x}_0\bar{x}_1 \vee \bar{z}_1x_1\bar{x}_0 \vee \bar{z}_1\bar{z}_0\bar{x}_0 \vee \bar{z}_1z_0x_1x_0 \\
&= \bar{x}_1x_0 \vee \bar{z}_1x_1 \vee \bar{z}_0\bar{x}_1
\end{aligned}
$$

The resulting automaton can be seen in 30. As you can see in table 3, situation 7 leads to a cycle, by stepping to configuration 11, which steps back to configuration 7, and so on. So to have a state stable design in multistate machines, it is also necessary to regard the transitions between the states. A rule has to be found, to filter these race conditions. In the following transformation the auomaton of subsection b is stabilized, by changing δ_{z_1}.

3.3 Stable 2-dimensional example

The state transfer function δ_{z_1} has now been changed, to make it a stable automaton. For $x = (x_1, x_0)$, $y = (y)$ and $z = (z_1, z_0)$ the transformation of the Mealy machine $(X, Y, Z, \delta, \lambda)$ with the state transfer functions and output function

$$\delta_{z_0}(z,x) = \bar{x}_0 \vee z_1x_1 \qquad \delta_{z_1}(z,x) = z_0x_1x_0 \vee z_1x_1$$

$$\lambda(z,x) = z_1x_1 \vee z_0\bar{x}_1\bar{x}_0$$

will be executed. The machine before the transformation is outlined in figure 29.

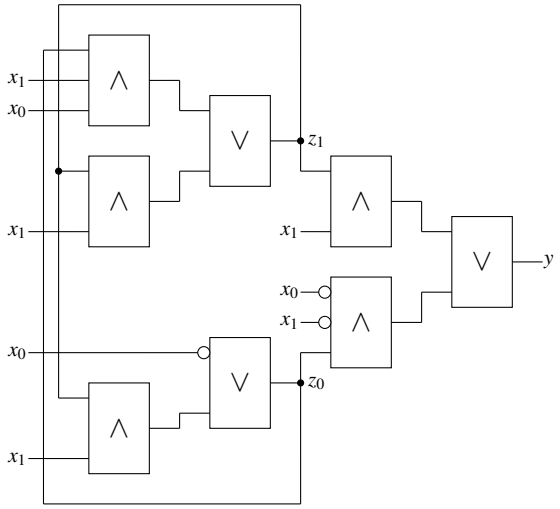


Figure 29: Automaton before the transformation

3.3.1 Design of $\underline{\delta}_{z_0}$

The KV-diagram of figure 31 will be lossless compacted given by the pointers in the diagram to the diagram of figure 32.

#	z_1	z_0	x_1	x_0	δ_{z_1}	δ_{z_0}	$\underline{\delta}_{z_1}$	$\overline{\underline{\delta}}_{\bar{z}_1}$	$\underline{\delta}_{z_1}$	$\underline{\delta}_{z_0}$	$\overline{\underline{\delta}}_{\bar{z}_0}$	$\underline{\delta}_{z_0}$	λ	A	B	C	D	$\overline{E}$	$\overline{F}$	$\overline{G}$	$\overline{H}$	y	$\bar{y}$	y
0	0	0	0	0	0	1	*	1	0	1	*	1	0	*	*	*	*	*	1	*	*	*	1	0
1	0	0	0	1	0	0	*	1	0	*	1	0	0	*	*	*	*	1	*	*	*	*	1	0
2	0	0	1	0	0	1	*	*	0	1	*	1	0	*	*	*	*	*	1	*	*	*	1	0
3	0	0	1	1	0	0	*	1	0	*	1	0	0	*	*	*	*	1	*	*	*	*	1	0
4	0	1	0	0	0	1	*	1	0	1	*	1	1	*	1	*	*	*	*	*	*	1	*	1
5	0	1	0	1	0	0	*	1	0	*	1	0	0	*	*	*	*	1	*	*	*	*	1	0
6	0	1	1	0	0	1	*	*	0	1	*	1	0	*	*	*	*	*	1	*	*	*	1	0
7	0	1	1	1	1	0	1	*	1	*	1	0	0	*	*	*	*	*	*	1	*	*	1	0
8	1	0	0	0	0	1	*	1	0	1	*	1	0	*	*	*	*	*	1	*	*	*	1	0
9	1	0	0	1	0	0	*	1	0	*	1	0	0	*	*	*	*	1	*	*	*	*	1	0
10	1	0	1	0	1	1	*	*	1	1	*	1	1	*	*	*	1	*	*	*	*	1	*	1
11	1	0	1	1	0	1	*	1	0	1	*	1	1	*	1	*	*	*	*	*	*	1	*	1
12	1	1	0	0	0	1	*	1	0	1	*	1	1	*	1	*	*	*	*	*	*	1	*	1
13	1	1	0	1	0	0	*	1	0	*	1	0	0	*	*	*	*	1	*	*	*	*	1	0
14	1	1	1	0	1	1	*	*	1	1	*	1	1	*	*	*	1	*	*	*	*	1	*	1
15	1	1	1	1	1	1	1	*	1	1	*	1	1	*	*	*	1	*	*	*	*	1	*	1

Table 3: Truth table of the transformed automaton

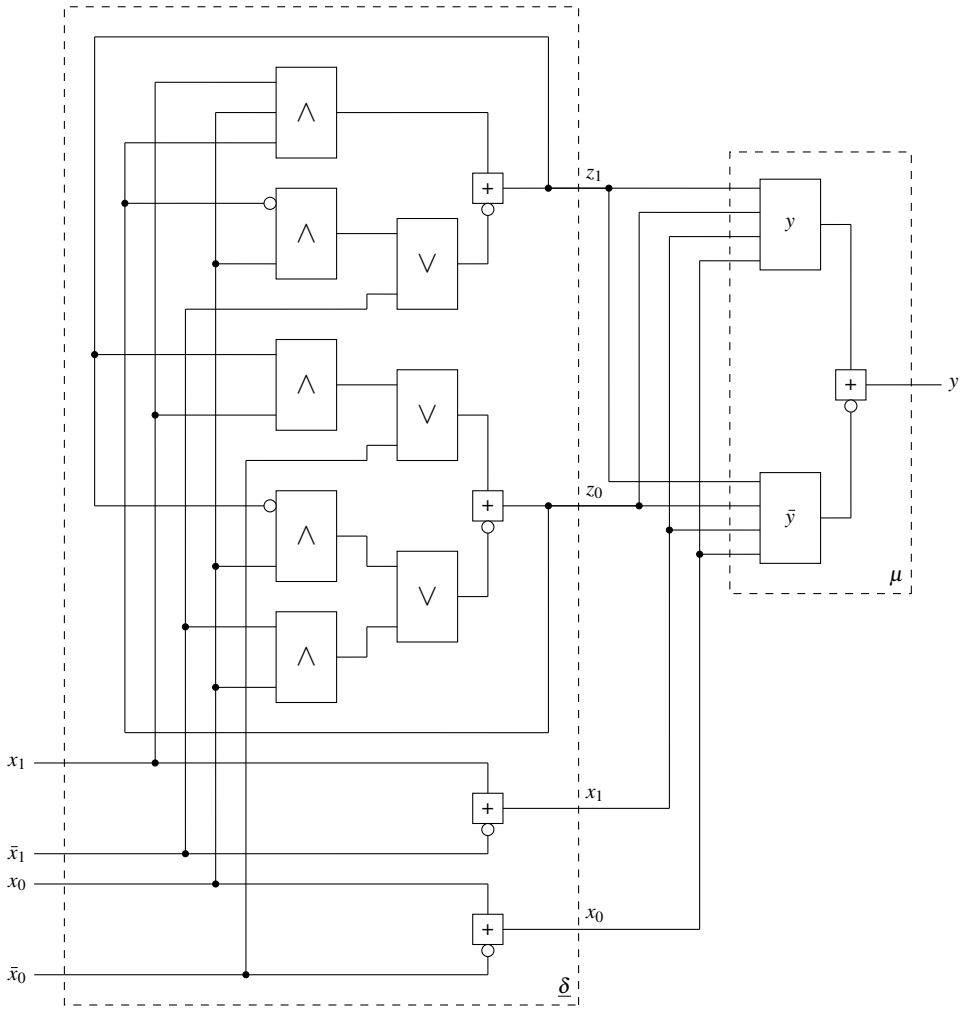


Figure 30: Resulting automaton

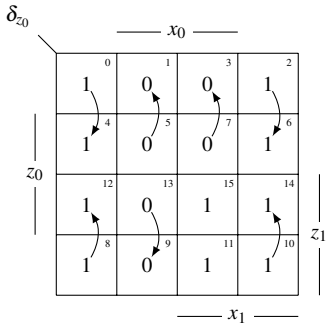


Figure 31: KV-diagram of δ_{z_0}

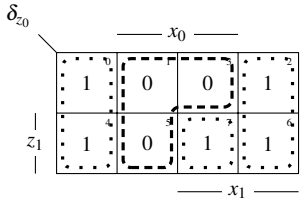


Figure 32: compacted KV-diagram of δ_{z_0}

With the compacted KV-diagram the RS-buffer for δ_{z_0} can be coded by specifying the functions $\underline{\delta}_{z_0}$ and $\overline{\underline{\delta}}_{\bar{z}_0}$, see figure 33.

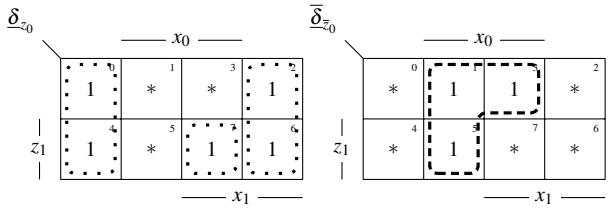


Figure 33: $\underline{\delta}_{z_0}$ and $\overline{\underline{\delta}}_{\bar{z}_0}$

State instable parts are not allowed in the state transfer function:

$$_{\bar{z}_0}\delta_{z_0} \wedge {}_{z_0}\overline{\delta}_{\bar{z}_0} = (\bar{x}_0 \vee z_1 x_1) \wedge (\bar{z}_1 x_0 \vee \bar{x}_1 x_0) = 0$$

Now the resulting $\underline{z}_0$ can be calculated:

$$\underline{\delta}_{z_0}(z,x) = {}_{z_0}\delta_{z_0} \wedge {}_{\bar{z}_0}\delta_{z_0} = \bar{x}_0 \vee z_1 x_1$$
$$\overline{\underline{\delta}}_{\bar{z}_0}(z,x) = {}_{z_0}\overline{\delta}_{\bar{z}_0} \wedge {}_{\bar{z}_0}\overline{\delta}_{\bar{z}_0} = \bar{z}_1 x_0 \vee \bar{x}_1 x_0$$
$$\underline{z}_0 := \underline{\delta}_{z_0} + \neg(\overline{\underline{\delta}}_{\bar{z}_0})$$
$$\underline{z}_0 := \bar{x}_0 \vee z_1 x_1 + \neg(\bar{x}_1 x_0 \vee \bar{z}_1 x_1)$$

3.3.2 Design of $\underline{\delta}_{z_1}$

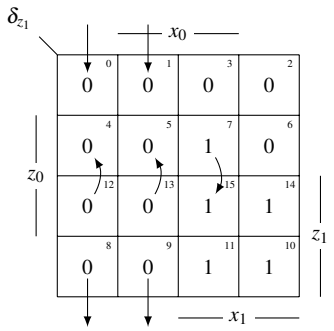


Figure 34: KV-diagram of δ_{z_1}

For $\underline{\delta}_{z_1}$ the same approach applies. First the KV-diagram, see figure 34, will be compacted:

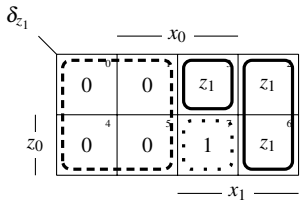


Figure 35: compacted KV-diagram of δ_{z_1}

Now the functions $\underline{\delta}_{z_1}$ and $\overline{\underline{\delta}}_{\bar{z}_1}$ can be deduced:

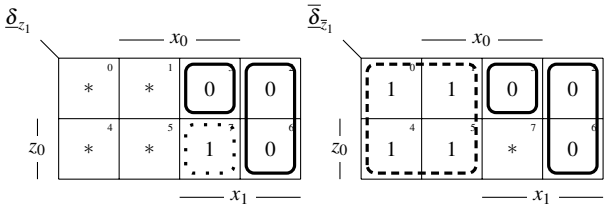


Figure 36: KV-diagrams for $\underline{\delta}_{z_1}$ and $\overline{\underline{\delta}}_{\bar{z}_1}$

First δ_{z_1} is checked for instable parts:

$$_{\bar{z}_1}\delta_{z_1} \wedge _{z_1}\overline{\delta}_{\bar{z}_1} = z_0x_1x_0 \wedge \bar{x}_1 = 0$$

No instable parts, so the RS-buffer can be coded for δ_{z_1} by determining the functions $\underline{\delta}_{z_1}$ and $\overline{\underline{\delta}}_{z_1}$:

$$\underline{\delta}_{z_1}(z,x) = _{z_1}\delta_{z_1} \wedge _{\bar{z}_1}\delta_{z_1} = z_0x_1x_0$$
$$\overline{\underline{\delta}}_{\bar{z}_1}(z,x) = _{z_1}\overline{\delta}_{\bar{z}_1} \wedge _{\bar{z}_1}\overline{\delta}_{\bar{z}_1} = \bar{x}_1$$
$$\underline{z}_1 := \underline{\delta}_{z_1} + \neg(\overline{\underline{\delta}}_{\bar{z}_1})$$
$$\underline{z}_1 := z_0x_1x_0 + \neg(\bar{x}_1)$$

3.3.3 Design of λ

The KV-diagram of λ can be seen in 37. Because of the two-dimensional state, λ has to be divided into eight parts. They exist as follows:

(z_1,z_0,y)	$(z_1,z_0,\bar{y})$
$(001) = A$	$(001) = \overline{E}$
$(011) = B$	$(011) = \overline{F}$
$(101) = C$	$(101) = \overline{G}$
$(111) = D$	$(111) = \overline{H}$

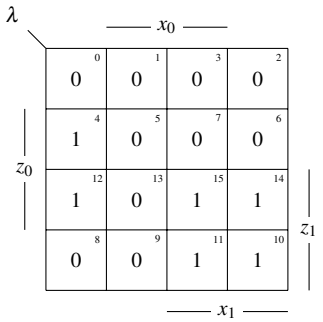


Figure 37: KV-diagram of λ

The parts of λ are set up in the KV-diagram, see figure 38.

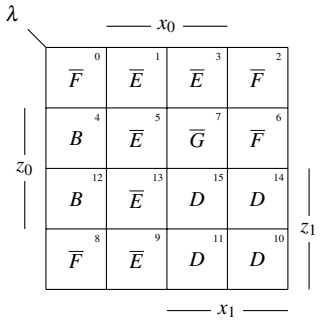


Figure 38: Parts of λ

The output functions y and $\bar{y}$ can be inferred from this:

$$\begin{aligned} y &= A \vee B \vee C \vee D \\ &= z_0\bar{x}_1\bar{x}_0 \vee z_1x_1 \\ \bar{y} &= \overline{E} \vee \overline{F} \vee \overline{G} \vee \overline{H} \\ &= \bar{x}_1x_0 \vee \bar{z}_1\bar{z}_0x_0 \vee \bar{z}_0\bar{x}_1\bar{x}_0 \vee \bar{z}_1x_1\bar{x}_0 \vee \bar{z}_1z_0x_1x_0 \\ &= \bar{x}_1x_0 \vee \bar{z}_1x_1 \vee \bar{z}_0\bar{x}_1 \end{aligned}$$

4 Test of Theory

To confirm the theory presented in the article, the state transfer function of the one dimensional example was analyzed making it a Medvedev. To simulate the circuit it was modelled with NAND gates and simulated in LT-Spice. The modeled NAND structure then was realized on the circuit board and the output was captured via an oszilloscope. The Medvedev machine $(X,Y,Z,\delta,1)$ with $x = (x_1,x_0)$, $y = (y)$ and $z = (z)$ with the state transformation function and output function be given:

The table and circuit:

#	z_1	z_0	x_1	x_0	δ_{z_1}	δ_{z_0}	$\underline{\delta}_{z_1}$	$\overline{\underline{\delta}}_{\bar{z}_1}$	$\underline{\delta}_{z_1}$	$\underline{\delta}_{z_0}$	$\overline{\underline{\delta}}_{\bar{z}_0}$	$\underline{\delta}_{z_0}$	λ	A	B	C	D	$\overline{E}$	$\overline{F}$	$\overline{G}$	$\overline{H}$	y	$\bar{y}$	y
0	0	0	0	0	0	1	*	1	0	1	*	1	0	*	*	*	*	*	1	*	*	*	1	0
1	0	0	0	1	0	0	*	1	0	*	1	0	0	*	*	*	*	1	*	*	*	*	1	0
2	0	0	1	0	0	1	*	*	0	1	*	1	0	*	*	*	*	*	1	*	*	*	1	0
3	0	0	1	1	0	0	*	*	0	*	1	0	0	*	*	*	*	1	*	*	*	*	1	0
4	0	1	0	0	0	1	*	1	0	1	*	1	1	*	1	*	*	*	*	*	*	1	*	1
5	0	1	0	1	0	0	*	1	0	*	1	0	0	*	*	*	*	1	*	*	*	*	1	0
6	0	1	1	0	0	1	*	*	0	1	*	1	0	*	*	*	*	*	1	*	*	*	1	0
7	0	1	1	1	1	0	1	*	1	*	1	0	0	*	*	*	*	*	*	1	*	*	1	0
8	1	0	0	0	0	1	*	1	0	1	*	1	0	*	*	*	*	*	1	*	*	*	1	0
9	1	0	0	1	0	0	*	1	0	*	1	0	0	*	*	*	*	1	*	*	*	*	1	0
10	1	0	1	0	1	1	*	*	1	1	*	1	1	*	*	*	1	*	*	*	*	1	*	1
11	1	0	1	1	1	1	*	*	1	1	*	1	1	*	*	*	1	*	*	*	*	1	*	1
12	1	1	0	0	0	1	*	1	0	1	*	1	1	*	1	*	*	*	*	*	*	1	*	1
13	1	1	0	1	0	0	*	1	0	*	1	0	0	*	*	*	*	1	*	*	*	*	1	0
14	1	1	1	0	1	1	*	*	1	1	*	1	1	*	*	*	1	*	*	*	*	1	*	1
15	1	1	1	1	1	1	1	*	1	1	*	1	1	*	*	*	1	*	*	*	*	1	*	1

Table 4: Truth table of the transformed automaton

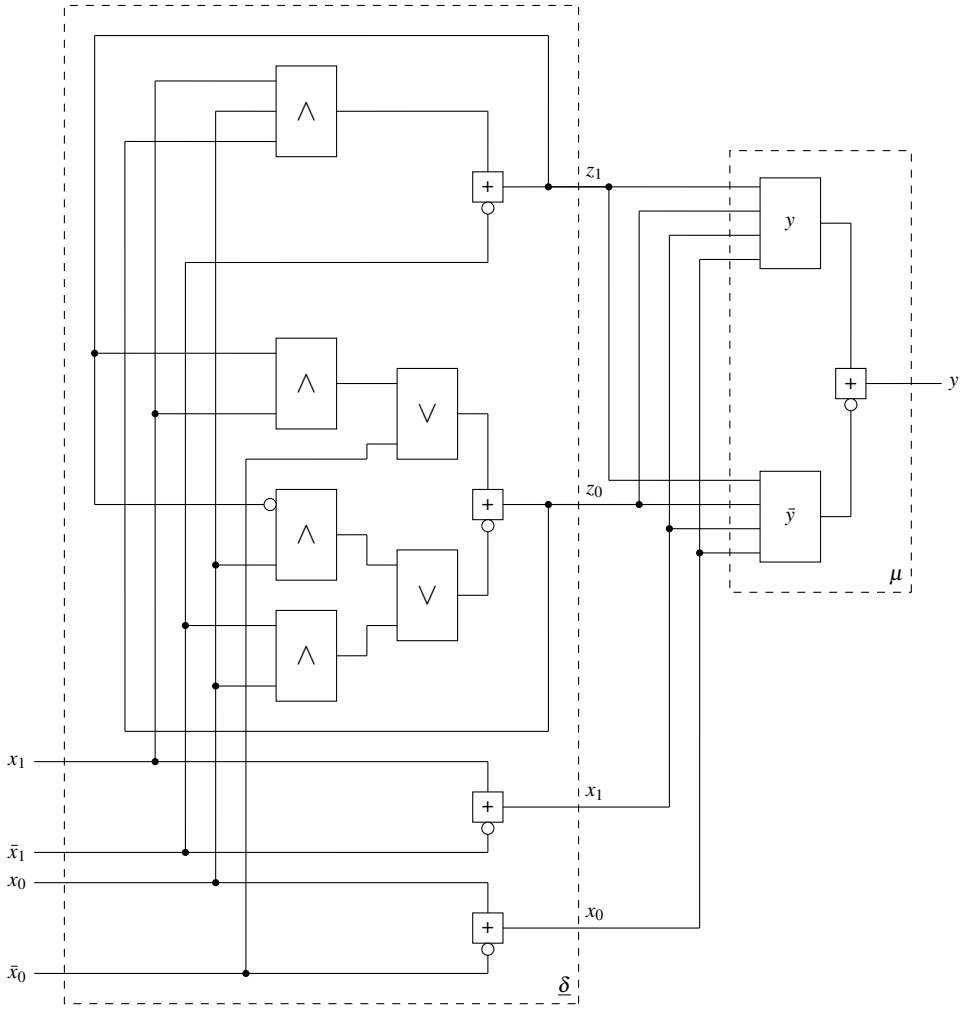


Figure 39: Resulting automaton

$$\delta(z,x) = zx_0\bar{x}_1 \vee x_1$$

The KV-Diagramm:

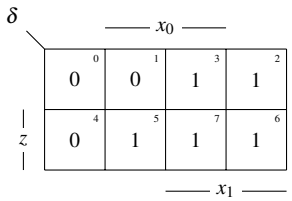


Figure 40: KV-diagram of δ

The table:

z	x_1	x_0	δ
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

Table 5: Truth table of the automaton

The graph of the automaton:

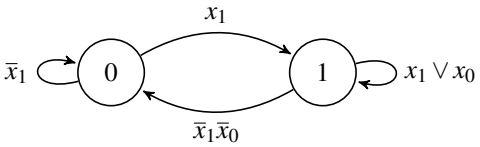


Figure 41: Graph of the automaton

NAND gates have been used to realize the structure. Therefore the DNF was transformed into a NAND strucutre by twice negating the DNF.

$$\delta(z,x) = \overline{\overline{zx_0\bar{x}_1 \vee x_1}} = \overline{\overline{zx_0\bar{x}_1} \wedge \bar{x}_1}$$

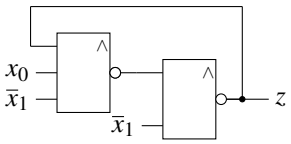


Figure 42: machine before the transformation

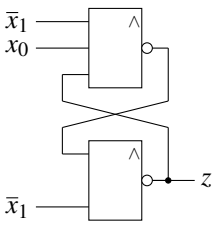


Figure 43: Different view of the automaton

Figure 43 shows two asynchronous feedbacks, which can lead to a race. On transition of x_1 from logical 1 to 0, while x_0 stays on logical 1, the feedback, which is first on logical 0, will make the other to a stable logical 1. Therefore on this transition, it can not be forecasted, whether the output will be on logical 1 or 0.

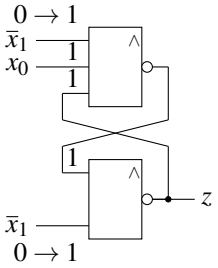


Figure 44: Race condition

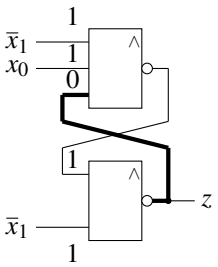


Figure 45: Race z = 0

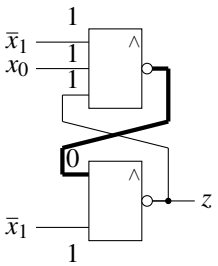


Figure 46: Race z =1

To simulate the race condition, the circuit was modelled and simulated in LT Spice. The simulation shows a race.

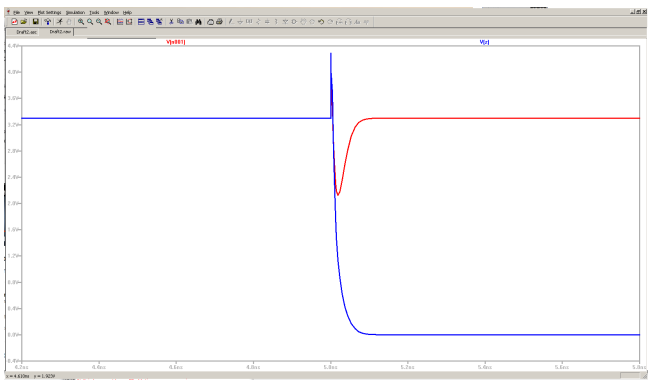


Figure 47: Simulated race situation

The circuit was realized in reality and the race was captured via an oscilloscope.

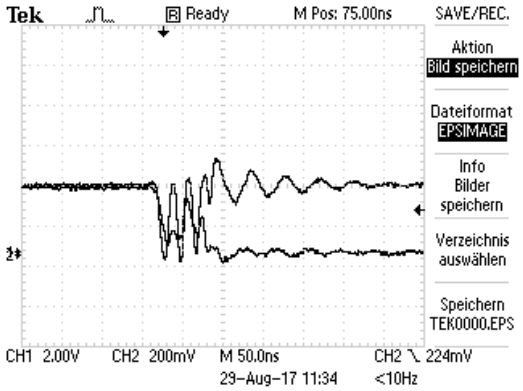


Figure 48: Race situation in reality

Then the transformation, which was presented in this journal, was made leading to the following structure, which does not have any contending feedbacks anymore. The only feedback this structure needs is nested in the RS-Buffer.

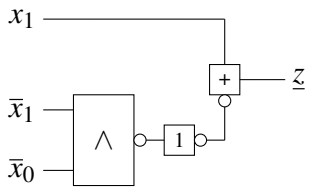


Figure 49: Transformed Moore

This structure does not show a race problem. To prove this statement the race condition of the old automaton has been set to the inputs. In figure 50 the state z and the rising edge of $\bar{x}_1$ is shown. It's clearly to see that the state doesn't change, not even a fluctuation of the signal can be recognized. Therefore the transformed machine is determined and stable.

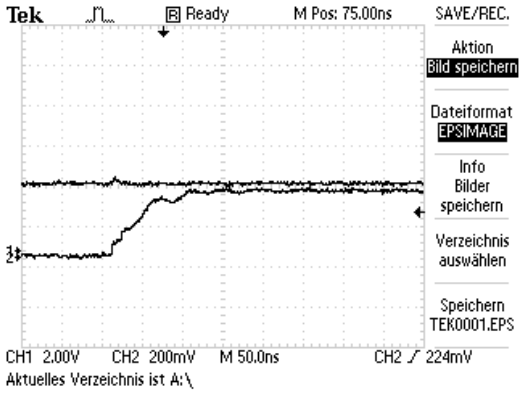


Figure 50: Race condition applied to the transformed automaton

5 Conclusion and Future Work

An asynchronous function stable Mealy was transformed into a state stable Moore under the use of the RS-buffer with suitable coding. The transformation provides general asynchronous design rules, e.q. the isolation of critical situations via the hold function. Due to the stabilisation by RS-buffer with the hold function, no races will be propagated through the circuit. In this journal three examples have been presented, a 1-dimensional and two 2-dimensional. Although the stable 2-dimensional transformation has produced an asynchronous feedback, this feedback leads either to a confirmation or a hold of the RS-buffer. The feedback has to be seen as a starting condition which gets triggered by the input and a change of the feedback will only affect the RS-buffer if the input changes. This design guideline can be used for all circuit designs, to create a stable, reliable system. The goal of the design is to remove the outer feedback and nest it in the RS-buffer. The inputs have been coded as pseudo states and stabilized via the RS-buffer. The output function has also been designed in dual-rail. This leads to a circuit, that looks at first sight as a combinational circuit. With this method you can simplify complex automata and combine them. The results have to be simulated to get a detailed look at the behavior of the transformed Moore. As it was shown, one dimensional machines get stabilized by this design, but for more dimensional machines it is necessary to design stable functions and regard the transitions between the states, not only concentrating on one state. So a rule has to be found, which can check for instabilities between states and stabilize machines in these constellations. Additionally the formula for the Transformation can be programmed, to create an automated process of transformation. Additionally the output function λ has to be checked for functional Hazards and addapted to guarantee that they are no longer existing.

References

[1] Birtwhistle, G.: Davis, A.: Asynchronous Digital Circuit Design, Springer London, 1995

[2] Van Berkel, K.: Burgess, R.: Kessels, J.L.W.: Peeters, A.: Roncken, M.: Schalij, F.: A fully asynchronous low-power error corrector for the DCC player, IEEE J. Solid-State Circuits, vol. 29, Dec. 1994. - P. 1429-1439

[3] Uygur, G.: Sattler, S.M.: A Real-World Model of Partially Defined Logic. 12th International Workshop on Boolean Problems 2016, Freiberg

[4] Mokhov, A.: Khomenko, V.: Sokolov, D.: Yakovlev, A.: On Dual-Rail Control Logic for Enhanced Circuit Robustness, ACSD '12 Proceedings of the 2012, 12th International Conference on Application of Concurrency to System Design, Pages 112-121, June 27 - 29, 2012, IEEE Computer Society Washington, DC, USA

[5] Shirvani, P.: Mitra, S.: Ebergen, J.: Roncken,M.: DUDES: a fault abstraction and collapsing framework for asynchronous circuits, in Advanced Research in Asynchronous Circuits and Systems (ASYNC), Proceedings, Sixth International Symposium on, 2000, pp. 73–82

[6] Özgül, M.: Deeg, F.: Sattler, S.M.: Mealy-to-Moore Transformation, 20th IEEE International Symposium on DDECS 2017, Dresden

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 164-173 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Improving Patient Outcomes Through Untethered Patient-Centered Health Records

Mohammed Abdulkareem Alyami[*,1], Majed Almotairi[1], Alberto R. Yataco[2], Yeong-Tae Song[1]

[1]*Dept. of Computer & Information Sciences, Towson University, 21252, USA*

[2]*Medical Director Get Well Immediate Care Towson, 21204, USA*

A R T I C L E I N F O

Article history:
Received: 16 November, 2017
Accepted: 11 March, 2018
Online: 27 March, 2018

Keywords:
Personal Health Records (PHRs)
Personal Health Records Systems (PHRS)
Dublin Core Metadata
Cloud storage
Standard Medical Codes

A B S T R A C T

Patient generated data, or personal clinical data, is considered an important aspect in improving patient outcomes. However, personal clinical data is difficult to collect and manage due to its distributed nature. For example, they can be located in multiple places such as doctors' offices, radiology centers, hospitals, or some clinics. Another factor that can make personal clinical data difficult to manage is that it can be heterogeneous data types such as text, images, charts, or paper-based documents. In case of emergencies, this situation makes personal clinical data retrieval very difficult. In addition, since the amount and types of personal clinical data continue to grow, finding relevant clinical data when needed is getting more difficult if no action is taken. In response to such scenarios, we propose an untethered patient health record system that manages personal health data by utilizing meta-data that enables easy retrieval of clinical data. We incorporate cloud-based storage for easy access and sharing with caregivers to implement continuity of care and evidence-based treatment. In emergency cases, we make critical medical information such as current medications and allergies available to relevant caregivers with valid license numbers only. Clinical data needs to be stored or made accessible from one place for easy sharing and retrieval. Well-managed personal cloud space could outlive the lifetime of personal health records system (PHRS) since the discontinuity of the service does not affect the data stored in the cloud space. In our approach, we separate the clinical data from applications in order to make the data independent from the application. Also, the users can have alternative applications for their clinical data. Such independence motivates users to use PHRS with flexibility.

1. Introduction

For most people, healthcare is considered important as there has been significant increase in chronic diseases such as heart disease, cancer, diabetes and asthma. This requires continuous treatment, reduces quality of life, and increases overall medical expenses (The Growing Crisis of Chronic Disease in the United States) [1]. According to the Centers for Disease Control and Prevention (CDC), in the U.S. about 610,000 people die of heart disease every year. In addition, 26 million people suffer from Type I or Type II Diabetes, around 14 million have severe chronic respiratory problems such as Chronic Obstructive Pulmonary Disease (COPD), and 68 million have been diagnosed with hypertension [2]. However, many of these diseases can be prevented and managed through early detection, physical activities, a balanced diet and treatment therapy. Adopting PHRS could help improve patient outcomes. Recently, there has been more focus on preventive care and proactive measures such as monitoring and controlling patients' symptoms. Nowadays, there are many mobile health applications and sensors such as blood pressure sensors, electrocardiogram sensors, blood glucose measuring devices, and others that are used for monitoring and controlling personal health. These apps and sensors produce personal health data that can be used for treatment purposes. If managed and handled properly, it can be considered patient-generated data. There are other types of personal health data that are available from various sources such as hospitals, doctor's offices, clinics, radiology centers or any other caregivers. Aforementioned health documents are deemed as a personal health record (PHR). According to American Health Information Management Association (AHIMA) [3], PHR can be defined as an electronic, lifelong resource of health information

[*]Mohammed Abdulkareem Alyami, Email: aljakman2006@hotmail.com

needed by individuals to make health decisions. However, it is not easy to collect all the relevant personal health data because of the fact that they are in different data types, available from different sources, and stored in different media and devices. To overcome such difficulties, it is desirable to have personal health data in one place where users have full control over their own clinical data. In order to be useful, the clinical data should be sharable when needed for diagnosis and treatment. Without proper clinical information (medical history, allergies, current medications, and adverse reactions) medical mistakes could occur when making medical decisions due to insufficient information. Even if a patient has a complete medical history and all the necessary clinical data, if it is not shared properly among caregivers at the time of need, discontinuity in care may occur. In order to meet the needs of such scenario, PHRS should have the following properties: robust and private storage, easy retrieval and maintenance, secure, sharable, and able to handle emergency situations.

There are two types of PHRS: untethered and tethered. Untethered PHR is an independent PHRS where patients have full control over their own personal health records. They can collect, manage, and share their health records. On the other hand, tethered PHRS are linked to a specific healthcare providers' EHR system, where the users typically gain easy access to their own records through secure portals and see their own clinical information such as test results, immunization records, family history, and other relevant information. They can also utilize secure messaging with their collaborating clinicians. The participating patients need to share the cost and the information with their care provider. However, these tethered records may not be complete since the information sources are from one provider only. Despite all the benefits PHRS provide, the adoption rate of PHR by the general public still remains low in the U.S. [4-6]. In our previous work [7], we identified six barriers (usability, ownership, interoperability, privacy and security, portability and motivation) that cause the slow adoption of PHRS. One of the main concerns for not having PHRS is the ownership issue. For example, one incident happened on January 1, 2012. Google Health™ System, one of the biggest PHRS providers, stopped its service and asked their registered customers to move their records to their computers or other PHRS vendors, which exacerbated ownership concerns by the public.

As an attempt to overcome some of the barriers, we propose an untethered PHRS that utilizes personal cloud storage, offers simplicity in organizing various kinds of clinical data by utilizing Dublin Core (DC) metadata, and provides easy access to emergency clinical data to paramedics or clinicians in case of emergency. DC metadata has been successfully applied in many areas, but since it is not specifically designed for clinical data, there are some limitations in its expressive power in the healthcare domain. In this research, we simplified the categorization of clinical data by human body part for easy retrieval of clinical data using DC, so users can manage their own clinical data without in-depth knowledge about clinical information. As a proof of our concept, we developed a system called My Clinical Record System (MCRS) to help users store, organize, retrieve, and share their clinical data with caregivers when needed, including in emergency situations. In an emergency situation, clinicians (e.g. physicians, paramedics, nurses, and others) can access patient's data using their license numbers and the patient's name and date of birth. Emergency information consists of current medication lists, known allergies, and side effects. By having complete medical history, MCRS users may be able to reduce medical errors and improve patients' outcomes. It also ensures continuity of care by sharing personal clinical data among healthcare providers when needed. The remainder of this paper is organized as follows: Section 2 discusses related work and in section 3 we focus on the clinical data's type and format. In Section 4, we discuss how to solve the data organization and retrieval issues using DC metadata to facilitate better and more accurate data retrieval. Section 5 discuses cloud-based storage. In section 6, we discuss the current situation in the healthcare industry (AS-IS). Section 7 uses a scenario as an example. Section 8 discusses the solution domain. In section 9, we introduce MCRS as a proof of concept and finally we conclude our study and discuss our future work in section 10.

2. Literature Review

In [8], Fearon defined Meta-data as structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. However, metadata standards have not been employed by many repositories and most of the meta-data was generally descriptive, rather than administrative or for preservation [9]. In [10], Greenberg investigated the ability of resource authors to create acceptable metadata in an organizational web site at the National Institute of Environmental Health Sciences in the U.S. The findings of their study indicate that resource authors can create better quality metadata when working with Dublin Core. In some circumstances, they may be able to create quality metadata better than a metadata professional. In [11], Talha developed their own metadata tool called Metadata Management System (MMS) to facilitate the creation, maintenance and storage of metadata. MMS supports two well-known metadata models, Dublin Core and SCORM 1.2 (IEEE Learning Object Metadata). The authors implemented their metadata tools in the Malaysia Grid for Learning (MyGfL). The results of the study indicate that MMS can substantially improve the discovery, retrieval management and control of web resources and learning objects in their MyGfL portal. In the healthcare domain, meta-data has been utilized as a method for confidentiality tags that indicate data sensitivity levels. This enables patients to give consent to the exchange of some parts of their health records (e.g. the medical diagnosis), while withholding consent for the exchange of other areas (e.g. a mental health counseling session) [12]. This work can help in increasing patients' privacy and allow them with some freedom in exchanging their health records. However, their work focused on EHR, but did not incorporate PHRS, which limited the scope of their study. Other researchers have adopted the ontology approach to quickly search and access relevant and meaningful information among large numbers of CDA documents within healthcare providers' systems (Electronic Health System), which in turn enables semantic interoperability[5, 13]. However, these studies were limited to one health data type (CDA), whereas health records include other data types such as images (e.g. x-rays, scanned documents, ultrasounds, and others), and observed symptoms noted by patients and clinical sensors' data. In [14] Patel developed a system called TrialX, on top of PHR, where patients not only can search by keywords, as in ClinicalTrials.gov, but also by demographics (e.g. age, gender, city and study site). This system enables patients to match their health condition to clinical trials. This system can accelerate and improve

the search results among health records. However, the authors did not take into account how to retrieve relevant clinical data since the amount and types of personal clinical data continue to grow, which makes it difficult to find such data. In [15], Appelboom reviewed the literature on smart wearable body sensors and found that these sensors are accurate and have clinical utility, but still are underutilized in the healthcare industry. These devices can be used to monitor physiological, cardiovascular and many other factors of health variables and transmit data either to a personal device or to an online storage site. The smart wearable body sensors are placed on different parts of the user's body based on the purpose of the sensor device. For instance, the physical therapy sensor is placed on the ankle; the cardiopulmonary sensor can be placed on the wrists, fingers, arms or thighs.

In [16], Zhang developed an application to apply meta-data efficiently on clinical trial data. The authors chose Microsoft Excel due to the wealth of built-in features (e.g. spell checking, sorting, filtering, finding, replacing, importing and exporting data capabilities), which contribute to the ease of use, power, and flexibility of the overall meta-data application. They focused on the analysis process in a drug development environment such as adverse clinical events (ACE), Electrocardiogram (ECG), laboratories (LAB), and vital signs (VITAL), where the raw data is stored in the clinical trial database and then the data can be manipulated.

Another study in [17], Teitz developed a website called HealthCyberMap with the goal of mapping Internet health information resources in novel ways for enhanced retrieval and navigation. They used Protégé-2000 to model and populate a qualified DC RDF meta-data base. They also extended the DC elements by adding quality and location elements. Also, the W3C RDFPic project extends the DC schema by adding its own elements such as camera, film, lens and film development date for describing and retrieving digitized photos. In [18], Ekblaw built a system (RedRec) to enable patients to access their medical health records across health providers (e.g. pediatrician, university physician, dentist, employer health plan provider, specialists, and others). Their system applies novel, blockchain smart contracts to create a decentralized content-management system for healthcare data across providers. RedRec governs medical records access while providing the patient with the ability to share, review, and post new records via flexible interface. The raw medical record content is kept securely in providers' existing data storage. However, when the patient wants to retrieve data from their provider's database, their Database Gatekeeper checks authentication. If it is approved, the Gatekeeper retrieves the relevant data for the requester and allows a sync with the local database.

In order to solve interoperability problems in exchanging clinical data, in [19], Dogac proposed archetypes to overcome the interoperability problems. They provided guidelines on how ebXML Registries can be used as an efficient medium for annotating, storing, discovering and retrieving archetypes. They also used ebXMLWeb services to retrieve data from clinical information systems. An archetype is defined as "a reusable, formal expression of a distinct, domain-level concept such as "blood pressure", "physical ex-amination", or "laboratory results", expressed in the form of constraints on data whose instances conform to some reference model"[19]. However, these studies are not comprehensive. Their limitations come from: focus only on one health data type, do not take into account PHR, lack of retrieving relevant clinical data, and overlook emergency access to medical records. Our proposed approach overcomes these limitations. Our study is comprehensive, which covers many different clinical and nonclinical documents such as images (e.g. x-rays, scanned documents, ultrasounds, and others), text (e.g. CDA, CCR, CCD, and others), and observed symptoms noted by patients and their clinical sensors' data. This can organize these various data types in a way that can help in storing and retrieving such data in an efficient way.

3. Clinical Data

In this section, we describe the types, formats, and sources of clinical data.

3.1. Measurements data from portable medical devices, sensors, or mobile application

One way to collect measurement data is through explicit clinical sensors. A clinical sensor is a device that responds to a physical stimulus and transmits a resulting impulse for interpretation or recording. Some sensors are designed to work outside the human body, while others can be implanted within the human body [20]. In this research, we are referring to clinical sensors for homecare settings, such as blood oxygen monitors, thermometers for body temperature, heart rate sensors, blood pressure sensors, and others. In addition to these textual data types, there can be non-textual data generated from sensors such as electrocardiogram measurement devices. The clinical sensors play a major role in healthcare, including early detection of diseases, diagnosis, disease monitoring and treatment monitoring.

Another method to collect measurement data is through mobile apps. For instance, most smartphones (e.g. Android or IOS) offer health and fitness apps that help users monitor their daily activities and health (e.g. track diet and nutrition calories, track vital signs, track fitness progress, share health data with their doctor electronically, and others). The data collected from these applications can be sent as a message or an email attachment to whom the users want to share it with. For interoperability, the collected data needs to be in standard format, such as HL7 CDA or in standard code such as SNOMED-CT.

3.2. Observed Symptoms

Patients sometimes experience particular symptoms (e.g. chest pain, nausea, vomiting, shortness of breath and others). If the patient notices such symptoms, they should be recorded and shared with their physician for proper treatment. If these symptoms are not shared with their physician, due to incomplete information, misdiagnosis could occur. When patients are recording, the observed symptoms should be described in standardized code such as SNOMED-CT. This will allow semantic interoperability, since the same symptoms can be described in multiple ways. Without codified descriptions, there can be discrepancies about the perceptions regarding symptoms between patients, nurses, or physicians [21].

3.3. Images

Most of the medical imaging machines produce standard image format called Digital Imaging and Communications in Medicine (DICOM). DICOM is defined as the international standard for medical images and related information (ISO 12052). There are two types of clinical data images: images that are based on DICOM standard (e.g. x-rays, Computed Tomography (CT), Magnetic Resonance (MR), and ultrasound devices) and scanned documents. The DICOM –format combines images and meta-data that describes the medical imaging procedure. Accessing data in DICOM files becomes as easy as working with TIFF or JPEG images [22]. On the other hand, the scanned documents (e.g. PDF/ JPEG) are difficult to retrieve because the content is not searchable. For example, some physicians write notes on clinical forms while diagnosing their patients and then type them on the computer. They also sometimes scan the notes and upload them to the patient records. Either way is time consuming, difficult to retrieve in a timely manner, and consumes relatively large storage space. In addition, the patient may have more than one doctor or may have been treated by many healthcare providers, which in turn fragments his/her records. So when patients obtain their records, they mostly receive them either printed out or sent as an email attachment. This makes it difficult to retrieve scanned documents because its content cannot be retrieved by computers. To alleviate such issues, we have utilized meta-data to describe such medical documents so computerized retrieval and systematic organization are possible.

3.4. Clinical Document

Clinical documentation (CD) is a computerized record describing a medical treatment, medical trial or clinical test, which can be exchanged among healthcare providers [23]. EHR data may be collected from healthcare providers. There are four types of clinical document formats: Continuity of Care Record (CCR), Clinical Document Architecture (CDA), Continuity of Care Document (CCD), and Consolidated CDA (C-CDA). All of which allow healthcare providers to exchange clinical information summary about a patient. However, CCR was excluded from the 2014 edition of EHR Certification, which is a standard certification criteria for EHRs that was established by The Centers for Medicare & Medicaid Services (CMS) and the Office of the National Coordinator for Health Information Technology (ONC), as a valid way to send summary of care documents. Hence, the content from a CCR was merged into a CDA format and called Continuity of Care Document (CCD). C-CDA includes nine different types of commonly used CDA documents such as CCD, consultation notes, discharge summary, imaging integration, DICOM diagnostic imaging reports, history and physical, operative note, progress note, procedure note, and unstructured documents. Each C-CDA Document Template was designed to satisfy a specific information exchange situation.

In the following subsection, we will discuss the difference between these formats (CCR, CDA, and CCR):

- CCR documents provide a snapshot of treatment and basic patient information such as diagnosis and reason for referral between healthcare providers. It also uses only specified XML code and does not support narrative text (free-text), which hinders physicians from writing notes if needed, and it is not electronically acceptable by all systems.
- CDA documents are a flexible standard which can be read by the human eye or processed by machine due to the use of XML language. It is based on the HL7 Reference Information Model (RIM), and uses HL7 V3 data types. CDA can be transported using different methods such as HL7 V2, HL7 V3, DICOM, MIME-encoded attachments, HTTP, or FTP. CDA documents can include text, images and even multimedia.
- CCD documents are not a complete medical history of the patient but it includes only the information critical to effectively provide continuity of care. Its primary purpose is exchanging patient information between different healthcare providers. It is based on XML standard and can be displayed on a web browser using style sheet. It also allows narrative text which is an advantage over other standards like CCR [24].

When using untethered PHRS, patients are responsible for collecting clinical data from their healthcare providers or from their own patient-generated measurement data and keeping it in their own personal cloud space. For example, CDA, CCR and CCD can be obtained from healthcare providers, X-rays can be obtained from the radiology department, and lab test results can be obtained from the test lab or doctor's office. Patients can share their health records with their clinicians by either electronically transmitting or granting access to their storage through the PHRS. If electronic sharing is not allowed, the patient may download the file and make hardcopies or store them in a USB, CD, or other mediums for sharing [25].

3.5. Meta-data

There are two different methods of storing meta-data. In the first method, meta-data can be embedded in the data (e.g. in the header of a digital file). The advantages of this option are ensuring that the meta-data will not be lost, eliminating the need for linking data and meta-data, and updating the data and meta-data together. In the second method, meta-data can be stored separately. The advantage of this option is that it can simplify the management of meta-data and can expedite the retrieval of the data [8]. In our approach, we employed the latter method to accelerate the retrieval of clinical data and to enhance expressive power. However, in this method, there can be inconsistencies between meta-data and clinical data. This can occur when transitioning to a new platform, integration between different systems or sharing data across multiple systems [12]. There are many meta-data formats that have been accepted internationally including: Dublin Core (DC), Federal Geographic Data Committee (FGDC), Encoded Archival Description (EAD), and Government Information Locator Service (GILS) [26]. Some of the metadata standards/schemas are generic, while others are domain-specific. For example , SCORM 1.2 (IEEE LOM) is used for educational approaches and rights management; Friend of a Friend (FOAF) is used for people and organizations; Simple Knowledge Organization (SKOS) is used for concept collections; Asset Description Metadata Schema (ADMS) is used for describing semantic interoperability assets; and Dublin Core is used for published material (text, images).

Generic metadata formats, such as Dublin Core, tend to be easy to use and widely adopted. In this study, we adopted DC metadata because it was developed for author-generated metadata, supports resource sharing and interoperability among information systems, has the broadest level of commonality of elements, extensibility, international acceptability and the flexibility it provides for extensions to the basic elements.

DC metadata solves one of the major issues in using meta-data among healthcare systems, which lacks interoperability [7]. Therefore, among these different metadata formats, DC metadata is the most appropriate format that aligns with our approach.

Meta-data benefits personal health record management in many ways. These benefits include the following:

- Consistency in definitions: properly defined tags provide structured information about the clinical data stored.
- Clarification of the relationships: meta-data can be used to clarify the relationships among the clinical items by defining categories and associated relationships within the category. We have defined each tag in DC for clinical purposes. When the data is uploaded or modified, the corresponding meta-data is required to update as well.

3.6. Meta-data Management

Meta-data management ensures that the data is associated with the datasets and utilized efficiently throughout and across organizations. Data governance is needed for successful meta-data implementation so it can provide trustworthy, timely and relevant information to decision makers, as well as personal users. For successful implementation, data governance must be aligned with the intended purposes of the users or organizations. This means that a data governance program starts by specifying its strategy, goals and the scope of its success. An organization needs to define the three data governance pillars including policies, people (and people skills), and processes. Once the above steps are in place, the organization can determine the best tools and technology to implement its data governance initiative [27].

3.7. Dublin Core Meta-data for Clinical Use

The DC Meta-Data Initiative (DCMI), is an open organization supporting innovation in meta-data design and best practices across the meta-data ecology. The DC Meta-data consists of 15 optional elements including: title creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage and rights.

In this study, we defined the usage of DC Meta-data elements for clinical purposes to describe and retrieve clinical data efficiently as shown in Table 1. Some of the meta-data elements - title subject, description, type, data and resource - are required for the integrity of data. These elements must be present for every clinical data item. The optional fields can be skipped, but if it has been filled, the metadata quality will be increased.

Table 1: Meta-data schema for PHR

Entity	Description
DC. Title	The title of the document
DC. Creator	The author of the document
DC. Subject	Subject of the document
DC. Description	Description of the document
DC. Relation	- One of the body parts (Thorax, Abdomen, Heart, extremities, Integumentary, Head, Urinary or Reproductive) - This element is linked to the subject element
DC. Date	Date of meta data creation
DC. Type	Type will be used for lab tests (blood work, urinalysis, fecal sample, nasopharyngeal sample, oropharyngeal sample and others) or images (x-ray, cat scan/ CT, Ultrasound, Magnetic resonance/MR, Scanned Document, Electrocardiogram, EKG/ or ECG, CDA, C-CCD and others)
DC. Format	PDFs, Text, JPEGs , TIFFs, HL7 CDA, and others
DC. Identifier	Optional Document ID
DC. Language	English and other languages
DC. Coverage	Geographical and time-related information
DC. Rights	Copyright and access rights (secured or unsecured)
DC. Source	Data source

4. Healthcare Processes

4.1. AS-IS Healthcare Processes

Some of the issues that the current healthcare industry is having include discontinuity of care and unacceptably high rates of medical mistakes due to unavailable patient medical records at the time of need. Some of the factors that cause these issues are listed below and illustrated in the Figure 1. (the numbers listed below correspond to the numbers in the Figure 1)

- Personal clinical data is difficult to collect and manage because they are located over multiple places such as doctor's offices, radiology centers, hospitals, or some clinics (1).
- Heterogeneous data types such as text, images, charts, or paper- based documents (2) [8].
- Discontinuity of care due to lack of communication among caregivers. This is caused by distributed and fragmented medical information (3).
- Lack of evidence-based treatment due to limited access to medical records (4).
- Medical errors due to incomplete medical history or access to emergency health information (e.g. allergies, current medication list, side effects, and others) at the time of need (5). According to Sunyaev. [28], most PHRS do not offer built-in emergency access to the record, except through third-party services that are available for HealthVault. For example, Microsoft Health Vault provides users with

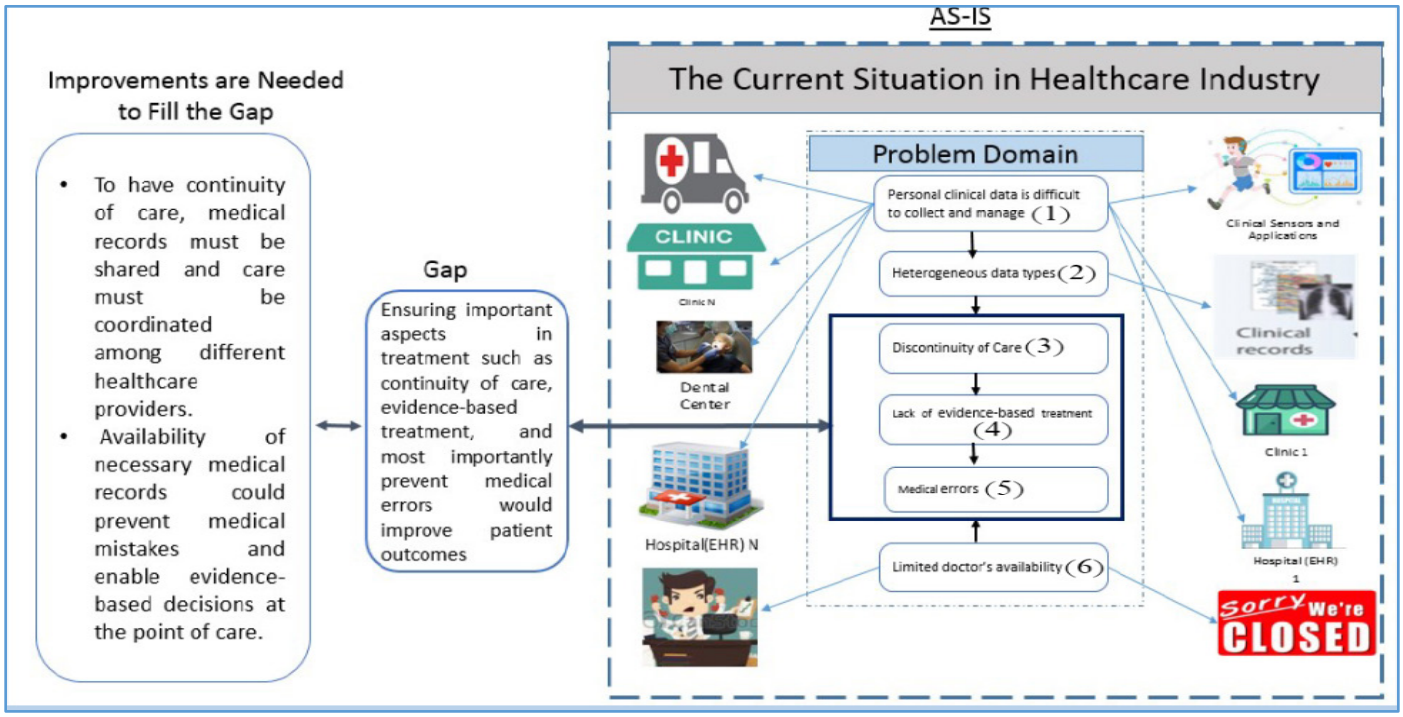


Figure 1: Problem Domain in the Current Healthcare Industry

access codes that can be given to emergency responders and other people they trust to allow access to the emergency information.

- Limited doctor availability. For instance, patients may need to be seen on the weekend or on a holiday when their doctor's office is closed (6).

4.2. Scenario for the Need of PHRS

John was suffering from tiredness and lack of energy for the past 4 weeks. He has several chronic conditions (type 2 diabetes, chronic kidney disease, chronic lower back pain, generalized anxiety disorder, depression, bipolar disorder, dyslipidemia, hypothyroidism, coronary artery disease and congestive heart failure) and is on multiple medications prescribed by his PCP, cardiologist, psychiatrist and pain management doctor. John had recently requested to become one of Dr. Smith's patients. After multiple attempts, Dr. Smith was able to obtain some of his previous medical records. Dr. Smith also was interested in reviewing his previous medical diagnoses and prior/current medications. Unfortunately, there was no integration of the records and medications taken. After several interviews with John and his wife, Dr. Smith was able to determine that the patient was using the same type of long-acting insulin twice a day because one had the generic name and the other the commercial or brand name. The patient was supposed to use this insulin once a day only. This patient's error kept his glucose at very low levels in blood which led to constant tiredness and lack of energy. Once the dose was corrected, the patient felt better. The immediate access to medical and prescription information would have allowed Dr. Smith to identify the error faster and provide him with the ability to take prompt corrective measures.

4.3. Solution Concept (TO-BE) Processes

To have continuity of care, medical records must be shared and care must be coordinated among different healthcare providers. Availability of necessary medical records could help prevent medical mistakes and enable evidence-based decisions at the point of care. It would be convenient to have clinical data stored in the same place for easy sharing and retrieval. Well-managed personal cloud space could outlive the lifetime of PHRS since clinical data is stored independently. In our approach, we separate the clinical data from applications to make the data independent from the application. Also, the users can have alternative applications to access their clinical data. Such independence helps clinical data outlive its applications. Our proposed solution concepts are illustrated in the Figure 2. In the Figure, the clinical data is separated from the application for data independence. The numbers in the Figure 2 correspond to the problems listed in the Figure 1.

5. My Clinical Record System (MCRS): A Proof of Concept

As mentioned in the introduction, there are a number of obstacles in collecting and maintaining personal health data. In an attempt to remove such obstacles, we developed a web-based system called My Clinical Record System (MCRS) that can help users to upload, organize, share, and retrieve relevant health data.

5.1. Overview of MCRS

Some of the main features of MCRS are listed below:

- Users can search their cloud storage not only by keyword, but also by any of the meta-data elements using document finder search features. They also can search by two elements such as subject and date in order to filter data by

showing more relevant data. Additionally, they can find a group of records based on a date range they specify.

- MCRS allows users to share their health records with their physicians using the share document feature as shown in the Figure 3.
- MCRS helps users to create meta-data for any documents (textual and non-textual data) and upload them to their cloud storage. For the non-textual data generated from sensors such as electrocardiogram (ECG) measurement devices, the users can scan them or take a picture and then upload them as an image or use a plotted number. Thus, the user will be able to include such data into their PHR and share it with their healthcare providers. MCRS also helps to retrieve those documents easily and can direct the users to its location if more information is needed.
- To overcome the ownership barrier, we separate the clinical data from applications, which will give the users more freedom by not limiting them to one provider or application. Also, their data is saved on their own storage, thus we do not have to store it in our system.
- To overcome the interoperability barrier, we used DC standards to describe any clinical data using our tag definition. The DC meta-data content for doctor visit summary document is shown in the Figure 4 and in the Figure 5.

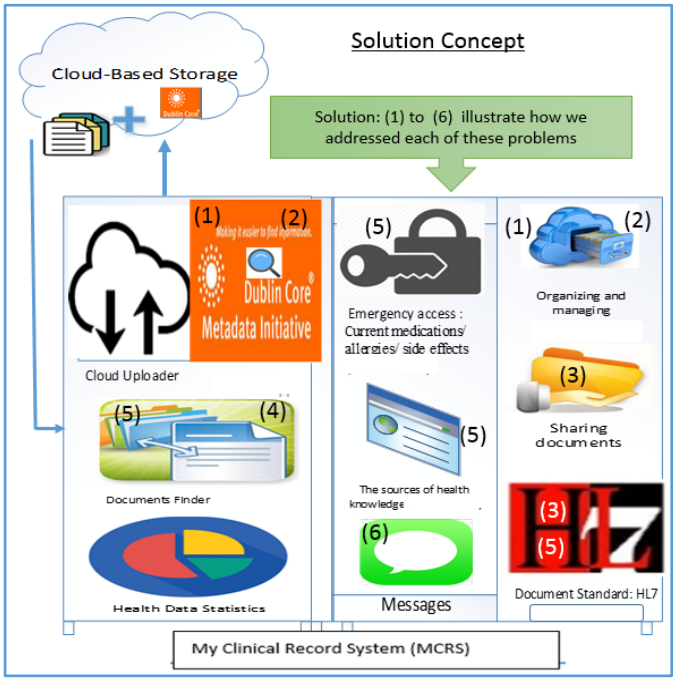


Figure 2: Solution Concept

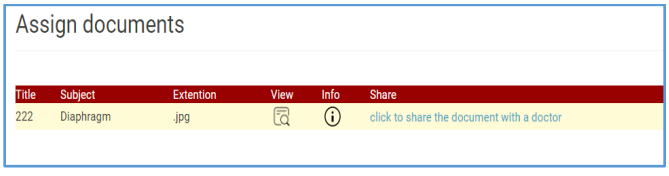


Figure 3: Sharing Health Records with Physician

```
<?xml version="1.0" encoding="utf-8"?>
<Metadata xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:mcr="http://MyClinicalRecords.net/MyClinicalRecords.aspx"
xmlns:dcterms="http://purl.org/dc/terms/">
  <Report Id="Mohammed1262017233939">
    <dc:Title>doctor visit summary</dc:Title>
    <dc:Creator>Dr. Yoataco</dc:Creator>
    <dc:Subject>nose</dc:Subject>
    <dc:Description>cough, runny nose</dc:Description>
    <dc:Publisher></dc:Publisher>
    <dc:Contributor></dc:Contributor>
    <mcr:Findings>Acute bronchitis with bronchospasm</mcr:Findings>
    <mcr:Treatment>Mucinex Sinus-Max Day &amp;amp; Night Relief </mcr:Treatment>
    <dc:Type>Report</dc:Type>
    <dc:Format>PDF</dc:Format>
    <dc:Identifier></dc:Identifier>
    <dc:Source>GetWell</dc:Source>
    <dc:Language>Eng</dc:Language>
    <dc:Relation>Head</dc:Relation>
    <dc:Coverage></dc:Coverage>
    <dc:DateOfVisit>12/12/2016</dc:DateOfVisit>
  </Report>
</Metadata>
```

Figure 4: DC Meta-data Content for Doctor Visit Summary

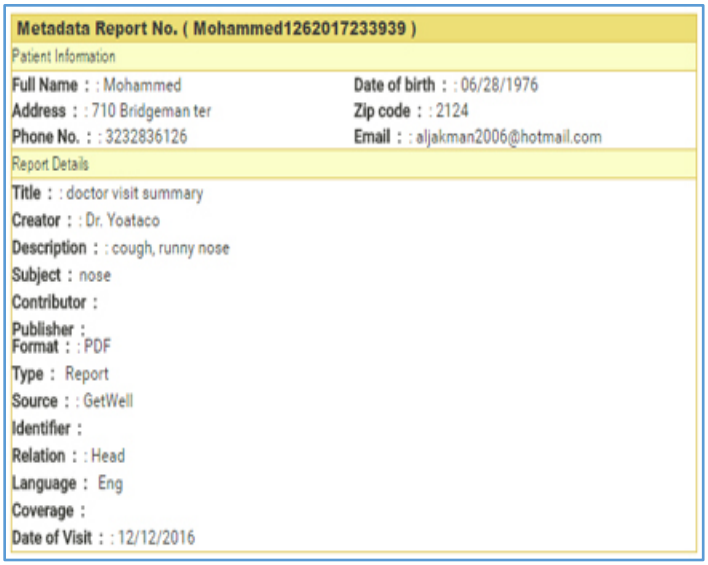


Figure 5: Visualization of the Meta-data

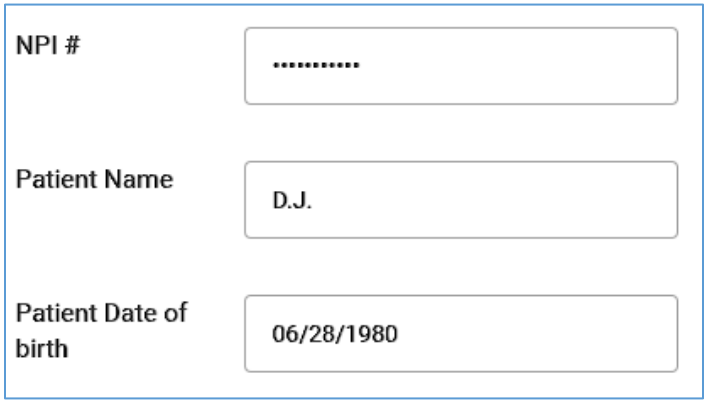


Figure 6: Emergency Access with Valid License Number

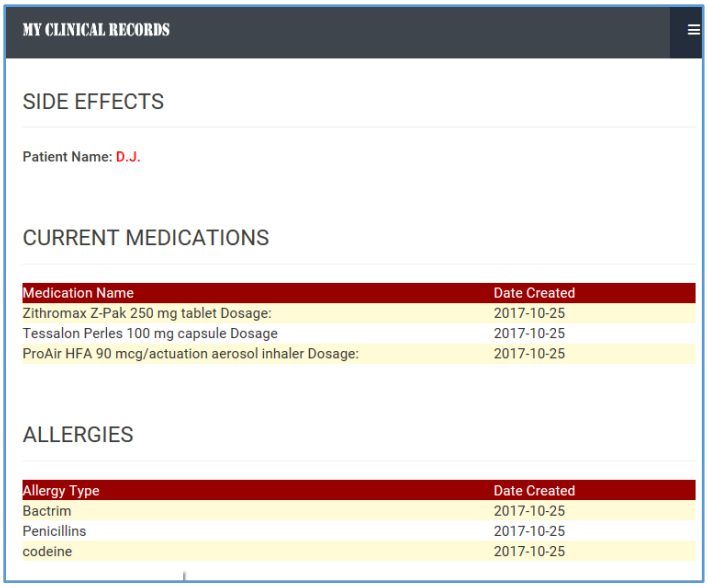


Figure 7: Emergency Information

- The easy access to users' health data and the ability to contribute to their record enhances users' motivation to use PHR.
- MCRS enables emergency clinical data access by emergency crew only with valid license number. We use the National Provider Identifier (NPI), patient name, and date of birth for the emergency medical information access as shown in the Figure 6. MCRS uses the Application Programming Interface (API) that was provided by NPPES NPI Registry in order to verify the NPI. Emergency information contains allergies, current medication list, and side effects. This information is updated regularly by patients as shown in the Figure 7. It also contains any references to the time of the last update.
    - Healthcare providers apply for NPI using the National Plan and Provider Enumeration System (NPPES) [29].
    - NPI can be validated through NPPES NPI Registry.

5.2. Using MCRS

We use patient's Dropbox access token to allow the connection between Dropbox and MCRS, so patients can have their own storage and have the ability to provide access to their storage through MCRS when needed. This allows users to keep their own data without binding to any specific application. MCRS contains no clinical data as they are stored in the patient's cloud storage. Patients need accounts for Dropbox and MCRS separately.

5.3. Personal Cloud Storage

Cloud storage is a place where users can store their data and access it anytime, from anywhere, and from any device via the Internet. It is maintained, managed and operated by cloud storage service providers such as Google, Amazon, or Microsoft Cloud storage services have many advantages such as cost savings, ease of use, ability to share data, accessibility, and sustainability. Personal Cloud Storage (PCS) is getting more popular because of the aforementioned convenience. Any cloud storage service provider may be used (SugarSync, Carbonite, IDrive, Dropbox, Google Drive, and others) – for storing personal clinical data as long as they provide required functionality and security. For testing purposes, we chose Dropbox™ as cloud storage. However, for practical use, secure cloud storage services that are HIPAA complaint can be used for privacy and security purposes, such as Dropbox (Business), Box, Google Drive, Microsoft OneDrive, and Carbonite [30]. The contents of each storage are described by DC meta-data for interoperability. In the case that data has embedded meta-data, we create another layer of meta-data so entire contents can be retrieved through our DC meta-data.

5.4. Managing Health Data

MCRS categorizes clinical data based on human body parts. There are eight categories: abdomen, heart, head, thorax, extremities, integumentary, urinary, and reproductive, as shown in Table 2. Any clinical data will be stored and linked based on these categories using the relation and subject tag elements of the DC meta-data. We kept the categories to a minimum so it can be simple enough to be used by patients. Users can specify the category (the human body part of interest) when searching for relevant clinical data, so it can show only the clinical documents (e.g. doctor visit summary, x-ray, and others) that are related to that part.

When using the relationships between the resource (DC subject) and target resource (DC relation), it is possible to combine the result to a greater scope, e.g. instead of eyes and ears, it can be categorized by head. This can be done by predefining each part of the human body and associating it with its related category in the system. Also, we have constrained the DC subject to a small core set that can be selected from a drop-down menu (all possible parts of the human body) to best describe the subjects (as shown in the Figure 8). When the users select the subject element, the DC relation field will be populated automatically with the associated part of its related category. For example, when a user searches by keyword (e.g. head) and chooses the element (e.g. relation), the search results will be filtered and show only all clinical documents that are relevant to the head (e.g. eyes, ears, brain, mouth, teeth, nose, and chin) as shown in the Figure 9. This is a less time-consuming method to filter the data instead of showing all documents as shown in the Figure 10. Also, the user can filter the search by date if they need to specify a period of time to find clinical documents.

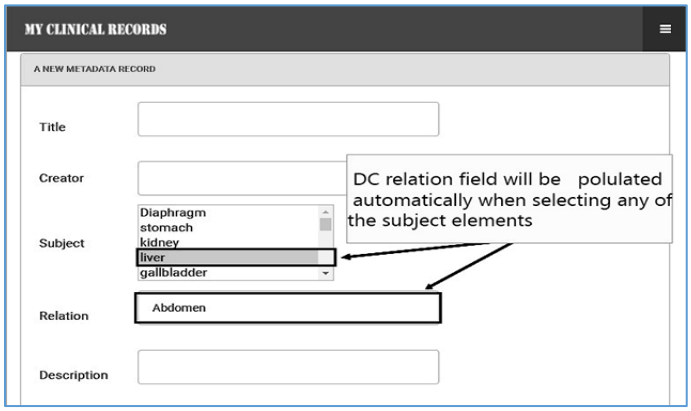


Figure 8: Create DC Meta-data

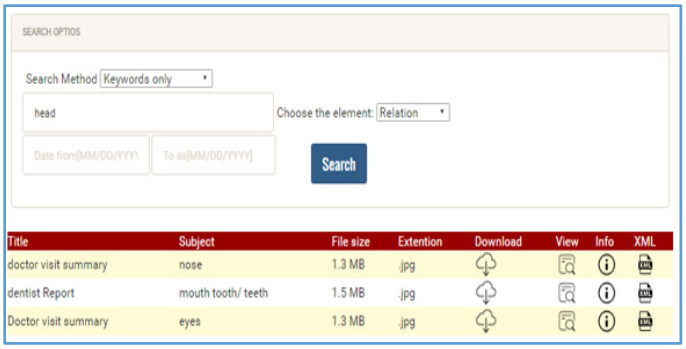


Figure 9: Example of Retrieving DC Meta-data for only Related Documents

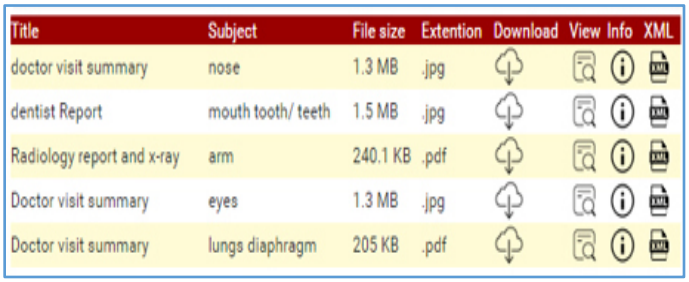


Figure 10: Example of DC Meta-data for all Documents

Table 2: Human Body Categories

Body categories	Body parts
The abdomen	Contains diaphragm, stomach, liver, gallbladder, pancreas, small intestine, large intestine, cava, spleen, and others.
Heart	Contains superior vena cava, pulmonary artery, pulmonary veins, pulmonic valve, tricuspid valve, inferior vena cava, right atrium, right ventricle, left ventricle, aortic valve, mitral valve, left atrium, aorta and others.
Head	Contains eyes, ears, brain, mouth, teeth, nose, chin, spinal cord, tonsil, uvula, gullet, meninges, pharynx and others.
Thorax	Contains lungs, diaphragm/pleura, nasopharynx/oral, cavity, trachea/Larynx, ribs, capillaries, bronchial tube, windpipe/trachea, chest, esophagus and others.
Extremities	Contains arms, elbows, hands, wrists, shoulders, hips/thighs, fingers, thumbs, legs, knees, toe, vertebral column, neck, ankles, breast, back pain, feet and others.
Integumentary	Skin and associated structures such as hair, nails, sweat glands, and oil glands
Urinary	Kidneys, ureters, urinary bladder, and urethra
Reproductive	Gonads (testes or ovaries) and associated organs; in females: uterine tubes, uterus, and vagina; in males: epididymis, ductus deferens, prostate gland, and penis

6. Conclusion

As the medical industry is going through a paradigm shift from clinician-centered to patient-centered, readily available complete personal medical history is becoming crucial. This will help ensure the three major goals in medical industry: evidence-based treatment, continuity of care, and prevention of medical mistakes. In this paper, we proposed an untethered personal health record system to help achieve such goals. Our proposed system, MCRS, provides a method to collect and organize heterogeneous personal health data using DC meta-data. The retrieval of the organized data was completed by the reorganized DC tags, which allowed the organization of clinical data by body parts for easy retrieval. Finally, we allowed access to personal emergency clinical data to emergency crew only by their license number at the time of need. In the future, we plan to expand the usage of clinical data collected from the application to analyze and identify the regional characteristics in health mapping to build better public policy for the nation. Our experiment was limited to one healthcare provider using the clinical data from that provider. Also, the data collection was limited due to the patient privacy regulations in the healthcare industry.

References

[1] This paper is an extension of work originally presented in 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, 2017.

[2] Key Developments in the Connected Health Markets https://www.parksassociates.com/whitepapers.

[3] AHIMA e-HIM Personal Health Record Work Group. "Defining the Personal Health Record." Journal of AHIMA 76, no.6 (June 2005): 24-25.

[4] Richard Brandt and Rich Rice, "Building a better PHR paradigm: Lessons from the discontinuation of Google Health™", Health Policy and Technology Volume 3, Issue 3, 2014, Pages 200-207, ISSN 2211-8837, https://doi.org/10.1016/j.hlpt.2014.04.004. (http://www.sciencedirect.com/science/article/pii/S221188371400032X)

[5] Morgan Price, Paule Bellwood, Nicole Kitson, Iryna Davies, Jens Weber and Francis Lau, "Conditions potentially sensitive to a Personal Health Record (PHR) intervention, a systematic review". BMC Medical Informatics and Decision Making 15, 32 (2015), https://doi.org/10.1186/s12911-015-0159-1

[6] Price, Margaux M. and Pak, Richard and M{\"u}ller, Hendrik and Stronge, Aideen, "Older adults' perceptions of usefulness of personal health records" "Universal Access in the Information Society 12, 191-204 (2013), doi="10.1007/s10209-012-0275-y

[7] Alyami, M.Aalyami and Song, Yeong-Tae: Removing Barriers in using Personal Health Record Systems. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). DOI: 10.1109/ICIS.2016.7550810

[8] Fearon. James D. and Laitin, David D. "Ethnicity, Insurgency, and Civil War" American Political Science Review volume={97} pages={75–90}, (2003). DOI={10.1017/S0003055403000534

[9] Sweet Lauren E. and Moulaison Heather Lea. " Electronic Health Records Data and Metadata: Challenges for Big Data in the United States" Big Data. January 2014, 1(4): 245-251. https://doi.org/10.1089/big.2013.0023

[10] Greenberg, Jane, Maria Cristina Pattuelli, Bijan Parsia, and W. Davenport Robertson. "Author-generated Dublin Core metadata for web resources: a baseline study in an organization." In International Conference on Dublin Core and Metadata Applications, pp. 38-45. 2001.

[11] Talha, Norhaizan Mat. "Metadata management system (MMS)." In International Conference on Dublin Core and Metadata Applications. 2004.

[12] Haugen, Mary Beth and Herrin, Barry and Slivochka, Sharon and Tolley, Lori McNeil and Warner, Diana and Washington, Lydia,"Rules for handling and maintaining metadata in the EHR" Journal of AHIMA 84, no.5 (May 2013): 50-54. URL: http://europepmc.org/abstract/MED/23755436

[13] P. Dhivya and S. Roobini and A. Sindhuja, "A. Symptoms Based Treatment Based on Personal Health Record Using Cloud Computing" Procedia Computer Science 47, 22-29 (2015), doi = "https://doi.org/10.1016/j.procs.2015.03.179", url = http://www.sciencedirect.com/science/article/pii/S1877050915004470

[14] Chintan Patel and Karthik Gomadam and Sharib Khan and Vivek Garg, "TrialX: Using semantic technologies to match patients to relevant clinical trials based on their Personal Health Records" Web Semantics: Science, Services and Agents on the World Wide Web, 8, 342-347 (2010) doi = "https://doi.org/10.1016/j.websem.2010.08.004", url = http://www.sciencedirect.com/science/article/pii/S1570826810000636

[15] Appelboom, Geoff and Camacho, Elvis and Abraham, Mickey E. and Bruce, Samuel S. and Dumont, Emmanuel LP et al. "Smart wearable body sensors for patient self-assessment and monitoring" Archives of Public Health 72, 28 (2014). doi="10.1377/hlthaff.2012.1216", url=https://doi.org/10.1377/hlthaff.2012.1216

[16] Zhang J., Chen D.& Wong T. 'Metadata application on clinical trial data in drug development,' Proceedings of the Twenty Eighth Annual SAS® Users Group International Conference, Seattle, WA (2003).

[17] Tal Teitz and Dwayne G. Stupack and Jill M. Lahti, "Halting Neuroblastoma Metastasis by Controlling Integrin-Mediated Death" Cell Cycle 5, 681-685 (2006). doi = {10.4161/cc.5.7.2615}, URL = http://dx.doi.org/10.4161/cc.5.7.2615

[18] Ariel Ekblaw, Asaf Azaria, Thiago Vieira, Andrew Lippman. (2016). MedRec: Medical Data Management on the Blockchain. PubPub, https://www.pubpub.org/pub/medrec version: 57e013615dbf3f3300152554].

[19] Asuman Dogac , Gokce B. Laleci , Yildiray Kabak , Seda Unal , Sam Heard , Thomas Beale , Peter Elkin , Farrukh Najmi , Carl Mattocks , David Webber et al. "Exploiting ebXML registry semantic constructs for handling archetype metadata in healthcare informatics" International Journal of Metadata, Semantics and Ontologies 1, 21-36 (2006). https://doi.org/10.1504/IJMSO.2006.008767

[20] Poeggel, Sven, Daniele Tosi, DineshBabu Duraibabu, Gabriel Leen, Deirdre McGrath, and Elfed Lewis. "Optical fibre pressure sensors in medical applications." Sensors 15, no. 7 (2015): 17115-17148.

[21] Guner C, Akin S, Durna Z. Comparison of the symptoms reported by post-operative patients with cancer and nurses' perception of patient symptoms. Eur J Cancer Care 2014;23:523-30. doi: 10.1111/ecc.12144

[22] Jeff Mather, MathWorks "Accessing data in DICOM files" Technical Articles and Newsletters Published 2002http://www.mathworks.com/company/newsletters/articles/accessing-data-in-dicom-files.html?requestedDomain=www.mathworks.com (accessed October 2016)

[23] Meltzer, Eli O., Daniel L. Hamilos, James A. Hadley, Donald C. Lanza, Bradley F. Marple, Richard A. Nicklas, Allen D. Adinoff et al. "Rhinosinusitis: developing guidance for clinical trials." Journal of Allergy and Clinical Immunology 118, no. 5 (2006): S17-S61.

[24] LaConte, Grace M. "Documentation roles in an electronic health record environment." PhD diss., The College of St. Scholastica, 2011.

[25] T.Brnson, Principles of Health Interopersability HL7 ans SNOMED CT, Health Information Technology Standards, 2012.

[26] Chao-chen Chen, Hsueh-hua Chen, Kuang-hua Chen. "The Design of Metadata Interchange for Chinese Information and Implementation of Metadata Management System" BULLETIN of the American Society for Information Science and Technology Vol. 27, No. 5 June / July 2001, url: https://www.asis.org/Bulletin/Jun-01/chen.html

[27] Informatica Metadata Management for Holistic Data Governance. (July 2013).at https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/white-paper/metadata-management-data-governance_white-paper_2163.pdf

[28] Sunyaev, Ali, Dmitry Chornyi, Christian Mauro, and Helmut Krcmar. "Evaluation framework for personal health records: Microsoft HealthVault vs. Google Health." In System Sciences (HICSS), 2010 43rd Hawaii International Conference on, pp. 1-10. IEEE, 2010.

[29] Bindman AB. Using the National Provider Identifier for health care workforce evaluation. Medicare Medicaid Res Rev. 2013;3(3):1-10.

[30] C. N. Daily, "Top 20 Best Cloud Storage Providers – Reviews and Comparison of the Top Secure Solutions and Services. Find Unlimited Cloud Based Data Storage Services and Options," 2017.

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 168-190 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Linear algebra as an alternative approach to the synthesis of digital devices of automation and control systems

Nikolay Chernov[1], Nikolay Prokopenko[*,2,3], Vladislav Yugai[1], Nikolay Butyrlagin[2]

[1]*Systems of automation control, Southern Federal University, SFedU, Taganrog, 347928, Russia*

[2]*Information systems and radioengineering, Don State Technical University, DSTU, Rostov-on-Don, 344000, Russia*

[3]*Institute for Design Problems in Microelectronics of Russian Academy of Sciences, IPPM RAS, Zelenograd, 124681, Russia*

ARTICLE INFO

Article history:
Received: 18 January, 2018
Accepted: 18 January, 2018
Online: 30 January, 2018

Keywords :
multivalued component base
multivalued logic
linear logic synthesis

ABSTRACT

The article considers linear algebra as an alternative mathematical tool of logic synthesis of digital structures to Boolean algebra and synthesis methods of digital electronic component base (ECB) on its ground. The methods of solving the applied problems of logic synthesis are shown, including the expansion of an arbitrary logic function by means of monotonic functions. The proposed mathematical apparatus actually provides the creation of digital structures on the principles of analog circuitry. It can find application in the design of multivalued digital ECB, specialized system-on-chip and analog-digital sensors with current output. The examples of synthesis of the combinational and sequential two-valued and multivalued digital devices are given. In conclusion, the advantages of linear algebra in comparison with Boolean algebra are formulated.

1. Introduction

This article is a continuation of the studies presented at the conference SIBCON-2017 [1].

Boolean algebra is known [2] as a leading mathematical tool for logic synthesis of two-valued digital structures. Almost all existing methods of logic synthesis are formed on its ground. The success of Boolean algebra is caused, among other things, by the fact that the Boolean representation of the realized logic function turned out to be rather technological: the circuit implementation of logic elements was relatively simple. TTL, C-MOS and ESL and other technologies occurred to be the most preferable for this purpose.

Despite this, the history of the development of digital microelectronics knows the attempts to replace both the approach to logic synthesis (the use of spectral representations [3 - 6] using the arithmetic polynomials [7 - 9]) and technological realization (I^2L, I^3L, ...).

The most significant contribution to the alternative theory of logic synthesis has been made by the threshold interpretation of Boolean algebra, called the threshold logic.

*Corresponding Author: Nikolay Prokopenko, Email: prokopenko@sssu.ru

Threshold logic related to one of the directions of synthesis of the digital structures [10], is constantly evolving. We know a significant number of publications devoted to this problem [11 – 14]. For two-valued functions, the threshold synthesis problems have already been solved beginning from the well-known papers of M. Dertouzos [10]. One of the advanced approaches of synthesis for the *k*-valued threshold functions is considered in this article [11].

The threshold logic was initially implemented in a two-valued version, but many publications on the multivalued threshold logic have recently appeared [14 - 17]. Why are they dedicated to the multivalued logic? The fact is that the multivalued logic is currently considered as a means of improving the quality characteristics of LSI (the ratio of "number of elements / number of links", "total area / real-estate" of the crystal, etc.), which doesn't require a drastic change in LSI technology. The authors of this article also follow this approach.

To develop this ideology we proposed an alternative approach to the logic synthesis of digital devices - the replacement of the mathematical tool of Boolean algebra by another body of mathematics - linear algebra. Such a replacement entailed fundamental changes in various aspects of the implementation of digital structures:

- the potential realization of Boolean logic was replaced by the current linear implementation;

- the key operating mode of the elements was replaced by the linear one;

- Boolean values of the variables took a quantitative form instead of the qualitative (logic) one;

- the value of logic was determined not by the scheme, but by the significance of signals;

- the output signal was presented by the difference of signals of two circuits operating in parallel, which improves the performance of digital structures.

Thus, the use of linear algebra as a mathematical tool of the logic synthesis ensured the development of alternative methods for the logic synthesis of current digital circuits and their practical implementations.

The authors published a number of papers [18-32] devoted to the synthesis of logic (nonthreshold) two-valued and multivalued digital structures. This paper considers the use of linear algebra as a mathematical tool for the logic synthesis of two-valued and multivalued, logic and threshold digital structures.

The fuzzy concept of the "threshold synthesis" can be interpreted in two ways:

- as a normal logic synthesis of digital structures with circuit implementation on threshold logic elements (any logic function can be implemented in this way);

- as a logic synthesis of threshold logic functions (an arbitrary logic function can be implemented by a network of threshold logic elements).

The purpose of this article is to propose logical and threshold current hardware components for constructing digital structures within the two specified areas. Within the framework of this goal, the authors' solutions for the following tasks are given below:

- since linear algebra is used as a mathematical tool for the logic synthesis of current logic structures, the article presents the main definitions and aspects of the practical use of linear algebra;

- as there is a close relationship between the threshold and monotonic functions, the article gives a definition of the monotonic function and explains the ways of representing arbitrary functions by monotonic linear functions;

- some versions of transformation of two-valued and multivalued monotonic functions into a threshold form in linear algebra are analyzed;

- versions of circuit implementation of linear threshold logic elements are considered.

2. Linear Algebra

A. Definition of linear algebra. Let $\boldsymbol{P} \to \langle P; +, \cdot, 0.1\rangle$ – field, $\langle A; +, \cdot, \theta\rangle$ – algebra with two binary operations and one nullary operation. The system $\boldsymbol{\Lambda} \to \langle A; +, \cdot, \theta, \mathbf{P}\rangle$ is called *linear algebra*, if the following conditions are met:

– the system $\boldsymbol{\Lambda} \to \langle A; +, \cdot, \theta, \boldsymbol{P}\rangle$ – linear (vector) space over the field $\boldsymbol{P}$;

– distributivities of operations + and ·

$$\forall(a, b, c \in A)\ (a + b)c = ac + bc \wedge c(a + b) = ca + cb;$$

– associativities of vector multiplication by elements of the field $\boldsymbol{P}$

$$\forall(a, b \in A) \wedge \forall(k \in P)\ k(ab) = (ka)b = a(kb).$$

B. Extension of linear algebra.

Let $\boldsymbol{A} \to \langle A; +, \cdot, \theta\rangle$ – vector space of linear algebra $\boldsymbol{\Lambda}$, $\boldsymbol{P} \to \langle P; \Omega = \{\omega_k | k \in P\}, 0.1\rangle$ – field of linear algebra $\boldsymbol{\Lambda}$, which contains the operations ω_k, which in general case do not necessarily coincide with the operations of linear space $\boldsymbol{A}$. Then the system $\boldsymbol{\Lambda'} \to \langle \{A; +, \cdot, \theta\}, \{P; \Omega, 0, 1\}\rangle$ is called *the extension of linear algebra* $\boldsymbol{\Lambda}$.

When interpreting this algebraic system in a certain way, we can obtain algebras with different properties. For example, interpreting A as a set of terms of Boolean functions $f(x_1,\ldots, x_n)$, the operations "+" and "·" - as $max(x_1,\ldots, x_n)$, $min(x_1,\ldots, x_n)$, we obtain algebra $\boldsymbol{A} \to \langle A; max, min, \theta; \boldsymbol{P}\rangle$. Similarly, leaving the semantics of operations in the original form (i.e., defining them as ordinary arithmetic operations), we can consider the reduced system as linear algebra on the set A of vectors in a linear space. The reduced algebraic system is considered below in this form.

C. Creation of bases. To form the bases from logic variables, it is possible to construct different constructions of linearly independent vectors with given properties. The choice of operations for creation of the bases is made independently on the operations of linear space and can be determined by various (mathematical, circuit, technological and other) requirements. In the applied plan, this allows obtaining the ideologically unified (based on operations of linear space) circuit solutions of functional elements (from different implementations based on field operations).

A basis is a system of m linearly independent vectors $\{F\} = \{\varphi_0, \varphi_1,\ldots, \varphi_{m-1}\}$, which enables to describe any vector $a \in \boldsymbol{A}$ in the linear form

$$a = \sum_{i=0}^{m-1} a_i \varphi_i\ , \qquad (1)$$

with respect to these vectors.

Each vector of the basis $\{F\} = \{\varphi_0, \varphi_1,\ldots, \varphi_{m-1}\}$ is uniquely determined by the set of coordinates $\varphi_i = \{\varphi_{i0}\ , \varphi_{i1},\ldots, \varphi_{i,m-1}\}$, with the help of which we can make a square matrix of order m:

$$F = \begin{bmatrix} \varphi_{10} & \varphi_{11} & \cdots & \varphi_{1,m-1} \\ \varphi_{20} & \varphi_{21} & \cdots & \varphi_{2,m-1} \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_{m-1,0} & \varphi_{m-1,1} & \cdots & \varphi_{m-1,m-1} \end{bmatrix},$$

corresponding to the given basis {F}.

Two bases $\{F\} = \{\varphi_1, \varphi_2, \ldots, \varphi_{m-1}\}$ and $\{Y\} = \{\gamma_1, \gamma_2, \ldots, \gamma_{m-1}\}$, the matrices F and Y of which are inverse to each other, are reciprocal (dual). Besides,

$$F \cdot Y = E,$$

where E – diagonal matrix of order m with $\varepsilon_{ij} = 1$, when $i = j$ and $\varepsilon_{ij} = 0$, if $i \neq j$, which is an orthonormal basis $\{E\}$.

Since the resolution of an arbitrary vector a over the basis $\{F\}$ has the form

$$a = a_0\varphi_0 + a_1\,\varphi_1 + \cdots + a_{m-1}\,\varphi_{m-1},$$

then, multiplying both parts of this resolution by γ_i, $i = [0, m-1]$, we obtain:

$$a_i = a\cdot\gamma_i = a\cdot(Y\cdot\varepsilon_i) = a\cdot(F^{-1}\cdot\varepsilon_i).$$

Then the procedure for representing (i.e. obtaining the values of the expansion coefficients) of the arbitrary vector a in the given basis {F} is reduced to performing the following operations:

– construction of the basic matrix F;

– construction of the inverse basic matrix F^{-1};

– multiplication of the row-vector a by the column-vector of the matrix F^{-1} and computation of the expansion coefficients of the vector a over the basis $\{F\}$;

– writing of the expression for the vector a in the linear form (1) with respect to the basis $\{F\}$.

Example 1. Get a representation of the conjunction operation of two arguments $x_1\&x_2$ of the value 2 in the basis

$$F = \begin{bmatrix}1 & 1 & 1 & 1\\ 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 1\\ 0 & 1 & 1 & 0\end{bmatrix},\quad (F)^{-1} = \frac{1}{2}\begin{bmatrix}2 & -1 & -1 & -1\\ 0 & 1 & -1 & 1\\ 0 & -1 & 1 & 1\\ 0 & 1 & 1 & -1\end{bmatrix}.$$

Solution.

a) we represent the sequence of values of the two-valued logic function by the row-vector

$$x_1\&x_2 = [0, 0, 0, 1];$$

b) we multiply the resulting row-vector by the columns of the inverse basic matrix $(F)^{-1}$ and obtain the row-vectors of the expansion coefficients of the represented logic function with respect to the given basis and the expression of the conjunction operation

$$x_1\&x_2 = [0, 0, 0, 1]\cdot\frac{1}{2}\begin{bmatrix}2 & -1 & -1 & -1\\ 0 & 1 & -1 & 1\\ 0 & -1 & 1 & 1\\ 0 & 1 & 1 & -1\end{bmatrix} =$$

$$= [0, 0, 0, 1]\cdot\frac{1}{2}[0, 1, 1, -1] = \frac{x_1 + x_2 - |x_1 - x_2|}{2}.$$

It is noteworthy that the last expression has been known since 1953 [3], but the author didn't describe the method of its obtaining (more precisely, it was described later).

The authors of this article propose three approaches to the creation of the basis vectors of linear bases, depending on the operations used for this purpose:

- based on Boolean operations ∨ , &, ⊕, ~, etc .;

- based on the truncated difference operation ∸;

- based on the comparison operation ≥.

When using Boolean operations, the upper and lower "cuts" [14, 18] of variables of different orders are used as operands for creating basis vectors:

– variables of the first order $x^{(i)} = x\&i,\ x_{(i)} = x \vee i$;

– variables of the second order

$$x^{(ij)} = x\&i - x\&j,\ x_{(ij)} = x \vee i - x \vee j$$

and so on.

When using the *truncated difference* operation

$$x_1 \dot{-} x_2 = \begin{cases} x_1 - x_2 & \text{when } x_1 \geq x_2;\\ 0 & \text{when } x_1 < x_2,\end{cases}$$

all logical operations are replaced by some combinations of this operation on variables involving operations of linear space. For example, the upper and lower "cuts" of the first and second orders are replaced by the expressions

$$x^{(i)} = i \dot{-} (i \dot{-} x) = x \dot{-} (x \dot{-} i);$$

$$x_{(i)} = i + (x \dot{-} i) = x + (i \dot{-} x);$$

$$x^{(ij)} = (x \dot{-} j) \dot{-} (x \dot{-} i);$$

$$x_{(ij)} = (i - j) \dot{-} (x - j).$$

Similarly, Boolean operations on two variables are replaced by the following expressions:

$$x_1\&x_2 = x_1^{(x_2)} = x_2 \dot{-} (x_2 \dot{-} x_1) = x_1 \dot{-} (x_1 \dot{-} x_2);$$

$$x_1 \vee x_2 = x_{1(x_2)} = x_2 + (x_1 \dot{-} x_2) = x_1 + (x_2 \dot{-} x_1);$$

$$x_1 \oplus x_2 = x_1 + x_2 - k\{1 \dot{-} [k \dot{-} (x_1 + x_2)]\};$$

$$x_1 \ominus x_2 = x_1 - x_2 + k\{1 \dot{-} [1 \dot{-} (x_2 \dot{-} x_1)]\};$$

and others.

When using the *comparison* operation, the above expressions are reduced to the following form

$$x^{(i)} = \sum_{j=0}^{i-1}(x > j);$$

$$x^{(i)} - x^{(j)} = \sum_{t=j}^{i-j}(x > t);$$

$$x_{(i)} = x + \sum_{j=0}^{i}(j > x) = i + \sum_{j=i+1}^{k-2}(x > j);$$

$$x_{(i)} - x_{(j)} = (i-j) - \sum_{t=j}^{i-1}(x > t).$$

It should be noted that the most interesting result of the studies is the fact of constructing logical structures based on *truncated difference* and *comparison* operations other than Boolean ones. Naturally, there are certain dependencies between the *truncated difference* and *comparison* operations, some of which are given below:

-"*truncated difference – comparison*":

$$x \dot{-} i = \sum_{j=i+1}^{k-1}(x > j);$$

$$i \dot{-} x = \sum_{j=1}^{i}(j > x);$$

$$j \dot{-} (x \dot{-} i) = \sum_{p=1}^{j}[(i+p) > x];$$

$$j \dot{-} (i \dot{-} x) = (j \dot{-} i) + \sum_{p=i-1}^{i \dot{-} j}(x > p);$$

- "*comparison - truncated difference*":

$$x \le i = 1 \dot{-} \{1 \dot{-} [(i+1) \dot{-} x]\};$$

$$x < i = 1 \dot{-} [1 \dot{-} (i \dot{-} x)];$$

$$x \ge i = 1 \dot{-} [1 \dot{-} (x \dot{-} i)];$$

$$x > i = 1 \dot{-} [(i+1) \dot{-} x];$$

$$x_1 > x_2 = 1 \dot{-} [1 \dot{-} (x_1 \dot{-} x_2)];$$

$$x_1 \ge x_2 = 1 \dot{-} \{1 \dot{-} [(x_1 \dot{-} x_2) + 1]\};$$

$$x_1 < x_2 = 1 \dot{-} \{1 \dot{-} [(x_2 \dot{-} x_1) + 1]\}.$$

3. Monotonic Logic Functions

Regarding the close connection between threshold and monotonic functions, we present some known results obtained in linear algebra in a simpler and more obvious way.

A. Construction of sequences of nondecreasing componentes. Suppose we have an arbitrary vector $a = (a_0, a_1, , \dots, a_{m-1}) \in Z^m$. We renumber the components of the vector with k-ary n-bit numbers. The vector a$\in$ Z^m is called monotonically increasing (decreasing), if for a bitwise comparison of k-ary number codes of the components a_i and a_j we have:

$$\forall (i, j \in Z^m) i \ge j \Rightarrow a_i \ge a_j \wedge i \le j \Rightarrow a_i \le a_j. \qquad (2)$$

The bitwise comparison of k-ary number codes of the components enables to single out the sequences of nondecreasing (nonincreasing) components.

The necessity to create these sequences is that for the monotonicity of the vector, condition (2) must be met in each sequence. Hence follows the first factor of simplicity of the analysis results in linear algebra: to verify the vector by monotonicity it is sufficient to establish its monotonicity within each sequence. This reduces the amount of computation and the overall complexity of the verification process.

The sequences of nondecreasing components are constructed as follows. The set G of components (constituents 1) of logic functions of the chosen number of n variables is divided into n groups $G = \{g_0, \dots, g_{n-1}\}$. Each group g_{it} includes the sequences of nondecreasing components with starting numbers t, $i \le t \le n-1$, determined by the following relation

$$k^i - k^{i-1} \le j \le k^{i+1} - k^i - 1,$$

in this case we take $(k^i - k^{i-1})\big|_{i=0} = 0$. The starting numbers are component numbers that should be compared with the values of the other elements in the sequence. Each sequence contains components with numbers the decimal equivalents of which are determined as $t + k^t, t + k^{t+1}, \dots, t + k^{n-1}$. As a result, the entire sequence generally takes on the form

$$g_{it} = \{g_t, g_t + k^t, g_t + k^{t+1}, \dots, g_t + k^{n-1}\}.$$

Thus, it is possible to construct the sequences of nondecreasing components for the given values of k and n. The sequence graphs of nondecreasing components for $k = 2$, $k = 3$ and $n = 3$ are shown in Figure 1. The digits in the circles denote the decimal numbers of the vector components, and the sequence itself includes a certain vertex and the nearest right vertexes connected with it by the lines.

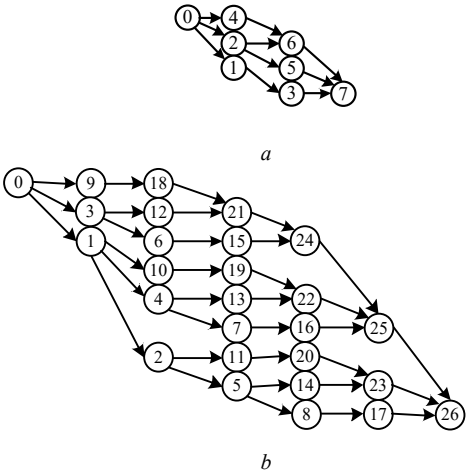


Figure 1: Sequence graphs of the nondecreasing components for $k = 2$, $k = 3$ and $n = 3$.

Example 2. Construct sequences of nondecreasing components for $k = n = 3$. Since $n = 3$, there are 3 groups of sequences with starting numbers 0, 1 and 2 for the initial data under consideration.

For group "0", the value of t_0 varies in the range

$$3^0 - 3^{-1} \le t_0 \le 3^1 - 3^0 - 1,$$

or $\quad 0 \le t_0 \le 1.$

Therefore, this group includes two sequences with the initial elements $t_{00} = 0$ and $t_{01} = 1$. The first sequence g_{00} of the group "0" contains the elements with numbers t_{00}, $t_{00} + 3^0$, $t_{00} + 3^1$, $t_{00} + 3^2$, i.e. 0, 1, 3, 9. The second sequence g_{01} of the group "0" comprises the elements with numbers t_{01}, $t_{01} + 3^0$, $t_{01} + 3^1$, $t_{01} + 3^2$, i.e. the elements with numbers 1, 2, 4, 10.

Similarly, for group 1, the value of t_1 is

$$3^1 - 3^0 \le t_1 \le 3^2 - 3^1 - 1,$$

or

$$2 \le t_1 \le 5.$$

Consequently, this group includes sequences with the initial elements $g_{1t} = 2 ... 5$, and each sequence of the group comprises the elements with numbers t_{1j}, $t_{1j} + 3^1$, $t_{1j} + 3^2$.

Finally, for group 2, the value of t_2 is

$$3^2 - 3^1 \le t_2 \le 3^3 - 3^2 - 1,$$

or

$$6 \le t_2 \le 17,$$

i.e. this group contains sequences with the initial elements $g_{2t} = 6 ... 17$, and each sequence includes the elements with numbers t_{2j}, $t_{2j} + 3^2$. Thus, we obtain the sequence structure of nondecreasing components in the form shown in Figure. 1*b*.

The monotonicity property of vectors from Z^m enables to obtain various representations of an arbitrary vector by means of monotonic vectors. Let's consider such representations in the following three variants:

– the difference of two monotonic vectors that have a value greater than the significance of the represented vector;

– the algebraic sum of monotonic vectors of the same value as the original vector;

– the algebraic sum of monotonic vectors of the value, the smaller value of the represented vector.

B. The problem of representing the logic function by the difference of two monotonic vectors of greater significance is solved by the following algorithm [18]:

Algorithm 1.

1. i=0. $b_0 = a_0$, $c_0 = b_0 - a_0$;

2. $1 \le i \le m - 1$.

$$b_i = \begin{cases} b_{i-1} \text{ when } a_i < a_{i-1} \\ b_{i-1} + a_i - a_{i-1} \text{ when } a_i \ge a_{i-1} \end{cases}, \quad c_i = b_i - a_i;$$

3. $i = m$, the end.

Here i and i–1 are neighboring indices of the elements in the sequence of nondecreasing components, $m = k^n$. The validity of the algorithm results from the following elementary arguments.

1. Let $b_0 = a_0$, then $c_0 = b_0 - a_0 = 0$.

2. Let the pair of components a_δ and a_γ satisfies the condition of monotonicity, i.e. $a_\delta \le a_\gamma$. In this case

$$a_\delta - a_\gamma \ge 0,$$

and, consequently, from the identity

$$b_\gamma = b_\delta + a_\gamma - a_\delta,$$

it follows that

$$b_\gamma \ge b_\delta,$$

i.e. b_γ also meets the monotonicity condition.

3. If the pair of components a_γ and a_δ doesn't fulfill the condition of monotonicity, i.e. $a_\delta \ge a_{\gamma\gamma}$, then, taking $b_\gamma = b_\delta$ we remain the monotonicity condition for the components of the vector b again.

Example 3. Construct a representation of the vector a = {0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0} as a difference of two monotonic vectors for $k = 2$, $n = 4$, $m = k^n = 16$.

The sequence of nondecreasing components for this case has the following form:

Table 1

g	Sequences	Values of components
0	0, 1, 2, 4, 8	0, 1, 1, 1, 1
1	1, 3, 5, 9,	1, 0, 0, 0
2	2, 6, 10	1, 0, 1
	3, 7, 11	0, 1, 1
3	4, 12	1, 1
	5, 13	0, 1
	6, 14	0, 1
	7, 15	1, 0

We construct the components of the vectors b and c:

$$b_0 = 0 \qquad c_0 = 0$$
$$b_1 = b_1 + a_1 - a_0 = 1 \qquad c_1 = 0$$
$$b_2 = b_1 + a_2 - a_1 = 1 \qquad c_2 = 0$$
$$b_3 = b_1 = 1 \qquad c_3 = 1$$
$$b_4 = b_2 + a_4 - a_2 = 1 \qquad c_4 = 0$$
$$b_5 = b_1 = 1 \qquad c_5 = 0$$
$$b_6 = b_2 = 1 \qquad c_6 = 1$$
$$b_7 = b_3 + a_7 - a_3 = 2 \qquad c_7 = 1$$
$$b_8 = b_4 + a_8 - a_4 = 1 \qquad c_8 = 0$$
$$b_9 = b_5 + a_9 - a_5 = 1 \qquad c_9 = 1$$

$$b_{10} = b_6 + a_{10} - a_6 = 2 \qquad c_{10} = 1$$
$$b_{11} = b_7 + a_{11} - a_7 = 2 \qquad c_{11} = 1$$
$$b_{12} = b_8 + a_{12} - a_8 = 1 \qquad c_{12} = 0$$
$$b_{13} = b_9 + a_{13} - a_9 = 2 \qquad c_{13} = 1$$
$$b_{14} = b_6 + a_{14} - a_6 = 1 \qquad c_{14} = 1$$
$$b_{15} = b_7 = 2 \qquad c_{15} = 2$$

The described algorithm demonstrates one more fact, which confirms the simplicity of analysis in linear algebra.

As it can be seen from the example, when using the described algorithm in the general case, the value of the original vector does not coincide with the significance of the resolution vectors.

C. The expansion of the arbitrary logic function of n arguments into the algebraic sum of monotonic logic functions of the same value of the following form

$$f(\tilde{x}^{(n)}) = (-1)^i \sum_{i=0}^{p} \varphi_i(\tilde{x}^{(n)}),$$

where $p \le kn$, $\varphi_i(\tilde{x}^{(n)})$ – monotonic expansion functions with the property $\varphi_1 \supset \varphi_2 \supset \ldots$, are made in accordance with the following algorithm [18].

Algorithm 2.

1. We choose a minimal summation of the positive summands of the arithmetic-logical representation of the logic function $f(\tilde{x}^{(n)})$ covering (in the logical sense) all other positive summands, and combine it by the operation $\vee$. Thus, we form the first expansion function $\varphi_1(\tilde{x}^{(n)})$. To remain the equality, all possible logical products of summands from $\varphi_1(\tilde{x}^{(n)})$ with the signs defined as $(-1)^j$, where $1 < j < s$, s is a number of summands in $\varphi_1(\tilde{x}^{(n)})$, are added to $\varphi_1(\tilde{x}^{(n)})$. Then we reduce similar terms and represent the initial logic function $f(\tilde{x}^{(n)})$ in the following form $f(\tilde{x}^{(n)}) = \varphi_1(\tilde{x}^{(n)}) - f_1(\tilde{x}^{(n)})$, where $f_1(\tilde{x}^{(n)})$ is a remainder of the initial function after reduction of similar terms.

2. We repeat clause 1 for $f_1(\tilde{x}^{(n)})$. As a result, the initial logic function is represented in the following form

$$f(\tilde{x}^{(n)}) = \varphi_1(\tilde{x}^{(n)}) - \varphi_2(\tilde{x}^{(n)}) + f_2(\tilde{x}^{(n)}).$$

3. We repeat clause 2 until the remainder of the initial logic function becomes zero.

Since the number of arguments of the logic function is n, and the violation of monotonicity is possible in each of the k values of each argument, then, the maximum number of the expansion functions doesn't exceed $2n(k–1)$.

The process convergence follows from the fact that each successive resolution vector eliminates some violation of monotonicity in the original vector and doesn't introduce new monotonicity violations, since it is monotonic itself.

The problem of representing the arbitrary logic function by the algebraic sum of monotonic functions of the same value can be solved analytically. As is known [13], the minimal disjunctive normal form (DNF) of the monotonic logic function doesn't contain inversions over variables. Consequently, the given logic function must be reduced to the representation in the form of the algebraic sum of such functions. To do this, we must perform the following actions:

– conversion of the Boolean expression of the logic function into the linear one using the following identities

$$x_1 \vee x_2 = x_1 + x_2 - x_1 x_2;$$

$$\bar{x} = 1 - x;$$

– inverse transformation to the given (that doesn't contain inversions over variables) form.

Example 4. Obtain the two-valued logic function mapping of three arguments

$$f(\tilde{x}^{(3)}) = (0,1,1,0,1,0,0,1),$$

into the algebraic sum of monotonic logic functions.

Solution. The logic function is represented by a vector of values. Its linear representation has the following form

$$f(\tilde{x}^{(3)}) = x_1 + x_2 + x_3 - 2x_1x_2 - 2x_1x_3 - 2x_2x_3 + 4x_1x_2x_3.$$

Now we obtain the mapping of this function into the algebraic sum of monotonic functions. The first expansion function is formed from the first three summands, covering in aggregate all the remaining positive terms of sum:

$$f(\tilde{x}^{(3)}) = (x_1 \vee x_2 \vee x_3 + x_1x_2 + x_1x_3 + x_2x_3 - x_1x_2x_3) -$$
$$-2x_1x_2 - 2x_1x_3 - 2x_2x_3 - 4x_1x_2x_3 =$$
$$= \varphi_1(\tilde{x}^3) - x_1x_2 - x_1x_3 - x_2x_3 + 3x_1x_2x_3.$$

The subsequent expansion functions are obtained in the same way:

$$f(\tilde{x}^{(3)}) = \varphi_1(\tilde{x}^{(3)}) - (x_1x_2 \vee x_1x_3 \vee x_2x_3 + x_1x_2x_3 +$$
$$+x_1x_2x_3 + x_1x_2x_3 + x_1x_2x_3 - x_1x_2x_3) + 3x_1x_2x_3 =$$
$$= \varphi_1(\tilde{x}^{(3)}) - \varphi_2(\tilde{x}^{(3)}) + \varphi_3(\tilde{x}^{(3)}).$$

To obtain the threshold representation of the initial function, it suffices to transform the monotonic functions of the resulting expansion into the threshold form and to perform their algebraic addition.

Each conjunctive term of any logical expansion function is transformed into the threshold form in accordance with the following identical equation

$$x_i x_j \ldots x_s = P[(x_i + x_j + \cdots + x_s) > t - 1],$$

where t - a number of arguments in the term, P - the predicate symbol. Then all the terms are reduced to the common value of the right-hand side with the introduction of the corresponding

coefficients for the variables, then they are added together and the threshold value of the received sum is determined.

Let's explain the transformation of the monotonic logic function into the threshold form with the function from the previous example.

With the transformation of $\varphi_1(\tilde{x}^{(3)})$ everything is simple:

$$\varphi_1(\tilde{x}^{(3)}) = x_1 \vee x_2 \vee x_3 = P_1[(x_1 + x_2 + x_3) > 0];$$

The transformation of $\varphi_2(\tilde{x}^{(3)})$ looks somewhat more complicated:

$$\varphi_2(\tilde{x}^{(3)}) = x_1x_2 \vee x_1x_3 \vee x_2x_3 =$$
$$= P_2[(x_1 + x_2) > 1 + (x_1 + x_3) > 1 + (x_2 + x_3) > 1] =$$
$$P_2[(2x_1 + 2x_2 + 2x_3) > 2] = P_2[(x_1 + x_2 + x_3) > 1];$$

The transformation of $\varphi_3(\tilde{x}^{(3)})$ is produced just as easily as $\varphi_1(\tilde{x}^{(3)})$:

$$\varphi_3(\tilde{x}^{(3)}) = x_1x_2x_3 = P_3[(x_1 + x_2 + x_3) > 2].$$

Thus, to realize the considered logic function, three threshold logic elements are required. The simplicity of the transformations in linear algebra becomes obvious.

D. Representation of the arbitrary vector by the vectors of lower significance. This problem has many solutions. The variant considered below assumes the solution of this problem in two stages:

- obtaining a representation of the arbitrary vector from Z^m by the vectors of lower significance;

- transformation of the obtained representation into the representation by means of monotonic vectors.

Let us first consider the solution of the first stage of the problem. The representation of the arbitrary vector by the lower-valued vectors can be obtained by weighting (each resolution vector is included into the final representation with some weight coefficient) or unitary (each resolution vector is included into the final representation with a unitary weight coefficient) coding.

1. It is known from the theory of numbers [33] that any number a can be uniquely represented in the following form:

$$\sum_1^s a_{s-i}k^{s-i}, \qquad (3)$$

where $s = [\log_k a]$ – the closest integer to $\log_k a$, a_{s-i} – the values of the expansion coefficients of the number a in the k-valued number system. Due to the uniqueness, this relation determines the isomorphism between any number a and the described representation of this number.

The quantity a_{s-i} can be determined from the following relation:

$$a_{s-i} = \frac{a\,(mod\ k^{s-i+1}) - a\,(mod\ k^{s-i})}{k^{s-i}}.$$

This operation is linear

$$a + b = \sum_{i=1}^{s} a_{s-i}\,k^{s-i} + \sum_{i=1}^{s} b_{s-i}\,k^{s-i} =$$
$$= \sum_{i=1}^{s}(a_{s-i} + b_{s-i})\,k^{s-i};$$
$$\lambda a = \sum_{i=1}^{s} \lambda a_{s-i}\,k^{s-i} = \lambda \sum_{i=1}^{s} a_{s-i}\,k^{s-i},$$

thus, it is applicable to the vectors of the linear space. If now we associate the vector a to the number a in the linear space Z^m, and the weighted sums of the vectors a_{s-i} of the spaces K_1^m and K_2^m to the weighted sums $\sum_{i=1}^{s_1} a_{s_1-i}\,k^{s_1-i}$ and $\sum_{i=1}^{s_2} a_{s_2-i}\,k^{s_2-i}$ correspondingly, then the expression (3) is equivalent to the representation of the vector $a \in Z^m$ by the weighted sum of the vectors from K_1^m or K_2^m respectively. Since in both spaces the vector a is uniquely represented, both representations are isomorphic. Let's call such a representation by coding, and the corresponding linear operator - by *the encoding operator.*

Let the vector $a \in Z^m$ in the canonical basis be represented as

$$a = a_1e_1 + a_2e_2 + \cdots + a_me_m.$$

Using expression (3), we expand each coordinate of the vector a with respect to the bases *k1* and *k2*. As a result, we obtain its representation in the following form

$$a_i = \sum_{i=1}^{s_1} a_{i,s_1-i}\,k_1^{s_1-i},$$

or

$$a_j = \sum_{j=1}^{s_2} a_{j,s_2-j}\,k_2^{s_2-j},$$

where $a_{i,s-i}$ – the expansion components of the number a_i (i – e vector components a_i of the vector representation a).

Then the vector representation a in each of the spaces can be described in the following form

$$a = \left(a_{1,s-1}k^{s-1} + a_{1,s_1-2}k^{s-2} + \cdots + a_{1,0}k^0\right) +$$
$$+\left(a_{2,s-1}k^{s-1} + a_{2,s_1-2}k^{s-2} + \cdots + a_{2,0}k^0\right) + \cdots$$
$$+ \left(a_{m,s-1}k^{s-1} + a_{m,s_1-2}k^{s-2} + +a_{m,0}k^0\right),$$

or

$$a = \left(a_{1,s-1} + a_{2,s-1} + \cdots + a_{m,s-1}\right)k^{s-1} +$$
$$+\left(a_{1,s-2} + a_{2,s-2} + \cdots + a_{m,s-2}\right)k^{s-2} + \cdots$$
$$+\left(a_{1,0} + a_{2,0} + \cdots + a_{m,0}\right)k^0.$$

Thus, the arbitrary vector a can be written in the following form

$$a = \sum_{j=1}^{s}\left(\sum_{i=1}^{m} a_{i,s-j}\right)k^{s-j},$$

i.e. a is represented by the weighted sum s of the vectors from K_1^m or K_2^m

Example 5. Let $k_1 = 10$, $n = 1$. The problem is to represent the vector $a = \{0, 1, 3, 9, 7, 6, 9, 4, 3, 8\}$ by its k_2-weighted and k_2-unitary sums of the vectors when $k_2 = 3$.

Solution. Since $\lceil log_{k_2} k_1 \rceil = \lceil log_3 10 \rceil = 2$, then the vector a can be represented by the weighted sum of three vectors a_2, a_1 and a_0. The coordinates of the highest weight vector a_2 are determined as

$$a_2 = \frac{a\ mod 3^3 - a\ mod 3^2}{3^2} =$$

$$= \frac{(0139769348) - (0130760438)}{9} = (0001001000).$$

The components of the subsequent vectors a_1 and a_0 are determined similarly:

$$a_1 = \frac{a\ mod 3^2 - a\ mod 3^1}{3^1} =$$

$$= \frac{(0130760348) - (0100100102)}{3} = (0010220112);$$

$$a_0 = \frac{a\ mod 3^1 - a\ mod 3^0}{3^0} =$$

$$= \frac{(0100100102) - (0000000000)}{1} = (0100100102).$$

Thus,

$$a = \sum_0^3 a_i 3^i = (0001001000)3^2 + (0010220112)3^1 + (0100100102)3^0.$$

2. We now represent the arbitrary number a as follows:

$$a = \sum_{i=0}^{s-1}(a - \sum_{i=0}^{s-1} a_i)\ \&(k_2 - 1), \qquad (4)$$

where s - the nearest larger integer

$$s = \left\lceil \frac{k_1 - 1}{k_2 - 1} \right\rceil.$$

For the reasons stated above, this representation is also an isomorphism.

If we associate each coordinate of the vector a with its representation in form (4), we obtain:

$$a = \left[a_1\&(k_2 - 1) + \left(a_1 - a_{11}\&(k_2 - 1)\right) + \cdots + a_1 - \sum_{i+1}^{s-1} a_{1i}\right]\&(k_2 - 1) +$$

$$+ \left[a_2\&(k_2 - 1) + \left(a_2 - a_{21}\&(k_2 - 1)\right) + \cdots + a_2 - \sum_{i+1}^{s-1} a_{2i}\right]\&(k_2 - 1) + \cdots$$

$$+ \left[a_m\&(k_2 - 1) + \left(a_m - a_{m1}\&(k_2 - 1)\right) + \cdots + a_m - \sum_{i+1}^{s-1} a_{mi}\right]\&(k_2 - 1).$$

Thus,

$$a = \sum_0^{s-1} a_i,$$

where $a_i = (a - \sum_{i+1}^{s-1} a_i)\&(k_2 - 1)$, i.e. the arbitrary vector $a \in Z^m$ is represented by the k_2-unitary sum of vectors from Z^m.

Example 6. Present the original vector from the previous example using unitary coding.

Solution. Since under the same initial conditions $s = \left\lceil \frac{10-1}{3-1} \right\rceil = 4$, the vector a can be represented by the unitary sum of five vectors a_4, a_3, a_2, a_1 and a_0, where

$a_4 = a\&(3 - 1) = (0139769438)\&(3 - 1) = (0122222222)$;

$a_3 = (a - a_4)\&(3 - 1) = (0012222212)$;

$a_2 = (a - a_4 - a_3)\&(3 - 1) = (0002222002)$;

$a_1 = (a - a_4 - a_3 - a_2)\&(3 - 1) = [0002102002]$;

$$a_0 = (a - a_4 - a_3 - a_2 - a_1)\&(3 - 1) =$$

$$= [0001001000].$$

Thus,

$$a = \sum_0^3 a_i = (0122222222) + (0012222212) +$$

$$+(0002222002) + (0002102002) +$$

$$+0001001000.$$

The solution of the second stage of the problem, i.e. the transformation of the obtained representation into the representation by means of monotonic lower-valued vectors, can be obtained on the base of algorithm 2.

Thus, the body of the linear space theory determines easy-to-use means of representing arbitrary vectors of any given value by monotonic vectors of greater, similar or lesser significance, and also converting monotonic functions to the threshold form.

4. Synthesis of Logic and Threshold Elements

In linear algebra, the logic function is represented by a weighted algebraic sum of terms. As a whole the variables, terms and function take values from the set of numbers 0, 1, ..., k - 1. The algebraic sum is realized by operations of the linear space, and individual terms - depending on the choice made - by other operations chosen for technical, technological or operational reasons. In this article, *"Truncated Difference"* and *"Comparison"* are used as such operations. In the future, the developers of linear LSIs can be motivated to choose other operations. Thus, the process and the possibilities of logic synthesis will be further demonstrated on the bases based on the *truncated difference* and *comparison* operations.

To solve the problems of logic synthesis in linear algebra, first, it is necessary to solve the problem of the formation of bases.

In Boolean algebra, 17 previously defined functionally complete systems of logic functions of two arguments are known [34]. Selecting any of them, you can get its linear analog and

perform a logic synthesis of the circuit that implements the given logic function according to the established algorithm. Therefore, it is possible to construct bases as a mapping of Boolean functionally complete systems to a linear space. For *truncated difference* and *comparison* operations, the bases can be formed directly.

To form the bases, we need:

- a set of operations;

- a set of generating functions.

The following operation are considered as the basic ones:

- logical operations in their traditional understanding;

- operation *"truncated difference";*

- operation *"comparison".*

Constants 1, logical variables x_1, x_2, …, x_n and their totals subjected to the effect of basic operations are used as generators for the formation of bases.

The following methods of logic synthesis of digital structures are considered:

- direct synthesis in a given basis (logic synthesis in the linear space);

- synthesis with preliminary transformation of the value of the given function (for example, a multivalued two-digit implementation);

- synthesis with preliminary expansion of the given function into the algebraic sum of monotonic functions (threshold synthesis).

The first method is the most obvious and consists in the representation of the realized logic function in the chosen basis.

The second method presupposes a preliminary representation of the realized function in the form of an algebraic sum of logic functions larger, smaller or equal to the original value. The methods of this transformation have been described in the previous section.

Finally, the third method consists of the preliminary representation of the realized function by the algebraic sum of monotonic logic functions. In this case, each monotonic function is realized by a single threshold element.

Note that the synthesis of digital structures with increasing of their complexity, as usual, requires the involvement of methods of system engineering.

A. Basic operations of logic synthesis. The set of operations most often used to represent logic functions will be considered as basic operations of logic synthesis. In Boolean algebra, such operations can be considered as operations of the basic functional system AND (&), OR (∨), ($\bar{}$), and also the operations "modulo 2 sum" (⊕), "logic equivalence" (~) "dual stroke" (↓), "Sheffer stroke" (|), "prohibition with respect to x_1 (x_2)" $(x_1\bar{x}_2)$, "implication from x_1 to x_2)" $(x_1 \vee \bar{x}_2)$, majority operation $x_1 \# x_2 \# x_3$.

By *basic operations of linear algebra* we mean arithmetic operations (operations of the linear space), as well as operations used to form bases. We classify the current realizations of logical operations: the difference module, *the truncated difference*, the arithmetic sum and the difference, the multiplication by a constant coefficient, the operations of changing the sign to them. Perhaps there are other arithmetic operations suitable for the use in linear logic synthesis, but they are not considered in this article.

Since our task is logic synthesis in linear algebra, we start with the logic synthesis of basic operations of linear algebra.

Algebraic sum. The operation is realized by the assembly connection of conductors, over which the currents flow. The conditional graphic representation of the operation of the algebraic sum is shown in Figure 2.

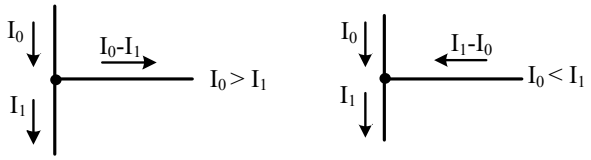


Figure 2: Conditional graphical representation of the algebraic addition of currents.

The arithmetic sum differs from the algebraic one in that in the latter case the signals of only one sign (i.e. only inflowing or only flowing out) are fed to the input of the element. From the output, the sum of the signals is removed. The positive direction of the input currents is the current direction to the connection point Σ_1, i.e. the positive current flowing into the connection point. The negative direction of the input currents is the direction of the current from the connection point of the conductors Σ_1, i.e. negative - the current flowing from the connection point. For the output (resulting) current, the positive direction is the direction from the connection point Σ_1, and the negative direction - to the connection point Σ_1,

The number of positive and negative inputs is determined by the function being implemented. Since the logical variables take values from the positive semi-axis of the numerical axis, when subtracting, it is necessary to meet the condition that the total sum of the negative summands (current quanta) doesn't exceed the total sum of the positive terms of the sum. Otherwise, instead of the algebraic sum, the *truncated difference* operation is performed, in which the subtraction from zero is a logical component of this operation and which carries the meaning of the *comparison* operation with zero (Figure 3)

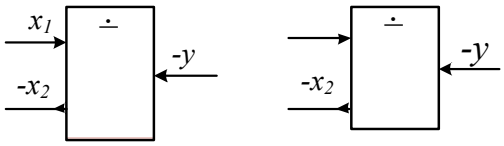
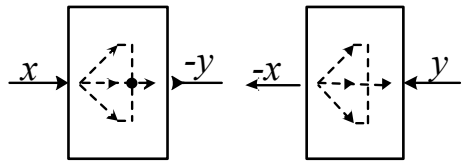


Figure 3: Conditional graphic representation o*f the truncated difference* operation.

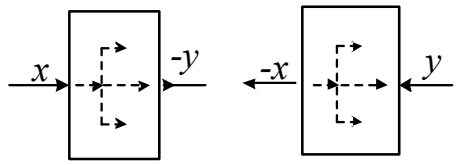
The operation of multiplying (dividing) by a constant coefficient consists in multiplying the input signal by several outputs and then combining the outputs of the multiplied signal (when multiplying) or outputting some part of the input signal (when dividing), as shown in Figure 4.

The comparison operation is used to determine the relationship between the compared variables. It can be performed in two forms: in the relative form, i.e. in the form of determining the difference in the values of the variables, or in the absolute form, i.e. in the form of the "more-less" comparison itself. The conditional graphic

symbols of the comparison elements in these forms are shown in Figure 5.

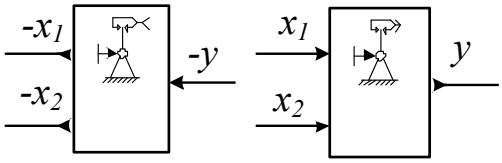


a

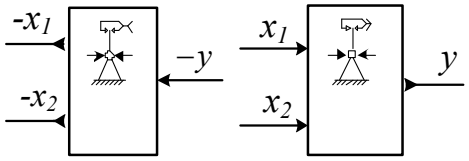


b

Figure 4: Conditional graphic representation of the operation of multiplication *(a)* and division *(b)*.



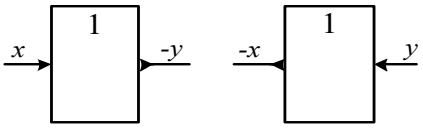
a



b

Figure 5: Conditional graphic representation of the *comparison* operation: relative *(a)* and absolute *(b)*.

The operation of *changing the sign* of terms allows changing the sign of the term, thereby transforming it from the summand to the subtrahend and vice versa. It is performed by the current inverter, the conditional graphic representation of which is shown in Figure 6.



a *b*

Figure 6: Conditional graphic representation of the operation of changing the sign.

The operation of changing the sign can functionally be combined with the operations of multiplication and division by the constant coefficient.

B. Formation of bases. The bases being formed are bases of the linear space, and according to the form they are sets of variables and their various combinations (terms) combined by the selected operations. We can approach to the formation of bases from different sides. For the two-valued synthesis, as it was shown above, it is possible, for example, to construct bases equivalent to 17 functionally complete systems (logical-arithmetic), and then create their analogs in linear algebra.

To form the multivalued bases, we can use the cuts of multivalued variables (10 combinations of such cuts are proposed in [18]), and then construct their analogs in linear algebra. Such a solution is also applicable for the two-valued bases. The problem of forming the bases is solved uniformly for the cuts of any type, so we demonstrate its solution on the basis of the upper cuts of the first level, which are analogs of the functionally complete system AND, NOT (OR, NOT).

The bases of this type for $k = 2$, $n = 2$ and $n = 3$, and for $k = 3$, $n = 2$ are given below. For each basis, the basic and inverse matrices are presented. Two-digit operations & and ∨ and their multivalued analogs $min(x_1x_2)$ and $max(x_1x_2)$ are used as generating operations. Therefore, the bases of this type are called logical-arithmetic.

The two-valued logical-arithmetic bases of two variables have the form:

- conjunctive (AND, NOT)

$$\&(\mathrm{B1},\mathrm{B1})^{(2)} = \begin{bmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \\ x_1^{(1)}x_2^{(1)} \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1\&x_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$

$$\left(\&(\mathrm{B1},\mathrm{B1})^{(2)}\right)^{-1} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & 1 & 1 \end{bmatrix};$$

- disjunctive (OR, NOT)

$$\vee\,(\mathrm{B1},\mathrm{B1})^{(2)} = \begin{bmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \\ x_1^{(1)} \vee x_2^{(1)} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix};$$

$$(\vee\,(\mathrm{B1},\mathrm{B1})^{2})^{-1} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & -1 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 1 & 1 & -1 \end{bmatrix}.$$

The two-valued logical-arithmetic bases of three variables have the similar form:

$$\&(\mathrm{B1},\mathrm{B1},B1)^{(2)} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_1x_2 \\ x_1x_3 \\ x_2x_3 \\ x_1x_2x_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

$$\left(\&(\mathrm{B1},\mathrm{B1},B1)^{(2)}\right)^{-1} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & -1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

$$\vee(\mathrm{B1},\mathrm{B1},B1)^{(2)}=\begin{bmatrix}1\\x_1\\x_2\\x_3\\x_1\vee x_2\\x_1\vee x_3\\x_2\vee x_3\\x_1\vee x_2\vee x_3\end{bmatrix}=\begin{bmatrix}1&1&1&1&1&1&1&1\\0&1&0&1&0&1&0&1\\0&0&1&1&0&0&1&1\\0&0&0&0&1&1&1&1\\0&1&1&1&0&1&1&1\\0&1&0&1&1&1&1&1\\0&0&1&1&1&1&1&1\\0&1&1&1&1&1&1&1\end{bmatrix};$$

$$\left(\vee(\mathrm{B1},\mathrm{B1},B1)^{(2)}\right)^{-1}=\begin{bmatrix}1&0&0&0&0&0&0&-1\\0&0&0&0&0&0&-1&1\\0&0&0&0&0&-1&0&1\\0&0&0&-1&0&1&1&-1\\0&0&0&0&-1&0&0&1\\0&0&-1&0&1&0&1&-1\\0&-1&0&0&1&1&0&-1\\0&1&1&1&-1&-1&-1&1\end{bmatrix}.$$

The three-valued logical-arithmetic analogs of these bases look like this:

- conjunctive

$$min(B1,B1)^{(3)}=\begin{bmatrix}1\\x_1^{(1)}\\x_1^{(2)}\\x_2^{(1)}\\min\left(x_1^{(1)},x_2^{(1)}\right)\\min\left(x_1^{(2)},2x_2^{(1)}\right)\\x_2^{(2)}\\min\left(2x_1^{(1)},x_2^{(2)}\right)\\min\left(x_1^{(2)},x_1^{(2)}\right)\end{bmatrix}=\begin{bmatrix}1&1&1&1&1&1&1&1&1\\0&1&1&0&1&1&0&1&1\\0&1&2&0&1&2&0&1&2\\0&0&0&1&1&1&1&1&1\\0&0&0&0&1&1&0&1&1\\0&0&0&0&1&2&0&1&2\\0&0&0&1&1&1&2&2&2\\0&0&0&0&1&1&0&2&2\\0&0&0&0&1&1&0&1&2\end{bmatrix};$$

$$\left(min(B1,B1)^{(3)}\right)^{-1}=\begin{bmatrix}1&-1&0&-1&1&0&0&0&0\\0&0&0&2&-2&0&-1&1&0\\0&0&0&-1&1&0&1&-1&0\\0&2&-1&0&-2&1&0&0&0\\0&0&0&0&2&-1&0&-1&1\\0&0&0&0&0&0&0&1&-1\\0&-1&1&0&1&-1&0&0&0\\0&0&0&0&0&1&0&0&-1\\0&0&0&0&-1&0&0&0&1\end{bmatrix};$$

- disjunctive

$$max(B1,B1)^{(3)}=\begin{bmatrix}1\\x_2^{(1)}\\x_2^{(2)}\\x_1^{(1)}\\max\left(x_2^{(1)},x_1^{(1)}\right)\\max\left(x_2^{(2)},2x_1^{(1)}\right)\\x_1^{(2)}\\max\left(2x_2^{(1)},x_1^{(2)}\right)\\max\left(x_2^{(2)},x_1^{(2)}\right)\end{bmatrix}=\begin{bmatrix}1&1&1&1&1&1&1&1&1\\0&0&0&1&1&1&1&1&1\\0&0&0&1&1&1&2&2&2\\0&1&1&0&1&1&0&1&1\\0&2&2&1&2&2&2&2&2\\0&0&0&0&1&1&0&2&2\\0&1&2&0&1&2&0&1&2\\0&1&2&2&2&2&2&2&2\\0&1&2&1&1&2&2&2&2\end{bmatrix};$$

$$\left(max(B1,B1)^{(3)}\right)^{-1}=\begin{bmatrix}1&0&0&0&-1&0&0&0&0\\0&0&0&0&2&0&0&-1&0\\0&-1&0&0&-1&0&0&1&0\\0&0&0&0&2&-1&0&0&0\\0&0&0&0&-2&1&0&1&-1\\0&2&-1&0&0&0&0&-1&1\\0&0&0&-1&-1&1&0&0&0\\0&0&0&2&0&-1&-1&0&1\\0&-1&1&-1&1&0&1&0&-1\end{bmatrix}.$$

To obtain linear analogs of the above logical-arithmetic bases, we can use the relations that establish the connection between logical operations, on the one hand, and with the operations used to form the bases of linear space, on the other hand. For the bases of the considered type these relations are as follows:

- logical operations - *truncated difference* operation

$$x^{(i)}=i\dot{-}(i\dot{-}x)=x\dot{-}(x\dot{-}i);$$

$$x^{(i)}-x^{(j)}=x^{(ij)}=(x\dot{-}j)\dot{-}(x\dot{-}i);$$

$$min(x_1,x_2)=x_1\dot{-}(x_1\dot{-}x_2)=x_2\dot{-}(x_2\dot{-}x_1);$$

$$max(x_1,x_2)=x_1+(x_2\dot{-}x_1)=x_2+(x_1\dot{-}x_2).$$

- logical operations - *comparison* operation

$$x^{(i)}=\sum_{j=0}^{i-1}(x>j);$$

$$x^{(i)}-x^{(j)}=x^{(ij)}=\textstyle\sum_{t=1}^{i-j}P_t(x>t);$$

$$min(x_1,x_2)=P\{[(x_1>0)+(x_2>0)]>1\}+$$
$$+P\{[(x_1>1)+(x_2>1)]>1\};$$

$$max(x_1,x_2)=P\{[(x_1>0)+(x_2>0)]>0\}+$$
$$+P\{[(x_1>1)+(x_2>1)]>0\}.$$

Using the above relations, we can easily obtain linear analogs of the above bases:

- conjunctive $k = 2$, $n = 2$ and $n = 3$

$$\&(\mathrm{B1},\mathrm{B1})^{(2)}=\begin{bmatrix}1\\x_1^{(1)}\\x_2^{(1)}\\x_1^{(1)}x_2^{(1)}\end{bmatrix}=\begin{bmatrix}1\\x_1\\x_2\\(x_1+x_2)\dot{-}1\end{bmatrix}=\begin{bmatrix}1&1&1&1\\0&1&0&1\\0&0&1&1\\0&0&0&1\end{bmatrix};$$

$$\left(\&(\mathrm{B1},\mathrm{B1})^{(2)}\right)^{-1}=\begin{bmatrix}1&-1&-1&1\\0&1&0&-1\\0&0&1&-1\\0&1&1&1\end{bmatrix};$$

$$\&(\mathrm{B1},\mathrm{B1},B1)^{(2)}=\begin{bmatrix}1\\x_1\\x_2\\x_3\\x_1x_2\\x_1x_3\\x_2x_3\\x_1x_2x_3\end{bmatrix}=\begin{bmatrix}1\\x_1\\x_2\\x_3\\(x_1+x_2)\dot{-}1\\(x_1+x_3)\dot{-}1\\(x_2+x_3)\dot{-}1\\(x_1+x_2+x_3)\dot{-}2\end{bmatrix}=$$

$$=\begin{bmatrix}1&1&1&1&1&1&1&1\\0&1&0&1&0&1&0&1\\0&0&1&1&0&0&1&1\\0&0&0&0&1&1&1&1\\0&0&0&1&0&0&0&1\\0&0&0&0&0&1&0&1\\0&0&0&0&0&0&1&1\\0&0&0&0&0&0&0&1\end{bmatrix};$$

$$\left(\&(\mathrm{B1},\mathrm{B1},B1)^{(2)}\right)^{-1} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & -1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

- disjunctive $k = 2$, $n = 2$ and 3

$$\vee(\mathrm{B1},\mathrm{B1})^2 = \begin{bmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \\ x_1^{(1)} \vee x_2^{(1)} \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1 + (x_2 \div x_1) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix},$$

$$(\vee(\mathrm{B1},\mathrm{B1})^2)^{-1} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & -1 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 1 & 1 & -1 \end{bmatrix},$$

$$\vee(\mathrm{B1},\mathrm{B1},B1)^{(2)} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_1 \vee x_2 \\ x_1 \vee x_3 \\ x_2 \vee x_3 \\ x_1 \vee x_2 \vee x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ 1 \div [1 \div (x_1 + x_2)] \\ 1 \div [1 \div (x_1 + x_3)] \\ 1 \div [1 \div (x_2 + x_3)] \\ 1 \div [1 \div (x_1 + x_2 + x_3)] \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

$$\left(\vee(\mathrm{B1},\mathrm{B1},B1)^{(2)}\right)^{-1} =$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & 0 & 1 & -1 \\ 0 & -1 & 0 & 0 & 1 & 1 & 0 & -1 \\ 0 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{bmatrix}.$$

Similarly, we obtain linear analogs of multivalued bases:

- conjunctive, $k = 3$, $n = 2$

$$min(B1,B1)^{(3)} = \begin{bmatrix} 1 \\ 1 \div (1 \div x_1) \\ x_1 \\ 1 \div (1 \div x_2) \\ [1 - (1 \div x_1)] - (1 \div x_2) \\ x_1 \div 2(1 \div x_2) \\ x_2 \\ 2[1 - (1 \div x_1)] - (2 \div x_2) \\ x_1 \div (x_1 \div x_2) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 2 \end{bmatrix};$$

$$\left(min(B1,B1)^{(3)}\right)^{-1} =$$

$$= \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & -1 & 0 & -2 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix};$$

- disjunctive

$$max(B1,B1)^{(3)} = \begin{bmatrix} 1 \\ 1 \div (1 \div x_1) \\ x_1 \\ 1 \div (1 \div x_2) \\ 1 \div (x_1 + x_2) \\ x_2 + \{2[1 \div (1 \div x_1)] \div x_2\} \\ x_2 \\ x_1 + \{2[1 \div (1 \div x_2)] \div x_1\} \\ x_1 + (x_2 \div x_1) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 0 & 2 & 2 & 1 & 2 & 2 & 2 & 2 & 2 \\ 0 & 1 & 2 & 1 & 1 & 2 & 2 & 2 & 2 \end{bmatrix};$$

$$\left(max(B1,B1)^{(3)}\right)^{-1} =$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & -2 & 2 & 0 & 1 & -2 \\ 0 & 0 & 0 & 2 & 0 & -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & -1 & 1 & 0 & 1 & -1 \\ 0 & 2 & -1 & 0 & 0 & -1 & 0 & -1 & 2 \\ 0 & -1 & 1 & -1 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

C. Logic synthesis of basic functions. The resulted conditional graphic designations of operations of the linear space can be used for graphic representation of basic Boolean logic operations in linear algebra and construct functional schemes of logic elements on their base. For this it is sufficient:

- to choose the basis from the listed above;

- to use the previously described technique of logic synthesis in linear algebra.

We choose a conjunctive version of the chosen basis as an object of logic synthesis in which we perform a logic synthesis of the two-valued and three-valued functions of circular shift

$$(x_1 \& x_2) \oplus 1 = [1\ 1\ 1\ 0],$$

and

$$min(x_1, x_2) \oplus 1 = [1\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 0],$$

and

$$min(x_1, x_2) \ominus 1 = [2\ 2\ 2\ 2\ 0\ 0\ 2\ 0\ 1].$$

In accordance with the above, the first action is to obtain the expansion vector of the value vector of the function in terms of the basis. Multiplying the vectors written above by the columns of the inverse matrix of the bases $\&(\mathrm{B1}, \mathrm{B1})^{(2)}$ and $min(B1, B1)^{(3)}$ results in the following:

$$w[(x_1 \& x_2) \oplus 1] = [0\ 1\ 1\ -2];$$

$$w[min(x_1, x_2) \oplus 1] \Rightarrow [1\ 0\ 0\ 0\ 3\ 0\ 0\ 0\ -2];$$

$$w[min(x_1, x_2) \ominus 1] \Rightarrow [2\ 0\ 0\ 0\ -3\ 0\ 0\ 01].$$

Weighing the basis vectors with respect to the obtained coefficients, we obtain expressions of the logic functions in the given basis

$$(x_1 \& x_2) \oplus 1 = 1 - (x_1 + x_1 - 2x_1 \& x_1);$$

$$min(x_1, x_2) \oplus 1 = 1 + 3min\left(x_1^{(1)}, x_2^{(1)}\right) - 2min\left(x_1^{(2)}, x_2^{(2)}\right);$$

$$min(x_1, x_2) \ominus 1 = 2 - 3min\left(x_1^{(1)}, x_2^{(1)}\right) + min\left(x_1^{(2)}, x_2^{(2)}\right).$$

The functional schemes corresponding to these expressions are shown in Figure 7 *a, b, c.*

To obtain the equivalent expressions of the functions in linear analogs of the logical-arithmetic basis under consideration, one can proceed in two ways:

- replace the used basis vectors with their linear analogs in the expressions obtained;

- use the above relations between logic and linear operations and convert the expressions written above into the linear form.

In all three cases, we obtain the following results:

$$(x_1 \& x_2) \oplus 1 = 1 - \{x_1 + x_1 - 2[(x_1 + x_1) \div 1]\};$$

$$min(x_1, x_2) \oplus 1 = 1 + 3\{[1 \div (1 \div x_1)] \div (1 \div x_2)\} - -2\{[(x_1 \div 1) + (x_2 \div 1)] \div 1\};$$

$$min(x_1, x_2) \ominus 1 = 2 - 3\{[1 \div (1 \div x_1)] \div (1 \div x_2)\} + +\{[(x_1 \div 1) + (x_2 \div 1)] \div 1\}.$$

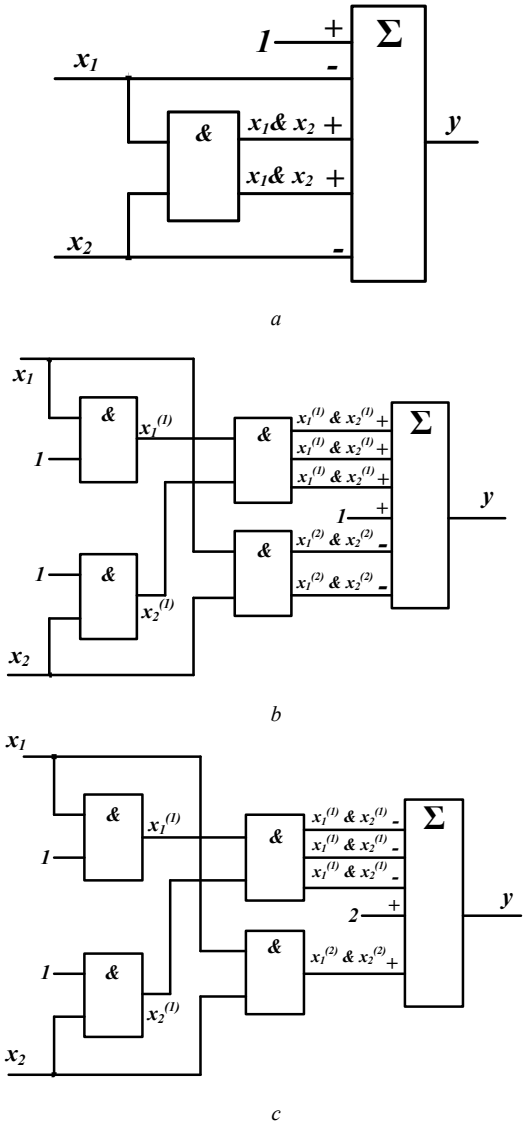


Figure 7: Functional schemes of the logical-arithmetic realization of the circular shift operation: a - two-valued, b - three-valued with a shift to the right, c - three-valued with a shift to the left.

The functional schemes corresponding to these expressions are shown in Figure 8, *a, b, c.*

The difference between the linear representations obtained from the arithmetic-logical representation is the possibility of physical realization. The authors of the article have obtained more than 25 patents for the circuit implementation of two-valued and three-valued logic circuits.

D. Expansion of the arbitrary function into the algebraic sum of monotonic functions of the same value. To obtain the representation of the functions under consideration by the algebraic sum of monotonic functions, we expand them into the algebraic sum of monotonic functions within each value. For the two-valued function, the sequence graph of nondecreasing components has the form shown in Figure 9.

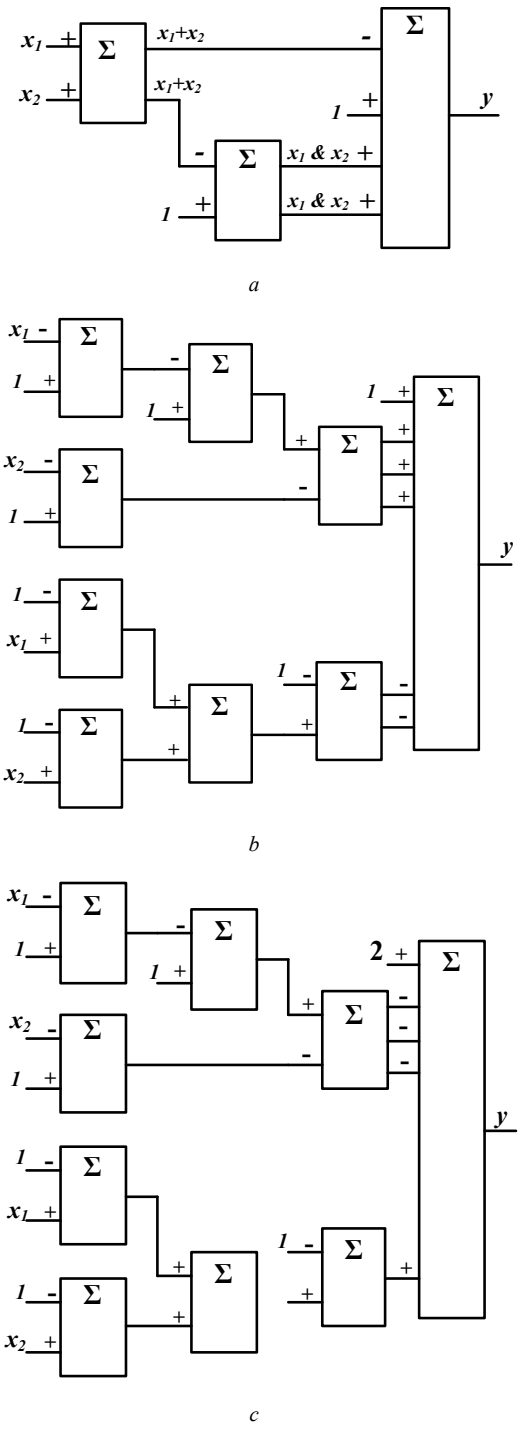


Figure 8: Functional schemes of the linear realization of the circular shift operation based on the *truncated difference*: *a* - two-valued, *b* - three-valued with a shift to the right, *c* - three-valued with a shift to the left.

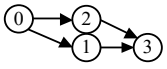


Figure 9: Sequence graph of nondecreasing components of the two-valued functions of two variables.

The structure of the expansion function of the equivalence $\bar{x}_1\bar{x}_2 \vee x_1x_2 = (x_1\&x_2) \oplus 1$ into the algebraic sum of monotonic functions is given in Figure 10.

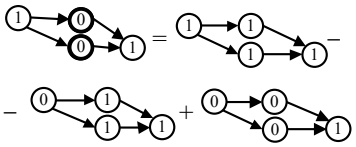


Figure 10: The structure of the expansion of the two-valued function $\bar{x}_1\bar{x}_2 \vee x_1x_2$ into the algebraic sum of monotonic functions.

Here and below, the double circle marks the terms the values of which violate the monotonicity of the function.

It follows from the structure of the expansion that the representation of the function under consideration by the algebraic sum of monotonic functions has the following form

$$f(\tilde{x}^{(2)}) = (x_1\&x_2) \oplus 1 = 1 - x_1 \vee x_2 + x_1\&x_2.$$

Carrying out the similar transformations for the three-valued functions, the sequence graph of nondecreasing components of which is shown in Figure 11.

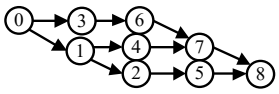


Figure 11: Sequence graph of nondecreasing components of the three-valued functions of two variables.

we obtain the following representations of the functions under consideration by the algebraic sum of monotonic functions:

- for $min(x_1, x_2)\oplus 1$.

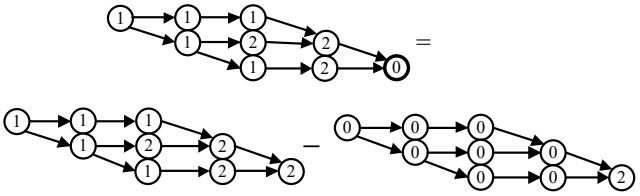


Figure 12: The structure of the expansion of the three-valued function $min(x_1, x_2)\oplus 1$ into the algebraic sum of monotonic functions.

$$min(x_1, x_2)\oplus 1 = 1 - min\left(x_1^{(1)}, x_2^{(1)}\right) -$$
$$-2min\left(x_1^{(2)}, x_2^{(2)}\right) + min\left(x_1^{(1)}, x_2^{(1)}\right);$$

– for $min(x_1, x_2) \ominus 1$.

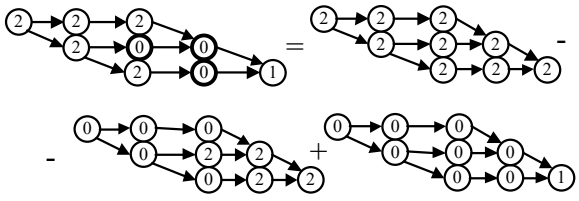


Figure 13: The structure of the expansion of the three-valued function $min(x_1, x_2) \ominus 1$ into the algebraic sum of monotonic functions.

$$min(x_1, x_2) \ominus 1 = 2 - 2min\left(x_1^{(1)}, x_2^{(1)}\right) +$$
$$+min\left(x_1^{(2)}, x_2^{(2)}\right).$$

To obtain the equivalent representations of the expansion functions in linear analogs of the basis under consideration, we can proceed in the same way as with the functions themselves.

Expansion of the arbitrary function into the algebraic sum of smaller value.

We consider this operation with the help of the above transformation of the three-valued function $min(x_1, x_2)\oplus 1$, by unitary and weighted coding of the values of variables and functions.

In these cases, the encoding of variables and functions looks like this:

- unitary coding

x	x_2	x_1
0	0	0
1	0	1
2	1	1

Hence, it follows that for the unitary coding

$x = x_1 + x_2;$

– weighted coding

x	x_2	x_1
0	0	0
1	0	1
2	1	0

It follows that at the weighted coding

$$x = x_1 + 2x_2.$$

Using the basic and its inverse matrix, we obtain the expressions for the two-valued expansion functions from the multivalued arguments:

- for the unitary coding

$$f_1 = 1 + min\left(x_1^{(1)}, x_2^{(1)}\right) - min\left(x_1^{(2)}, x_2^{(2)}\right);$$

$$f_2 = 2min\left(x_1^{(1)}, x_2^{(1)}\right) - min\left(x_1^{(2)}, x_2^{(2)}\right);$$

$$f = f_1 + f_2;$$

- for the weighted coding

$$f_1 = 1 - min\left(x_1^{(1)}, x_2^{(1)}\right);$$

$$f_2 = 2min\left(x_1^{(1)}, x_2^{(1)}\right) - min\left(x_1^{(2)}, x_2^{(2)}\right);$$

$$f = f_1 + 2f_2.$$

The representation of the three-valued function $min(x_1, x_2)\oplus 1$ by the two-valued ones for both versions of coding is given in Table 2.

Table 2 – The expansion of the three-valued function $min(x_1, x_2)\oplus 1$ into the sum of the two-valued ones

x_2	x_1	Unitary		Weighted	
		f_2	f_1	f_2	f_1
0	0	0	1	0	1
0	1	0	1	0	1
0	2	0	1	0	1
1	0	0	1	0	1
1	1	1	1	1	0
1	2	1	1	1	0
2	0	0	1	0	1
2	1	1	1	1	0
2	2	0	0	0	0

Replacing the three-valued variables with the two-valued ones in accordance with the rules given above, we obtain the two-valued expansion functions in the following form:

- for the unitary coding

$$f_1 = 1 - x_{11}x_{12}x_{21}x_{22};$$

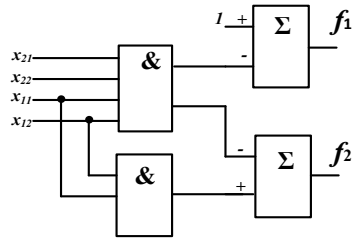
$$f_2 = x_{11}x_{21} - x_{11}x_{12}x_{21}x_{22};$$

- for the weighted coding

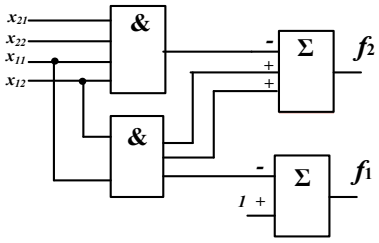
$$f_1 = 1 - x_{11}x_{21};$$

$$f_2 = 2x_{11}x_{21} - x_{11}x_{12}x_{21}x_{22}.$$

The functional schemes of the linear realization of the circular shift operation by the expansion into the sum of the two-valued functions are shown in Figure 14.



a



b

Figure 14: Functional schemes of linear realization of the circular shift operation by the expansion into the sum of the two-valued functions: *a* - on the basis of the unitary coding, *b* - on the basis of the weighted coding.

For the threshold representation and the subsequent threshold realization, it suffices to transform the equations obtained above into the threshold form. After completing this transformation, we get:

- for the unitary coding

$$f_1 = 1 > [(x_{11}+x_{12}+x_{21}+x_{22}) > 3];$$

$$f_2 = [(x_{11}+x_{21}) > 1] - [(x_{11}+x_{12}+x_{21}+x_{22}) > 3];$$

- for the weighted coding

$$f_1 = 1 - [(x_{11}+x_{21}) > 1];$$

$$f_2 = 2[(x_{11}+x_{21}) > 1] - [(x_{11}+x_{12}+x_{21}+x_{22}) > 3].$$

The functional schemes shown in Figure 15 correspond to these equations.

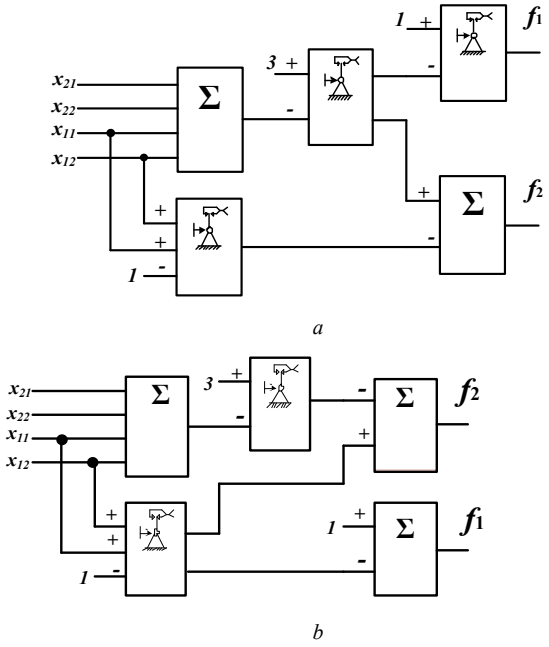


Figure 15: Functional schemes of linear realization of the circular shift operation by the expansion into the sum of the threshold two-valued functions: *a* - on the basis of the unitary coding, *b* - on the basis of the weighted coding.

The considered complex of transformations of logic functions in linear algebra proves useful in the design of digital structures for various applications.

5. Circuitry of the Linear Logic and Threshold Elements

A. Basic functional nodes of digital logic elements. The implementation of the mathematically predetermined set of linear operations over the current signals requires the corresponding set of hardware. The circuit implementation of digital signal transformation functions based on the mathematical tool of linear algebra can be reduced to performing a relatively simple set of operations over the current signals. These operations include:

- conversion of the standard logic signals into the binary current signals matched with a reference current quantum I_0;

- formation of the multivalued (non-binary) algebraic sums of current signals;

- performing the *comparing* operations of the received sums with the given levels of the reference currents.

These operations are typical for analog microcircuitry, therefore the main nodes of various functional elements can be constructed on the basis of the widely used analog structures. In addition, TTL circuitry and CMOS circuitry of these operations completely coincide.

The reference signal driver. It is designed to generate voltages that provide operational modes of elements of digital circuits synthesized in linear algebra. The schematic configuration of the driver is shown in Figure 16.

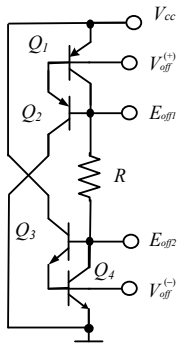


Figure 16: Reference signal driver.

The symmetric structure of the reference signal source is necessary for realizing the operations of converting "inflowing" and "flowing out" currents when creating the algebraic sums of variables in the mathematical tool of linear algebra. The only current-stabilizing two-terminal network, in particular case the resistor *R*, determines the levels of all reference signals. $V_{off}^{(+)}$ and $V_{off}^{(-)}$ are reference for setting the operating modes of current mirrors, and E_{off1} and E_{off2} - for setting the offset in the comparators of the currents. It can also be replaced by some semiconductor structure, and then the circuit becomes completely homogeneous and highly technological. It is also possible to make the two-terminal network *R* as an external element, which will allow changing the power consumption and the associated characteristics of the LSI during the debugging process.

The reference current source (RCS, Figure 17). The operating mode of the RCS is set by the reference voltages $V_{off}^{(+)}$ for the current sources and $V_{off}^{(-)}$ for the current sinks. The problem of constructing the RCS for digital circuits synthesized in linear algebra is completely similar to their problems in analog circuitry.

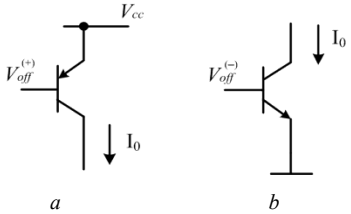


Figure 17: RCS of: a – the current current, b – the current sink.

The current follower (logic interpretation - direction converter). It is intended for the coordination of current directions at their algebraic summation. The schematic configurations of some versions of current followers in TTL circuitry are shown in Figure 18.

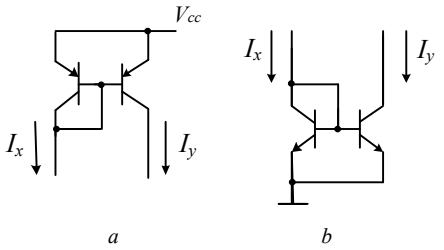


Figure 18: Current direction converter of: *a* – the sink current, *b* - the source current.

Circuitry of CMOS converters of the current direction is similar.

Algebraic current adder. It is an wired pack of the outputs of current mirrors with current directions determined by the mathematical representation of the realized logic function. To agree the operating modes of subsequent elements, it is provided with a buffer stage. The schematic configuration of the algebraic adder is shown in Figure 19.

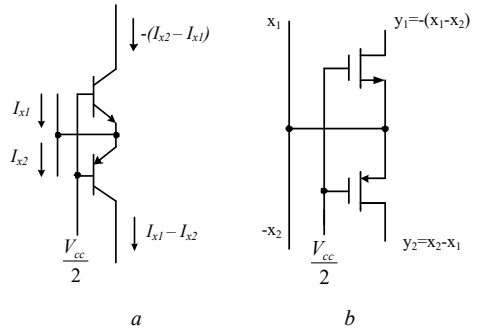


Figure 19: Algebraic current adder:
a - TTL circuit, *b* - CMOS circuit.

Current comparators. They are designed to determine the excess value of one of the input currents over the other. The schematic configurations of current comparators are given in Figures 20 and 21.

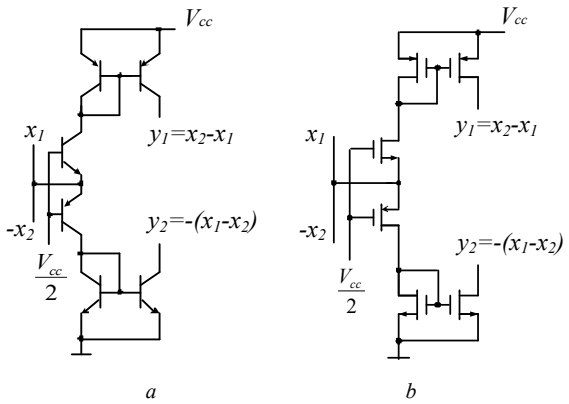


Figure 20: Comparator based on the *truncated difference*:
a - bipolar circuit, *b* - CMOS circuit.

In the circuit in Figure 21, the excess of one of the input currents (I_{x1}) over the other (I_{x2}) is determined by subtracting the second current from the first. One of the compared currents must be a source current and the second one – a current sink. Such a comparator is applicable for any logic value.

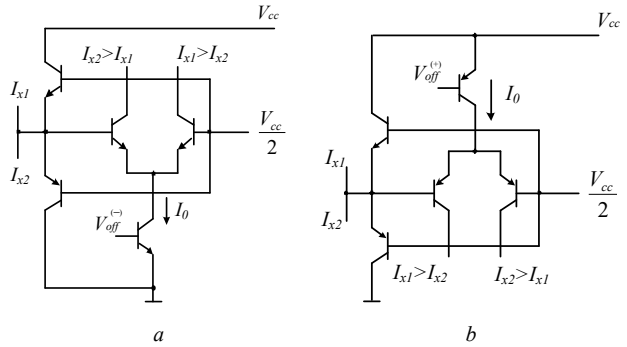


Figure 21: Current comparator based on the comparison: *a* - for current sinks, *b* - for source currents.

In the circuit of Figure 21, *a* at $I_{x1} > I_{x2}$ the left-hand transistor of the differential stage (DS) is closed and the current of the current mirror "leaves" through the right transistor of the DS. At $I_{x1} < I_{x2}$, the right transistor of the DS is closed and the current of the current mirror "goes" through the left transistor. The output current of the DS I_0 is the current sink.

In the circuit in Figure 21, *b* everything appears in a similar way, but with other current directions. Such a scheme is applicable for implementation of the two-valued digital structures.

In the previous section, the difference between the two-valued and multivalued implementations of the *comparison* operation is shown. In the latter case, it is possible to determine not only the fact, but the magnitude of the excess of one input signal over another one. In Figure 22 there is a circuit of the three-valued comparator performing such a modified *comparison* operation.

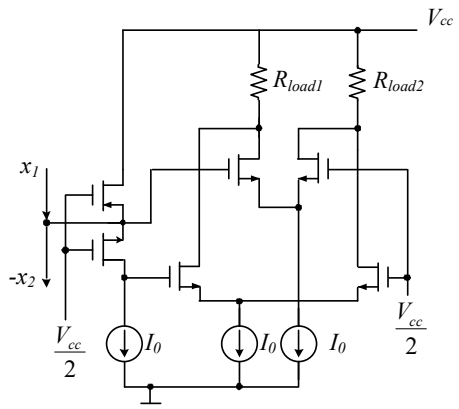


Figure 22: Modified multivalued comparator based on the comparison.

Such a scheme can be constructed for any value, increasing accordingly the number of parallelly operating DSs available to work with different values of currents.

Logic elements AND, OR, NOT. The representations of the operations of the two-valued basic functionally complete system in linear algebra using the *truncated difference* have the following form

$$x_1 \& x_2 = (x_1 + x_2) \div 1;$$

$$x_1 \vee x_2 = 1 \div [1 \div (x_1 + x_2)];$$

$$\bar{x} = 1 - x.$$

The schematic configuration of these elements are shown in Figures 23, 24 and 25, correspondingly.

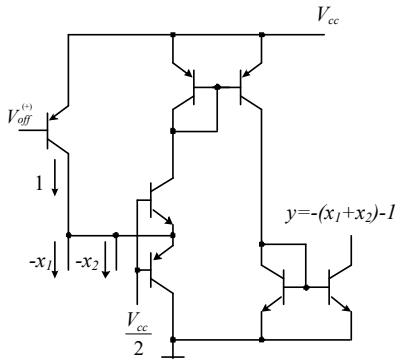


Figure 23: Schematic configuration of the element AND.

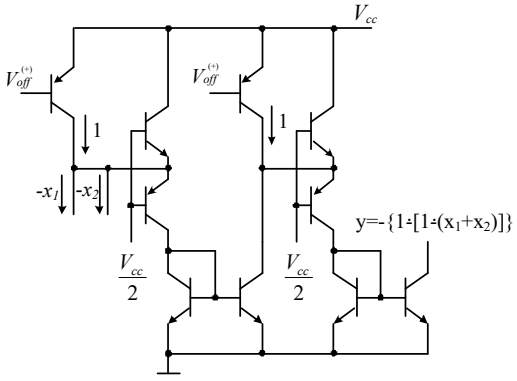


Figure 24: Schematic configuration of the element OR.

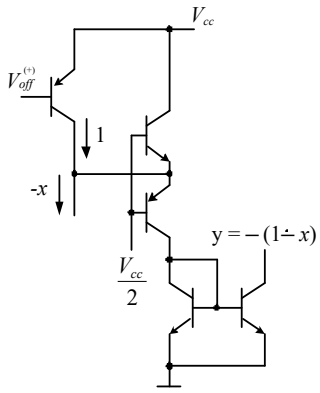


Figure 25: Schematic configuration of the element NOT.

Comparing the schemes shown in Figures 24 and 25, it is easy to see that removing the circuit of in Figure 25 from the scheme of Figure 24, we obtain the OR-NOT element.

Logic element "Inhibition". The representation of the two-valued operation in linear algebra using the *truncated difference* has the following form

$$x_1 \& \bar{x}_2 = x_1 \div x_2.$$

For the multivalued version, the two-valued inversion operation (in accordance with the accepted generalization ideology) should be replaced with the direct $min(x_1, x_2 \oplus 1)$, or inverse min $min(x_1, x_2 \ominus 1)$ cycle operation, or left unchanged; i.e. in the following form $min(x_1, 1 \div x_2)$. Other ideologies of generalization are also possible.

When using the *comparison* operation, the above expressions for operations of the functionally complete system take the following form

$$x_1 \& x_2 = (x_1 + x_2) > 1;$$

$$x_1 \vee x_2 = (x_1 + x_2) > 0;$$

$$\bar{x} = 1 > x.$$

The implementation of these functions on the basis of the *comparison* operation can be performed with the help of the universal logic element (ULE), the schematic configuration of which is shown in Figure 26.

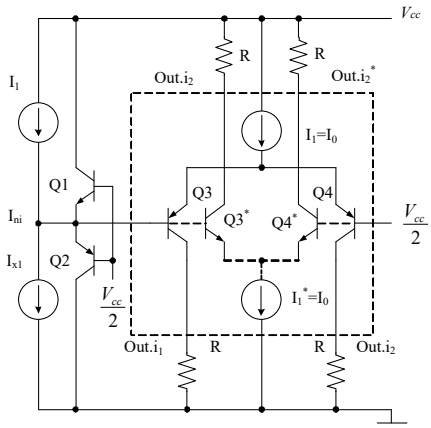


Figure 26: Basic scheme of the universal logic element.

At the inequality $I_1 < I_{x1}$, the source difference current is generated at the node In.i of the ULE. It will "go" to the emitter of the transistor Q1, increasing the voltage at the first input (In.1) of the voltage comparator up to the value $V_{OFF} + U_{be1}$, where $U_{be1} \approx 0.7V$ - voltage of the open emitter junction of the transistor Q1. In this case, the input transistors Q3 and Q4 of the voltage comparator (VC) switch to the inverse states - the collector current of the transistor Q3 becomes zero, and the transistor Q4 starts to transmit the current of the reference current source $I_2 = I_0$ to the second current output (Out.i_2) of the ULE.

Thus, depending on the difference in the numerical values of the currents I_1 and I_{x1}, the output currents of the ULE take one of two values: either it is the current of the reference current source $I_1 = I_0$ or "zero" (no current). Since the current I_1 (I_1^*) is equal to the current quantum I_0, then in one of the current outputs of the voltage comparator a standard current signal I_0 of one of the logic levels is generated, and in its second output - an inverse logic level signal. Depending on the numerical values of I_1 ($I_1 = 0{,}5I_0$, $I_1 = I_0$, $I_1 = 1{,}5I_0$) and the methods of forming the input current signals of the ULE (Figure 26), various logic functions can be performed, for example

– $\bar{x} = 1 - x$ at $y_1 = \begin{cases} I_0 \; at \; I_1 > I_x \\ 0 \; at \; I_1 \le I_x \end{cases}$;

– $x_1 \& x_2 = (x_1 + x_2) > 1$ when $y_1 = \begin{cases} I_0 \; at \; 1{,}5I_1 > I_x \\ 0 \; at \; 1{,}5I_1 \le I_x \end{cases}$;

– $x_1 \vee x_2 = (x_1 + x_2) > 0$ if $y_1 = \begin{cases} I_0 \; at \; 0{,}5I_1 > I_x \\ 0 \; at \; 0{,}5I_1 \le I_x \end{cases}$.

In the last two expressions $I_x \equiv x_1 + x_2$.

The element considered above can be used as a logic element, or as a threshold one; i.e. the generalization of the ordinary and threshold logics in the linear representation is very close. In the first case, unitary variables are fed to the inputs of the element. Besides, the number of inputs must correspond to the number of the variables. In the second case, the weighted sum of the variables should be fed to the "positive" input, and the constant equal to the calculated threshold value should be fed to the "negative" input. It is noteworthy that the described element can serve as an element of the homogeneous matrix, which can be used for matrix synthesis of the current digital structures. In addition, it is possible to construct universal current logic modules on its basis.

Similarly, the Ban operation can be expressed in terms of the *truncated difference* as

$$min(x_1, \{x_2 + 1 \dot{-} k[1 \dot{-} (x_2 \dot{-} 1)]\});$$

$$min(x_1, \{x_2 - 1 + k[1 \dot{-} x_2]\}),$$

and through the *comparison* - as

$$min(x_1, [x_2 + 1 \dot{-} k(x_2 > 2)]);$$

$$min(x_1, \{x_2 - 1 + k[1 > x_2]\}),$$

Cut formers. In fact, the cut former is an input signal limiter at the given level. As a former, the schemes that implement the operations $min(x_1, x_2)$ (upper cuts) and $max(x_1, x_2)$ (lower cuts) can be used.

Buffer output stage. The combination of the algebraic adder and the current follower (converter) can be applied as a buffer output stage. With the help of the latter, the given number of the output circuits can be arranged to provide the required output branching factor.

B. The synthesis of logical schemes. Let's consider it through the example of the two-valued and multivalued circular shift elements discussed above. They can be represented as a *single operation* in the form $min(x_1, x_2 \oplus 1)$ or $m(x_1, x_2 \ominus 1)$, or as a *compound operation*, i.e. as the sequential combination of operations working on one another $min(x_1, x_2)$ or $max(x_1, x_2)$ and operations $\oplus 1$ or $\ominus 1$, that is, in the form $min(x_1, x_2) \oplus 1$ or $min(x_1, x_2) \ominus 1$. We confine ourselves to the synthesis of the circuits using the operation $min(x_1, x_2)$.

In accordance with [17-32], the synthesis of the functional scheme corresponding to any logic function consists in multiplying the row-vector of the values of the function by the inverse basic matrix and obtaining the expansion vector of the function from the selected basis, and then recording the representation of the function as a weighted sum of the basis vectors. We perform the synthesis of the selected schemes using the basis presented below.

In the two-valued case, the element AND-NOT has an arithmetic-logical representation, described by the expression

$$\overline{x_1 \& x_2} = 1 - x_1 \& x_2,$$

which can be represented by the *truncated difference* in the following form

$$\overline{x_1 \& x_2} = 1 - [(x_1 + x_2) \dot{-} 1],$$

and by *comparison* - in the following form

$$\overline{x_1 \& x_2} = 1 > [(x_1 + x_2) > 1].$$

The schematic configurations of the elements can be constructed directly according to these expressions. Figure 27 shows an element scheme based on the *truncated difference*.

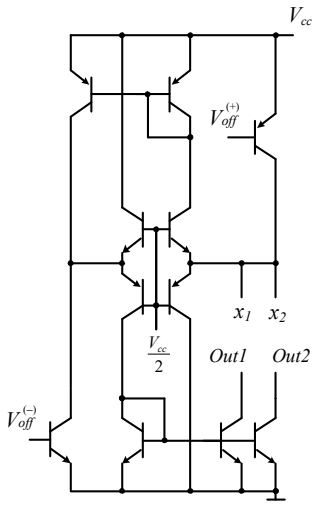


Figure 27: Schematic configuration of the element of AND-NOT based on the *truncated difference.*

We proceed similarly for the three-valued schemes in the first case (by the single operation). Using the basis

$$min(B1, B1)^{(3)} = \begin{bmatrix} 1 \\ 1 \dot{-} (1 \dot{-} x_1) \\ x_1 \\ 1 \dot{-} (1 \dot{-} x_2) \\ [1 - (1 \dot{-} x_1)] - (1 \dot{-} x_2) \\ x_1 \dot{-} 2(1 \dot{-} x_2) \\ x_2 \\ 2[1 - (1 \dot{-} x_1)] - (2 \dot{-} x_2) \\ x_1 \dot{-} (x_1 \dot{-} x_2) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 2 \end{bmatrix}$$

with the inverse matrix

$$\left(min(B1,B1)^{(3)}\right)^{-1} =$$

$$= \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & -1 & 0 & -2 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix},$$

we obtain the following results for the *single operation*:

- resolution vectors

$$w(min(x_1,x_2)\oplus 1) = [1\ 0\ 0\ 0\ 3\ 0\ 0\ 0\ -2];$$

$$w(min(x_1,x_2) \ominus 1) = [2\ 0\ 0\ 0\ -3\ 0\ 0\ 0\ 1];$$

- linear expressions of the functions

$$min(x_1,x_2) \oplus 1 = 1 + 3[1 - (1 \div x_1) - (1 \div x_2)] -$$

$$-2[x_1 \div (x_1 \div x_2)] .$$

$$min(x_1,x_2) \ominus 1 = 2 - 3[1 - (1 \div x_1) - (1 \div x_2)] +$$

$$+[x_1 \div (x_1 \div x_2)].$$

The schematic configuration of the element of the right circular shift, synthesized directly from the above expression, is shown in Figure 28.

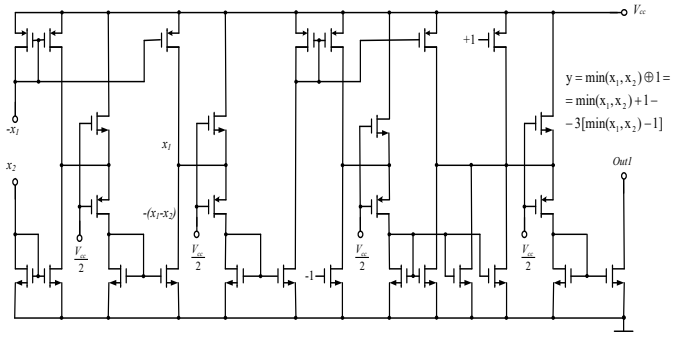


Figure 28: CMOS element of the right circular shift.

The schematic configuration of the element of the left circular shift, synthesized directly from the expression given above, is given in Figure 29.

Similarly, for the *compound operation*:

- the resolution vector $min(x_1,x_2)$

$$w\big(min(x_1,x_2)\big) = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1],$$

- linear expression of the function $min(x_1,x_2)$

$$min(x_1,x_2) = x_1 \div (x_1 \div x_2). \quad (5)$$

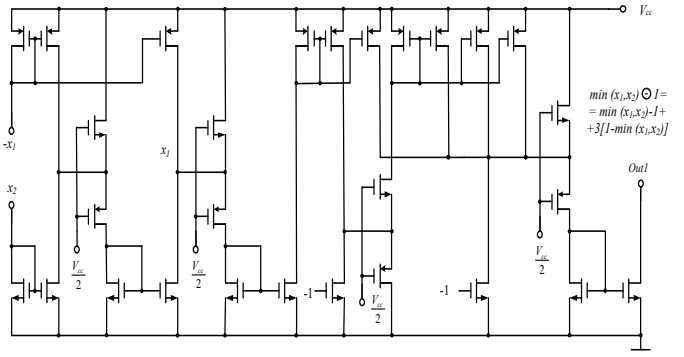


Figure 29: CMOS element of the left circular shift.

Similarly, for the *compound operation*:

- the resolution vector $min(x_1,x_2)$

$$w\big(min(x_1,x_2)\big) = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1],$$

- linear expression of the function $min(x_1,x_2)$

$$min(x_1,x_2) = x_1 \div (x_1 \div x_2). \quad (5)$$

- resolution vectors of the operations $x\oplus 1$ or $x \ominus 1$:

$$w(x\oplus 1) = [1\ 3\ -2];$$

$$w(x \ominus 1) = [2\ -3\ 1].$$

- linear expressions of the functions:

$$x \oplus 1 = 1 + [1 \div (1 \div x)] - 2x;$$

$$x \ominus 1 = 2 - 3[1 \div (1 \div x)] + x.$$

The schematic configuration of the element $min(x_1,x_2)$, synthesized directly from expression (5) discussed above, is shown in Figure 30.

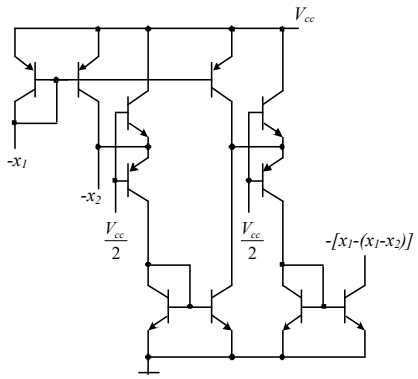


Figure 30: Schematic configuration of the element $min(x_1,x_2)$.

The schemes of the elements of the left and right cycles are shown in Figures 31 and 32.

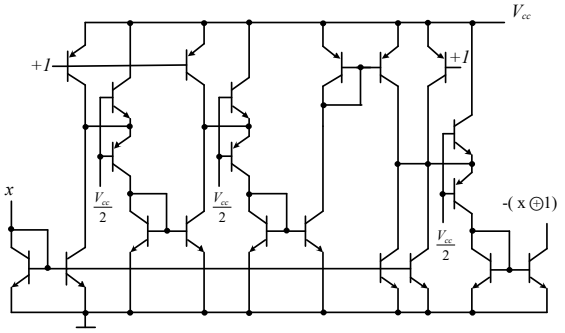


Figure 31: Element of the right cycle.

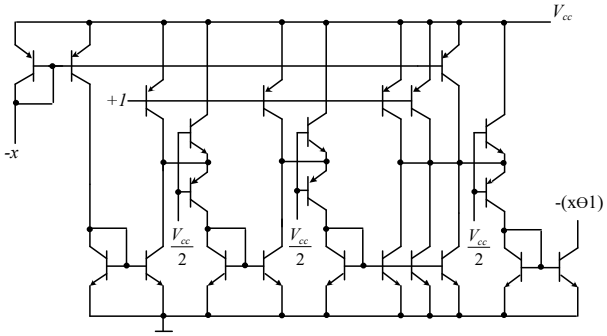


Figure 32: The element of the left cycle.

To obtain the final expression in the procedure for synthesizing a particular logic element, it is sufficient to substitute $min(x_1, x_2)$, for x in the last expression, and for the circuit implementation, - use the above implementation of the function $min(x_1, x_2)$, then connect the circuit $(x \oplus 1)$ or $(x \ominus 1)$ to its output.

C. To demonstrate the possibilities of the *sequential circuit synthesis*, we use the results of the logic element synthesis obtained above.

Boolean approach to the logic synthesis of triggers consists in supplying the memory element with a control circuit that provides a specified law for the operation of a specific type of the trigger. In the two-valued case, the memory element is a scheme of two 2AND-NOT elements covered by the positive feedback (Figure 33):

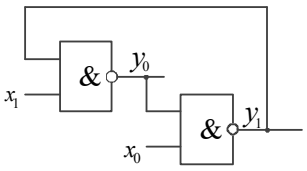


Figure 33: Two-valued trigger (memory element).

The linear synthesis of the two-valued triggers in linear algebra does not differ fundamentally from Boolean synthesis [17]. The schematic configuration of the trigger can be constructed in the same way as it is done in Boolean logic (for example – Figure 34):

When moving to higher values while remaining the general idea of synthesis, it is necessary to use the elements and operations that are a generalization of the two-valued operations and logic elements. The operations *min* (x_1,x_2) and *max* (x_1,x_2) are generalization of the operations & and ∨ is. As for the inversion operation, to generalize it for a multi-valued case it is convenient to represent it in the form of $\bar{x} = x \oplus 1 = x \ominus 1$ and generalize it by a circular shift of Post (left or right). The general functional configuration of the memory element based on three-valued elements of the direct circular shift is shown in Figure 35.

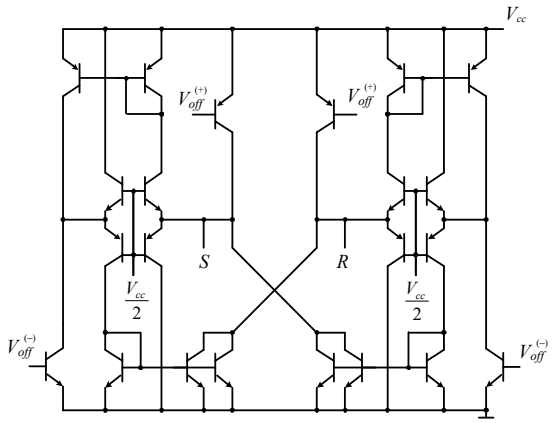


Figure 34: Two-valued RS-flip-flop in the studied basis based on the *truncated difference.*

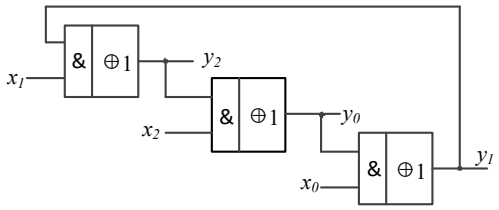


Figure 35: Three-valued trigger on the elements of the direct circular shift.

The general functional configuration of the memory element based on the three-valued elements of the inverse circular shift is given in Figure 36.

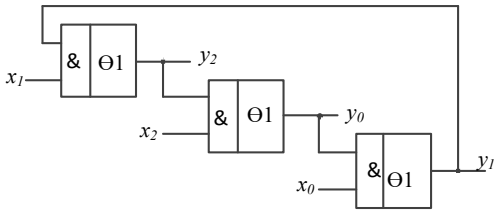


Figure 36: Three-valued trigger on the elements of the inverse circular shift.

The functional schemes of the memory elements of higher significance look similar.

Thus, to construct any multivalued memory element, it is necessary to synthesize the elements of direct and inverse circular shifts. To create a trigger of the given type (D-, RS-, JK-, etc.), it is required to equip the memory element with the corresponding control circuit.

The control circuits of the memory elements are constructed on the base of the verbal description of the operation of a specific type of the trigger. The result of the analysis of this description is the detection of logic control functions for each input of the memory element.

For example, the RS-flip-flop operates according to the following algorithm:

- the values of the input signals $S = R = k-1$ correspond to the storage mode;

- the signal *S* increases the trigger state index relatively to the current state towards the state *k*-1, and the signal R reduces the state index towards the state "0";

- the value of the state index change is equal to the value of the input signal: the signal equal to 1 can increase (or decrease) the current state index by 1, the signal equal to 2 - by 2, etc .; besides, the signal S increases the trigger state index to the state *k*-1, and the signal *R* reduces the state index to "0"; the state change on the cycle is impossible (we leave this for the universal triggers!);

- to change the trigger state index relatively to the current state *i*, the signal *S* can take values from 1 to *k*-1-*i* (larger values are equivalent to the value of *k*-1-*i*). Similarly, the signal *R* can take the values from 1 to *i* -1 (larger values are equivalent to the value *i* -1);

- all combinations are inhibited combinations of values of the input signals, except 0 – *k*-1, 1 – *k*-1, *k*-2 – *k*-1, *k*-1 – *k*-1, *k*-1 – 0, *k*-1 – 1, …, *k*-1– *k*-2.

Now it is possible to synthesize the logic function of the control circuit of the *RS*-flip-flop. Its output signal is a multi-valued signal *x* that sets the next state of the memory element, and the input signals are multivalued signals *S* and *R*, as well as the current trigger state index.

The truth table of the control functions of the three-valued RS-flip-flop, compiled on the basis of this description, has the following form:

S	R	Q_{t+1}	x_0	x_1	x_2
0	2	Q_2	0	1	1
1	2	Q_1	1	0	1
2	2	Q_t	1	1	1
2	1	Q_1	1	0	1
2	0	Q_0	1	1	0

From the last table it follows that the logic functions for controlling the state of the memory element are described as follows:

$$x_0 = P(S > 0) = 1 \div (1 \div S);$$

$$x_1 = 1 > [1 > (1 > S)] > [1 > (1 > R)] =$$

$$= 1 \div [1 \div |1 \div S|] \div [1 \div |1 \div R|];$$

$$x_2 = P(R > 0) = 1 \div (1 \div R).$$

The combination of the memory element circuit and the control circuit results in the implementation of the trigger of the given type.

Similarly, arguing, we can obtain the logical control functions of the three-valued D-flip-flop:

$$x_0 = (1 \div C)\&(1 \div D) = 1 \div [(1 \div D) \div C] =$$

$$= 1 \div [(1 \div C) \div D];$$

$$x_1 = (1 \div C)\&[D \div 2(D \div 1)] = 1 \div \{D \div [2(D \div 1) \div C]\} =$$

$$1 \div [1 \div (|D \div 1| + C)];$$

$$x_2 = (1 \div C)\&(D \div 1) = 1 \div [(D \div 1) \div C] =$$

$$= D \div [(1 \div C) \div 1].$$

In the same way, we can obtain the description of the control system of any other multivalued trigger, the number of types of which is certainly greater than the two-valued one.

Combining the control circuit and the memory element in series, it is possible to obtain the schematic configuration of the trigger of any type and any value.

6. Conclusion

1. The mathematical tool of linear algebra can be newly applied in problems of logic synthesis and circuit implementation of the current digital structures.

2. Linear algebra:

- enables to create not only a two-valued, but also a really properly functioning multivalued element base for digital signal processing devices;

- conduces the design of the digital element base with the improved technological, technical and operational characteristics (in comparison with the potential logic based on Boolean algebra);

- can serve as a basis for creating LSI on the basis of matrix fields of homogeneous elements (as in modern Altera design systems, etc.);

- improves the reliability of current digital LSIs synthesized on its basis, under extreme operating conditions (temperature, radiation, in-phase interferences, etc.).

3. Two-valued and three-valued triggers based on cyclic shift elements are considered. It is shown that in the transition to greater significance, while preserving the general idea of synthesis, it is necessary to use operations that are generalizations of two-valued operations and the corresponding logical elements.

4. The presents a basic set of current logic elements for the devices of automation, which allows solving the problems of transformation of the current signals in a different and more efficient way.

5. In the schemes of the developed class is provided a differential representation of the output signal that minimizes the effect of temperature and radiation on their basic parameters.

Conflict of Interest

The authors declare that there is no conflict of interests regarding of publication of this paper.

Acknowledgment

The research is carried out at the expense of the Grant of the Russian Science Foundation (project № 16-19-00122).

References

[1] N. N. Prokopenko, N. I. Chernov, V. Ya. Yugai, N. V. Butyrlagin, "The Element Base of the Multivalued Threshold Logic for the Automation and Control Digital Devices," on International Siberian Conference on Control and Communications, SIBCON-2017, Astana, Kazakhstan, 29-30 June, 2017.

[2] A. S. Karpenko, "Multi-valued logics (monograph)," in series "Logics and computer", Moscow, issue 4, 1997, 223 p. (in Russian).

[3] L. A. Zalmanson, "Fourier transform, Walsh, Haar and their application in management, communications and other fields," Moscow, Nauka, Fizmatlit publ., 1989, 496 p. (in Russian).

[4] V. S. Vykhovanets, V. D. Malugin, "Spectral methods in logical management," Proceedings of the 2nd international scientific-technical conference "Modern methods of digital processing of signals in spelthorne, monitoring, diagnosis and control (OS 1998)", Minsk, 1998, pp. 56-59. (in Russian).

[5] M. Thornton, "Mixed representation of the multi-Boolean function spectra and critical charts," Automatics and telemechanics, 2004, vol.6, pp. 188-200. (in Russian).

[6] G. I. Ivchenko, V. A. Mironov, "Some questions of spectral analysis of random Boolean functions with constraints," Discrete mathematics, 2013, vol. 1, pp. 90-110. (in Russian).

[7] V.N. Kondratyev, A.A. Shalyto, "Implementation of systems of Boolean functions using arithmetical polynomials," Automatics and telemechanics, 1993, vol. 2, pp. 114-122. (in Russian).

[8] V. D. Malyugin, "The parallel logic computation by the arithmetic of polynomials," Moscow, Nauka, Fizmatlit publ., 1997, 192 p. (in Russian).

[9] A. V. Sokolov, O. N. Zhdanov, A. A. Ayvazyan, "Methods for the synthesis of the algebraic normal forms of functions of multivalued logic," System analysis and applied Informatics, No. 1, 2016, p. 69-76. (in Russian).

[10] M. Dertouzos, "Threshold logic," Moscow: Mir publ., 1967. (in Russian).

[11] V. G. Nikonov, "Threshold representations of Boolean functions," Review in applied and industrial mathematics, Series discrete mathematics, 1994, vol. 1, No. 3, pp. 402-457. (in Russian).

[12] E. A. Butakov, "Methods of synthesis of relay devices," Moscow: Energy publ., 1970, 328 p. (in Russian).

[13] E. N. Vavilov, etc. The Synthesis of threshold circuits for the elements – Moscow: Soviet Radio publ., 1970, 368 p. (in Russian).

[14] S. Muroga, "Threshold logic and its applications," New York: Wiley, 1971.

[15] W. H. Piers, "Failure Tolerant computer dezign," New York and London: Academic Press, 1965.

[16] V. G. Nikonov, N. Nikonov, "Features of threshold concepts k-valued functions," Tr. on discr. mod., 2008, volume 11, issue 1, pp. 60-85. (in Russian).

[17] J. Hastad, "On the size of weights for threshold gates." SIAM J. Discr. Math. 1994.

[18] N. I. Chernov, "Foundations of Logic Synthesis of Real Numbers Field Digital Structures", Taganrog: TRTU, 2000, p.146 (in Russian).

[19] N. I. Chernov, "Boolean Linear Space as an Algebraic Structure for Logic Synthesis of Digital Units," The News of TRTU, 2003, No.1, pp. 215-220. (in Russian).

[20] N. I. Chernov, "Structural Synthesis of Digital Units within Boolean Linear Spaces," The News of TRTU, 2003, No.2, pp. 73-76. (in Russian).

[21] N. I. Chernov, "Efficiency of Application of the Body of Linear Spaces within Logic Synthesis," Conferences on Artificial Intelligence Systems (IEEE AIS' 05)' and Intelligent CAD (CAD-2005, vol.1, pp. 420–424. (in Russian).

[22] N. I. Chernov, "Logic Design of Digital Structures on Controlled Current Generator," The News of TREU, 2005, No.11, pp. 77- 83. (in Russian).

[23] N. I. Chernov, "The effectiveness of the apparatus of linear spaces in the logic synthesis of digital structures," Proceedings of the international scientific and technical conferences "Intelligent systems (IEEE AIS'05)" and "Intelligent CAD (CAD-2005",vol.1, pp. 420 –424. (in Russian).

[24] N. I. Chernov and V. Ya. Yugai, "Nonclassical Synthesis of Digital Structures by Tools of Analogous Circuits Engineering," Problems of Today's Analogous Circuits Engineering: The Collected Articles of IX International Scientific-Practical Seminar edited by N.N. Prokopenko, Shakhty, Rostov-on-Don Region: FSBEU HPE "SRSUES" Publishers, 2012, pp. 138 – 143. (in Russian).

[25] N. N. Prokopenko, N. I. Chernov, V. Ya. Yugai, "Base Concept of Linear Synthesis multi-Valued Digital Structures within Linear Spaces," Proceedings of The IS&IT13 Congress, The Scientific Edition in four volumes, Moscow: PhisMathLit publ., 2013, v. 1, pp. 284−289. (in Russian).

[26] N. N. Prokopenko, N. V. Butyrlagin, N. I. Chernov, V. Ya. Yugay, "Synthesis of binary triggers in the apparatus of linear algebra," Izvestiya SFedU. Technical Sciences. No. 2. 2015, pp. 115-125. (in Russian).

[27] N. N. Prokopenko, N. I. Chernov, V. Ya. Yugai, N. V. Butyrlagin, "Linear synthesis of k-valued digital components of the base current logical signals: the principle of generalization", Problems of development of perspective micro and nanoelectronic systems - 2016. Proceedings / under the General editorship of academician RAS A.L. Stempkovskogo. Moscow: IPPM RAS, 2016. (in Russian).

[28] N. I. Chernov, V. Y. Yugai, N. N. Prokopenko, N. V. Butyrlagin, "Basic concept of linear synthesis of multivalued digital structures in linear spaces," Proceedings of IEEE East-West Design and Test Symposium, EWDTS 2014, art. no. 7027045. DOI: 10.1109/EWDTS.2014.7027045

[29] N. N. Prokopenko, N. V. Butyrlagin, N. I. Chernov, V. Ya. Yugai, "The linear concept of logic synthesis of digital IP-modules of control and communication systems," 2015 International Siberian Conference on Control and Communications, SIBCON 2015 - Proceedings, art. no. 7147182. DOI: 10.1109/SIBCON.2015.7147182.

[30] N. N. Prokopenko, N. V. Butyrlagin, N. I. Chernov, V. Ya. Yugai, "Basic linear elements of k-Valued digital structures," 2016 International Conference on Signals and Electronic Systems, ICSES 2016 - Proceedings, pp. 7-12. DOI: 10.1109/ICSES.2016.7847763.

[31] N. N. Prokopenko, N. I. Chernov, V. Yugai, N. V. Butyrlagin, "The element base of the multivalued threshold logic for the automation and control digital devices," Proceedings of 2017 IEEE International Siberian Conference on Control and Communications, SIBCON 2017, art. no. 7998508. DOI: 10.1109/SIBCON.2017.7998508.

[32] N. N. Prokopenko, N. I. Chernov, V. Yugai, N. V. Butyrlagin, "The multifunctional current logic element for digital computing devices, operating on the principles of linear (not boolean) algebra," Proceedings of 2016 IEEE East-West Design and Test Symposium, EWDTS 2016, art. no. 7807723. DOI: 10.1109/EWDTS.2016.7807723.

[33] V. I. Nechaev, "Numeric system," Moscow: Education publ.,1975, 199 p. (in Russian).

[34] V. A. Gorbatov, "Foundations of discrete mathematics," Moscow: Higher school publ., 1986, 311 p. (in Russian).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 174-183 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Model for Optimising the Deployment of Cloud-hosted Application Components for Guaranteeing Multitenancy Isolation

Laud Charles Ochei[*,1], Christopher Ifeanyichukwu Ejiofor[2]

[1]*School of Computing and Digital Media, Robert Gordon University, United Kingdom*

[2]*Department of Computer Science, University of Port Harcourt, Nigeria*

ARTICLE INFO

Article history:
Received: 25 November, 2017
Accepted: 24 February, 2018
Online: 27 March, 2018

Keywords:
Cloud-hosted application
Multitenancy
Degree of Isolation
Simulated Annealing
Metaheuristic
Optimization model
Component
Cloud deployment

ABSTRACT

Tenants associated with a cloud-hosted application seek to reduce running costs and minimize resource consumption by sharing components and resources. However, despite the benefits, sharing resources can affect tenant's access and overall performance if one tenant abruptly experiences a significant workload, particularly if the application fails to accommodate this sudden increase in workload. In cases where a there is a higher or varying degree of isolation between components, this issue can become severe. This paper aims to present novel solutions for deploying components of a cloud-hosted application with the purpose of guaranteeing the required degree of multitenancy isolation through a mathematical optimization model and metaheuristic algorithm. Research conducted through this paper demonstrates that, when compared, optimal solutions achieved through the model had low variability levels and percent deviation. This paper additionally provides areas of application of our optimization model as well as challenges and recommendations for deploying components associated with varying degrees of isolation.

1. Introduction

Designing and planning component deployment of a cloud-hosted application with multiple tenants demands special consideration of the exact category of components that are to be distributed, the number of components to be shared, and the supporting cloud resources required for component deployment. [1] This is because there are different or varying degrees of multitenancy isolation. For instance, in components providing critical functionality, the degree of isolation is higher compared to components that only require slight re-configuration prior to deployment [2].

A low degree of isolation actively encourages tenants to share resources and components, resulting in lower resource consumption and reduced operating costs, however, there are potential challenges in both security and performance in the instance where one component sees a sudden workload surge. A high degree of isolation tends to deliver less security interference, although there are challenges instigated by high running costs and resource consumption in view that these tenants are not sharing resources [2]. Consequently, the software architect's main challenge is to first identify solutions to the opposing trade-off of high degrees of isolation (including excessive resource consumption issues and high operating costs), versus low degrees of isolation (including performance interference issues).

Motivated by these key challenges, this paper presents a model for the deployment of components which provides exemplary solutions specific to cloud-based applications and aims to do so in a way that secures the segregation of multitenancy. The approach for this research includes creating an optimization model which is mapped to a Multichoice Multidimensional Knapsack Problem (MMKP) before solutions are tested using a metaheuristic. The approach is analysed through comparing the different optimal solutions achieved which then collectively compose an exhaustive search tool to analyse the solutions capacity specifically for minor problem occurance, in its entireity.

This paper and its research questions: "***How can we optimize the way components of a cloud-hosted service is deployed for guaranteeing multitenancy isolation?***". It is possible to guarantee the specific degree of isolation essential between

[*]Laud Charles Ochei, Aberdeen, United Kingdom. l.c.ochei@rgu.ac.uk

tenants, whilst efficiently managing the supporting resources at the same time through component deployment optimization of the cloud-based application.

This paper expands on the previous work conducted in [3]. The core contributions of this article are:
1. Mathematical optimization model providing optimal component deployment solutions appropriate for cloud-based applications to guarantee multitenancy isolation.
2. Mathematical equations inspired by open multiclass queuing network models to determine the average request totals for granting component and resource access.
3. Variants of metaheuristic solutions to deliver optimization model resolutions attributed to simulated annealing.
4. Guidelines and recommendations for component deployment in cloud-hosted applications seeking to guarantee required levels of multitenancy isolation.
5. Application areas of a cloud-hosted service where it is possible for the optimization model to be directly applied to component deployment with the aim to guarantee required degree of multitenancy isolation.

The remainder of this paper is structured as outlined below: Section II focuses on the challenge of identifying and delivering near-optimal component deployment solutions specific to cloud-hosted applications that guarantee the essential levels of multitenancy isolation; Section III presents the Optimisation Model; the following section presents the Metaheuristic Solution; Section V considers the Open Multiclass Queuing Model; Section VI evaluates the results and presents the experimental setup; Section VII discusses the results; Section VIII discusses the model's application areas for the model; and Section IX conclude with recommendations of future work.

2. Optimising the Deployment of Components of Cloud-hosted Application with Guarantee for Multitenancy Isolation

This section examines multitenancy isolation, the conflicting trades-offs in delivering optimal deployment influenced by the varying degrees of multitenancy isolation, and other associated research on cloud resources and optimal allocation of such resources.

2.1. Multitenancy Isolation and Trade-offs for Achieving Varying Degrees of Isolation

In a multitenant architecture (also referred to as *multitenancy*), multiple tenants are able to access a single instance of a cloud service. These tenants have to be isolated when there are changes in workload. In the same way that it is possible to isolate multiple tenants, it is also possible to isolate multiple components of a cloud-hosted application.
In this paper, we define "Multitenancy Isolation" as a way of ensuring that other tenants are not affected by the required performance, stored data volume, and access privileges of one of the tenants. accessing the cloud-hosted application [3] [4].

A high degree of isolation is achieved when there is little or no impact on other tenants when a substantial increase in workload occurs for one of the tenants, and vice versa. The three cloud patterns that describe the varying degrees of multitenancy isolation are:
(i) Dedicated Component: tenants cannot share components; however, a component may be associated with one or more tenants or resources;
(ii) Tenant-isolated Component: tenants can share resources or components, and isolation of these is guaranteed; and
(iii) Shared Component: tenants can share resources or components, but these remain separate from other components.

If components required a high degree of isolation between them, then each tenant requires that each component is duplicated. This can be expensive and also lead to increased resource consumption. On the other hand, there could also be a need for a low degree of isolation which could, in contrast, reduce cost and resource consumption. However, any changes in workload levels that the application cannot cope with risk interference [4]. The question, therefore, is how optimal solutions can be identified to resolve trade-offs when conflicting alternatives arise.

2.2. Related Work on Optimal Deployment and Allocation of Cloud Resources

Research on optimal resource allocation in the cloud is quite significant, however, much fewer studies focus on optimal solutions in relation to component deployment across cloud-based applications in a way which guarantees the required degree of multitenancy isolation. Researchers in [5] and [6] aim to keep cloud architecture costs to a minimum by implementing a multitenant SaaS Model. Other authors [7] concentrated on bettering execution times for SaaS providers whilst reducing resource consumption using evolutionary algorithms opposed to traditional heuristics. A heuristic is defined in [8] for the capacity planning purposes for the SaaS inspired by a utility model. The aim of the utility model was to generate profit increases and so it largely concentrated on business-related aspects of delivering the SaaS application.

It is explained in [9] how optimal configuration can be identified for virtual servers, such as using certain tests to determine the required volumes of memory for application hosting. Optimal component distribution is discussed and analyzed by [10] in relation to virtual servers. Research conducted in [11] is of a similar nature to this paper in that it attempts to reduce costs (using a heuristic search approach is inspired on hill climbing), specifically in relation to the use of VMs from an IaaS provider with limitations around SLAs response time.

The studies noted above predominantly focus on scaling back costs associated with cloud architectural resources. The use of metaheuristics is not considered in these studies in delivering optimal solutions that can guarantee the required degree of multitenancy isolation. Additionally, previous research involving

optimization models have operated with one objective; an example is where [11] look to minimize VM operational costs. For this paper's model, a bi-objective case is used (i.e., maximising the required degree of multitenancy isolation and number of requests permitted to access a component). Thereafter, a modern metaheuristic inspired by simulated annealing is used to solve the model.

3. Problem Formalisation and Notation

This section defines the problem and explains the process of mapping it to a Multichoice Multidimensional Knapsack Problem (MMKP).

3.1. Description of the Problem

Assuming a tenant has multiple components associated with the same supporting cloud infrastructure. A team may represent a team or department, a company with a responsibility to design a cloud-based application, its components, and underlying processes. Components varying in size and function has to integrate with their cloud-hosted application to achieve effective deployment in a multitenant style. It is also possible to define component categories based on different features, such as function, for example, processing or storage. Within these categories, different components are likely to have differing degrees of isolation enabling some components to deliver the same function which can hence be accessed and used by multiple tenants, opposed to other components which may be solely allocated to certain tenants or departments.

Every component within an application needs a particular allocation of resources from the cloud infrastructure in order to support the volume of requests received. In instances where one component in the application experiences surges in workload, then it must be considered how the designer can choose components to deliver optimal deployment to effectively respond to the sudden changes in such a manner that: (i) maximises component degrees of isolation through ensuring they behave in the same way as components of other tenants, thus, isolating against one and other; and (ii) maximises the total requests permitted to access and use each components.

3.2 Mapping the Problem to a Multichoice Multidimensional Knapsack Problem (MMKP)

The above mentioned optimal component deployment problem can be closely linked to a 0-1 Multichoice Multidimensional Knapsack Problem (MMKP). An MMKP is a variant of the Knapsack Problem commonly depicted as a member of the NP-hard class of problems. For the purpose of this paper, the problem of focus can be formally defined as:

Definition 1 (Optimal Component Deployment Problem): Consider that there are N groups of components ($C_1,\ldots,C_N$) with each group having a_i ($1 \le i \le N$) components useful for designing (or integrating with) a cloud-hosted application. Each component of the application is affiliated with: (i) the degree of isolation that is required between components (I_{ij}); (ii) the rate at which requests arrive to the component λ_{ij}; (iii) the service demand of resources required to support the component D_{ij}; (iii) the average request totals permitted to access the component Q_{ij} and (iv) resources for supporting the component, $r_{ij} = r_{ij}^1, r_{ij}^2, \ldots, r_{ij}^n$. For the cloud to properly support all components, a certain volume of resources are required; the total number of resources needed can be calculated as $R = (R^1, R^2, \ldots, R^n)$.

The aim of an MMKP is to choose one component present in each category for deployment to the cloud in a manner that ensures that if one component sees sudden increases in load, then the: (a) the degree of isolation of other components is maximized; (b) the total requests permitted to access the component and application is maximised without using more resources than are actually available.

Definition 1 identifies two objectives within the problem. An aggregation function is used to convert the multi-objective problem into a single-objective because of merging the two objective functions merge (i.e., g1=degree of isolation, and g2=number of requests) into one single objective function (i.e., g=optimal function) in a linear way. Because the optimal function of this study is linear, a *priori single weight* strategy is employed to aid defining the weight vector selected based on the decision maker's individual preferences [12]. This paper also adopts the approach discussed in [12] for computing the absolute percentage difference (see section 7.1), the target solutions used to compare against the optimal solution (see section 7.2), and the use of the number of optimal function evaluations as an alternative to measuring the computational effort of the metaheuristic.

Therefore, the purpose is redefined as follows: to deliver a near-optimal solution for component deployment to the cloud-based application that also fulfils system requirements and achieves the best value possible for optimal function, G.

Definition 2 (Optimal Function): For a cloud-hosted application architect, the main issues impacting the optimal deployment of components are changes to workload, which can be expressed as:

$$\sum_{i=1}^{N} \sum_{j \in C_i}^{n} g_{ij} . a_{ij}$$

Subject to:

$$\sum_{i=1}^{N} \sum_{j \in C_i}^{n} r_{ij}^{\alpha} . a_{ij} \le R^{\alpha} (\alpha = 1, 2, \ldots, N) \quad (1)$$

$$\sum_{j \in C_i}^{N} a_{ij} = 1$$

$a_{ij} \in 0,1$ (i = 1, 2 ,…, N), j $\in C_i$

where (i) a_{ij} is fixed at 1 if component j is chosen from group C_i and 0 otherwise; (ii) g_{ij} is determined by a weighted calculation of parameters involving the degree of isolation, average requests permitted to access a component, and penalty for constraint violations.

$$g_{ij} = (w1 \times I_{ij}) + (w2 \times Q_{ij}) - (w3 \times P_{ij}) \quad (2)$$

Specific weight values are allocated to w1, w2, and w3; namely 100, 1 and 0.1 respectively. The allocation of weights is done using a method that provides preference to the required degree of isolation. The penalty, P_{ij}, imposed for components that surmount resource cap is expressed as:

$$P_{ij} = \sum_{i=0}^{n} max\left\{0, \left(\frac{R^k - R^k_{max}}{R^k_{max}}\right)\right\}^2 \quad (3)$$

For every component ($\boldsymbol{g}$), the degree of isolation, I_{ij}, is assigned either 1, 2, or 3 indicating either shared, tenant-isolated or dedicated components, respectively. The sum described as: $r_{ij} = r^1_{ij}, r^2_{ij}, \ldots, r^n_{ij}$ refers to the resource consumption in group C_i. for each individual application component j. Total resource consumption r^{α}_{ij} for all application components needs to be lower than the total available resources in the cloud infrastructure $R = R^{\alpha}, (\alpha = 1,\ldots,m)$.

It is presumed that the service demands at the CPU, RAM, Disk I/O, and the supporting bandwidth of each component can be identified and/or measured readily by the SaaS supplier or customer. This assumption enables us to calculate the number of requests, Q_{ij} that may be permitted access for each component through analysis of an open multiclass QN Model [13]. The following section expands further on the open multiclass network.

4. Queuing Network (QN) Model

Queueing network modelling is one modelling approach through which the computer system is depicted as a network of queues that can be solved in an analytic fashion. In its most basic form, a network of queues is an assembly of service centers representative of system resources, and customers representative of business activity, such as transactions [13]. Service centers are basically supporting resources for the components, such as CPU, RAM, disk and bandwidth.

Assumptions: For the purpose of this paper, the following component assumptions are made:
(i) components cannot support other applications or alternative system requirements, and is therefore exclusively deployed to one cloud-application;
(ii) component arrival rates are separate to the main system state and so component requests may have significantly varied behaviours.
(iii) it is possible to identify and effortlessly measure the service demands at the CPU, RAM, Disk, and Bandwidth supporting each component by both the SaaS provider and/or customer.
(iv) sufficient resource is available to support each component during changes to workload, particularly during significant surges of new incoming requests. Ensuring sufficient resource means that there are no overloads during peak times where all components are operating.

The assumptions noted above allow the study to utilise an open multiclass queuing network (QN) model for the purpose of calculating average requests permitted to reach the component, whilst simultaneously ensuring the required degree of isolation, as well as system requirements. The magnitude and intensity of workload volume in an open multiclass QN is determined by request arrival rates. The arrival rate is not typically reliant on the system state, and so is not reliant on the volume of other tenants in the system either [13].

Definition 3 (Open Multiclass Queuing Network Model): Assuming there is a total of N classes, where every class c is an open class with arrival rate λ_c. The arrival rates are symbolised as a vector by $\vec{\lambda} = (\lambda_1, \lambda_2, \ldots \lambda_N)$. The use of each component in class c at the center k is given by:

$$U_{c.k}(\vec{\lambda}) = \lambda_c D_{c.k} \quad (4)$$

To solve the QN model, assumptions are made, such as that a component stands for a single open class system hosting four service centers otherwise referred to as supporting resources, such as CPU, RAM, disk capacity and bandwidth. At any one service center (e.g., CPU), the average request totals for a specific component is:

$$Q_{c,k}(\vec{\lambda}) = \frac{U_{c.k}(\vec{\lambda})}{1 - \sum_{i=1}^{N} U_{i.k}(\vec{\lambda})} \quad (5)$$

Consequently, to determine the average amount of requests accessing the particular component, the length of the queue of all requests reaching all service centers (i.e., components' supporting resources such as CPU, RAM, disk capacity and bandwidth) would be totaled.

$$Q_c(\vec{\lambda}) = \sum_{i=0}^{n} Q_{c,k}\ \vec{\lambda} \quad (6)$$

5. Metaheuristic Search

The optimisation problem as explained in the section before is an NP-hard problem renowned for its feasible search capacity and exponential growth [14]. The number of potential and feasible solutions that may achieve optimal component deployment and solve the problem can be determined using this equation:

$$\left\{\binom{n}{r}\right\}^N \quad (7)$$

The above equation, Equation 4, signifies the different ways one or more (**r**) components can be chosen from each group (comprising of **n** components) from a pool of numerous (N) groups of components, for the purpose of creating and integrating them into a cloud-hosted application upon receipt of updates or changes to workload by the component. Thus, to manage such changes to workloads, the specific number of different ways that one component can be selected (i.e., r=1) from each of the 20 different groups (i.e., N=20), comprising of 10 items per group (i.e., n=10), approximately 10.24 x 10^{12} possible solutions can be identified. Contingent on the number of changes to workload and also the regularity of these, a cloud-hosted service quite large in size could experience a much greater volume of possible solutions.

Therefore, to obtain an optimal solution for the identified optimisation problem, it is essential to use an efficient metaheuristic. In addition, this should be done in real-time with the SaaS customer or cloud architect. Two versions of a simulated annealing algorithm are implied: (i) *SAGreedy*, incorporates greedy principles in conjunction with a simulated annealing algorithm; (ii) *SARandom*, employs randomly propagated solutions in conjunction with a simulated annealing algorithm. Both of these versions can be effectively utilised to achieve near-optimal solutions for component deployment. Additionally, an algorithm was developed for this study to generate an extensive search of the full solution area for a small problem size. Algorithm 1 includes the algorithm for *SA(Greedy)*. However, SA(Random) only needs a minor change to this algorithm, which will be described further in the following section. An extensive breakdown of Algorithm 1 can be viewed below:

Algorithm 1 SA(Greedy) Algorithm

```
1:  SA (Greedy) (mmkpFile, N)
2:  Randomly generated N solutions
3:  Initial temperature fixed to T0 to st. dev. of all
    optimal solutions
4:  Create greedySoln a1 with optimal value g(a1)
5:  optimalSoln = g(a1)
6:  bestSoln = g(a1)
7:  for I = 1, N do
8:  Create neighbour soln a2 with optimal value
    g(a2)
9:  Mutate the soln a2 to improve it
10: if a1 < a2 then
11:     bestSoln = a2
12: else
13:     if random[0,1) < exp(-(g(a2) – g(a1))/T) then
14:         a2 = bestSoln
15:     end if
16:     end if
17:     Ti+1 = 0.9 * Ti
18: end for
19: optimalSoln = bestSoln
20: Return (optimalSoln)
```

5.1. The SAGreedy for Near-optimal Solution

The first algorithm is a combination of simulation annealing and greedy algorithm which is used to obtain a near-optimal solution for the optimisation problem modelled as an MMKP. First, the algorithm extracts the key details from the MMKP problem instance before populating the encompassing variables (i.e., collections of different dimensions storing isolation values of isolation; average request totals; and the resource consumption of components). A basic linear cooling schedule is used, where $T_{i+1} = 0.9Ti$. The method for prescribing and fixing the preliminary temperature T_0 will be to randomly generate an optimal solution whose number is equivalent to the total number of groups (n) in the problem instance, multiplied by the number of iterations (N) used in the experimental settings before running the simulated annealing element of the algorithm.

When the problem instance and/or the total iterations is low, the magnitude of optimal solutions created may be limited by the number of groups (n) in the problem instance, multiplied by total iterations (N) used in the experimental settings. Next, the initial temperature T_0 is determined for the standard deviation of all optimal solutions (Line 2-3) through random generation. The algorithm then uses the greedy solution as the preliminary solution (Line 4) which is assumed as the best current solution. The simulated annealing process enhances the greedy solution further providing a near-optimal solution for cloud component deployment.

The execution of the algorithm in its most basic form for the instance C(4,5,4) is explained as follows: let us imagine that the total number of iterations is 100, 400 (i.e., 4 groups x 100 iterations) optimal solutions are randomly generated before calculating the standard deviation for all solutions. Assuming a value of 50.56, T_0 is identified as 50. It is also assumed that the algorithm creates a foremost greedy solution with $g(a^1) = 2940.12$, before a current random solution with $g(a^2) = 2956.55$. The solution a^2 will substitute a^1 with probability, P =exp(-16.43/50)=0.72, because $g(a^2) > g(a^1)$. In lines 14 to 16, a random number (rand) is generated between 0 and 1; if rand <0.72, a^2 replaces a^1 and we proceed with a^2. Alternatively, the study continues with a^1. Now, the temperature T is reduced providing $T_1 = 45$ (Line 17). Iterations continue until N (that is, the identified number of iterations set to enable the algorithm to function), is reached, thus the search converges with a high probability of near-optimal solution.

5.2. The SA(Random) for Optimal Solutions

Considering the SA(Random) metaheuristic version, a solution is also randomly generated before being encompassed within the simulated annealing process to provide a preliminary solution. It can be seen in Line 4 that rather than creating a greedy solution, a random solution is created. An optimal solution representative of a set of components with the highest total isolation value and number of requests permitted to reach the component access is

then output by the algorithm. Every time a variance in workload is experiences, the optimal solution alters to respond to this.

6. Evaluation

We describe in the section how each instance was generated as well as the process, procedure and set up of the experiment.

6.1. Instance Generation

Reflective of different capacities and sizes, a number of problem instances were randomly generated. Instances were divided into two categories determined by those cited frequently in current literature: (i) OR benchmark Library [15] and other standard MMKP benchmarks, and (ii) the new irregular benchmarks used in [16]. All these benchmarks were used for single objective problems. This study edited and modified this benchmark to fit into a multi-objective case through assigning each component with one of two profit values: isolation values and average number of requests [17].

The values of the MMKP the instance, were produced as follows: (i) random generation of isolation values in the interval [1-3]; (ii) values of component consumption of CPU, RAM, disk and bandwidth (i.e., the weights) were generated in the interval [1-9]; (iii) individual component resource limits (i.e., knapsack capacities for CPU, RAM, disk and bandwidth) were created by halving the maximum resource consumption possible (see Equation 7).

$$c_k = \frac{1}{2} \times m \times R \tag{8}$$

An identical principle has been employed to create instances for OR Benchmark Library, as well as for instances used in [18] [19]. This research considers the total resources/constraints as four (4) for each group, which reflects the minimal resource requirement to deploy a component to the cloud. The notation for each instance is: C(n,r,m), representing the number of groups, the amount of components in each group, and resource totals.

6.2. Experimental Setup and Procedure

For consistency, all experiments were set up and operated using Windows 8.1 on a SAMSUNG Laptop with an Intel(R) CORE(TM) i7-3630QM at 2.40GHZ, 8GB memory and 1TB swap space on the hard disk. Table I outlines the experimental parameters. The algorithm was tested using different sized instances of different densities. In relation to large instances, it was not possible to conduct an exhaustive search due to a lack of memory resource on the machine used.

As a result of this limitation, the MMKP instance was implemented first, C(4,5,4), to provide a benchmark for analysis to enable algorithm comparison.

Table 1. Parameters used in the Experiments

Parameters	**Value**
Isolation Value	[1,2,3]
No. of Requests	[0,10]
Resource consumption	[0,10]
No. of Iterations	N=100 (except Table 4)
No. of Random Changes	5
Temperature	T_0 = st.dev of N randomly generated solns.
Linear Cooling Schedule	$T_{i+1} = 0.9Ti$

7. Results

Section 7 discusses the experiment results.

7.1 Comparison of the Obtained Solutions with the Optimal Solutions

The results delivered by algorithms SA(Greedy) and SA(Random) were initially compared with the optimal solutions generated by an exhaustive search for a small problem instance in the entire solution space (i.e., C(4,5,4)). Table 2 and Table 3 portray the findings. The instance id used is noted in Column 1 of Table 2. The second, third and fourth columns respectively highlight the optimal function variables as (FV/IV/RV), representing the optimal function value, isolation value, and number of permitted requests, for Optimal, SA(Random) and SA(Greedy) algorithms. The first and second columns of Table 3 depict a proportion of the optimal values for SA(Random) and SA(Greedy) algorithms, respectively. The final two columns note the absolute percentage difference, indicative of the solution quality, for SA(Random) and SA(Greedy) algorithms, which is measured as follows:

$$\frac{|f(s) - f(s^*)|}{f(s^*)} \tag{9}$$

where s is the obtained solution and s* is the optimal solution generated by the exhaustive search.

It is clear that SA(Greedy) and SA(Random) provide very similar results. Solutions identified for SA(Random) are almost 100% close most of the time to their optimal solution, and over 83% in a smaller proportion of cases whilst in just one occurrence it was below 66%. SA(Greedy) also delivered nearly 100% close solutions created to the optimal solution in a considerable number of cases, and more than 83% in others. Overall, SA(Greedy) delivered better results than SA(Random) when considering percent deviation from the optimal solution.

7.2 Comparison of the Obtained Solutions to a Target Solution

It was not possible to run instances greater than C(4,5,4) because of hardware limitations (i.e., CPU and RAM).

Consequently, the results were compared to a target solution. The target solution for percent deviation and performance rate was determined as (n x max(I) x w1) and ((n x max(I) x w1) + (0.5 x (n x max(Q) x w2))), respectively. So, for instance C(150,20,4), the target solution for computing percent deviation sit at 45,000.

Table III, IV, and V demonstrates average solution behaviour: (i) on a significant selection of varied instances using the same parameters; (ii) over various runs on the same instance (with differing quantities of optimal evaluation); and (iii) over different runs on the same instance. The robustness of solutions in relation to their behaviour on varying types of instances using the same parameters was measured. Table III shows this measure and that solutions are strong when considering average deviation of solution behaviour for both the SA(Greedy) and SA(Random), as implied through their low variability scores.

The average percent deviation and standard deviation (of the percent deviations), for SA(Greedy) is marginally greater than SA(Random) as a result of the significant absolute difference between some solutions and the reference solution. For example, the percent deviations of SA(Greedy) for the instances C(100,20,4) and C150,20,4) are higher than SA(Random). The results show how SA(Random) performs much better than SA(Greedy) in reference to small instances up to C(80,20,4).

Table IV compares solution quality with optimal function evaluations. It can be determined that the overall solution quality is good when both algorithms are tested on large instances. Once again, the standard deviation for SA(Greedy) is notably lower than SA(Random) in addition to great percent deviation stability. Table V highlighted that SA(Greedy) is stronger than SA(Random) evidenced by the average optimal values and low solution variability. The performance rate (PR) was computed by determining the reference solution as a function of the quantity of optimal evaluations. The PR of SA(Greedy) is marginally higher than SA(Random).

Figure I depict the relationship of solution quality relating to optimal values (i.e., fitness value) and the volume of optimal function evaluations. It can be seen from the diagram that SA(Greedy) benefited a little from the preliminary greedy solution more than SA(Random) when optimal function evaluations are few. Nonetheless, solution quality for both algorithms are better as iterations increase. Once 100 optimal evaluations are reached, the optimal solution stabilizes, but thereafter fails to show any further noticeable improvement.

Figure 2 portrays correlations between solution quality relating to percent deviation and the total function evaluations. In line with expectations, SA(Random) reported a smaller percent deviation than SA(Greedy) in the majority of results, particularly in instances where function evaluations are few. An explanation for this could be the low function evaluation total used in the study. However, percent deviation for SA(Greedy) showed greater stability despite being greater than SA(Greedy)'s results.

Table 2. Comparison of SA(Greedy) and SA(Random) with the Optimal Solution

Inst-id	Optimal (FV/IV/RV)	SA(R) (FV/IV/RV)	SA(Greedy) (FV/IV/RV)
I1	1213.93/12/24	1218/12/18	1218/12/18
I2	1213.97/12/14	1208.99/12/9	1109/11/9
I3	1222.99/12/23	1120/11/20	1120/11/20
I4	1119.98/11/20	1023/10/23	1019/10/19
I5	1219.99/12/20	1017/10/17	1017/10/17
I6	1229.92/12/30	1020.96/10/21	1022.99/10/23
I7	1224.90/12/25	1018/10/18	1018/10/18
I8	1228.96/12/29	822/8/22	1224.99/12/29
I9	1021.97/10/22	912/9/12	912/9/12
I10	1236/12/36	1236/12/36	1236/12/36

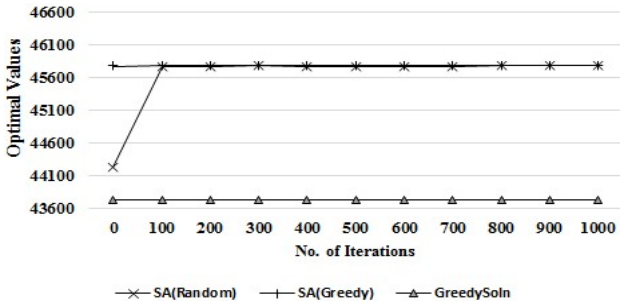


Figure 1. Relationship between Optimal Values and Function Evaluations.

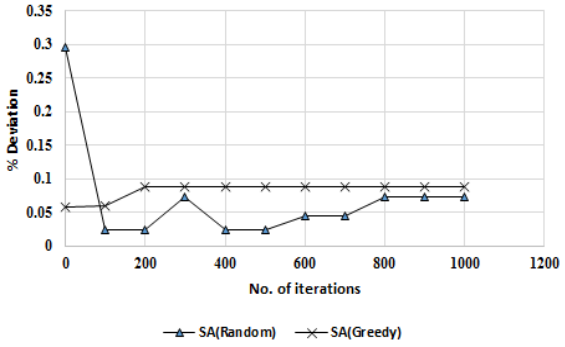


Figure 2. Relationship between Percent Deviation and Function evaluations.

8. Discussion

In Section 8 the implication of the results is discussed further and recommendations for component deployment of a cloud-hosted application that guarantees multitenancy isolation are considered and presented.

8.1 Quality of the Solutions

Solution quality was measured encompassing percent deviation from either the optimal or reference solution. Tables II,

III and IV note solutions generated by SA(Greedy) which demonstrate a low percent deviation. The results also indicate that SA(Greedy) performs efficiently for large instances. However, Table III clearly evidences SA(Random) outperforming SA(Greedy) on small instances, that is, from C(5,20,4) to C(80,20,4).

Table 3. Computation of a Fraction of the Obtained Solutions with the Optimal Solution.

Inst-id	SA-R/Opt	SA-G/Opt	%D (SA-R)	%D (SA-G)
I1	0.995	0.995	0.48	0.48
I2	0.996	0.914	0.41	8.65
I3	0.916	0.916	8.42	8.42
I4	0.913	0.910	8.66	9.02
I5	0.834	0.834	16.64	16.64
I6	0.830	0.832	16.99	16.82
I7	0.831	0.831	16.89	16.89
I8	0.669	0.997	33.11	0.32
I9	0.892	0.892	10.76	10.76
I10	1	1	0	0
ST.DEV	0.097	0.064	9.73	6.43

Table 4. Comparison of the Quality of Solution with Instance size

Inst (n=20)	SA-R	SA-G	SA-R	SA-G
C(5)	1528/15/28	1517/15/17	1.87	1.13
C(10)	3058/30/58	3057/30/57	1.93	1.9
C(20)	6082/60/82	6002/59/102	1.37	0.03
C(40)	12182/120/182	12182/120/182	1.52	1.52
C(60)	18317/180/317	18317/180/317	1.76	1.76
C(80)	24388/240/388	24380/240/380	1.62	1.58
C(100)	30252/298/452	30505/300/505	0.84	1.68
C(150)	45648/449/748	45793/450/793	1.44	1.76
Avg. %D			1.54	1.42
STD of Avg. %D			0.33	0.57

8.2. Computational Effort and Performance Rate of Solutions

Hardware limitations of the computer used for the study restricted evaluation and full consideration of computational effort for metaheuristics. On the other hand, it was still possible to compute algorithm performance rates as a function of optimal evaluation totals. Typically, the number of optimal evaluations is considered a computational effort indicator that is independent to the utilised computer system. Table V indicates that the performance rate of SA(Greedy) is marginally higher than SA(Random) when we put into consideration the total reference solutions achieved.

Table 5. Comparing Solution Quality vs. No. of Optimal Evaluations

No. of Iteration	SA(R)	SA(G)	%D (SA-R)	%D (SA-G)
1	44242.1	45776.8	3.30	0.06
100	45760.7	45776.9	0.02	0.06
200	45760.7	45790	0.02	0.09
300	45783	45790	0.07	0.09
400	45760.7	45790	0.02	0.09
500	45760.7	45790	0.02	0.09
600	45770.1	45790	0.04	0.09
700	45770.1	45790	0.04	0.09
800	45783	45790	0.07	0.09
900	45783	45790	0.07	0.09
1000	45783	45790	0.07	0.09
Worst	44242.1	45776.8	0.02	0.06
Best	45783	45790	3.30	0.09
Average	45632.46	45787.61	0.34	0.08
STD	439.77	5.07	0.93	0.01

8.3. Robustness of the Solutions

Low variability was seen from the SA(Greedy) algorithm particularly when run with large instances, therefore suggesting it delivers stronger solutions than SA(Random) making SA(Greedy) the most suitable algorithm for real-time use in a fast-paced environment experiencing regular changes to workload levels. Figure 1 demonstrates instability in SA(Random) for a small amount of evaluations, although following 1,000 evaluations it was seen to converge eventually at the optimal solution. In contrast, the SA(Greedy) showed steady improvement as evaluations increased.

8.4. Required Degree of Isolation

The study's optimisation model assumes that each component for deployment (or group of components) is linked to certain degree of isolation. By mapping the problem to a Multichoice Multidimensional Knapsack Problem linking each component to a profit values: either isolation value; or number of requests permitted to access a component, this was achieved. As a result of this, it was possible to monitor each component separately and independently, responding to each of their unique demands. Where limitations are faced, such as through cost, time or effort used to tag each component, a different algorithm could enable us to perform this function dynamically. In our research prior to this [20], an algorithm capable of dynamically learning the properties of current components in a repository was developed with this

information then used to associate each component with the required degree of isolation.

Table 6. Robustness of Solutions over Different Runs on the Same Instance.

Runs	SA(R)	SA(G)	%D SA-R	%D SA-G
1	45767	45658	1.70	1.46
2	45793	45744	1.76	1.65
3	45663	45793	1.47	1.76
4	45460	45658	1.02	1.46
5	45567	45658	1.26	1.46
6	45562	45793	1.25	1.76
7	45569	45658	1.26	1.46
8	45567	45658	1.26	1.46
9	45682	45658	1.52	1.46
10	45663	45767	1.47	1.70
Worst	45460	45658	1.02	1.46
Best	45793	45793	1.76	1.76
Avg	45629	45705	1.40	1.57
STD	97.67	58.40	0.22	0.13
Perf. Rate	2.0E-04	3.0E-04		

9. Application Areas for Utilizing the Optimization Model

In this section, we discuss the various areas where our optimization model can be applied to the deployment components of a cloud-hosted service for guaranteeing the required degree of multitenancy isolation.

9.1. Optimal Allocation in a resource constrained environment

Our optimization model can be used to optimize the allocation of resources especially in a resource constrained environment. or where there are frequent changes in workload. This can be achieved by integrating our model into a load balancer/manager. Many cloud providers have auto-scaling programs (e.g., Amazon Auto scaling) for scaling applications deployed on their cloud infrastructure, usually based on per-defined scaling rules. However, these scaling programs do not have a functionality to provide for guaranteeing the isolation of tenants (or components) associated with the deployed service. It is the responsibility of the customer to implement such a functionality for individual service deployed to the cloud. In an environment where there are frequent workload changes, there would be a high possibility of performance interference. In such a situation, our model can be used to select the optimal configuration for deploying a component that maximizes the number of request that can be allowed to access the component while at the same maximizing the degree of isolation between tenants (or components).

9.2. Monitoring Runtime Information of Components

Another area where our model can be very useful is in monitoring the runtime information of components. Many providers offer monitoring information on network availability and utilization of components deployed on their cloud infrastructure based on some pre-defined configuration rules. However, none of this information can assure that the component is functioning efficiently and guarantees the required degree of multitenancy isolation on the application level. It is the responsibility of the customer to extract and interpret these values, adjust the configuration, and thus provide optimal configuration that guarantees the required degree of multitenancy isolation.

Our model can be implemented in the form of a simple web service-based application to monitor the service and automatically change the rules, for example, based on previous experiences, user input, or once the average utilization of resources exceeds a defined threshold. This application can be either be deployed separately or integrated into different cloud-hosted services for monitoring the health status of a cloud service.

9.2. Managing the Provisioning and Decommissioning of Components

The provisioning and decommissioning of components or functionality offered to customers by many cloud providers is through the configuration of pre-defined rules. For example, a rule can state that once an average utilization of a system resource (e.g., RAM, disk space) exceeds a defined threshold then a component shall be started. When runtime information of components is extracted, and made available as stated earlier, they can be used to make important decisions concerning the provisioning of required components and decommissioning of unused components.

10. Conclusion and Future Work

Within this research, to provide a further contribution to the current literature on multitenancy isolation and optimized component deployment, we have developed an optimisation model alongside a metaheuristic solution largely inspired by simulated annealing for the purpose of delivering near-optimal component deployment solutions specific to cloud-based applications that guarantee multitenancy isolation. First, an optimisation problem was formulated to capture the implementation of the required degree of isolation between components. Next, the problem was mapped against an MMKP to enable early resolution through the use of a metaheuristic.

Results show greater consistency and dependability from SA(Greedy), particularly when functioning on large instances, whereas SA(Random) operates sufficiently on small instances. When considering algorithm strength, SA(Greedy) portrayed low variability and thus generates higher quality solutions in dynamic environments experiencing regular changes in workload levels. SA(random) appears to have greater sensitivity and instability to small deviations of input instances compared to SA(Greedy), more so in relation to large instances.

We intend on testing the MMKP problem instances further using various metaheuristic types (e.g., genetic and estimated distribution algorithms) and combinations (e.g., simulated

annealing merged with genetic algorithm) with the aim of identifying the most efficient metaheuristic to generate optimal solutions to operate in a variety of cloud deployment situations. For future research, a decision support system will be developed capable of creating or integrating an elastic load balancer for runtime information monitoring in relation to single specific components in order to deliver near-optimal component deployment solutions specific to responses required from cloud-hosted applications to changes in workload. This would be of immense value to cloud designers and SaaS customers for decision making around the provision and decommission of components.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research was supported by the Tertiary Education Trust Fund (TETFUND), Nigeria and IDEAS Research Institute, Robert Gordon University, UK.

References

[1] C. Fehling, F. Leymann, R. Retter, W. Schupeck, and P. Arbitter, *Cloud Computing Patterns*. Springer, 2014.

[2] R. Krebs, C. Momm, and S. Kounev, "Architectural concerns in multi-tenant saas applications." *CLOSER*, vol. 12, pp. 426–431, 2012.

[3] L. C. Ochei, A. Petrovski, and J. Bass, "Optimizing the Deployment of Cloud-hosted Application Components for Guaranteeing Multitenancy Isolation," 2016 International Conference on Information Society (i-Society 2016).

[4] L. C. Ochei, J. Bass, and A. Petrovski, "Implementing the required degree of multitenancy isolation: A case study of cloud-hosted bug tracking system," in *13th IEEE International Conference on Services Computing (SCC 2016)*. IEEE, 2016.

[5] F. Shaikh and D. Patil, "Multi-tenant e-commerce based on saas model to minimize it cost," in *Advances in Engineering and Technology Research (ICAETR), 2014 International Conference on*. IEEE, 2014, pp. 1–4.

[6] D. Westermann and C. Momm, "Using software performance curves for dependable and cost-efficient service hosting," in *Proceedings of the 2nd International Workshop on the Quality of Service-Oriented Software Systems*. ACM, 2010, p. 3.

[7] Z. I. M. Yusoh and M. Tang, "Composite saas placement and resource optimization in cloud computing using evolutionary algorithms," in *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*. IEEE, 2012, pp. 590–597.

[8] D. Candeia, R. A. Santos, and R. Lopes, "Business-driven long-term capacity planning for saas applications," *IEEE Transactions on Cloud Computing*, vol. 3, no. 3, pp. 290– 303, 2015.

[9] M. L. Abbott and M. T. Fisher, *The art of scalability: Scalable web architecture, processes, and organizations for the modern enterprise*. Pearson Education, 2009.

[10] F. Leymann, C. Fehling, R. Mietzner, A. Nowak, and S. Dustdar, "Moving applications to the cloud: an approach based on application model enrichment," *International Journal of Cooperative Information Systems*, vol. 20, no. 03, pp. 307–356, 2011.

[11] A. Aldhalaan and D. A. Menascé, "Near-optimal allocation of vms from iaas providers by saas providers," in *Cloud and Autonomic Computing (ICCAC), 2015 International Conference on*. IEEE, 2015, pp. 228–231.

[12] E.-G. Talbi, *Metaheuristics: from design to implementation*. John Wiley & Sons, 2009, vol. 74.

[13] D. Menasce, V. Almeida, and D. Lawrence, *Performance by design: capacity planning by example*. Prentice Hall, 2004.

[14] F. Rothlauf, *Design of modern heuristics: principles and application*. Springer Science & Business Media, 2011.

[15] J. E. Beasley, "Or-library: distributing test problems by electronic mail," *Journal of the operational research society*, vol. 41, no. 11, pp. 1069–1072, 1990.

[16] Z. Eckart and L. Marco. Test problems and test data for multiobjective optimizers. Computer Engineering (TIK) ETH Zurich. [Online]. Available: http://www.tik.ee.ethz.ch/sop/.../testProblemSuite/

[17] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach," *IEEE transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 257–271, 1999.

[18] R. Parra-Hernandez and N. J. Dimopoulos, "A new heuristic for solving the multichoice multidimensional knapsack problem," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 35, no. 5, pp. 708–717, 2005.

[19] N. Cherfi and M. Hifi, "A column generation method for the multiple-choice multi-dimensional knapsack problem," *Computational Optimization and Applications*, vol. 46, no. 1, pp. 51–73, 2010.

[20] L. C. Ochei, A. Petrovski, and J. Bass, "An approach for achieving the required degree of multitenancy isolation for components of a cloud-hosted application," in *4th International IBM Cloud Academy Conference (ICACON 2016)*, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 6, 175-185 (2017)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Data Exchange Strategies for Aggregating Geographical Distribution of Demands for Location-Dependent Information Using Soft-State Sketches in VANETs

Akihiro Yamada[*,1], Susumu Ishihara[2]

[1] *Graduate School of Integrated Science and Technology, Shizuoka University, 4328011, Japan*

[2] *College of Engineering, Academic Institute, Shizuoka University, Hamamatsu, 4328011, Japan, ishihara.susumu@shizuoka.ac.jp*

ARTICLEINFO

Article history:
Received: 30 October, 2017
Accepted: 02 December, 2017
Online: 14 December, 2017

Keywords:
VANET
Data Dissemination
Aggregation

ABSTRACT

When we drive a car, we often want to know the circumstances of roads between the current location and the destination. Ishihara et al. proposed a Vehicular ad-hoc NETworks (VANET) -based information sharing system, called the Real-Time Visual Car Navigation System, in which a driver can obtain Location-Dependent Information (LDI), such as photographs or videos of his/her Point of Interest (PoI), by telling the onboard device, e.g., a car navigation device, the PoI. The simplest way to provide LDI to vehicles is to flood the LDI to all vehicles, but unneeded LDI may be disseminated, and network resources may be wasted. To disseminate LDI to many vehicles that demand it with low communication traffic, we have proposed a data dissemination scheme based on Demand maps (Dmaps). In this scheme, each vehicle has a Dmap, a data set representing the geographical distribution of the strength of demands for LDI. To keep the Dmap up to date, each vehicle exchanges a subset of data constituting a Dmap (Dmap Information: DMI) with other vehicles. Vehicles preferentially send new LDI or forwarded LDI strongly demanded to the area where the LDI is frequently demanded based on Dmaps. If vehicles frequently send DMI, the accuracy of the Dmap becomes high, however, communication traffic becomes large. In this paper, we propose strategies for controlling the frequency of sending DMI and strategies for selecting DMI to be sent to improve the Dmap accuracy with low communication traffic. Simulation results show that one of the proposed strategies, the LDI Transfer Zone (LTZ) strategy, achieves highly accurate Dmaps with a small amount of communication traffic.

1 Introduction

If a driver can visually know the current circumstances of roads on a route to his/her destination, such as traffic jams, roadwork, accidents, and beautiful sights, he/she may select a better (faster, easier, or more fun) route to the destination. Currently, there are some applications that inform a driver of traffic information to make driving more comfortable, such as Google maps and ETC2.0 [1] in Japan, etc. In addition, Pioneer has provided a cellular network-based system that allows a driver to visually know the current circumstances of roads predefined by the system [2]. However, to the best of our knowledge, there is no application that visually informs a driver of the current circumstances of roads that are designated by the driver.

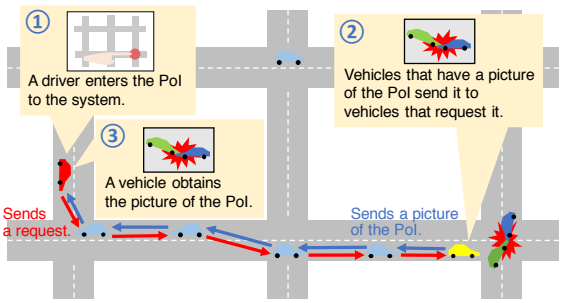


Figure 1. Real-Time Visual Car Navigation System

[*]Corresponding Author: Akihiro Yamada, Shizuoka University, 3-5-1 Johoku, Naka-ku Hamamatsu, Japan, akihiro@ishilab.net

Ishihara et al. proposed a Vehicular Ad-hoc NETworks (VANET)-based road traffic information sharing system, called the *Real-Time Visual Car Navigation System* [3–5]. This system allows driver to obtain photographs/videos of his/her Point of Interest (PoI) that is inputted to an onboard device, e.g., a car navigation device. In this system, when a driver inputs his/her PoI to an onboard device, the vehicle sends a request message to vehicles in the vicinity of the PoI. Then, vehicles that have a photograph/video of the PoI send it to the requesting vehicle (Figure 1).

In this system, we assume that vehicles use only VANET to send request messages to vehicles in the vicinity of a PoI and to send photographs/videos of PoI to requesting vehicles. If we use a simple pull-based strategy for sending the photographs/videos in the *Real-Time Visual Car Navigation System*, a requesting vehicle sends request messages to vehicles in the vicinity of a PoI, and vehicles that received a request message respond by sending photographs/videos of the PoI to the requesting vehicle. With this strategy, it is difficult for vehicles to certainly send the photographs/videos to the requesting vehicles with low latency because the connectivity between vehicles is intermittent due to the high mobility of vehicles. Furthermore, if multiple vehicles that are close together demand photographs/videos of the same PoI at the same time, the same data and the same request messages are repeatedly sent to the same location at the same time. Consequently, network resources may be wasted. Cooperative caching techniques such as ones proposed in [6] can be used to mitigate this problem. Another problem of pull-based schemes is that it is difficult to constantly deliver newly generated Location-Dependent Information (LDI) of PoI items to vehicles having interest in the PoI.

On the other hand, if all vehicles that have photographs/videos of the PoI send them based on a simple push-based strategy for certainly sending the photographs/videos to requesting vehicles, all vehicles send the photographs/videos whenever vehicles receive or generate them. Thus, unneeded data may be disseminated and limited network resources are wasted.

To ensure that newer LDI is sent to requesting vehicles with low communication traffic, Ishihara et al. have proposed a data dissemination scheme based on *Demand maps* (Dmaps) [4]. A Dmap is a data set representing the geographical distribution of the strength of demands for LDI. This scheme assumes that every vehicle has a Dmap. Vehicles grasp the geographical distribution of the strength of demands for LDI using a Dmap, and they send LDI to areas where it is strongly demanded via an efficient route. To keep Dmaps on vehicles up to date, each vehicle broadcasts several data constituting its Dmap (Dmap Information: DMI) with a beacon that it sends periodically. When vehicles receive other DMI, they update their Dmap by merging the received DMI with their own.

Since vehicles send LDI in accordance with Dmaps, the accuracy of Dmaps, i.e. the reproducibility of real geographical distribution of the strength of demands in Dmaps, should be high. Each vehicle should frequently update its Dmap by frequently exchanging DMI with other vehicles to make the accuracy of Dmaps high. However, if each vehicle frequently sends DMI, communication traffic becomes high. Therefore, strategies to increase the accuracy of Dmaps and control communication traffic are needed. In [4], Ishihara, et al. design the basic specification of Dmap-based data dissemination, however, they did not design the details of how to send DMI.

In this paper, to solve the tradeoff between Dmap accuracy and communication traffic, we propose strategies that control the frequency of sending DMI and strategies that each vehicle selects DMI to be sent. In addition, we evaluate the accuracy of Dmaps and communication traffic in the case that vehicles use the proposed strategies through simulations.

The contributions of this paper are summarized as follows.

1. We propose a strategy for controlling the frequency of broadcasting DMI. In this strategy, each vehicle determines whether it sends DMI based on the frequency of updating its DMI.

2. We propose three strategies for selecting DMI to be sent: i) the Recently Update First (RUF) strategy, which considers how new each DMI is, ii) the LDI Transfer Zone-based (LTZ) strategy, which considers the relation between each PoI and each location where requests occur, and iii) a hybrid strategy.

3. We evaluated the effectiveness of the proposed strategies through simulations and confirmed that the LTZ strategy most effectively reduces communication traffic and makes the accuracy of Dmaps high.

This paper is an extension of work originally presented in 2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA) [7]. In this paper, we conduct simulations with more appropriate parameters, and provide a more detailed explanation of the proposed strategies and Dmap-based data dissemination. In addition, we improve the implementation of our proposed strategies in an existing simulator, and obtain more reliable simulation results. The description of related work is updated.

The remainder of this paper is composed as follows. In the next section, we review related work. In Section 3, we recapitulate the data dissemination scheme based on Dmaps and introduce the data structure of Dmap based on soft-state sketches. In Section 4, we propose strategies for sending DMI to improve the accuracy of Dmaps with low communication traffic. Then, in Section 5, we evaluate the effectiveness of the proposed strategies. Finally, we conclude this paper in Section 6.

2 Related Work

In Dmap-based data dissemination, we assume that a vehicle that has generated LDI appends the location where it generated the LDI. In addition, vehicles can know locations where the LDI is demanded based on their Dmap. Vehicles carry out the routing of LDI by referring to a location where the LDI is generated and locations where the LDI

is demanded. This concept is similar to Named Data Networking (NDN). In NDN, vehicles give a name, e.g. a location where data item is generated, to each data item and deliver the data item to vehicles requesting the data item associated with that name.

In [8], Grassi et al. proposed Navigo, an interest forwarding scheme in NDN. In Navigo, each vehicle that has received or generated an interest forwards it based on the shortest path between the location where it exists and the location where many vehicles having the demanded data exist. This scheme aims to deliver interests to locations where many vehicles having the required data are. In Navigo, vehicles send requests (interests) to many vehicles having the required data by sending the requests to the locations of these vehicles. In the Dmap-based scheme, vehicles register their requests to their Dmap instead of sending requests to the PoIs. In addition, vehicles having LDI can know not only a request for the LDI, but also the geographical distribution of the strength of demands for the LDI, and utilize this information for efficiently routing the LDI.

In [9], Ahmed et al. proposed CODIE as a data dissemination scheme using NDN. In CODIE, vehicles that forward an interest append the hop count of the interest from the requesting vehicle, and vehicles having the required data forward the data in accordance with the hop count. In this scheme, the same LDI may be disseminated to the same region at the same time, and network resources may be wasted. On the other hand, in the Dmap-based scheme, each vehicle grasps regions where each LDI is highly desired as well as its distribution. Therefore vehicles may be able to disseminate demanded LDI to many vehicles with a small number of the LDI transfers. In addition, Bian et al. introduced the NDN architecture to VANETs and proposed forwarding and caching schemes [10].

In Dmap-based data dissemination, requesting vehicles register their requests to their Dmaps. In the publish/subscribe (pub/sub)-based approach [14], e.g. [15] and [16] , requesting vehicles register their interests and their IDs to broker nodes. We regard that Dmap-based data dissemination is a kind of pub/sub based approach. A Dmap can be regarded as information on subscription and subscriber, and locations where requesting vehicles exist can be regarded as subscribers. In the Dmap-based scheme, each requesting vehicle does not register their IDs to a Dmap. However, locations where it requests LDI and locations where the LDI is generated are registered to a Dmap when it demands LDI. By disseminating LDI based on locations where requesting vehicles exist, the Dmap-based scheme can efficiently disseminate LDI to requesting vehicles compared with disseminating LDI based on IDs of requesting vehicles.

In our scheme, each vehicle sends DMI with a beacon that is sent periodically. In this paper, we discuss controlling the frequency of sending DMI. Many schemes for controlling the frequency of sending beacons have been proposed for sharing traffic information with low communication traffic. Dynamic beaconing (DynB) [11], UV-CAST [12], and DV-Cast [13] control the frequency of sending beacons adaptively to the density of vehicles and communication condition. In this paper, we propose a strategy that vehicles control the frequency of sending DMI adaptively to the relationship between the frequency of updating their DMI and the frequency that neighboring vehicles update their DMI and sharing a Dmap with low communication traffic. Strategies that use density of vehicles and communication condition, such as DynB, UV-CAST, and DV-Cast can be applied to our strategy.

Dmap-based data dissemination proposed in [4] and [17] can be categorized into a kind of data dissemination system for sharing Location-Dependent Information (LDI) in VANETs. Traffic information is typical LDI treated in VANETs. Various applications for sharing traffic information in VANETs, such as SOTIS [18], Trafficview [19], CASCADE [20], and Zone Flooding [21] have been proposed so far. They have a function for aggregating LDI obtained in a distributed manner. For example, in SOTIS, vehicles periodically broadcast a packet including traffic information, e.g. their location, velocity, and direction. When a vehicle receives information from other vehicles, it aggregates all the known information on each road segment to one average value. In those applications, unneeded information may be sent because vehicles broadcast traffic information without considering how strongly the traffic information is required. On the other hand, Dmap-based data dissemination aims to reduce redundant information dissemination by disseminating strongly required information in accordance with a Dmap.

In this work, a Dmap is represented with soft-state sketches proposed in [22]. A soft-state sketch is a data structure to stochastically estimate the number of unique data for aggregating data and reducing communication traffic. A soft-state sketch is represented by an array of non-negative values. In [22], soft-state sketches are used for counting the number of available parking spaces every load segment and sharing this information among vehicles. Vehicles use a set of soft-state sketches for estimating the number of available parking spaces in a load segment. In [22], each vehicle has a set of soft-state sketches every load segment. When each vehicle finds available parking spaces, it inputs the initial TTL of available parking spaces information to a soft-state sketch corresponding to the load segment where the vehicle exists. In addition, vehicles append sets of soft-state sketches to a beacon that is sent periodically, and update their information by merging their own soft-state sketches with received ones.

On the other hand, in this paper, we use soft-state sketches for counting the number of demands for LDI and sharing this information among vehicles. In our scheme, each vehicle has a set of soft-state sketches for every pair of each location where LDI is generated and each location where LDI is required. Then, each vehicle inputs the initial TTL to a soft-state sketch corresponding to its request when it requests LDI, and sends sets of soft state sketches as DMI with a beacon.

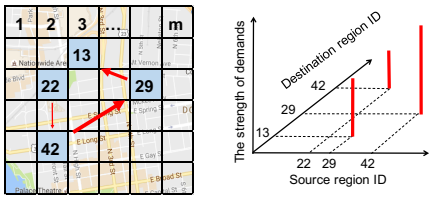
In [22], to reduce communication traffic for sending sets of soft-state sketches, a method for compressing soft-state sketches is proposed. However, we aim to reduce communication traffic by controlling the frequency of sending sets of soft-state sketches and selecting sets of soft-state sketches that vehicles should send.

3 A data dissemination scheme based on Demand map

In this section, we describe the details of the data dissemination scheme based on a Demand map (Dmap). In this scheme, each vehicle has a Dmap, which is a data set representing the geographical distribution of the strength of demands for LDI. Each vehicle forwards strongly demanded LDI items preferentially in accordance with its Dmap.

3.1 Demand map (Dmap)

A Dmap represents the geographical distribution of the strength of demands for LDI. We assume that the strength of demands for LDI is measured by the number of demands generated by vehicles. If multiple vehicles in geographically distant locations demand LDI of the same PoI, e.g. vehicles in the east end of a city and vehicles at the west end of the city require LDI of the city center, the routes for delivering the LDI to the vehicles in distant locations will be different. Thus, we divide an area into multiple regions, and a Dmap manages the strength of demands for LDI of each region every each region where demands are generated.



(a) Regions in a Dmap (b) Concept of a Dmap

Figure 2. Demand map

A Dmap is created as follows. Vehicles divide an area supported by their Dmap into multiple small regions and give a unique ID to each small region as shown in Figure 2 (a). We refer to a region where a requesting vehicle exists as a source region, and refer to a region where the requested LDI is generated as a destination region. For instance, in Figure 2, if a vehicle in region 29 requires LDI of region 13, we call region 29 a source region and call region 13 a destination region. A Dmap records the number of demands, i.e. the strength of demands, every pair of each source region and each destination region. For instance, in Figure 2 (a), if the start point of an arrow is a source region, and the end point of an arrow is destination region, and width of an arrow indicates the strength of demands corresponding to the arrow, a Dmap is represented as shown in Figure 2 (b).

3.2 How vehicles use a Demand map for disseminating Location-Dependent Information (LDI)

A Dmap on each vehicle is used for decision of routing of LDI. When a Dmap is updated or each vehicle generates new LDI or receives LDI, each vehicle refers to its Dmap and knows the number of demands per pair of a source region and a destination region. In addition, the vehicle refers to LDI time stamps, and it preferentially selects LDI that was required many times from sets of LDI that are current. If vehicles that are far from a source region send LDI, it may become outdated during delivery, even though LDI should be new. Therefore, vehicles that are far from the source region do not send LDI.

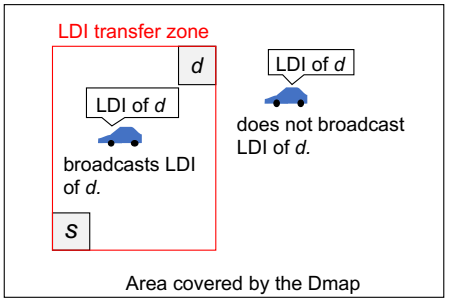
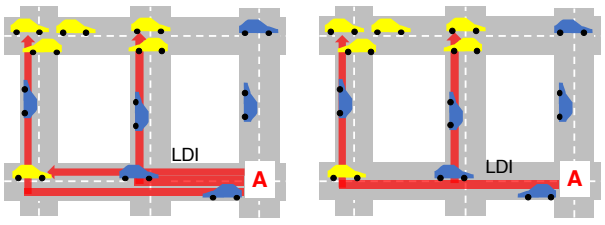


Figure 3. Disseminating LDI of s to d in accordance with a Dmap

In addition, LDI should be delivered from the destination region to source regions via a route as short as possible. To realize this, we defined a zone where LDI is delivered as an LDI Transfer Zone (LTZ). An LTZ is a rectangular zone that consists of a diagonal whose ends are a source region and a destination region (Figure 3). When a vehicle sends LDI, it determines whether it sends the LDI based on whether it is in the LTZ corresponding to the LDI. A vehicle sends an LDI item only when it is in the LTZ corresponding to the pair of the destination region and the source region of the LDI (Figure 3). If the vehicle is in one of the LTZs corresponding to LDI that is strongly demanded from many source regions, it aggressively broadcasts the LDI item.

LDI is delivered via many source regions. Thus, in the case that LDI is demanded from multiple source regions at the same time, if duplicate routes exist, LDI is not passed through the duplicate route multiple times, and redundant dissemination may be reduced compared with simple pull-based strategy. For instance, in Figure 4 (a), if yellow cars request LDI of A, the LDI is sent via routes represented by red arrows multiple times and duplicate dissemination occurs. However, in Figure 4 (b), if all cars have the same content of a Dmap, the LDI is sent via a route represented by red arrows once, then, duplicate disseminations do not occur.



(a) Simple pull-based strategy (b) Dmap-based strategy

Figure 4. Comparison of Dmap-based strategy and simple pull-based strategy

We assume that vehicles send new LDI demanded by many vehicles to the area where the demanding vehicles exist so that the LDI can be delivered via a route satisfying the aforementioned requirements. A routing strategy based on a propagation function proposed in [23] can be used for

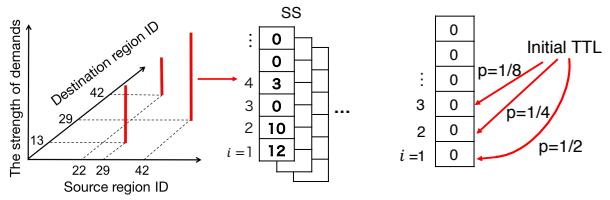
this purpose.

By this operation, vehicles that want the LDI will be able to obtain the LDI with a low data delivery delay and a small amount of traffic. For achieving high reliability and low data traffic for delivering LDI, the behavior of neighboring vehicles that have the same LDI and potentially broadcast the LDI should be considered. The detailed design of data forwarding algorithms based on a Dmap and behavior of the neighboring vehicles is beyond the scope of this paper.

3.3 Data structure of Demand map

One main scheme to create a Dmap is to collect all log data concerning demands that are generated by all vehicles, e.g. source region ID, destination region ID, car ID, and Time to Live (TTL) of demands, etc. However, if vehicles create their Dmap in accordance with this main scheme, all vehicles broadcast all log data concerning demands, and data size that vehicles have to broadcast and manage in their local storage increase in proportion to the number of demands and the number of vehicles.This leads to low scalability. To reduce the required communication traffic for maintaining Dmaps and the required storage on vehicles, a Dmap should be represented by a suitable data structure and needs algorithms to handle the data that can aggregate many demands generated by different vehicles in different locations.

Therefore, in [4] and [17], Ishihara et al. proposed to represent a Dmap by using soft-state sketches proposed in [22], a data structure to aggregate data and stochastically count the number of unique data. In [22], soft-state sketches are used for managing observation of free parking spaces in the same area by multiple vehicles. Meanwhile, we use soft-state sketches for counting the number of demands for the same pair of source region and destination region.



(a) Dmap with soft-state sketch (b) Putting TTL

Figure 5. Soft-state sketches in a Dmap

To manage a Dmap, each vehicle has a set of h ($h \geq 1$) sketches for a pair of source region and destination region (Figure 5 (a)). A sketch is an array of non-negative integers as shown in Figure 5 (a). Hereafter, we refer to a set of sketches as an SS for simplicity. Since an SS is used for representing DMI for a pair of source region and destination region, we will use these words interchangeably.

When a vehicle requires LDI of a PoI, it randomly selects a sketch from the SS corresponding to the pair of the vehicle's current region, i.e. the source region of the demand, and PoI, i.e. the destination region of the demand (Figure. 6). Then, it puts the initial TTL for the demand into i-th element from the smallest index of elements in the selected sketch with a probability of $1/2^i$ (Figure 5 (b)). The TTL value in each element is decremented as time progresses until the TTL value becomes 0. Since the probability that the TTL value is put into an element with a small index i is high and the probability that the TTL value is put into an element with a large index i is low, elements with a small index easily have non-zero values, but ones with a large index are hardly filled with a non-zero value. Therefore, the more demands occur, the more easily elements with large indexes are filled with non-zero values and the longer the run length of non-zero values from an element with the smallest index of a sketch is. Thus, the number of generating demands for a pair of source region and destination region, i.e. the strength of demands for the pair, is estimated based on the run length of non-zero values from an element with the smallest index of each sketch in an SS for the pair.

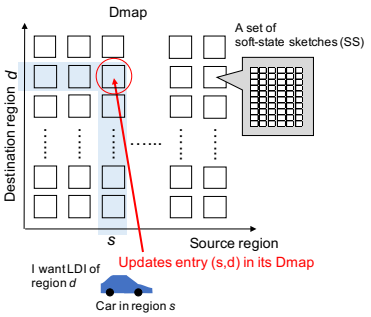


Figure 6. Behavior of a vehicle when it demands LDI

To keep Dmaps on vehicles up to date and share Dmaps among vehicles, each vehicle selects several SSs and sends them with a beacon that is sent periodically. When a vehicle receives SSs from another vehicle, it merges each of the received SSs with a corresponding SS in its Dmap. To merge the received SSs with the corresponding SSs in a local Dmap, the vehicle selects a larger value in the same element of the same index of sketches in each of received SSs and the corresponding SS in the local Dmap. Then the selected value is set to the element of the sketch of the SS in the receiver vehicle's Dmap as shown in Figure 7.

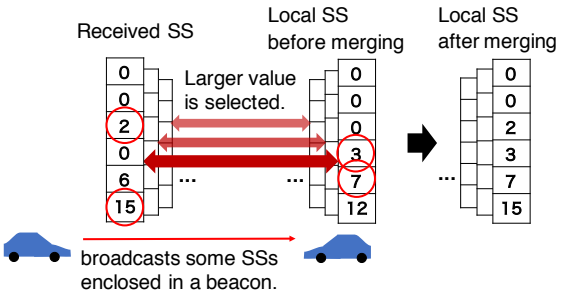


Figure 7. Merging with local SSs and other's

The number of demands for a pair of a source region and a destination region recorded in an SS consisting of m sketches $(S_1, \ldots, S_m)$ is estimated by the following formula presented in [22],

$$C(S_1, \ldots, S_m) := m \cdot \frac{2^{\sum_{j=1}^{m} Z(S_j)/m} - 2^{-\kappa \cdot \sum_{j=1}^{m} Z(S_j)/m}}{\varphi}, \quad (1)$$

with $\kappa \sim 1.75$, $\varphi \sim 0.77351$. S_j is a sketch in an SS, and m is the number of sketches composing an SS. $Z(S_j)$ is the run length of non-zero values from an element with the smallest index in S_j. In addition, the standard error of the estimate is $0.78/\sqrt{m}$. Therefore, the larger the number of sketches that compose an SS, the higher the accuracy of the estimate. However, data size that each vehicle has and sends increases in proportion to the number of sketches that compose an SS.

4 Strategies for efficiently sending DMI

In this section, we propose a strategy that controls the frequency of sending DMI and strategies that select DMI that vehicles should send for sending DMI with making the accuracy of Dmap high and making communication traffic low. To reflect current demands in a Dmap certainly, i.e. to improve the accuracy of a Dmap, it is desirable for each vehicle to frequently receive SSs for all pairs of all source regions and all destination regions from many vehicles and make its local Dmap up to date. However, if all vehicles frequently send all SSs that they have, the communication resource will be wasted because unneeded SSs may be sent and vehicles may send duplicate SSs. Therefore, we propose strategies for making the accuracy of Dmap high with low communication traffic.

4.1 Strategy controlling the frequency of sending DMI — Frequent strategy

We propose *selecting frequently updating vehicles* strategy, a strategy that controls the frequency of sending DMI (= SSs), to reduce the communication traffic for sending SSs without deteriorating the accuracy of Dmaps. Hereafter, we refer to this strategy as *Frequent* strategy for simplicity. Based on this strategy, when a vehicle sends a beacon, it determines whether it sends SSs with a beacon. The key idea of this strategy is that vehicles that frequently receive and update SSs may have SSs that certainly reflect the actual number of demands. On the other hand, vehicles that rarely receive and update SSs may have SSs that incorrectly reflect the actual number of demands. In addition, vehicles that frequently generate demands and update their SSs may have many SSs that are still not reflected in the SSs of other vehicles. Therefore the vehicles should send their SSs frequently. Therefore, in the Frequent strategy, vehicles that frequently update their SSs have a high priority to send SSs. Here, *updating* an SS means that one of the elements in a sketch of the SS is overwritten by putting a larger value due to a new demand generation or receiving a corresponding SS that has at least one larger value for any of elements in sketches in the receiver vehicle's SS from a neighboring vehicle.

In this strategy, each vehicle counts the total number of updates to its SSs during a certain past period, τ_u. We refer to the number as the Total Count of Updating Sketches (TCUS), for simplicity. In this strategy, vehicles whose TCUS is large are regarded as vehicles that frequently update their SSs. Vehicles whose TCUS is larger than one of their neighbor vehicles have a higher priority to send their SSs.

The operation of vehicles in this strategy is summarized follows. Each vehicle sends its TCUS with a beacon periodically. When a vehicle receives a beacon from another vehicle, it stores the received TCUS with the ID of the vehicle that has sent the TCUS. When a vehicle sends a beacon, it sorts TCUSs received from vehicles in the vicinity and its TCUS in descending order. Then, if its TCUS is in the top-k group of TCUSs, it decides to send SSs with a beacon (Figure 8). Otherwise, it does not send any SSs.For instance, as shown in Figure 8, if TCUS of car A is 4th from the largest TCUS and $k > 4$, car A sends SSs.

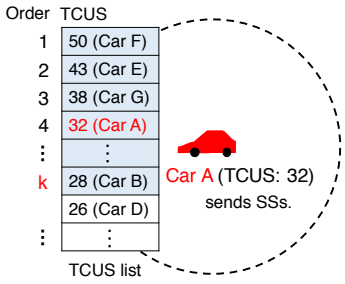


Figure 8. Frequent strategy

4.2 Strategies for selecting DMI that vehicles send

Each vehicle should select SSs to send so that Dmaps on vehicles, including itself, can achieve high accuracy with low communication traffic by eliminating redundant information that will not update the SSs of other vehicles. To achieve this, we propose three strategies, the Recently Updated First (RUF) strategy, the LDI Transfer Zone (LTZ) strategy, and the RUF + LTZ Hybrid strategy.

4.2.1 Recently Updated First (RUF) strategy

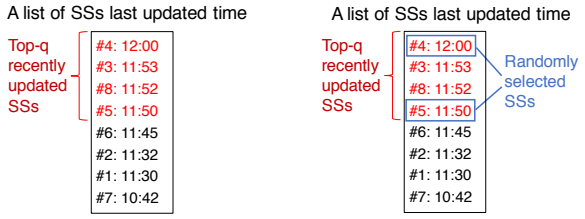
Recently updated SSs may have a large TTL value and may update SSs on vehicles that receive them, while SSs that are updated long ago may have a small TTL value and may not update SSs of the receiver vehicles. Sending SSs that will not update SSs of other vehicles wastes limited communication resources.

Therefore, in the RUF strategy, we give a high priority to recently updated SSs. When a vehicle updates an SS, it records the current time with a pair of a source region ID and a destination region ID. If a vehicle decides to send SSs with a beacon, it selects top-q recently updated SSs from its local SSs that have been updated so far (Figure 9 (a)).

However, if each vehicle selects SSs only based on the last update time of them, SSs that have the same pair of source region and destination region may frequently be sent and SSs that have other pairs will lose a chance to be sent. For instance, if an SS for a pair of source region s and destination region d is the most recently updated SS in a vehicle, the vehicle sends the SS. Then, when neighbor vehicles receive and update the SS, the SS for a pair of s and d becomes the most recently updated SS in the vehicles and the

vehicles will send it. The SS for a pair of s and d is repeatedly sent and SSs for other pair are rarely sent. Consequently, for some pairs of a source region and a destination region, the transition of the number of demands on a vehicle may not be reflected in SSs, and the Dmap accuracy may become low.

To provide the opportunity for not only the most recently updated SSs to be sent, in this strategy, each vehicle randomly selects up to n SSs from the top-q recently updated SSs and sends them (Figure 9 (b)).



(a) Selecting recently updated SSs (b) n SSs from the top-q SSs

Figure 9. RUF strategy

4.2.2 LDI Transfer Zone (LTZ) strategy

In Dmap-based data dissemination, we assume that vehicles send LDI for a pair of source region s and destination region d only when they are in the LTZ corresponding to s and d, as mentioned in Section 3.2. Therefore, vehicles in the LTZ corresponding to s and d should certainly know the number of demands for LDI for the pair of s and d. On the other hand, vehicles that are not in the LTZ do not have to know the number of demands for LDI. If the SSs for the pair of s and d are sent by only vehicles in the LTZ corresponding to s and d, the SSs are exchanged among only vehicles in the LTZ. As a result, the accuracy of the number of demands only for a pair of s and d composing an LTZ containing a vehicle holding the Dmap becomes high because noisy unneeded SSs are not sent and SSs that affect other vehicles SSs for a pair of s and d composing the LTZ are sent.On the other hand, exchanging SSs among vehicles outside the LTZ is not effective because vehicles do not need SSs that they will not use when they disseminate LDI.

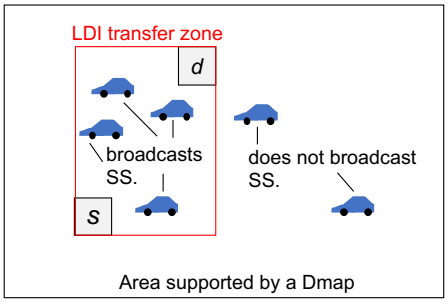


Figure 10. LTZ strategy

In the LTZ strategy, vehicles send SSs for the pair of s and d only when they are in the LTZ corresponding to s and d, and vehicles do not send the SSs when they are not in the LTZ. To this end, when each vehicle decides to send SSs, it first selects only SSs whose pair of source region s and destination region d satisfies $P \in A_{s,d}$. Here, P denotes the vehicle's current region, and $A_{s,d}$ denotes the LTZ for the pair of s and d (Figure 10). If the number of selected SSs is larger than the maximum number of SSs that can be included in a beacon packet, the vehicle randomly selects up to n SSs from the selected SSs.

4.2.3 RUF + LTZ Hybrid strategy

To leverage the advantages of both the RUF and LTZ strategies, we propose the RUF + LTZ Hybrid strategy. In this strategy, each vehicle sends more recently updated SSs only when it is in the LTZ corresponding to the SSs. Each vehicle selects and sends SSs as described below. If a vehicle v decides to send SSs with a beacon, it selects the top-q recently updated SSs from its local SSs that have been updated so far, like in the RUF strategy. Note that a vehicle selects SSs only based on whether SSs are more recently updated in this phase. We refer to the set of these SSs as $\boldsymbol{L}_v$. Then, it selects SSs whose pair of source region s and destination region d satisfies $P \in A_{s,d}$ from $\boldsymbol{L}_v$. We refer this set of selected SSs as $\boldsymbol{L}'_v$. Finally, v randomly selects up to n SSs from $\boldsymbol{L}'_v$ and broadcasts them. This algorithm is shown in Algorithm 1.

Algorithm 1 RUF + LTZ strategy

When vehicle v attempts to send some sets of sketches
$\boldsymbol{L}_v, \boldsymbol{L}'_v \leftarrow \phi$
$\boldsymbol{L}_v \leftarrow$ top q recently updated sets of sketches on v
for all (s,d) of a set of sketches in $\boldsymbol{L}_v$ **do**
  if $P_v \in A_{s,d}$ **then**
    add $S_{v,s,d}$ to $\boldsymbol{L}'_v$
  end if
end for
select n items randomly from $\boldsymbol{L}'_v$

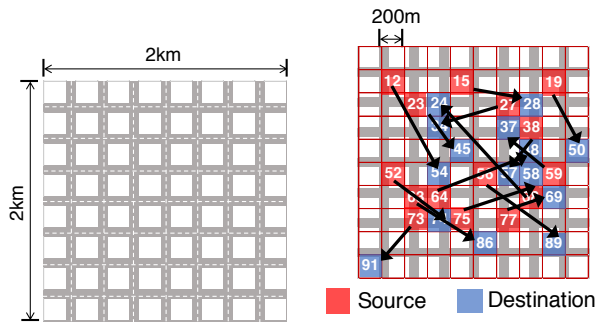
5 Evaluation of strategies for efficiently sending DMI

In this section, we evaluate the effectiveness of the proposed strategies through simulations. We focus on the accuracy of Dmaps and communication traffic required for sending DMI in this evaluation.

5.1 Simulation settings

5.1.1 Simulation scenario

In each scenario of our simulations, each vehicle broadcasts SSs to update and to share their Dmap using only VANET. We used Scenargie [24], a network simulator developed by Space-Time Engineering, and we implemented our proposed strategies and comparative strategies described later. In our scenarios, there are 14 double lane roads in a square area that has 2 km on each side as shown in Figure 11 (a). All intersections have traffic lights.



(a) Simulation area (b) Regions in Dmap

Figure 11. Simulation area and regions for a Dmap

All vehicles move on the roads in the simulation area by a mobility log generated using the SUMO road traffic simulator [25]. In addition, we conducted simulations of vehicle mobility by changing the inter-vehicle arrival rate. These vehicle mobility simulation logs are used for evaluating how the density of vehicles affects the accuracy of Dmaps and the effectiveness of the proposed strategies.

Table 1. Pairs of source region ID and destination region ID of demands generated in the simulations

Source	Destination	Source	Destination
12	54	23	45
19	50	15	28
52	74	63	86
56	89	27	34
38	57	73	91
75	58	77	69
59	37	64	48
68	24		

Table 2. Simulation parameters

Parameters	Values
Inter-vehicle arrival rate [/h·lane]	100, 300
Interval of sending beacons [sec]	1
Tx power [dBm]	10
Interval of generating demands [sec]	1
The effective period of a demand	63
Simulation time [s]	900
MAC Protocol	IEEE802.11p
Propagation model	ITU-R P.1411
App. max. data size [bytes]	500
Size per SS [bytes]	192
τ_u	30
k	10, 17
q	7
n	2
p	0.8

We divided the simulation area into small square regions 200×200m^2 and gave an ID to each small region for creating a Dmap of each vehicle Figure 11 (b). In Figure 11 (b), vehicles in blue regions demanded LDI of red regions corresponding to the blue regions. Table 1 also summarizes pairs of a source region and a destination region of demands that vehicles generated in the simulations. Vehicles in each source region shown in Table 1 demanded LDI of the destination region corresponding to the source region every second when they were moving towards the direction of the destination region. For instance, a vehicle in region 12 demanded LDI of region 54 every second only when it was moving eastbound or southbound.

We repeated simulation runs ten times with different random number seeds. Table 2 summarizes parameters used in the simulations. We determined the values of n, k and the interval of sending beacons based on preliminary simulations. We selected values of these parameters so that communication traffic could be as small as possible and the reproducibility of real demands in Dmaps (the ratio of coincidence described later) could be high based on the results of the preliminary simulations using the same scenarios.

5.1.2 Comparative strategies

To evaluate the effectiveness of proposed strategies, we used three comparative strategies, the *All* strategy, the *Random Timing* (RandT) strategy, and the *Random Selecting* (RandS) strategy. The All strategy and RandT strategy are for controlling the frequency of sending SSs and used for evaluating the effectiveness of the Frequent strategy. The RandS strategy is for selecting SSs to be sent, and used for evaluating the effectiveness of the LTZ strategy, RUF strategy, and RUF + LTZ Hybrid strategy.

In the All strategy, all vehicles do not control the frequency of sending SSs and send SSs with every beacon. In the RandT strategy, each vehicle decides to send SSs with a probability of p when it sends a beacon. In the RandS strategy, vehicles that send SSs randomly select n SSs to send from SSs that have been updated so far.

5.2 Simulation metrics

We evaluated the performance of the proposed strategies focusing on the accuracy of a Dmap in each vehicle and communication traffic required for sending DMI (= SSs). Here, the accuracy of a Dmap means the reproducibility of the actual number of demands in the Dmap. Since vehicles send LDI based on the estimated number of demands given by the Dmap, the reproducibility of the actual number of demands in the Dmap should be high. The reproducibility of the actual number of demands should be evaluated by using a relative metric because the influence of the error is different according to how the actual number of demands is. Therefore, for evaluating the accuracy of Dmaps, we use the ratio of coincidence between the actual number of demands and the estimated number of demands in Dmaps, M.

M for a pair of source region and destination region is defined as follows.

$$M = 1 - \min\{1, |R - D|/R\} \quad (0 \le M \le 1). \qquad (2)$$

where D denotes the estimated number of valid demands (TTL > 0) that are generated in a source region for a destination region, and R denotes the actual number of valid demands for the same combination of a source region and a destination region as D. In addition, we assume that R is not 0 in this formula. The closer M is to 1, the higher the accuracy of the Dmap. As mentioned above, only vehicles in an LTZ for the pair of source region s and destination region d should surely know the number of demands for LDI for the pair of s and d because we assume that vehicles send

the LDI only when they are in LTZ corresponding to s and d. Therefore, we evaluated M for the pair of source region s and destination region d in a Dmap in only vehicles that are in the LTZ for the pair of s and d at the time point of the data analysis.

We also evaluated the communication traffic required for sending SSs. It is the average value of the product of the data size of an SS and the total number of broadcasted SSs in each simulation.

5.3 Simulation results

5.3.1 The ratio of coincidence M

Figure 12 shows the CDF of M calculated 800 seconds after the start of the simulation with vehicle traffic of 300 vehicles/lane·hour when the All strategy is used for controlling the frequency of sending SSs, and we evaluated the performance of the proposed strategies for selecting SSs. The vertical axis is the ratio of vehicles that have an SS representing M below each M on the horizontal axis. Each curve in the graph corresponds to a pair of a source region and a destination region of demands shown in Table 1.

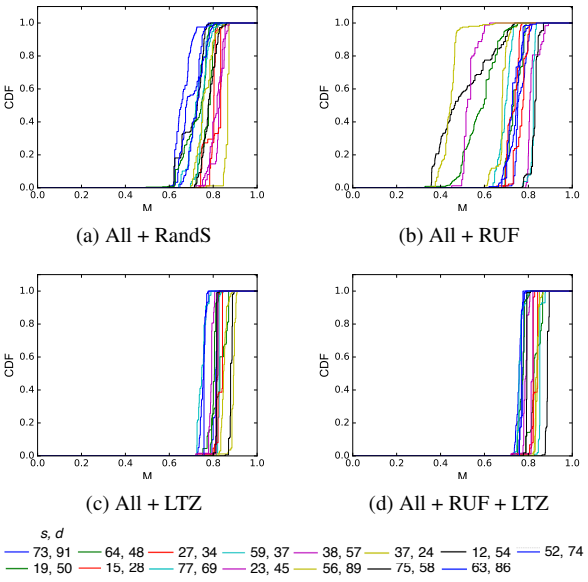


Figure 12. CDF of the Dmap accuracy (M) at 800 sec (inter-vehicle arrival rate: 300 vehicles/lane·hour)

As shown in Figure 12, the LTZ strategy and the RUF + LTZ strategy outperform the other strategies. When the RandS strategy is used, M of pairs of s and d (Hereafter, we call them sd-pairs) ranges from 0.6 to 0.86. With the LTZ strategy and the RUF + LTZ strategy, M of sd-pairs is improved compared with M when the RandS strategy is used. M of sd-pairs when the LTZ strategy or the RUF + LTZ strategy is used ranges from 0.7 to 0.86. This is because SSs for the pair of s and d are preferentially sent in LTZ corresponding to the SSs. The reason why M in the LTZ strategy is similar to M in the RUF + LTZ strategy is summarized as follows. SSs are frequently sent in their corresponding LTZ if the LTZ strategy is used, then the time stamp of the SSs on vehicles in the LTZ is frequently updated. Therefore, SSs are frequently sent in the LTZs when vehicles select recently updated SSs in RUF + LTZ strategy. Therefore, similar SSs are selected both in the LTZ strategy and the RUF + LTZ strategy. Consequently, M in the LTZ strategy is similar to M in the RUF + LTZ strategy. On the other hand, when only the RUF strategy is used for selecting SSs, M is deteriorated. The reason is that SSs of the same sd-pairs tend to be sent repeatedly and transmission of SSs of other sd-pairs is suppressed.

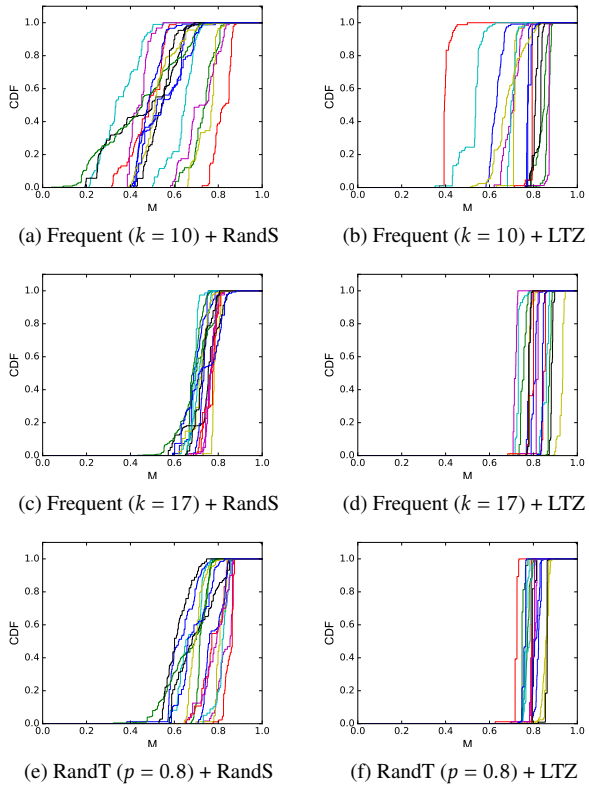
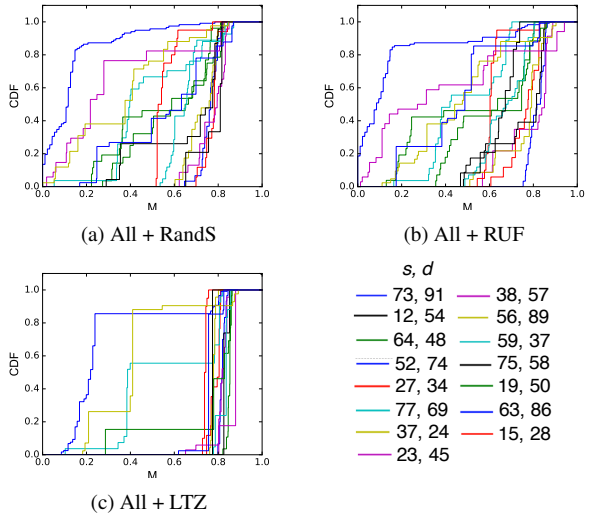


Figure 13. CDF of the Dmap accuracy (M) at 800 sec (inter-vehicle arrival rate: 300 vehicles/lane·hour)

Figure 13 shows the CDF of the Dmap accuracy (M) when the Frequent strategy and the RandT strategy were used for controlling the frequency of sending DMI (= SSs) with the RandS strategy or the LTZ strategy for selecting SSs. When vehicles use the RandT strategy and the RandS strategy, M is deteriorated compared to that in the case when the All strategy and the RandS strategy are used. This is simply because the number of exchanged SSs is decreased. However, when vehicles use the RandT strategy and the LTZ strategy, M is similar to one in the case when the All strategy and the LTZ strategy are used. The reason is that SSs are frequently exchanged in LTZ corresponding to them when the LTZ strategy is used even if the small number of vehicles send SSs. When the Frequent strategy with $k = 10$ is used, M is more deteriorated than that in the case when the RandT strategy is used. The reason is that the number of vehicles that send SSs is excessively restrained in the Frequent strategy with $k = 10$. On the other hand, when the Frequent strategy with $k = 17$ is used, M approximated M in the cases when the All strategy is used because the number of vehicles that send SSs was hardly not restrained.

Figure 14 shows CDF of M calculated at 800 seconds from the start of the simulation with 100 vehicles/lane·hour

vehicle traffic. Compared with Figure 12 (a)–(c), we can see that M is deteriorated when the vehicle density is low. This is because a sufficient number of SSs is not exchanged due to lack of connectivity between vehicles. Interestingly, even with poor connectivity, the LTZ strategy outperforms other strategies and most of the sd-pairs achieve high M around 0.8. The reason is that SSs are easily updated in LTZ corresponding to the SSs because SSs are frequently sent in the LTZ.



(a) All + RandS (b) All + RUF

(c) All + LTZ

Figure 14. CDF of M at 800 sec (inter-vehicle arrival rate: 100 vehicles/lane·hour)

5.3.2 Communication traffic

Figure 15 shows communication traffic in each combination of strategies when vehicle traffic is 300 vehicles/lane·hour. We can see that the LTZ strategy most effectively reduces communication traffic among strategies for selecting SSs. The reason is that each SS is sent only in the LTZ corresponding to the SS and is not sent outside the LTZ.

Concerning strategies for controlling the frequency of sending SSs, the Frequent strategy ($k = 10$) most effectively reduces the communication traffic. However, when the Frequent strategy (k=10) is used, M becomes low as shown in Figure 13 (a), (b) because SSs are rarely updated due to the limited chance to send SSs.

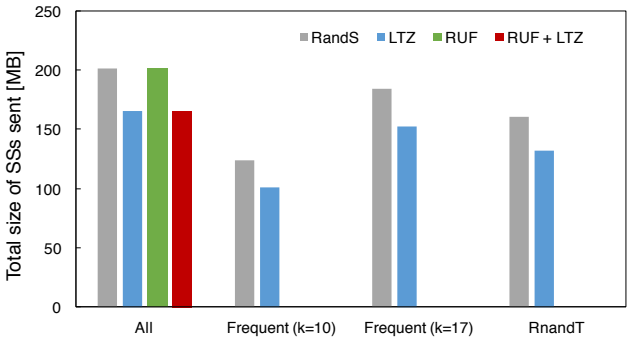


Figure 15. Communication traffic (Inter-vehicle arrival rate: 300 vehicles/lane·hour)

Summarizing the simulation results of the ratio of coincidence M and communication traffic, we can conclude that the LTZ strategy is the most useful to improve both the Dmap accuracy and communication traffic needed for exchanging SSs. The reason is that this strategy increases the frequency of SS exchange in their corresponding LTZ and reduces the frequency of exchanging them outside the LTZ. Concerning strategies for controlling the frequency of sending SSs, the Frequent strategy reduces the communication traffic. However, it deteriorates the Dmap accuracy. Therefore, the Frequent strategy does not solve the tradeoff between the Dmap accuracy and communication traffic. In addition, when both the LTZ strategy and RandT strategy are used, communication traffic and the Dmap accuracy is most effectively improved.

6 Conclusions

In this paper, we proposed a strategy for controlling the frequency of sending DMI (SSs) and strategies for selecting SSs that vehicles should send for improving the accuracy of a soft-state sketch-based Dmap on vehicles used for Dmap-based location dependent data dissemination services on VANETs, such as a real-time visual car navigation system. The simulation results revealed that one of the proposed strategies for selecting SSs, the LTZ strategy, effectively reduces communication traffic and achieves high Dmap accuracy, i.e. the reproducibility of the actual number of demands in Dmaps. The key feature of the LTZ strategy is that it is designed in consideration of where the exchanged information, i.e. SSs, may be used. In this strategy, vehicles send SSs in only locations where SSs may be used and do not send SSs in locations where SSs may not be used. Thus, it can reduce communication traffic and improve the accuracy of SSs in locations where SSs is used. On the other hand, the strategy for controlling the frequency of sending SSs, the Frequent strategy, which gives a high priority to vehicles frequently updating SSs, deteriorates the Dmap accuracy, while it drastically reduces communication traffic required for exchanging SSs.

The demand generation patterns used in the evaluation in the paper is limited. Thus, the performance evaluation of various demand generation patterns and realistic demand generation patterns are needed because there are many demand generation patterns in real life. For example, a demand generation pattern that frequency of demands change as time progress or change depending on a distance between a source region to a destination region should be considered. The mobility pattern of vehicles will also affect the performance of sharing Dmap information. For example, if an isolated vehicle has important Dmap information of a distant area, giving a high priority to exchanging the important Dmap information will contribute to that many vehicles knowing the distributions of demands for LDI. If many vehicles in the distant area want to obtain LDI of an area that has low connectivity from them in terms of store-carry-forward-based data delivery, such demand information will be carried by isolated vehicles.

For delivering LDI based on Dmaps, additional information is needed for each vehicle to determine whether it forwards an LDI item or not. For example, if its neighboring vehicle has sent the same LDI item recently, it should

not forward the LDI item to reduce the redundant data transmission without loss of the reliability of data delivery. Besides, if there are multiple isolated routes between the PoI and the destination of the LDI item, whether selecting one of the routes or using all the routes will affect the reliability of the LDI data delivery. In addition, we should design a criterion for selecting LDI that vehicles send, a method for calculating efficient forwarding routes based on Dmaps, etc. Our proposed strategies are important factors in the routing of LDI because vehicles disseminate LDI based on Dmaps in Dmap-based data dissemination. The design of a data forwarding algorithm based on Dmaps and additional information remains as our future work.

Acknowledgment This work was supported by JSPS KAKENHI Grant Number 15H02689.

References

[1] MLIT Ministry of Land, Infrastructure, Transport and Tourism: ITS, http://www.mlit.go.jp/road/road_e/p1_its.html#a2 (Accessed 14th September 2017)

[2] Pioneer news release: Pioneer introduces new CYBER NAVI car navigation systems for Japan Market, http://pioneer.jp/en/news/press/index/1617/May2013.

[3] J. Okamoto and S. Ishihara, "Distributing location-dependent data in VANETs by guiding data traffic to high vehicle density areas," in Vehicular Networking Conference, 2010. https://doi.org/10.1109/VNC.2010.5698248

[4] S. Ishihara, N. Nakamura, and Y. Niimi, "Demand-based Location Dependent Data Dissemination in VANETs" in the 19th annual international conference on Mobile computing & networking, 2013. https://doi.org/10.1145/2500423.2504580

[5] N. Nakamura, Y. Niimi, and S. Ishihara, "Live VANET CDN: Adaptive data dissemination scheme for location-dependent data in VANETs" in Vehicular Networking Conference, 2013. https://doi.org/10.1109/VNC.2013.6737595

[6] L. Yin and G. Cao, "Supporting cooperative caching in ad hoc networks" IEEE Trans. Mobile Computing, **5**(1), 77–89, 2006. https://doi.org/10.1109/TMC.2006.15

[7] A. Yamada and S. Ishihara. "Data Exchange Strategies for Aggregating Geographical Distribution of Demands for Location-Dependent Information Using Soft-State Sketches in VANETs" in Advanced Information Networking and Applications, 2017 IEEE 31st International Conference on, 2017. https://doi.org/10.1109/AINA.2017.42

[8] G. Grassi, D. Pesavento, and G. Pau, "Navigo: Interest forwarding by geolocations in vehicular Named Data Networking" in World of Wireless, Mobile and Multimedia Networks, 2015 IEEE 16th International Symposium on a, 2015. https://doi.org/10.1109/WoWMoM.2015.7158165

[9] S. H. Ahmed, S. H. Bouk, and M. A. Yaqub, "CODIE: Controlled Data and Interest Evaluation in Vehicular Named Data Networks" IEEE Trans. Vehicular Technology, **65**(6), 3954–3963, 2016. https://doi.org/10.1109/TVT.2016.2558650

[10] C. Bian , T. Zhao, X. Li , W. Yan, "Boosting named data networking for data dissemination in urban VANET scenarios" Vehicular Communications, **2**(4),195–207, 2015. https://doi.org/10.1016/j.vehcom.2015.08.001

[11] C. Sommer, S. Joerer, M. Segata, "How shadowing hurts vehicular communications and how dynamic beaconing can help" IEEE Trans. Mobile Computing, **14**(7), 1411–1421, 2015. https://doi.org/10.1109/TMC.2014.2362752

[12] W. Viriyasitavat, O.K. Tonguz, and F. Bai, "UV-CAST: an urban vehicular broadcast protocol" in Vehicular Networking Conference, 2010. https://doi.org/10.1109/VNC.2010.5698266

[13] O.K. Tonguz, N. Wisitpongphan, and F. Bai, "DV-CAST: A distributed vehicular broadcast protocol for vehicular ad hoc networks" IEEE Wireless Communications, **17**(2), 47–57, 2010. https://doi.org/10.1109/MWC.2010.5450660

[14] P. Th. Eugster, P. A. Felber, R. Guerraoui, and A. M. Kermarrec. "The many faces of publish/subscribe" ACM computing surveys, **35**(2), 114–131, 2003. https://doi.org/10.1145/857076.857078

[15] I. Leontiadis and C. Mascolo, "Opportunistic Spatio-Temporal Dissemination System for Vehicular Networks" in the 1st international MobiSys workshop on Mobile opportunistic networking, 2007. https://doi.org/10.1145/1247694.1247702

[16] T. Pandey, D. Garg, and M.M. Gore, "Publish/subscribe based information dissemination over VANET utilizing DHT" Frontiers of Computer Science, **6**(6), 713–724, 2012. https://doi.org/10.1007/s11704-012-1154-7

[17] Y. Niimi and S. Ishihara, "Demand map-based data dissemination scheme for location dependent data in VANETs" in Mobile Computing and Ubiquitous Networking, 2015 Eighth International Conference on, 2015. https://doi.org/10.1109/ICMU.2015.7061047

[18] L. Wischhof, A. Ebner, H. Rohling, "Information dissemination in self-organizing inter-vehicle networks" IEEE Trans. Intelligent Transportation Systems, **6**(1), 90–101, 2005. https://doi.org/10.1109/TITS.2004.842407

[19] T. Nadeem, S. Dashtinezhad, C. Liao, L. Iftode, "TrafficView: traffic data dissemination using car-to-car communication" ACM SIGMOBILE Mobile Computing and Communications Review, **8**(3), 6–9, 2004. https://doi.org/10.1145/1031483.1031487

[20] K. Ibrahim, M.C. Weigle, "Accurate data aggregation for VANETs" in the fourth ACM international workshop on Vehicular ad hoc networks, 2007. https://doi.org/10.1145/1287748.1287761

[21] J. Bronsted, L.M. Kristensen, "Specification and performance evaluation of two zone dissemination protocols for vehicular ad- hoc networks" in Specification and performance evaluation of two zone dissemination protocols for vehicular ad-hoc networks, 2006. https://doi.org/10.1109/ANSS.2006.43

[22] C. Lochert, B. Scheuermann, and M. Mauve, "A probabilistic method for cooperative hierarchical aggregation of data in VANETs" Ad Hoc Networks, **8**(5), 518–530, 2010. https://doi.org/10.1016/j.adhoc.2009.12.008

[23] D. Sormani, G. Turconi, P. Costa, D. Frey, M. Migliavacca, and L. Mottola, "Towards Lightweight Information Dissemination in Inter-Vehicular Networks" in the 3rd international workshop on Vehicular ad hoc networks, 2006. https://doi.org/10.1145/1161064.1161069

[24] Space-Time Engineering, https://www.spacetime-eng.com/ (Accessed 28th September 2017)

[25] SUMO — Simulation of Urban Mobility, http://sumo.dlr.de/wiki/Main_Page (Accessed 28th September 2017)

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 6, 186-193 (2017)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Fault Tolerant Control Based on PID-type Fuzzy Logic Controller for Switched Discrete-time Systems: An Electronic Throttle Valve Application

Wafa Gritli*, Hajer Gharsallaoui, Mohamed Benrejeb

Université de Tunis El Manar, LA.R.A Automatique, Ecole Nationale d'Ingénieurs de Tunis, BP 37, Le Belvédère, 1002 Tunis, Tunisie.

A R T I C L E I N F O

Article history:
Received: 09 November, 2017
Accepted: 01 December, 2017
Online: 14 December, 2017

Keywords:
Electronic throttle valve
Switched systems
Fault tolerant control
PID-type FLC
Unknown input observer
Lyapunov theory
LMIs

A B S T R A C T

This paper deals with the problem of Fault Tolerant Control (FTC) using PID-type Fuzzy Logic Controller (FLC) for an Electronic Throttle Valve (ETV) described by a switched discrete-time systems with input disturbances and actuator faults. In order to detect the faults, Unknown Input Observers (UIOs) are designed and formulated in terms of Lyapunov theory and Linear Matrix Inequalities (LMIs). This approach is designed in order to minimize the error between the desired flat trajectory generated using the flatness property and the estimated state provided from differents UIOs and to maintain asymptotic stability under an arbitrary switching signal, even in the presence of actuator faults. The simulation results have shown the effectiveness of the proposed approach.

1 Introduction

Actuator faults may cause undesired system behaviour and sometimes lead to instability, hence it is necessary to develop Fault Tolerant Control (FTC) methods against actuator faults of uncertain nonlinear systems. In the past decades some FTC design methods have been proposed for several classes of nonlinear systems with actuator faults [1-3]. It consists in computing control laws by taking into account the faults affecting the system in order to maintain acceptable performances and to preserve stability of the system in the faulty situations [4]. From the point of view of FTC strategies, the literature considers two main groups of techniques: passive and active ones. In passive FTC, the faults are treated as uncertainties. Therefore, the control is designed to be robust only to the specified faults [4]. Active FTC techniques consist in adapting the control law using the information given by the Fault Detection and Isolation (FDI) block [5]. The informations issued from the FDI block are used by the FTC module to reconfigure the control law in order to compensate the fault and to ensure an acceptable system performances.

The study of this problem was extended to switched systems in [6-8]. In [7], a switched discrete-time system with state delay has been considered. The design method is based on the construction of a filter and a fault estimation approach. In [8], an adaptive fuzzy tracking control method for a class of switched nonlinear systems with arbitrary switchings and with actuator faults has been proposed. The proposed control scheme guarantee the stability of the whole switched control system based on the common Lyapunov function stability theory and attenuate the effect of the actuator faults on the control performance by designing a new fuzzy controller to accommodate uncertain actuator faults. In [9], an observer has been built to detect the fault when it occurs. The problem of FTC for switchied linear systems is addressed by using a nominal control law designed in the absence of any fault, associated with fault detection, localization and reconfiguration techniques to maintain the stability of the system under an arbitrary switching signal in the presence of sensor faults. A state trajectory tracking has been proposed in [4] for actuator faults and observer bank based on controllers with switching mechanism for sensor faults has been also presented in [10]. A nonlinear observer based controller, adopting the so-

*Corresponding Author: Wafa Gritli, Email: wafa.gritli@enit.rnu.tn

https://dx.doi.org/10.25046/aj020623

called parallel distributed compensation structure, has been designed to choose an adequate state estimate to compensate the effects of the faults on the system in [11]. Switched systems are dynamical systems consisting of a collection of continuous-time subsystems. Switched systems have attracted more attention due to their significance in the modelling of many engineering applications, such as chemical processes, robot manipulators and power systems. Stability analysis and synthesis of switched systems have been made using Lyapunov function to ensure stability of the switched systems such in [12].

In this paper, in order to acheive the FTC for a switched discrete-time systems, a set of PID-type Fuzzy Logic Controller (FLC) is implemented to minimize the error between the desired flat trajectory and the estimated state and to maintain the stability of the system in the presence of actuator faults. The estimated state is provided from an Unknown Input Observer (UIO). Based on residual analysis, a switching strategy using stateflow is then designed. The global stability for the switched systems is studied by Lyapunov theory and expressed as a Linear Matrix Inequalities (LMIs).

The paper is organized as follows. In Section 2, the FTC problem statement is formulated. In Section 3, the fault detection method using UIO is introduced. In Section 4, the PID-type FLC is described. Section 5 deals with the flatness property. The proposed approach is applied to an ETV in Section 6. Finally, the conclusion is drawn in Section 7.

2 FTC problem statement

Consider the discrete time switched system, which can be formulated such that

$$x(k+1) = A_j x(k) + B_j u(k) + E_j d(k) + B_j f_a(k) \quad (1)$$

$$y(k) = C_j x(k)$$

$x(k) \in \mathbb{R}^n$ is the state vector, $u(k) \in \mathbb{R}^p$ the control input, $y(k) \in \mathbb{R}^o$ the measured output, $d(k) \in \mathbb{R}^p$ the unknown disturbance input and $f_a(k) \in \mathbb{R}^p$ the actuator faults. $A_j \in \mathbb{R}^{n\times n}$, $B_j \in \mathbb{R}^{n\times p}$, $C_j \in \mathbb{R}^{o\times n}$ and $E_j \in \mathbb{R}^{n\times p}$ are the known constant matrices for $j \in \psi = \{1,2,...,m\}$ and m the number of models, $m > 1$.

In this paper, a FTC structure based on PID-type FLC, given by Figure 1, is proposed to maintain the trajectory tracking and to preserve stability of ETV in the presence of both input disturbance $d(k)$ and actuator faults $f_a(k)$.

According to this structure, the FTC approach needs to detect actuator faults firstly and then to design the j^{th} control law, denoted $u_j(k)$ given by (2), in order to minimize the error between the desired flat trajectory, generated using the flatness property and the estimated state, provided from the j^{th} UIO.

$$u_j(k) = -K_{c,j}\hat{x}_j(k) + y_j^d(k) \quad (2)$$

$K_{c,j}$ is the j^{th} gain control law.

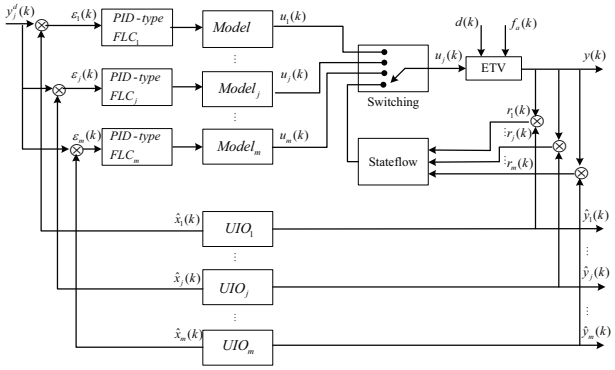


Figure 1: Proposed structure of FTC approach

In this case, LMI-based UIOs for the switched system (1) are designed using Lyapunov function for arbitary switching signal. Then, residuals $r_j(k)$ are determined by the UIOs and employed to achieve actuator faults detection.

3 Unknown input observers design and stability analysis

In this section, let us consider $f_a(k) = 0$, then the model (1) becomes

$$x(k+1) = A_j x(k) + B_j u(k) + E_j d(k) \quad (3)$$

$$y(k) = C_j x(k)$$

$j \in \psi$ denotes the j^{th} model. The structure of full rank UIOs can be formulated by

$$z(k+1) = F_j(k) + T_j B_j u(k) + K_j y(k) \quad (4)$$

$$\hat{x}(k) = z(k) + H_j y(k)$$

where $\hat{x}(k)$ is the estimated state vector $x(k)$, $z(k)$ is the state vector of full rank UIOs. F_j, T_j, K_j and H_j are unknown matrices which need to be designed.

Lemma: [13] Equation (4) is UIO of the switched system (3), if and only if the following conditions are satisfied

- $rank(C_j E_j) = rank(E_j)$
- $(C_j A_{j1})$ is observable

with

$$A_{j1} = A_j - E_j\left[(C_j E_j)^T C_j E_j\right]^{-1}(C_j E_j)^T C_j A_j \quad (5)$$

According to the above, the formulation of UIOs is constructed for each subsystem. In the next part, the multiple Lyapunov function will be used to realize the design of parameters for UIOs of switched system.

Theorem: [14] In the condition of arbitrary switching signal, for the system (3), if

$$(H_j C_j - I)E_j = 0 \quad (6)$$

$$T_j = (I - H_j C_j)$$

$$F_j = (A_j - H_j C_j A_j - K_{j1} C_j)$$
$$K_{j2} = F_j H_j$$

hold, and there exist symmetric matrices $P_j > 0$, $\forall j \in \psi$ such that

$$\begin{bmatrix} -P_j & P_j F_j \\ F_j^T P_j & -P_j \end{bmatrix} \leqslant 0, \ \forall j \in \psi \tag{7}$$

then the parameters of UIOs can be designed.

Proof: Define the state estimation error as $e(k) = x(k) - \hat{x}(k)$, the dynamic equation can be derived as

$$\begin{aligned} e(k+1) &= (A_j - H_j C_j A_j - K_{j1} C_j) e(k) \\ &- \left[F_j - (A_j - H_j C_j A_j - K_{j1} C_j)\right] z(k) \\ &- \left[K_{j2} - (A_j - H_j C_j A_j - K_{j1} C_j) H_j\right] y(k) \\ &- \left[T_j - (I - H_j C_j)\right] B_j u(k) - (H_j C_j - I) E_j d(k) \end{aligned} \tag{8}$$

with

$$K_j = K_{j1} + K_{j2} \tag{9}$$

In order to make the error decoupled from known control input $u(k)$, measured output $y(k)$ and unknown input $d(k)$, we should let

$$(H_j C_j - I) E_j = 0 \tag{10}$$

$$T_j - (I - H_j C_j) = 0$$
$$F_j - (A_j - H_j C_j A_j - K_{j1} C_j) = 0$$
$$K_{j2} - (A_j - H_j C_j A_j - K_{j1} C_j) H_j = 0$$

It can be concluded that

$$H_j = E_j \left[(C_j E_j)^T C_j E_j \right]^{-1} (C_j E_j)^T \tag{11}$$

$$A_{j1} = A_j - E_j \left[(C_j E_j)^T C_j E_j \right]^{-1} (C_j E_j)^T C_j A_j \tag{12}$$

$$F_j = A_j - H_j C_j A_j - K_{j1} C_j = A_{j1} - K_{j1} C_j \tag{13}$$

and the error dynamics is given by

$$e(k+1) = F_j e(k) \tag{14}$$

Equation (4) is UIOs of the switched system (3) if the estimation error tends asymptotically to zero despite the presence of an unknown input $d(k) \neq 0$.

Consider the following Lyapunov function

$$V_j(e(k)) = e(k)^T P_j e(k) \tag{15}$$

For $P_j = P_j^T > 0$, then, $V_j(e(k)) > 0$ holds, the ΔV given by (16) is negative

$$\begin{aligned} &V_i(e(k+1)) - V_j(e(k)) \\ &= e(k)^T F_j^T P_i F_j e(k) - e(k)^T P_j e(k) \\ &= e(k)^T (F_j^T P_i F_j - P_j) e(k); (i \in \psi, j \in \psi, i \neq j) \end{aligned} \tag{16}$$

if the following inequalities are satisfied

$$F_j^T P_i F_j - P_j \leqslant 0 \tag{17}$$

Thus, the error system (14) is stable asymptotically.

According to Schur complement lemma, the inequalities (17) can be rewritten as

$$\begin{bmatrix} -P_i & P_i F_j \\ F_j^T P_i & -P_j \end{bmatrix} \leqslant 0 \tag{18}$$

By substituting $F_j = A_{j1} - K_{j1} C_j$, the above matrix inequalities become

$$\begin{bmatrix} -P_i & P_i (A_{j1} - K_{j1} C_j) \\ (A_{j1} - K_{j1} C_j)^T P_i & -P_j \end{bmatrix} < 0 \tag{19}$$

and

$$\begin{bmatrix} -P_i & (P_i A_{j1} - W_{ij} C_j) \\ (P_i A_{j1} - W_{ij} C_j)^T & -P_j \end{bmatrix} < 0 \tag{20}$$

for $W_{ij} = P_i K_{j1}$.

Since $K_{j1} = P_j^{-1} W_{ij}$, we can obtain the value of K_{j1} then F_j, K_{j2} and $K_j = K_{j1} + K_{j2}$ from P_j and W_{ij} solutions of LMIs.

The aim of fault detection is to generate a residual signal $r_j(k)$, given by (21), which is sensitive to $f_a(k)$ in the presence of actuator faults which is the purpose of [15].

$$r_j(k) = y(k) - C_j \hat{x}_j(k) \tag{21}$$

4 PID-type fuzzy logic controller design

In this study, the FTC approach is based on PID-type FLC. To adjust the input and the output scaling factors of this controller, Genetic Algorithm (GA) optimization technique has been proposed in order to improve the performance of the controller.

4.1 PID-type fuzzy logic controller description

We consider a PID-type FLC structure as shown in Figure 2, [16], where $K_e \in \mathbb{R}^+$ and $K_d \in \mathbb{R}^+$ are the input scaling factors, $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^+$ the output scaling factors.

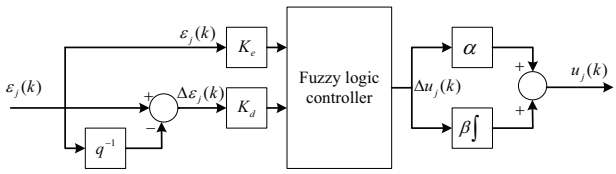


Figure 2: The PID-type FLC

The FLC inputs variables, the error $\varepsilon_j(k)$ between the desired flat trajectory $y_j^d(k)$ and the estimated state $\hat{x}(k)$, and error variation $\Delta\varepsilon_j(k)$ are given by the equations (22) and (23) where T_e is the sampling period.

$$\varepsilon_j(k) = -K_{c,j} \hat{x}(k) + y_j^d(k) \tag{22}$$

$$\Delta\varepsilon_j(k) = \frac{\varepsilon_j(k) - \varepsilon_j(k-1)}{T_e} \tag{23}$$

The output variable $\Delta u_j(k)$ of a such controller is the variation of the control law signal $u_j(k)$ which can be defined as (24).

$$\Delta u_j(k) = \frac{u_j(k) - u_j(k-1)}{T_e} \tag{24}$$

The output of the PID-type FLC is given by (25), [17]

$$\begin{aligned} u_j(k) &= \alpha \Delta u_j(k) + \beta \int \Delta u_j(k)\ \mathrm{d}t \\ &= \alpha(A + QK_e u_j(k) + DK_d \Delta u_j(k)) \\ &+ \beta \int (A + QK_e u_j(k) + DK_d \Delta u_j(k))\ \mathrm{d}k \\ &= \alpha A + \beta Ak + (\alpha K_e Q + \beta K_d D) u_j(k) \\ &+ \beta K_e Q \int u_j(k)\ \mathrm{d}k + \alpha K_d D \Delta u_j(k) \end{aligned} \tag{25}$$

where the terms Q and D are given by (26) and (27), [10].

$$Q = \frac{\Delta u_{(i+1)j} - \Delta u_{ij}}{\varepsilon_{i+1} - \varepsilon_i} \tag{26}$$

$$D = \frac{\Delta u_{i(j+1)} - \Delta u_{ij}}{\Delta\varepsilon_{j+1} - \Delta\varepsilon_j} \tag{27}$$

The fuzzy controllers with product−sum inference method, centroid defuzzification method and triangular uniformly distributed membership functions for the inputs and a crisp output proposed in [18], are used, in our study.

The linguistic levels, assigned to the variables $\varepsilon_j(k)$, $\Delta\varepsilon_j(k)$ and $\Delta u_j(k)$, are given in Table 1 as follows: NL: Negative Large; N: Negative; ZR: Zero; P: Positive; PL: Positive Large.

$\varepsilon_j(k)$ / $\Delta\varepsilon_j(k)$	N	ZR	P
N	NL	N	ZR
ZR	N	ZR	P
P	ZR	P	PL

Table 1: Fuzzy rules-base

4.2 Optimization of scaling factors using genetic algorithm

To adjust the input and the output scaling factors (K_e, K_d) and (α, β), GA is used in order to obtain their optimal values.

At first, an initial chromosome population is randomly generated. The chromosomes are candidate solutions to the problem. Then, the fitness values of all chromosomes are evaluated by calculating an objective function. So, based on the fitness of each individual, a group of the best chromosomes is selected through the selection process. The genetic operators, crossover and mutation, are applied to this 'surviving' population in order to improve the next generation solutions. Crossover is a recombination operator that combines subparts of two parent chromosomes to produce offspring. This operator extracts common features from different chromosomes in order to achieve even better solutions. Mutation is an operator that introduces variations into the chromosome. The modifications can consist of changing one or more values of a chromosome. Through the mutation operator the search space is explored by looking for better points. The process continues until the population converges to the stop criterion.

The most crucial step in applying GA is to choose the objective function that is used to evaluate the fitness of each chromosome. In this paper, the method of tuning PID-type FLC parameters using GA is based on minimizing the Integral of the Squared Error (ISE) used in [18].

5 Flatness and trajectory planning

The flatness approach is used in a discrete-time framework for system (1). Let the studied dynamic linear discrete system described by (28)

$$D_j(q) y_j(k) = N_j(q) v_j(k) \tag{28}$$

q is the forward operator, $v_j(k)$ and $y_j(k)$ are, respectively, the input and the output signals and $D_j(q)$ and $N_j(q)$ the polynomials defined by

$$D_j(q) = q^n + a_{j,n-1} q^{n-1} + \ldots + a_{j,1} q + a_{j,0} \tag{29}$$

$$N_j(q) = b_{j,n-1} q^{n-1} + \ldots + b_{j,1} q + b_{j,0} \tag{30}$$

where the parameters $a_{j,l}$ and $b_{j,l}$ are constants, $l = \{0, 1, ..., n-1\}$.

The system is considered as linear and controllable, consequently it is flat, [16].

The flat output $z_j(k)$, on which depend the output $y_j(k)$ and the control $v_j(k)$, can be seen as being the partial state of a linear system, [19]

$$v_j(k) = D_j(q) z_j(k) \tag{31}$$

$$y_j(k) = N_j(q) z_j(k) \tag{32}$$

The open loop control law can be determined by the following relations, [18]

$$v_j^d(k) = f(z_j^d(k), ..., z_j^{d^{(r+1)}}(k)) \tag{33}$$

$$y_j^d(k) = g(z_j^d(k), ..., z_j^{d^{(r)}}(k)) \tag{34}$$

f and g are vectorial functions and $z_j^d(k)$ is the desired trajectory that must be differentiable at the $(r+1)$ order.

In order to plan the desired flat trajectory $z_j^d(k)$, the polynomial interpolation technique is used.

Let consider the state vector: $Z_j^d(k) = (z_j^d(k)\ z_j{}^{d^{(1)}}(k)\ \ldots\ z_j^{d^{(r+1)}}(k))^T$ containing the desired flat output and its successive derivatives, [18].

The expression of $Z_j^d(k)$ can be given as following where k_0 and k_f are two moments known in advance.

$$Z_j^d(k) = M_{j,1}(k - k_0) c_{j,1}(k_0) + M_{j,2}(k - k_0) c_{j,2}(k_0, k_f) \tag{35}$$

such that

$$M_{j,1} = \begin{pmatrix} 1 & k & \cdots & \frac{k^{n-1}}{(n-1)!} \\ 0 & 1 & \cdots & \frac{k^{(n-2)}}{(n-2)!} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix} \tag{36}$$

$$M_{j,2} = \begin{pmatrix} \frac{k^n}{n!} & \frac{k^{n+1}}{(n+1)!} & \cdots & \frac{k^{2n-1}}{(2n-1)!} \\ \frac{k^{n-1}}{(n-1)!} & \frac{k^n}{n!} & \cdots & \frac{k^{(n-2)}}{(n-2)!} \\ \vdots & \ddots & \ddots & \vdots \\ k & \cdots & \frac{k^{n-1}}{(n-1)!} & \frac{k^n}{n!} \end{pmatrix} \quad (37)$$

$$c_{j,1} = Z_j^d(k_0) \quad (38)$$

$$c_{j,2} = M_{j,2}^{-1}(k_f - k_0)(Z_{j^d}(k_f) - M_{j,1}(k_f - k_0)Z_j^d(k_0)) \quad (39)$$

After planning a desired flat trajectory in discrete-time framework, the real output $y_j(k)$, to be controlled, follows the desired trajectory $y_j^d(k)$ such that (40), [18].

$$y_j^d(k) = N_j(q)z_j^d(k) \quad (40)$$

6 Application of an electronic throttle valve

6.1 Electronic throttle valve modeling

The proposed approach is applied here for the case of the electronic throttle valve, Figure 3, [15].

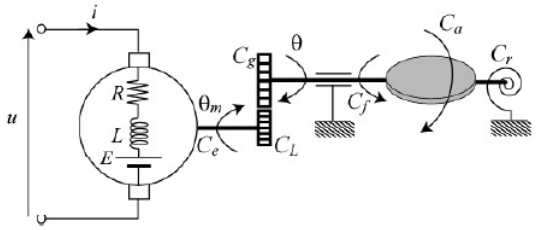


Figure 3: Electronic throttle system

The electrical part of this system is modeled by (41), [20,21].

$$u(t) = Ri(t) + L\frac{d}{dt}i(t) + k_v\omega_m(t) \quad (41)$$

L is the inductance, R the resistance, $u(t)$ and $i(t)$ the voltage and the armature current respectively, k_v a constant electromotive force and $\omega_m(t)$ the motor rotational speed.
The mechanical part of the throttle is modeled by a gear reducer, characterized by its reduction ratio γ such as (42)

$$\gamma = \frac{C_g}{C_L} \quad (42)$$

C_L is the load torque and C_g the gear torque.
The mechanical part is modeled according to (43) and (44), such that, [20,21]

$$J\frac{d}{dt}\omega_m(t) = C_e - C_f - C_r - C_a \quad (43)$$

$$\frac{d}{dt}\theta(t) = (180/\pi/\gamma)\omega_m(t) \quad (44)$$

$\theta(t)$ is the throttle plate angle, C_e the electrical torque, C_f the torque caused by mechanical friction, C_r the spring torque, C_a the torque generated by the airflow and J the overall moment of inertia.
The electrical torque is defined by

$$C_e = k_e i(t) \quad (45)$$

where k_e is a constant.
The electronic throttle valve involves two complex nonlinearities due to C_r and C_f, given by their static characteristics, [15]

- a dead zone in which the control voltage signal has no effect on the nominal position of the valve plate,
- two hysteresis combined with a saturation, due to the valve plate movement, limited by the maximum and the minimum angles.

The static characteristic of the nonlinear spring torque C_r is defined by

$$C_r = \frac{k_r}{\gamma}(\theta - \theta_0) + D\,\mathrm{sgn}(\theta - \theta_0) \quad (46)$$

for $\theta_{\min} \leqslant \theta \leqslant \theta_{\max}$, k_r is the spring constant, D a constant, θ_0 the default position and sgn (.) the following signum function

$$\mathrm{sgn}\,(\theta - \theta_0) = \begin{cases} 1, \text{ if } \theta \geqslant \theta_0 \\ -1, \text{ else} \end{cases} \quad (47)$$

The friction torque function C_f of the angular velocity of the throttle plate can be expressed as

$$C_f = f_v\omega + f_c\,\mathrm{sgn}(\omega) \quad (48)$$

where f_v and f_c are two constants.
By substituting in equation (43), the expressions C_e, C_f and C_r and by neglecting the torque generated by the airflow C_a, the two nonlinearities $\mathrm{sgn}(\theta - \theta_0)$ and $\mathrm{sgn}(\omega)$ and the two constants $\frac{k_r}{\gamma}\theta_0$ and f_v, the studied system is linear then it can be modelized by the following transfer function $H(s)$ (49), [21]

$$H(s) = \frac{(180/\pi/\gamma)k_e}{JLs^3 + JRs^2 + (k_ek_v + Lk_s)s + Rk_s} \quad (49)$$

with $k_s = (180/\pi/\gamma^2)k_r$ and s the Laplace operator; the identified parameters of this system are given in Table 2 at $25°C$, [20].

Parameters	Values	Units
R	2.8	Ω
L	0.0011	H
k_e	0.0183	$N.m/A$
k_v	0.0183	$v/rad/s$
J	4×10^{-6}	$kg.m^2$
γ	16.95	-

Table 2: Model's parameters

Recent work has shown that the ETV can be modeled by two linear models identified from the default position of the throttle plate for two values of the parameter k_s and for the sampling period $T_e = 0.002s$, [21]

- a model representing the position of the plate above the position by default for: $k_s = 1.877 \times 10^{-4} kg.m^2$, then the corresponding discrete-time transfer function $H_1(q^{-1})$ given by (50),

$$H_1(q^{-1}) = \frac{0.007480q^{-1} + 0.01334q^{-2} + 0.0007376q^{-3}}{1 - 1.948q^{-1} + 0.954q^{-2} - 0.006152q^{-3}} \tag{50}$$

- a model representing the position of the plate below the position by default for: $k_s = 1.384 \times 10^{-3} kg.m^2$, then the corresponding discrete-time transfer function $H_2(q^{-1})$ given by (51),

$$H_2(q^{-1}) = \frac{0.007479q^{-1} + 0.01333q^{-2} + 0.0007376q^{-3}}{1 - 1.946q^{-1} + 0.954q^{-2} - 0.006152q^{-3}} \tag{51}$$

6.2 Simulation results

In order to test the proposed fault tolerant control approach, tow models of an ETV in actuator fault and an unknown disturbance case are given as state space formulation as

$$x(k+1) = A_j x(k) + B_j u(k) + E_j d(k) + B_j f_a(k) \tag{52}$$

$$y(k) = C_j x(k)$$

where the parameter matrices and vectors of each model are given as

$$A_1 = \begin{pmatrix} 1.948 & -0.954 & 0.006152 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{53}$$

$$A_2 = \begin{pmatrix} 1.946 & -0.954 & 0.006152 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{54}$$

$$B_1 = B_2 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T \tag{55}$$

$$C_1 = \begin{pmatrix} 0.007480 & 0.01334 & 0.0007376 \end{pmatrix} \tag{56}$$

$$C_2 = \begin{pmatrix} 0.007479 & 0.01333 & 0.0007376 \end{pmatrix} \tag{57}$$

$$E_1 = E_2 = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}^T \tag{58}$$

The studied system fulfills the conditions of Lemma 1.

$$\begin{cases} rank(C_1 E_1) = rank(E_1) = 1 \\ rank(C_2 E_2) = rank(E_2) = 1 \end{cases} \tag{59}$$

then $(C_1 A_{11})$ and $(C_2 A_{21})$ are both observable.
By applying Theorem, the system is stable asymptotically for any switching signal for symmetric matrice P definite positive given by (60).

$$P = \text{e - 010} \begin{pmatrix} 0.1429 & -0.0857 & 0.1903 \\ -0.0857 & 0.5085 & 0.0494 \\ 0.1903 & 0.0494 & 0.4251 \end{pmatrix} \tag{60}$$

Then, the rest parameters of UIOs are calculated as below

$$F_1 = \begin{pmatrix} -0.0266 & 0.8566 & -1.5415 \\ -0.0078 & 0.1954 & -0.3515 \\ 0.0165 & -0.4329 & 0.7788 \end{pmatrix} \tag{61}$$

$$F_2 = \begin{pmatrix} -0.0233 & 0.8717 & -1.5349 \\ -0.0071 & 0.1998 & -0.3515 \\ 0.0148 & -0.4384 & 0.7715 \end{pmatrix} \tag{62}$$

$$T_1 = \begin{pmatrix} 0.9657 & -0.6187 & -0.3470 \\ -0.0343 & 0.3813 & -0.3470 \\ -0.0343 & -0.6187 & 0.6530 \end{pmatrix} \tag{63}$$

$$T_2 = \begin{pmatrix} 0.9672 & -0.6196 & -0.3476 \\ -0.0328 & 0.3804 & -0.3476 \\ -0.0328 & -0.6196 & 0.6524 \end{pmatrix} \tag{64}$$

$$K_1 = \begin{pmatrix} 0.0002 \\ 0.0031 \\ -0.0056 \end{pmatrix}, K_2 = \begin{pmatrix} 0.0182 \\ 0.0208 \\ -0.0389 \end{pmatrix} \tag{65}$$

$$H_1 = \begin{pmatrix} 46.3945 \\ 46.3945 \\ 46.3945 \end{pmatrix}, H_2 = \begin{pmatrix} 46.4707 \\ 46.4707 \\ 46.4707 \end{pmatrix} \tag{66}$$

In this study, the method of tuning PID-type FLC parameters using GA is based on minimizing the ISE. If $y_j^d(k)$ is the desired flat trajectory and $\hat{x}_j(k)$ the estimated state, then we have

$$\varepsilon_j(k) = -K_{c,j}\hat{x}_j(k) + y_j^d(k) \tag{67}$$

For the ISE defined by

$$ISE = \int_0^t \varepsilon_j^2(k)dt \tag{68}$$

In this paper, the considered fitness function is taken as inverse of this error, i.e. the following performance index

$$fitness\ value = \frac{1}{ISE} \tag{69}$$

Then, the obtained optimum values of the input/output scaling factors $(K_e, K_d, \alpha, \beta)$, using genetic algorithm are given as follows: $K_e = 0.2006$, $K_d = 0.8071$, $\alpha = 0.1146$ and $\beta = 0.2063$.
The desired discrete time flat trajectory $z_j^d(k)$, with $j = \{1, 2\}$ can be computed according to the following polynomial form

$$z_j^d(k) = \begin{cases} \frac{cst1}{B_j(1)}, \text{ if } 0 \leqslant k \leqslant k_0 \\ Poly_{1,j}(k), \text{ if } k_0 < k \leqslant k_1 \\ \frac{cst2}{B_j(1)}, \text{ if } k_1 < k \leqslant k_2 \\ Poly_{2,j}(k), \text{ if } k_2 < k \leqslant k_3 \\ \frac{cst1}{B_j(1)}, \text{ if } k > k_3 \end{cases} \tag{70}$$

where $cst1$ and $cst2$ are constant parameters, $k_0 = 3s$, $k_1 = 6s$, $k_2 = 10s$ and $k_3 = 15s$ are the instants of transitions, $B_j(1)$ is the static gain between the flat output $z_j(k)$ and the output signal $y_j(k)$ for each model and $Poly_{1,j}(k)$ and $Poly_{2,j}(k)$ are polynomials calculated using the technique of polynomial interpolation.
The desired trajectories $y_1^d(k)$ and $y_2^d(k)$ are then given in Figure 4.

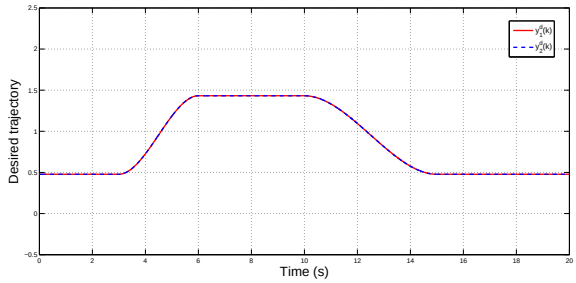


Figure 4: Desired trajectories $y_1^d(k)$ and $y_2^d(k)$

Let us consider the fault vector

$$f_a(k) = \begin{bmatrix} f_{a1}^T(k) \\ f_{a2}^T(k) \end{bmatrix} \tag{71}$$

such as $f_{a2} = 0$ and f_{a1} defined as follows

$$f_{a1}(k) = \begin{cases} 0, & k \in [0, 10] \\ -0.1\sin(0.314k), & k \in]10, 13] \\ 0, & k \in]13, 20] \end{cases} \tag{72}$$

The simulation results illustrated in Figure 5 and Figure 6, show some oscillations for the outputs signals and for the tracking error due to the unknown input. We remark that the system's responses with and without FTC track desired trajectories with disturbances rejection.

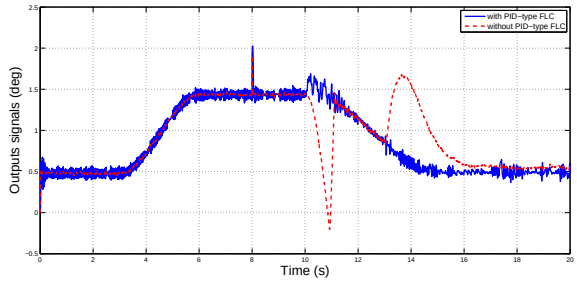


Figure 5: System outputs in actuator fault case with and without PID-type FLC

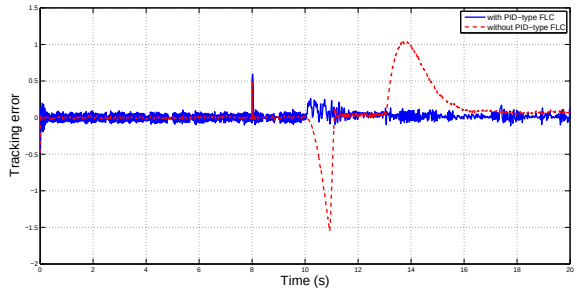


Figure 6: Tracking error in actuator fault case with and without PID-type FLC

At the time of actutor faults occurrence $k = 10s$, the system's behavior was changed. The system without FTC becomes unstable, whereas for the same reference input and by using the FTC, the system remains stable and the tracking error have a small deviation from zero which shows the effectiveness of the proposed FTC approach.

Figure 7, Figure 8 and Figure 9 show the residual values generated using UIOs and the switched signal. The switching between the two models is acheived based on residual values comparision.

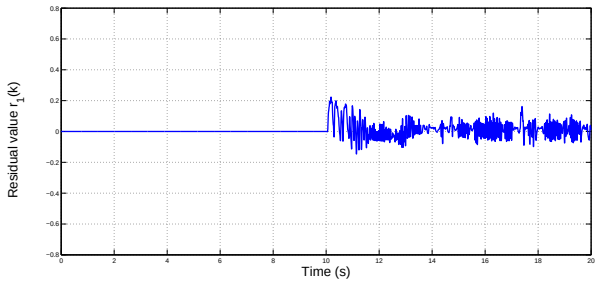


Figure 7: Residual value $r_1(k)$ in actuator fault case with PID-type FLC

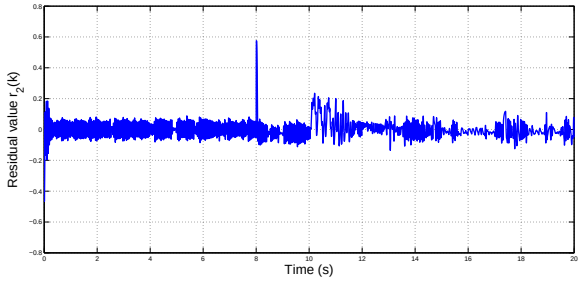


Figure 8: Residual value $r_2(k)$ in actuator fault case with PID-type FLC

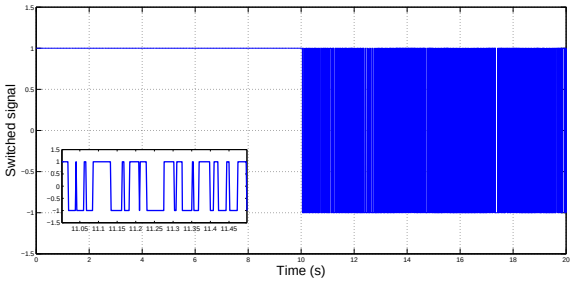


Figure 9: Switched signal in actuator fault case with PID-type FLC

7 Conclusion

In this paper, a new fault tolerant control law based on PID-type FLC is designed for the nonlinear complex system ETV, modelized by a multimodel structure. The approach is based on the use of a reference

model generated using the flatness property. The proposed control law is, then, designed to minimize the error between the desired flat trajectory and the estimated state, generated using UIOs, even in the presence of actuator faults based on minimizing the ISE.

References

1. B.Y. Huo, Y.M. Li and S.C. Tong, Fuzzy adaptive fault-tolerant output feedback control of multi-input and multi-output non-linear systems in strict-feedback form, IET Control Theory and Applications, **6**, 2704–2715, 2012.
2. Q.K. Shen, B. Jiang and V. Cocquempot, Fuzzy logic system-based adaptive fault-tolerant control for near-space vehicle attitude dynamics with actuator faults, IEEE Transactions on Fuzzy Systems, **21**, 289–300, 2013.
3. S.C. Tong, B.Y. Huo and Y.M. Li, Observer-based adaptive decentralized fuzzy fault-tolerant control of nonlinear large-scale systems with actuator failures, IEEE Transactions on Fuzzy Systems, **2**,1–15, 2014.
4. D. Ichalal, B. Marx, J. Ragot and D. Maquin, Fault tolerant control for Takagi-Sugeno systems with unmeasurable premise variables by trajectory tracking, IEEE International Symposium on Industrial Electronics (ISIE), Bari, 2010.
5. M. Blanke, M. Kinnaert, J. Lunze and M. Staroswiecki, Diagnosis and Fault-Tolerant Control, Springer-Verlag Berlin Heidelberg, 2006.
6. R. Nouailletas, D. Koeing and E. Mendes, LMI design of a switched observer with model uncertainty: Application to a hysteresis system, Proc. of the 46th IEEE CDC, New Orleans, 6298–6303, 2007.
7. D. Wang, W. Wang and P. Shi, Robust fault detection for switched systems with state delays, IEEE Trans. on Systems and Cybernetics-B, **39**, 800–805, 2009.
8. H. Yingxue, T. Shaocheng and L. Yongming, Adaptive fuzzy backstepping control for a class of switched nonlinear systems with actuator faults, International Journal of Systems Science, **47**, 3581–3590, 2015.
9. A. Benzaouia, M. Ouladsine, A. Naamane and B. Ananou, Fault detection for uncertain delayed switching discrete-time systems, International Journal of Innovative Computing, Information and Control, **8**, 1–11, 2012.
10. M. Oudghiri, M. Chadli and A. El Hajjaji, Robust observer-based fault tolerant control for vehicle lateral dynamics, International Journal of Vehicle Design, **48**, 173–189, 2008.
11. D. Ichalal, B. Marx, D. Maquin and J. Ragot, Nonlinear observer based sensor fault tolerant control for nonlinear systems, 8th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, Mexico, 2012.
12. M. Benrejeb, Stability Study of Two Level Hierarchical Nonlinear Systems, 12th IFAC Symposium on Large Scale Systems: Theory and Applications, **43** (8), 30–41 Villeneuve d'Ascq, 2010.
13. J. Chen, R.J. Patton and H.Y. Zhang, Design of unknown input observers and robust fault detection filters, International Journal of Control, **63**, 85–105, 1996.
14. W. Jiawei, S. Yi and Z. Miao, Fault Detection for Linear Switched Systems Using Unknown Input Observers, Proceedings of the 33rd Chinese Control Conference, Nanjing, 2014.
15. W. Gritli, H. Gharsallaoui and M. Benrejeb, Fault Detection Based on Unknown Input Observers for Switched Discrete-Time Systems, Int. Conf. on Advanced Systems and Electrical Technologies IC-ASET, Hammamet, 2017.
16. S. Bouallègue, M. Ayadi, J. Haggége and M. Benrejeb, A PID type fuzzy controller design with flatness-based planning trajectory for a DC drive, IEEE International Symposium on Industrial Electronics, Cambridge, 1197–1202, 2008.
17. M. Guzelkaya, I. Eksin and E. Yesil, Self-tuning of PID-type fuzzy logic controller coefficients via relative rate observer, Engineering Applications of Artificial Intelligence, **16**, 227–236, 2003.
18. W. Gritli, H. Gharsallaoui and M. Benrejeb, PID-type Fuzzy Scaling Factors Tuning Using Genetic Algorithm and Simulink Design Optimization for Electronic Throttle Valve, International Conference on Control, Decision and Information Technologies CoDIT, Malta, 2016.
19. M. Fliess, J. Lévine, P. Martin and P. Rouchon, On differentially flat non linear systems, In Proc. IFAC-Symposium NOLCOS, pp. 408-413, Bordeaux, 1992.
20. M. Lebbal, H. Chafouk, G. Hoblos and D. Lefebvre, Modelling and Identification of Non-Linear Systems by a Multimodel Approach: Application to a Throttle Valve, International Journal Information and Systems Science, **3**, 67–87, 2007.
21. C. Yang, Model-based analysis and tuning of electronic throttle controllers, Visteon Corporation, SAE 2004 World Congress & Exhibition, 63–67, 2004.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 194-202(2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Review on security issues in RFID systems

Mohamed El Beqqal*, Mostafa Azizi

University Mohamed first, ESTO, 60000, Morocco

ARTICLEINFO

Article history:
Received: 11 November, 2017
Accepted: 01 December, 2017
Online: 14 December, 2017

Keywords:
RFID System
Classification
Security
Privacy
Countermeasures

ABSTRACT

Radio frequency Identification (RFID) is currently considered as one of the most used technologies for an automatic identification of objects or people. Based on a combination of tags and readers, RFID technology has widely been applied in various areas including supply chain, production and traffic control systems. However, despite of its numerous advantages, the technology brings out many challenges and concerns still not being attracting more and more researchers especially the security and privacy issues. In this paper, we review some of the recent research works using RFID solutions and dealing with security and privacy issues, we define our specific parameters and requirements allowing us to classify for each work which part of the RFID system is being secured, the solutions and the techniques used besides the conformity to RFID standards. Finally, we present briefly a solution that consists of combining RFID with smartcard based biometric to enhance security especially in access control scenarios. Hence the result of our study aims to give a clear vision of available solutions and techniques used to prevent and secure the RFID system from specific threats and attacks.

1. Introduction

This paper is an extension of work originally presented in Wireless Technologies, Embedded and Intelligent Systems (WITS) [1]. It concerns the security in Radio Frequency Identification (RFID) systems. The implementation of this technology in many industrial fields has recently attracted a lot of attention. Due to its important role of identifying people as well as objects, and data tracking, RFID is becoming more and more popular as an important topic of research. In particular, with the emergence of the Internet of things (IoT) paradigm that refers to a kind of networks that links physical objects to the internet to exchange data [2] and where RFID plays a key role. Besides, RFID is considered to be the next generation of barcode technology which is the case in many countries [2]. The use of RFID technology in several domains including supply chain, healthcare, transportation, education and many other fields have significant advantages including the flexible way of deployment with low cost, the possibility of integration and cooperation with Wireless Sensor Network nodes [2]. However, despite its numerous advantages, RFID technology still brings out many challenges including technical problems, customer privacy issues, coexistence of heterogeneous RFID standards [3] and more importantly security and privacy concerns. In fact, an RFID system consists of tags that store the data, readers interacting with the tags and transmitting information to backend servers and finally a network canal that manages this communication [4]. This architecture multiplies the sources of security problems and threats.

Several types of attacks can target the RFID system depending on each part is more vulnerable to the attacker. For example, the modification of the data, Cloning technique, and impersonation can be applied at level tags, whereas the reverse engineering, Eavesdropping Side Channel Attacks can affect the communication between backend servers, readers and tags [5].

In the research field, many of solutions and techniques assuring security and privacy come with processing and cryptographic algorithms such as Elliptic Curve Cryptographic (ECC) technique or mutual authentication mode between all RFID system parts. Most of these techniques remain far from being implemented in practice due to the limited resources of storing in tags and the low capacity of processing and executing complex operations. Furthermore, RFID customers require some guaranties that data contained in tags or exchanged remain private and highly protected against tracking and customer identity reconnaissance. A guarantee must be provided to customers aiming to adopt RFID in their own goals by ensuring that their private data stored in the tags will remain private and protected against spying and tracking

*Corresponding Author: Mohamed El Beqqal, elbeqqal.mohamed@gmail.com

This work presents a clear vision about the concerns of security and privacy in an RFID system. Precisely we compare recent works dealing with security and privacy in an RFID system according to which layer the author is aiming to secure. To achieve this goal, we define some parameters and requirements such as CAI (Confidentiality, Availability and Integrity) based on the information suggested in [6]. We present for each work the category and the technique of solution used to deal with a typical defined attack. In addition, we verify for each work the application of RFID standards. Finally, we propose a device added scheme solution that we have used to enhance security in the access control scenario.

This paper is organized in 5 sections. The Section 2 summarizes the basics components of an RFID system, some examples of applications and a summarized review on most challenges and research directions where we concentrate on the problem of security. The section 3 presents our comparative study based on our parameters defining this study. The section 4 illustrates the smart card based biometrics solution to enhance RFID security and finally section 5 concludes this paper.

2. Background on RFID technology

In this part, we present the components of a typical RFID system, besides the related security challenges.

2.1. RFID Basics

An RFID system consists of 6 main components as shown in Figure 1:

- Tag: A tag is the data carrier and normally contains the ID number, and unique EPC code programmed into the tag.
- Reader & antenna: A reader captures the data provided by the tag when tags come in range of the area covered by the reader using its antenna.
- Middleware: The middleware can be a software as well as hardware dedicated to process data captured by the Reader, then dispatch this information to backend servers.
- Backend servers: The backend servers are the last station to make use of the data collected from RFID components. The information collected can be stored in database or sent to other systems for reporting and further analysis.
- Network infrastructure: Stands for the link between all RFID components which represents the main sources of RFID security and privacy threats.

2.2. Application domains

Animal tracking

RFID tagging systems are considered as powerful tools for animal identification management and tracking. Their benefits go beyond controlling the spread of diseases and ensuring quality-assured and safe food for consumers.
As an example of the application of RFID technology for animal tracking, an approach enabling an effective localization and tracking of small-sized laboratory animals was suggested in [7]

Logistics

The implementation of RFID technologies in logistics environments is increasingly attracting. The investment in this technology has shown that RFID is highly promising for identifying an object uniquely and providing the capability for "complete traceability" [8]. Moreover, it has been recognized as an effective solution to eliminate or reduce inventory inaccuracies [9].

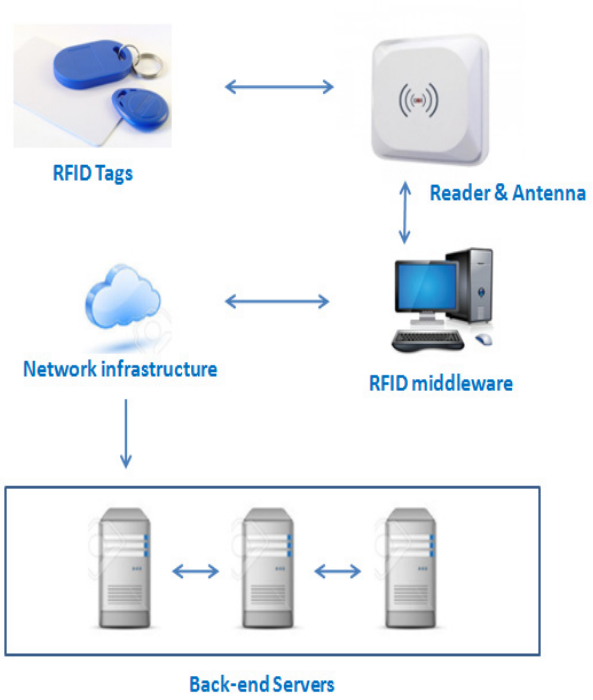


Figure 1. RFID system components

Intelligent transportation

The transport domain has widely benefited from the RFID adaptation due the ease of deployment of intelligence into each vehicle with a low cost approach [11].A vehicle emission inspection and notification system aiming to help daily monitoring of engine emissions through RFID devices was proposed in [12]. Also, a wireless system based on RFID technology was used to manage taxi fleets at the airports, train stations, bus terminals and all the taxis in cities [13]. RFID is also used in tolls collection as shown in Figure 2.

Healthcare

The use of RFID technology in the healthcare domain is making the patient's quality of care better. In fact, besides the collecting of information related to the patient as his presence inside a room and his health but also environment parameters can be detected such as temperature, humidity, and the presence of toxic gas [14].

In the same context, several efforts and research works have been made in this area. Exemplary, authors in [15] suggest a passive RFID platform for monitoring people during the night which deploys a long-range UHF RFID reader, wearable tags

integrated into clothes, and ambient tags dispersed in the environment, as well as a software engine for real-time processing and with warning modules [15].

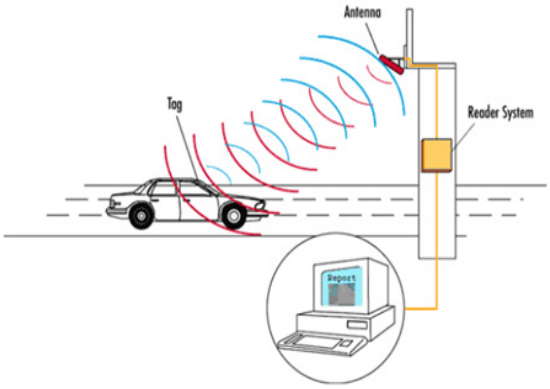


Figure 2. A toll collection using RFID [13]

2.3. RFID major issues

We address in this section some of the prominent issues related to RFID technologies and that still remain challenging topics of interest for a vast number of researchers.

Several works tried to study and gather the issues related the RFID to provide a baseline for researchers to understand the different sources as well as measures that can ensure the safety of sensitive data but more importantly to satisfy customers' needs and requirements in terms of efficiency, security and privacy. Among the most RFID challenges, we found technical problems, customer privacy issues and appliance to RFID standards [3]. In our work, we will focus on the last two problems which were taken into account in our comparative study (Section 3).

Technical problems

Among the biggest challenges during the implementation of an RFID system is studying the physical characteristics and capabilities of its components. In fact, the limited resource of storing information inside the tag, the low data processing and execution of complex operations can be a serious problem of performance and lack of mandatory features. In addition, due to the star architecture of the system, the reader can be the point of crush of the whole system, this problem occurs when a large number of items with RFID tags are energized by an RFID reader at the same time which leads to a confusion at the level of the reader and prevent it from tag scanning. A similar problem occurs when the coverage area of one RFID reader overlaps with another reader which leads to a signal interference and multiple reads of the same tag.

Standardization problems

Several efforts were dedicated to solve the problem of lacking of common RFID standards but the issue is still remaining in the industrial field. In fact, different standards coexist in parallel with divergent interests among which there is ISO 11784 that is utilized in the tracking of RFID cattle, the ISO 11785 for interface protocol, ISO 14443 for application of smart cards in payments and ISO 15693 that is applied to vicinity cards. The development of the electronic product codes (EPC) has been conceived to fill this need.

Obviously, the fact each of the existing standards is associated to a specific part of the RFID system make it different to deploy and protect data within an RFID system where tags standards can not comply with the communication standards for instance. Thus, it is of upmost importance to take into account the compliance to standards when implementing any RFID-based solution.

Security and privacy

Among the advantages of the deployment of RFID system is rendering the identification task automatically also gaining on performance and the ease of execution. All these benefits are not very useful is security and privacy are not guaranteed. For this reason, while designing an RFID system, many parameters should be taken into account to ensure a primary level of security, In fact, the data inside the tag, exchanged with the reader or dispatched to the back-end servers must be confidential, available and unchangeable (CIA characteristics). However, ensuring a full secure system which means keeping the whole system safe and operational, it remains a difficult task due to the several sources of attacks and threats that can coexist in parallel in some scenarios.

Table 1. RFID Typical attacks

Typical attack	The attack principle
Eavesdropping	This attack consists of listening to the network and sometimes recording for example all exchanges between tags and reader. The main goal of this attack is collecting data and taking advantage of the information gathered.
Denial of service	The denial of service consists of rendering equipment unavailable in the RFID system. An example of using this attack is spamming a reader with specific tags frequency requests which will make the reader out of service.
Cloning	The cloning attack is mainly to reproduce RFID tags. Using reverse engineering techniques to extract all the cart properties, precisely, the secrets keys, the tag can be modified and duplicated.
Tracking	The tracking attack consists of associating a tag to a person without his will and track his presence and movement as long as the attacker's reader is in range with the tag

During our research, many papers mentioned in their solutions some typical attacks that can take place in RFID systems, in the Table 1 below, we list some of most known attacks.

Many countermeasures and capabilities are proposed in literature to ensure security of the RFID system such as Pseudo-random based solution, Anonymous-ID scheme, symmetric and asymmetric cryptographic algorithms and others hashing based schemes [2].

Apart from the security aspect, the privacy is considered as a critical issue aiming to protect personal information. The expansion of the technology in several domains is not passing

without risks regarding the privacy the person since his identity information are associated to RFID chips that can be tracked and identified without his knowledge. As explained in table1 and due to the small size if RFID tags, they can be nested in personnel objects and clothes without being detected. The person movement and information are tracked and analyzed each time he come in range of the compatible reader area.

Some countermeasures and solutions were developed to protect personal data and reduce tracking capabilities. The table below illustrates the main techniques proposed in the literature studied for our classification.

Table 2. Some main countermeasures for privacy

Protection method	***How this technique improves privacy protection***
Delegation Tree	The principle of this method is delegating the control of reading tags depending on which privacy policy is assigned to each reader and tag. These specifications can be stored in database for example.
Protocol added *S*chemes	The technique consists of integrating a new coding scheme inside RFID tags and readers in order to create a specific protocol of communication; in this way, external readers cannot access to the network tags.
Tag killing	The tag killing method consists of destroying the content of tag after this one is no longer need to be used. The retailer can use the kill command after entering the right pin code. This technique can be used by the sellers after their products leave the store.
XOR encryption and PRNG	The XOR encryption and PRNG method is based on using a randomized protocol at each communication. This technique is powerful in the way to counter listening attacks such as Eavesdropping which will make the communication with tag a difficult task.
Blocker Tag	The Blocker tag technique aim to create an inductive field by the tag which will block communication between the tag and suspicious readers. Many occurrences of ID tag will be generated in order to hide the real one for the reader.

3. The comparative study

In this section, we will present the goal of our study besides the parameters defined for classification used to compare the selected academic works.

3.1. Motivation and design goal

Several levels of security can be applied to a typical RFID system, besides the nature of its architecture leads us to classify the solutions depending on each part of the system the attacks can take place. Furthermore, we refer to [6] to consider the three security main classes to classify the security threats in the reviewed works: Confidentiality, availability and integrity (CAI).

As defined in [6], the RFID system can be presented as three main layers: *1) RFID Edge Hardware Layer, 2) Communication Layer and 3) Backend Layer.* The possible threats and attacks related to each layer are mentioned based on CAI properties. In our work, besides classifying the attacks related to RFID security, we present also the solutions and techniques used to counter these problems. In [5], most of protection techniques can belong to specific categories as presented in Figure 3:

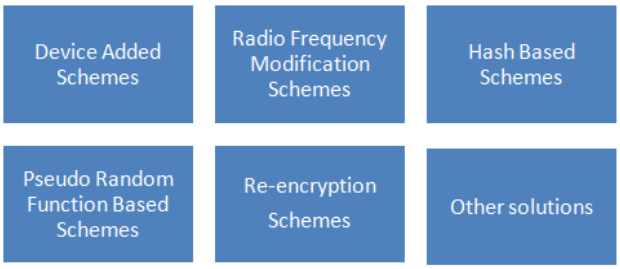


Figure 3. RFID security classes

In our study, we focus on recent works dealing with security and privacy, and we classify them according to the layers previously (attacks and solutions). In addition, we take into consideration other security requirements and parameters rather than CAI (untraceability, non-repudiation, forward secrecy, anonymity …).

Furthermore, we present the type of solutions and techniques dedicated for each layer and the attacks that can be related. At the end, we verify for each studied work the use and conformity to RFID standards. In this way, the following parameters are taken into account in our comparative study:

- *Privacy-focus / Security-focus*: Indicates whether the solution proposed take into account the security or consider privacy as the main concern.
- *Problem description and classification*: this parameter indicated which part of the system the problem can belong to. For this, we refer to the classification explained above.
- *Parameters of Security/Privacy considered:* Here we extract the security/privacy parameters considered by each author. Some papers focus on traditional security goals (confidentiality, integrity and availability) while other solutions take into account additional properties such as authorization, authentication, or forward secrecy.

Clearly, there are other properties to consider when authors focus in privacy as a main concern. These parameters include identity protection, anonymity and untraceability and. It is worth noting that these parameters are whether clearly cited in the paper or deducted.

- *Solution category*: As argued before, we have six mains classes of RFID solutions. For each of the twelve works studied we specify the class solution.
- *Techniques used:* For each class of solutions, we mention the techniques that have been exploited to ensure security and

privacy. Precisely, this includes cryptographic-based and no cryptographic-based methods.

- *Compliance to RFID standards?* Compliance to standards is considered as a primordial issue that need be highly highlighted. In this goal, for each studied work, we check the compliance to RFID or communication standards or if at least this issue is mentioned or an attention has been paid to by authors.
- *Examples of security attacks/ Threats considered:* this parameter indicates the security/ privacy attacks cited and treated in each paper. In the table below, the attack is defined for each part of the system it can occur and the solutions proposed to secure the system form this type of threats.

The objective of our comparative study as shown in Table 3, is not only presenting and classifying the security solutions and techniques regarding each of the recent studied paper but also making visible for researches and companies aiming to adopt an RFID system what are threats and attacks that should be taken into accounts while designing their systems.

As described before, the classification-based layer level helped us to locate the scope of each solution. Furthermore, during the study of the twelve works, we found that additional security / privacy parameters are considered rather than the traditional ones (CAI) in order to identify the several security/privacy threats. For Example, scalability is a primordial while aiming to enhance the system dimensions; some papers such as [24] consider scalability as a first topic while designing their solution. In fact, as mentioned in [22], this parameter remains among the real constraints of the majority of security solutions. Many techniques can be deployed to achieve scalability such as Checks handoff technique. As described in [24], this method was deployed to design a secure protocol supporting IoT.

Concerning techniques used, mutual authentication has taken more focus by authors as indicated in the works [20,16, 24] since, this security method is not only used in server's part but also between the reader and tags.

Concerning the security/privacy focus, many works try to ensure customer's privacy in their secured developed solutions and protocols. For instance, in [17, 20] intractability and an anonymous communication are taking into account to protect users' privacy. Thus, we noticed from the works presented in Table 3 that most of the solutions do not only focus on security or on privacy individually but these two concepts should be both take into consideration.

The second conclusion, we have subtracted from the Table 3 was that several works does not focus only on a single part of the system but they try to cover the whole layers (physical layer, communication layer, Back-end layer) that we have described in the section before.

For instance, authors in [18], propose two algorithms implemented by the Back-End server and a two-party protocol between the Back-End Server and a tag. In this way, two security methods were suggested in [20], the first contribution covers the backend layer by proposing two authentication (mutual and collaborative) schemes for both fixed and mobile readers, and the second contribution consisted at defining a secure channel between the reader and server in the communication layer. The main idea that we try to highlight is to not only paying attention to the data stored into tag or only securing communication between tags and reader but all the whole trajectory of data from data collecting to data storing in backend server.

As presented in Table 3, very few works we have studied focus on the device added schemes solution category. As an example, authors in [19] propose a solution that consist of changing the physical architecture of the RFID reader by dividing it into two different devices, an RF activator and a trusted shield device (TSD) playing an intermediary role between tags and reader for a secure communication [16].

In this way, the authors in [22] used the Physically Unclonable Functions (PUF) to secure data at level of tags (PUF is a security technology that seeks to introduce physical variation into individual devices taking part in cryptographic protocols [26]).

It is of upmost importance to tackle the issue of the RFID standardization. Indeed, and as illustrated in our comparative study, some works take into consideration the conformity of their solutions to RFID standards [12, 18, 19, 20, 20, 23, 24]. Intuitively, before designing any secure RFID solution it is crucial to verify whether this solution respects RFID standards or not.

Table 3: A comparative study of some security/privacysolutions in RFID systems

Characteristics / **Proposed Solution**	**Privacy-focus / Security -focus**	**Problem description and classification (Layer)**	**Parameters of Security/Privacy considererd**	**Solution category**	**Techniques used**	**Compliance to RFID standards?**	**Examples of security attacks/ Threat considered**
RFID-Tate: Efficient Security and Privacy Protection for Active RFID over IEEE 802.15.4 [16]	Both	Communication: A protocol for mutually authenticated (tag, reader) communication is suggested.	-Confidentiality -Availability - Authentication	Re-encryption	-Identity-based Encryption (IBE)	Yes : communication is performed over the standard IEEE 802.15.4f	- Cloning attacks. - Tag emulating, - Collision attack. - Spoofing - Replay attack -Sybil attack
A Secure Supply-Chain RFID System that Respects Your Privacy [17]	Both	Commiucation :p-Authors suggest a protocol between the reader and an individual tag	- Confidentiality - Integrity	Re-encryption	Public Key cryptography	Yes: (EPC tags standards)	-Reverse engineering

Secure Tag Search in RFID Systems Using Mobile Readers [18]	Both	Back-End: This paper suggests a search tag protocol; target backeEned Servers	- Confidentiality - Integrity - Anonymity - authorization	Pseudo-random based solution	XOR and 128bit PRNG operations	Yes : (EPC C1G2 compliance)	-Desynchronization, DoS attacks. -Side Channel Attacks: - Eavesdropping
Scalable RFID Security Framework and Protocol Supporting Internet of Things [24]	Security-Focus	Communication: A secure communication protocol and a framework supporting IoT	-Authentication. - Scalability - Confidentiality -Identity pretection -Adaptability	Other	Checks handoff (SCH) technique	Yes: EPCglobal standard	-Malware (Example SQLIA)
A New Security and Privacy Framework for RFID In Cloud Computing [25]	Both	Backend: A security and privacy model for RFID technology integrated to the Cloud computing (Servers) Was suggested.	- Confidentiality - Authentication	Re-encryption	Symmetric-key Cryptography	Not mentioned	No specified attacks was mentionned.
A secure ECC-based RFID authentication scheme integrated with ID-verifier transfer protocol [2]	Both	Both : Back-End: Two algorithms implemeted by the Back-End server are defined. Communication: A two-party protocol between the Back-End Server and a tag is proposed.	-Confidentiality - Availability - Forward security. -Anonymity Authorization, Authentication	Re-encryption	Public Key cryptography Elliptic curve cryptosystem	Mentioned but not applied	-Tag cloning attack. - Eavesdropping. - Replay attack. - Tag masquerade attack. -Denial-of-Service - Location tracking attack. - Server spoofing attack.
A Novel Coding Scheme for Secure Communications in Distributed RFID Systems [19]	Privacy-Focus	Physical Layer	Confidentiality - Integrity Anonymity	Device added scheme	- Reader changes architecture (New) Random Flipping and Random Jamming (RFRJ)	Only mentioned	- Random guessing attack. - Correlation attack. - Ghost-and-leech attack. - Eavesdropping.
A minimum disclosure approach to authentication and privacy in RFID systems [20]	Both	Both BackEnd: Two authentication (mutual and collaborative) schemes for both fixed and mobile reader are suggested. Communication: The channel between the reader and server is protected against attacks.	- Confidentiality - Untraceability Anonymity, Authentication, Forward secrecy	Pseudo-random based solution	Redundancy check (CRC) Pseudo random number generator (PRNG)	Yes (Conforme to EPC Class-1 Gen-2 spec-Ifications).	- Replay attack - Tag impersonation - Desynchronisation attacks.
Secure ownership transfer for multi-tag multi-owner passive RFID environment with individual-owner-privacy [12]	Both	Communication: Authors suggests multi-owner, multi-tag transfer protocol that achieves privacy-among-owners.	Confidentiality - Forward Untraceability Anonymity, Forward secrecy	Pseudo-random based solution	XOR and 128-bit pseudo-random number generators (PRNG)	Yes: complies EPC Global Class-1Gen-2 (C1G2)	-Tracking attacks - Replay Attacks - Denial of Service - Tag/Reader/Server Impersonation.
A Robust Grouping Proof Protocol for RFID EPC C1G2 Tags [21]	Security-focus	Communication: A grouping proof protocol ensuring a high level of security between tags and readers.	-Inegrity -Anonymity -Forward secrecy -Confidentiality	Pseudo-random based solution	(XOR) encryption and 128-bit pseudorandom number generators,	Yes: Complies EPC Global Class-1Gen-2 (C1G2)	-Replay Attacks -Denial of Service (DoS) Reader/Tag/Server impersonation Attack. -Active Attacks
Providing destructive privacy and scalability in RFID systems using PUFs [22]	Privacy-Focus	Communication: A private authentication protocol allowing a reader/tag communication.	- Confidentiality - Integrity - Authentication - Forward secrecy	Other	Physically Unclonable Functions (PUFs)	Not mentioned	- Compromising attack - Side-channel attacks
Computational Cost Analysis on Securing RFID Protocols Conforming to EPC Class-1 Generation-2 Standard [23]	Security-Focus	Both : A mutual authentication protocol seeking to achieve higher security level was proposed	-Confidentiality -Authentication - Forwad secrecy	Pseudo-random based solution	Pseudo-random number generators (PRNG)	Yes : Conform to EPC Class-1 Generation-2 standard (ISO 18000-6).	- DoS attacks - Reply attacks

4. RFID combining solution

4.1. Motivation

The comparative study presented in section 3, has shown many robust solutions and techniques that may counter several attacks, most of the techniques are based on cryptographic schemes and deal with direct attacks on physical layer of the RFID system components such as the tag and reader, whereas other techniques focus on the communication between readers and tags or reader and the backend servers to prevent the system from network attacks such as eavesdropping, man-in-the middle or desynchronization attacks.

Aiming to apply the RFID in the campus environment, we found that some scenarios are not totally secured. So, we propose to combine the RFID technology with other technologies to enhance security in the whole system.

4.2. Smart card based biometrics solution design

The RFID technology can be applied in many areas inside the university campus as shown below, in our work we focus on access control to exams rooms and sensitive areas like laboratories.

- Exams room
- Campus
- Research laboratories
- Library
- Staff offices
- Computer room

For instance, if we take the example of accessing to laboratories, the identity check must be verified before allowing the person access to the area. Based on the solutions proposed in Table 3, a cryptographic technique can be deployed in the inside the tag and the reader to secure the tag reading. However, the two main problems in our case are indicated in Figure 4.

For this reason, our method consists of enhancing the RFID technology by using smart card based biometrics solution. In fact, the RFID technology is seen as the fast-automatic technique of identification, whereas, the smart card check can support complex operations of cryptographic check as indicated in [27]. In addition, the biometric fingerprint will help us to check the true identity of the person aiming to access to the secured area. In Figure 5, we present the proposed system combining RFID with the smart card-based biometrics verification.

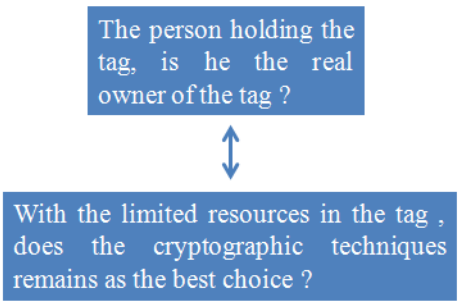


Figure 4. Access control issues in RFID system

As detailed in [27]. The step one consists of reading the RFID tag when it comes in range with the RFID reader. After that, in the step 2, the person's tag is sent to the local database to check if it belongs to the person's tags allowing him access the secure area. Once validated, the person inserts his smartcard and puts his fingerprint for validation. In this case, a first match-on-card is established comparing the user fingerprint with the other one stored in the smartcard. If the check is performed successfully, the fingerprint is sent to the database to verify that no changes were made on the couple (fingerprint, Tag ID). Finally, the door is opened if all the conditions are satisfied.

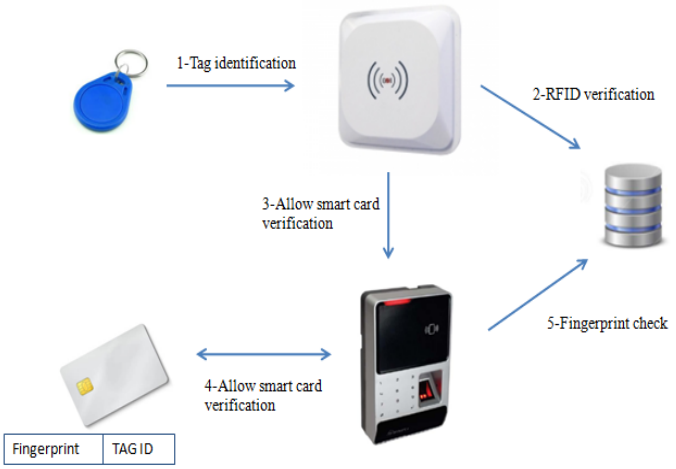


Figure 5. Components of the proposed system

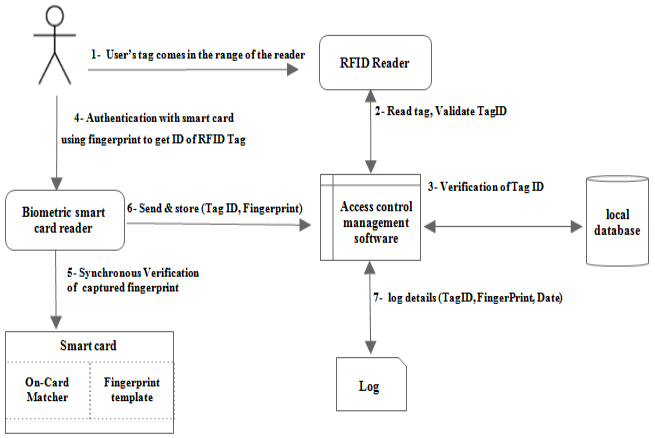


Figure 6. Flowchart of the access control system in sensitive areas [27]

In the proposed solution, we have chosen to combine smart card biometrics solutions with RFID in order to respect the limited resources and capabilities inside the tag for performing complex operations. Furthermore, even if the smart card is falsified or the person ask for authentication with a borrowed RFID tag, the verification explained above based multi-level of security will deny his access to the secured room. The flowchart in Figure 6 summarizes the sequence of actions during the authentication process. The system presented above meets the need presented in Figure 4 which are not guaranteed in the solutions presented in section 3.

5. Conclusion

In this paper, we have presented some recent works aiming to ensure security and Privacy in RFID systems. In particular, we tried to classify these solutions in order to provide a clear understanding of the different threats and risks related to security and privacy.

To achieve this goal, we compared some solutions according to the traditional security and privacy objectives. Also, we mentioned that additional parameters such as scalability and efficiency should be taken into account besides the traditional security and privacy objectives.

In addition, we proposed a combined solution using RFID and smartcard based biometrics to ensure performance and cover some security gaps detected while designing our RFID system. As future work, we plan to develop our device added scheme solution, by studying more the technical possibilities of its implementation inside the campus.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is performed inside the MATSI Lab., ESTO, University Mohammed First, Oujda (Morocco).

References

[1] M. El Beqqal and M. Azizi, "Classification of major security attacks against RFID systems," in 2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), 2017. https://doi.org/10.1109/WITS.2017.7934622

[2] Yi-Pin Liao, Chih-Ming Hsiao, "A secure ECC-based RFID authentication scheme integrated with ID-verifier transfer protocol", Journal of Ad Hoc Networks, Volume 18, July 2014, Pages 133–146,2013. https://doi.org/10.1007/978-3-642-35473-1_1

[3] AL-Kassab, J., W.C. Rumsch. 2008. Challenges for RFID cross-industry standardization in the light of diverging industry requirements. IEEE Systems Journal 2(2) 170-177. https://doi.org/10.1109/JSYST.2008.921291

[4] F. Kamoun," RFID System Management:State-of-the Art and Open Research Issues", IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, VOL. 6, NO. 3, SEPTEMBER 2009 . https://doi.org/10.1109/TNSM.2009.03.090305

[5] K. Sabaragamu Koralalage, J. Cheng, "A Comparative Study of RFID Solutions for Security and Privacy: POP vs. Previous Solutions", International Conference on Information Security and Assurance, 2008. https://doi.org/10.1109/ISA.2008.89

[6] A. Mitrokotsa, M. Beye and P. Peris-Lopez, "Classification of RFID Threats based on Security Principles", GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES, 2009

[7] L. Catarinucci, R. Colella, L. Mainetti, L. Patrono, S. Pieretti, Ilaria Sergi, and L. Tarricone, "Smart RFID Antenna System for Indoor Tracking and Behavior Analysis of Small Animalsin Colony Cages", IEEE SENSORS JOURNAL, VOL. 14, NO. 4, APRIL 2014. https://doi.org/10.1109/JSEN.2013.2293594

[8] M. Alamgir Hossain, "Development of an integrated model for RFID extension", BUSINESS PROCESS MANAGEMENT JOURNAL· AUGUST 2014. https://doi.org/10.1108/BPMJ-04-2013-0055

[9] T. Fan, Feng Tao, S. Deng, S. Li, "Impact of RFID technology on supply chain decisions with inventory inaccuracies", Int. J. Production Economics, 2014. https://doi.org/10.1016/j.ijpe.2014.10.004

[10] Oh-Keun Ha Yong-Seok Song Kyung-Yong Chung Kang-Dae Lee · Dongjoo Park, "Relation model describing the effects of introducing RFID in the supply chain: evidence from the food and beverage industry in South Korea",Journal of Pers Ubiquit Comput, 2014. https://doi.org/10.1007/s00779-013-0675-x

[11] Lei Xie, Yafeng Yin, Athanasios V. Vasilakos,Sanglu Lu, "Managing RFID Data: Challenges, Opportunities and Solutions", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 16, NO. 3, THIRD QUARTER 2014. https://doi.org/10.1109/SURV.2014.022614.00143

[12] Sundaresan, S. Doss, R., Wanlei Zhou, "Secure ownership transfer in multi-tag/multi-owner passive RFID systems", IEEE Global Communications Conference (GLOBECOM), 2013. https://doi.org/10.1109/GLOCOM.2013.6831513

[13] S. Samadi, "Applications and Opportunities for Radio FrequencyIdentification (RFID) Technology in Intelligent Transportation Systems: A Case Study", International Journal of Information and Electronics Engineering, Vol. 3, No. 3, May 2013. https://doi.org/10.7763/IJIEE.2013.V3.330

[14] Sara A, R. Lodato, S. Manzari, C. Occhiuzzi, and G. Marrocco, "RFID Technology for IoT-Based Personal Healthcare in Smart Spaces", IEEE INTERNET OF THINGS JOURNAL, VOL. 1, NO. 2, APRIL 2014. https://doi.org/10.1109/JIOT.2014.2313981

[15] C. Occhiuzz, Carmen Valleseb, S.Amendolab, S. Manzarib," NIGHT-Care: a passive RFID system for remote monitoring and control of overnight living environment", 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014). https://doi.org/10.1016/j.procs.2014.05.414

[16] Sadikin, M.F., Kyas, M., "RFID-tate: Efficient security and privacy protection for active RFID over IEEE 802.15.4, The 5th International Conference on Information, Intelligence, Systems and Applications, IISA 2014.in press. https://doi.org/10.1109/IISA.2014.6878787

[17] Arbit, A., Oren, Y. ; Wool, A, "A Secure Supply-Chain RFID System that Respects Your Privacy", Journal of Pervasive Computing, IEEE (Volume:13 , Issue: 2), 2014. https://doi.org/10.1109/MPRV.2014.22

[18] Sundaresan, S., Doss, R. ; Piramuthu, S. ; Wanlei Zhou, 'Secure Tag Search in RFID Systems Using Mobile Readers', Journal of Dependable and Secure Computing, IEEE Transactions on (Volume:12 , Issue: 2), 2014, in press. https://doi.org/10.1109/TDSC.2014.2302305

[19] Sakai, K, Sun, Min-Te ; Ku, Wei-Shinn ; Lai, T.H., "A Novel Coding Scheme for Secure Communications in Distributed RFID Systems",Computers, IEEE Transactions on (Volume:PP , Issue: 99), 2015, in press. https://doi.org/10.1109/TC.2015.2423671

[20] Robin Doss, , Wanlei Zhou, Saravanan Sundaresan, Shui Yu, Longxiang Gao , 'A minimum disclosure approach to authentication and privacy in RFID systems", journal of Computer Networks, Volume 56, Issue 15, Pages 3401–3416 pp. 544 - 549 , 2012. https://doi.org/10.1016/j.comnet.2012.06.018

[21] Sundaresan, S, Doss, R. ; Piramuthu, S. ; Wanlei Zhou," A Robust Grouping Proof Protocol for RFID EPC C1G2 Tags", Information Forensics and Security, IEEE Transactions on (Volume:9 , Issue: 6), 2014. https://doi.org/10.1109/TIFS.2014.2316338

[22] Mete Akgün, M. Ufuk Çağlayanb, "Providing destructive privacy and scalability in RFID systems using PUFs", Journal of Ad Hoc Networks, Volume 32, Pages 32–42, 2015, in press. https://doi.org/10.1016/j.adhoc.2015.02.001

[23] Avoine, G. Bingol, M.A. ; Carpent, X. ; Yalcin, S.B.O., "Privacy-Friendly Authentication in RFID Systems: On Sublinear Protocols Based on Symmetric-Key Cryptography". obile Computing, IEEE Transactions on (Volume:12 , Issue: 10), 2013. https://doi.org/10.1109/TMC.2012.174

[24] BR Ray, J Abawajy, M CHowdhury, "Scalable RFID security framework and protocol supporting Internet of Things", Computer Networks, vol. 67, pp. 89-103, 2014. https://doi.org/10.1016/j.comnet.2014.03.023

[25] Plos, T. Hutter, M. ; Feldhofer, M. ; Stiglic, M., "Security-Enabled Near-Field Communication Tag With Flexible Architecture Supporting Asymmetric Cryptography", Very Large Scale Integration (VLSI) Systems, IEEE Transactions on (Volume:21 , Issue: 11), 2013. https://doi.org/10.1109/TVLSI.2012.2227849

[26] D. Moriyama1, S. Matsuo1 and M.Yung, "PUF-Based RFID Authentication Secure and Private under Memory Leakage", IACR Cryptology ePrint Archive , 2013. https://doi.org/10.1016/j.adhoc.2015.02.001

[27] M. El Beqqal, M. A. Kasmi, and M. Azizi, "Access control system in campus combining RFID and biometric based smart card technologies," Adv. Intell. Syst. Comput., vol. 520, pp. 559–569, 2017. https://doi.org/10.1007/978-3-319-46568-5_56

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 195-205 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A decision-making-approach for the purchasing organizational structure in Moroccan health care system

Kaoutar Jenoui*, Abdellah Abouabdellah

Industrial Engineering laboratory, Team of Modeling, Optimization of Industrial Systems and Logistics (MOSIL) ENSA, University Ibn Tofail, Kenitra, Morocco.

ARTICLE INFO

Article history:
Received: 06 November, 2017
Accepted: 28 January, 2018
Online: 31 March, 2018

Keywords:
Healthcare organization
Centralization strategy
Hidden quality costs
Organizational structure

ABSTRACT

Excellence in hospitals supply management results in better quality, best prices, and good deliveries. One of the questions that come up as healthcare organization to capture the economies of scale in purchasing prices and process costs is whether their purchasing activities should be centralized or decentralized. In most cases, centralization strategy usually gives good supplier's service with a lower cost, but the consideration of supplier's cost in the hospital sector are mainly limited to visible ones. The high levels of hidden quality costs generated by suppliers and their unknown presence have serious consequences on the decisions made by the managers. However, the existence of this kind of costs has not been considered yet. Therefore, the main objective of this paper is to propose a decision-making-approach, integrating a new method of measuring supplier's hidden quality costs, in order to help managers to choose the appropriate purchasing organizational structure in the hospital sector.

1. Introduction

In Morocco, the ministry of health has undertaken several actions to make medicines and medical devices available and accessible to the population. It has a national list of medicines and essential medical devices, according to the last revision which was carried out at the end of 2011. The pharmaceutical products covered by this list benefit from an annual budget allocated by the ministry of health, to ensure their availability at the level of public hospitals and basic health care facilities. This budget increased by 67% between 2002 and 2012, reaching the sum of 1.6 billion Dirham [1-3]. Despite efforts made by the ministry, the access to medicines and medical devices in hospitals remains insufficient [2] [4]. The main factor that reduces the availability of medicines is a failure to achieve a good supplier's quality. Experiences have shown that it is possible to improve access to these products by making the best use of resources and streamlining management processes. Indeed, it is necessary to highlight the organizational purchasing structure, and improve the supplier's quality [5-7]. In this perspective, it is mandatory to study the relationship that exists between non-quality, costs and organizational structure. In fact, a change in an organizational structure impacts the entire system including customers, suppliers, and competitors. In the healthcare system, the suppliers are considered as the main stakeholders contributing to the improvement of the hospital's performance. The demand continues to support the growing supplies, which has changed the behavior of organizations to achieve good quality. This change includes both simple corrections and more complex changes for suppliers, at the level of their attitudes and their performances. The particularity of healthcare facilities is that the patient receives its service in real time. Though, the impact of the structure on the supplier's performance affects the quality of care and purchaser's satisfaction, this dissatisfaction could be estimated by a cost, which is considered as a hidden cost. Theses hidden quality costs may increase or decrease depending on the choice of the organizational structure.

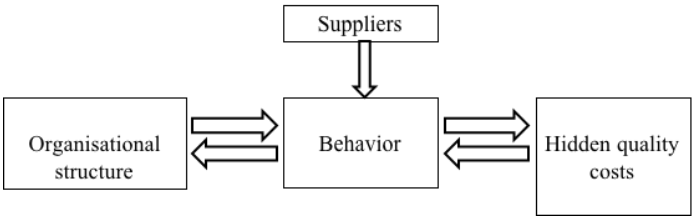


Figure 1. Relationship that exists between organizational structure and hidden costs

In light of the importance of hospital logistics, we propose in this paper to consider the supplier's hidden quality costs as one of the

*JENOUI Kaoutar, Email : k.jenoui@gmail.com

principle criteria to be based on for deciding the right purchasing strategy. This paper is an extension of a work originally presented in Logistiqua'17 conference [1]. It's organized as follows: Section II presents some definitions of the hospital sector, followed by section III that describes the problematic, and section IV that provides a literature review. In section V a deciding-making-approach is proposed and validated through a practical case study in section VI. Finally, section VII concludes the paper and states future work.

2. Hospital pharmacy presentation

This section provides important definitions that describe hospital pharmacy services, as well as the requirements for providing pharmaceutical care.

2.1. The hospital, hospital pharmacy and SEGMA hospital definition

The hospital is a healthcare unit composed of specialized scientific equipment, and certified healthcare professionals, working as a team for the common purpose of providing medical services to the population. While the hospital pharmacy defined as a department in a hospital, which is responsible for the supply of medical products, and headed by professionally competent and qualified pharmacist who directly supervises and ensures compounding and distribution of medication to in and out patients [8].

The status SEGMA stands for "State service managed in an autonomous way" and is defined as the services of the state whose expenses operations are executed by a head of the department attached to the superior accountant of the kingdom. In 2004, there was 38.4% of the total number of services dedicated to hospital care with SEGMA status. Currently 63% of the number of hospitals obtains the SEGMA status [9].

2.2. Goals of hospital pharmacy

The hospital pharmacy aims to provide safe and effective pharmaceutical care and services to users and health care facilities, through a global professional and ethical standards. Pharmaceutical services components are responsible for three main missions. Firstly, it's responsible for the procurement, distribution and control of pharmaceutical products, secondly, the evaluation and dissemination of comprehensive information about medical products and their use to the institution's employees and patients. And finally, it's responsible for the supplier's monitoring, evaluation, and assurance of the quality of drugs [8].

2.3. Hospital's pharmacy in Centralized and decentralized structure

Figure 2 shows the logistics flows for centralized and decentralized structure. The centralized structure aims to make orders for many hospitals independently when there is no grouping of commons orders. While decentralized structure is considered to refer to a practice model in which a pharmacist is responsible for orders purchasing and distribution services, including order validation and possibly order entry [5][8].

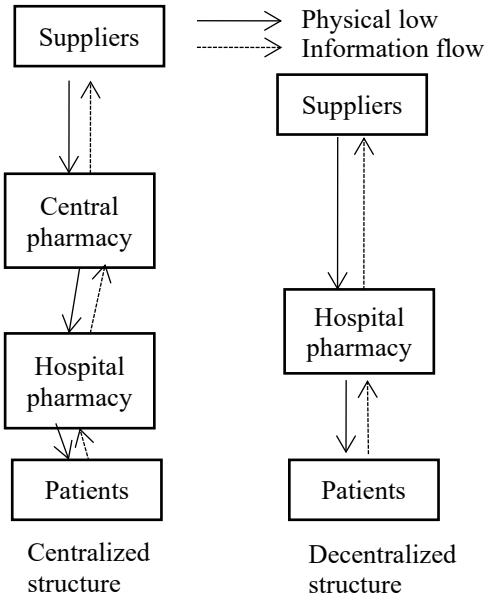


Figure 2. The information and physical flow in both structures

3. Problematic description

In 24 years, the Moroccan health care system switched many times between centralized and decentralized structure [9]. (See table 1)

In face of all these changes and developments of the organizational structure, Moroccan hospitals still suffer from inadequate quality, the insufficient availability of products and the increased costs [3].In decentralization structure, centers of storage and distribution cost up to one billion Dirhams per year, and operated by 200-250 people [1-4].Also, storage and distribution of medicines in Morocco cost the ministry of health over 30 million Dirhams per year. While in centralization structure, the rate of obsolescence of medicines decreased by 8% at the end of 2002, and the purchase prices of medicines through the centralized system in 2001 were lower than the prices paid in the decentralized system. This drop in prices allowed buying 50% more products. All these data show the supplier's bad quality and the inappropriate choice of the organizational structure, which indicates the importance of estimating poor quality costs generated by suppliers, also called hidden costs. Therefore, in this paper we propose a decision-making-approach, integrating a new method of measuring supplier's hidden quality costs, in order to help the managers to choose the appropriate purchasing organizational structure in the hospital sector.

4. Literature review

To discuss the literature related to this subject, this review will cover three principle topics, and we will start by a brief review of each of them:

4.1. Organizational structure

Decision-making can be performed in a centralized or a decentralized structure. In a centralized one, there exists a central responsibility for decision-making, whereas in a decentralized structure the individual entities can make their own decisions. In practice, each approach has its advantages and disadvantages. Most commonly, the strategic decisions are usually made centrally while operation decisions are decentralized. The performance of each approach depends on specific environment and particular decisions [9] [10]. Many researchers studied the effects of centralization and decentralization on the multi-item replenishment problem in a two-echelon supply chain. Chen et al. proposed both centralized and decentralized decision models and proved the optimal properties of both models to minimize costs [11] [12]. They found that it is beneficial to adopt centralized control and proposed a mechanism to coordinate the decentralized system so that each player in the chain benefit from it. Behdani et al. studied disruptions in a multi-plant company and considered alternative policies for coping with them. To model this complex system, they used an agent-based simulation model [13].

Table1. A brief history of the health care supply system

1980	Decentralization	With many difficulties faced by the central pharmacy, there has been an introduction of direct delivery system
1985-1986	Centralization	Worsening problems due to the small size of premises and storage spaces, implies return to centralization. There was also a need of the implementation of a new unit, and the improvement of the central pharmacy program, followed by a progressive implementation of semi-autonomous management for the regional hospitals which directly procure themselves.
1994	Decentralization	Establishment of the procurement division, reporting directly to the general secretariat of the ministry.
1995	Centralization	Commissioning of the storage unit to centralize procurement for medicines.
1997	Decentralization	Decentralization purchasing for SEGMA hospitals
2001	Centralization	Centralized purchasing (procurement, storage, distribution) for SEGMA regional hospitals by the procurement division.
2003	Decentralization	In view of difficulties of regular supply, it is envisaged to decentralize the supply of SEGMA hospitals.

However, no one has considered the hidden costs generated in each strategy. In our study, we will focus on this objective in a particular way.

Hidden quality cost:

In traditional systems, quality losses occur when the product deviates beyond the specification limits, and becomes unacceptable [14]. Taguchi proposed a narrower view of characteristic acceptability to indicate that any deviation from a characteristic's target value results in a loss, and that a higher quality measurement results in minimal variation from the target value [21-22]. Other authors analyze different aspects related to hidden quality cost in the construction business, and highlight the following examples of hidden quality costs: schedule delays, litigation and claims, loss of reputation and the subsequent impact on future business opportunities, loss of schedule and productivity, low operational efficiency, work inactivity from waiting or idle time, etc. [15-19]. Taguchi loss functions have been recently used for non-manufacturing applications. They were implemented to evaluate product quality as an aid to the selection of suppliers, been based on quantitative quality characteristics. These products are used for evaluation, comparison, and ranking process. There are several methods for the estimation of hidden costs [23-25]. However, none of them considers the quantification of hidden costs associated to qualitative criteria that presents subjectivity and uncertainty. The question which arises while measuring the cost associated to qualitative criteria is how to determine the target value, and the particular value specification that characterize the performance of each element.

4.2. Multi attributes decision making technique

There are many techniques developed for the supplier evaluation problem. Some of these techniques are categorical method, weighted point method [26], matrix approach [27], vendor performance matrix approach [28], vendor profile analysis (VPA) [29], analytic hierarchy process (AHP) [30-31], analytic network process (ANP) [32], mathematical programming [33-34] and multiple objective programming (MOP) [35-37]. However, most of these methods are more adapted to precise data. They don't have enough influence on factors such as imprecision preferences, qualitative criteria and incomplete information. The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is a multi-criteria decision analysis method based on the concept that the chosen alternative should have the shortest geometric distance from the positive ideal solution, and the longest geometric distance from the negative ideal solution [7]. Therefore, TOSPIS method is more appropriate to overcome problems in estimating hidden costs associated to qualitative characteristics. To sum up, the main contribution of this paper is the development of a method for the quantification of hidden quality costs based on Taguchi loss function and TOPSIS method, included in a decision-making-approach. We also compare the behavior of medicines suppliers under the centralized and decentralized strategies based on hidden quality costs. The suitability of the two strategies under two different scenarios is explored through a real case study in a Moroccan hospital.

5. The proposed deciding-making-approach

Based on the review of previous works and field interview with suppliers in Morocco, the organizational structure applied to purchase medical products for hospitals, helps to evolve the supplier's chain performance to flourish and vice versa. As indicated in figure 3, the organizational structure in the healthcare supply system is a critical component for any supplier's system improvement, and the decision to centralize or decentralize purchasing should be based on studies that focus on the actual supplier's hidden quality costs for each structure.

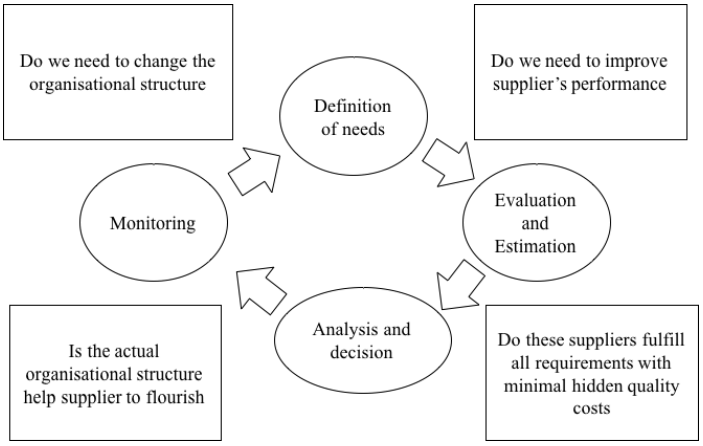


Figure 3. Deciding-making-approach lifecycle

5.1. Definition of needs

As indicated in the figure, targets of change include improving effectiveness at two principle levels: Supplier's performance and organizational structure. There is a strong relationship between these two elements that helps ensure solid communication. Therefore, in this phase, we should:

- Define the needs regarding supplier's specifications, functional and technical requirements.
- Recognize the problems inherent in managing the actual organizational structure and obstacles that overcome.
- Discover whether the actual resources and capabilities help increase hospital's ability to create value.

5.2. Evaluation and estimation of supplier's hidden quality costs

In this phase, the full capabilities of suppliers are being gauged against defined requirements. We believe that the supplier's quality loss occurs when his performance deviates from a target, the higher the deviation is, the worst the service is rendered, and the higher the purchaser is dissatisfied. This dissatisfaction could be estimated by a cost. According to Taguchi, as indicated in figure 4, the variation of this cost is represented by a quadratic curve. The curve is centered on the target value, which represents the best performance, and determining this best value is not usually a simple task.

Three types of loss functions have been presented in the Taguchi loss function [19], nominal the better (NTB), larger the better (LTB); and smaller the better (STB). Firstly, It is necessary to determine the criteria to be based on to estimate the supplier's hidden cost, the target value, then the upper and lower specification limits.

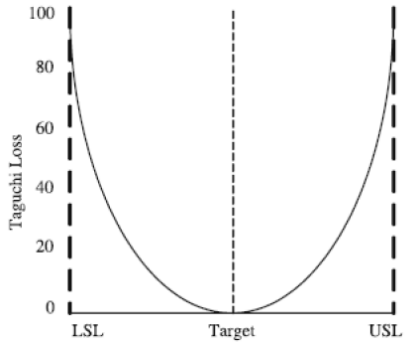


Figure 4. Two-sided loss function with specification preference

5.2.1. Supplier's hidden costs for quantitative quality characteristic

In this work, we adopt the form NTB for the quantitative quality characteristics which means that the nominal value is the best value. The loss function is given by Eq. 1.

$$L(y_i) = k(T-y_i)^2 \quad (1)$$

Where k is the loss coefficient, whose value is constant depending on the cost at the specification limits and the width of the specification, y_i is a particular value specification, that presents the value of the product characteristic, and $L(y_i)$ is the loss associated to y_i [19].

5.2.2. Supplier's hidden costs for qualitative quality characteristic

For qualitative criteria, we use a one-sided minimum (smaller-is-better) loss function of the form:

$$L(y_i) = k.(y_i)^2 \quad (2)$$

where the meanings of $L(y_i)$, y_i and k are the same as in equation 1. As indicated in figure 5, our philosophy aims to consider that the target value is presented as the best supplier's profile, which is characterized by the highest performance level at all quality attributes, and where the loss is equal to zero. Any supplier's distance from the ideal profile generates a cost; the smaller the distance, the lower the hidden cost is. This cost is estimated using TOPSIS method. This method measures this distance based on the supplier's performance level. It hypothesizes two artificial alternatives: The Ideal alternative that has the best level for all attributes considered, and the negative ideal alternative that has the worst attribute values [28]. TOPSIS allows determining the alternative that is the closest to the ideal solution and farthest from negative ideal alternative.

Let assumes that we have m alternatives, and n criteria. We have the score of each option with respect to each criterion. The supplier's distance from the ideal profile is calculated according to the following steps:

Let x_{ij} be the score of option i with respect to criterion j. We have a matrix X = (x_{ij}) with dimension m×n. Let J be the set of benefit attributes or criteria (LTB), and let J' be the set of negative attributes or criteria (SMB).

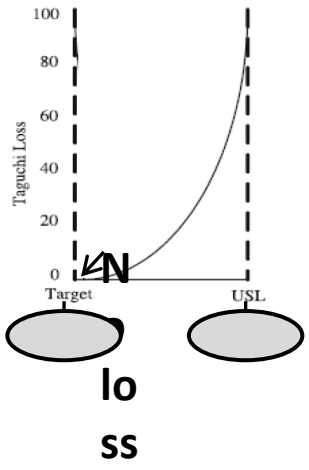


Figure 5. Loss function in case of qualitative criterion

Step 1: We construct normalized decision matrix

This step transforms various attribute dimensions into non-dimensional attributes, which allows comparisons across criteria. Scores or data are normalized as follows:

$$r_{ij} = x_{ij} / (\Sigma\, x^2_{ij}) \quad \text{for } i = 1, \ldots, m;\ j = 1, \ldots, n \tag{3}$$

Step 2: We construct the weighted normalized decision matrix.

In this step, we assume that we have a set of weights for each criteria w_j for j = 1,…,n. We multiply each column of the normalized decision matrix by its associated weight. An element of the new matrix is:

$$v_{ij} = w_j r_{ij} \tag{4}$$

Step 3: We determine the ideal and negative ideal solutions.

Ideal solution:

$$A^* = \{v_1^*, \ldots, v_n^*\}, \tag{5}$$

Where $v_j^* = \{ \max (v_{ij}) \text{ if } j \in J ;\ \min (v_{ij}) \text{ if } j \in J' \}$

Negative ideal solution:

$$A' = \{ v_1', \ldots, v_n' \}, \tag{6}$$

Where $v' = \{ \min (v_{ij}) \text{ if } j \in J ;\ \max (v_{ij}) \text{ if } j \in J' \}$

Step 4: We calculate the separation measures for each alternative.

The separation from the ideal alternative is:

$$S_i^* = [\Sigma (v_j^* - v_{ij})^2]^{1/2} \qquad i = 1, \ldots, m \tag{7}$$

Similarly, the separation from the negative ideal alternative is:

$$S'_i = [\Sigma (v_j' - v_{ij})^2]^{1/2} \qquad i = 1, \ldots, m \tag{8}$$

Step 5: We calculate the relative closeness to ideal solution y_i^*

$$y_i^* = S_i^* / (S_i^* + S'_i),\ 0 < y_i^* < 1 \tag{9}$$

The supplier's distance from the ideal profile determines *yi*, which allows calculating the loss occurred by equation (2).

5.3. Analysis and decision

There are significant costs associated with organizational structure management and supplier's performance. The choice of centralization or decentralization structure may reduce supplier's flexibility and therefore makes it difficult to be responsive to changing purchasing management. Furthermore, the structure chosen changes the distance and the number of touch points between hospitals entry and suppliers. Poor planning and oversight in such a highly complex demand leads to unnecessary expenses. For example, when many "hands" are involved, the right products in the right mix and volumes may not be in the right places at the right time. Therefore, this method allows us to evaluate our satisfaction with supplier's quality in the current structure, and to be taken into account before taking any decision of change.

5.4. Supplier's Monitoring

The movement of the healthcare system organization from one form to another often requires change within the organization itself. Therefore, the organization shall operate procedures for approval and continued monitoring of all its suppliers whose products or services may affect product safety. Monitoring a supplier's performance will provide an indication of changes within the suppliers' systems and procedures that could lead to potential issues. The results of evaluations and follow-up actions shall be recorded, for future decisions.

6. Practical case study: Military hospital of Rabat

We choose in this study the Mohamed V Military Hospital of Rabat/Morocco. This hospital carries the SEGMA status, and has been subject to the change of the structure from centralization to decentralization many times from the year 2000. The objective of this study is to report the experience of this hospital in the involvement of different types of structure in the management of medical products, materials and methods from 2001 to 2003. Taking into account the total number of suppliers who participated during this period, as shown in figure 6, we find that 47% of these suppliers participated during the decentralization period, 36% of them participated during centralization, and only 17% participated in both strategies. Our study will be focused on the estimation of hidden quality costs of suppliers, who have been involved in the purchase of medical products, in both strategies: centralization in 2001, and decentralization in 2003, in order to study the impact of organizational structure on supplier's behavior.

6.1. Definition of needs

As a result of the team efforts, and as recent data are confidential, we accessed to the suppliers' evaluation data in 2001 (centralization), and those in 2003 (decentralization). We highlight that old data will not have a significant impact on our

approach. In 2001, the hospital suffered from the unavailability of medical products, and the poor quality delivered by the suppliers, which was evaluated on several bases (see figure 7). The question which arises when changing the organizational structure from centralization to decentralization is whether this change will help suppliers to improve their performance or to get it worse. Therefore, in this section we will estimate the hidden quality cost behind the non-quality provided by suppliers in both strategies.

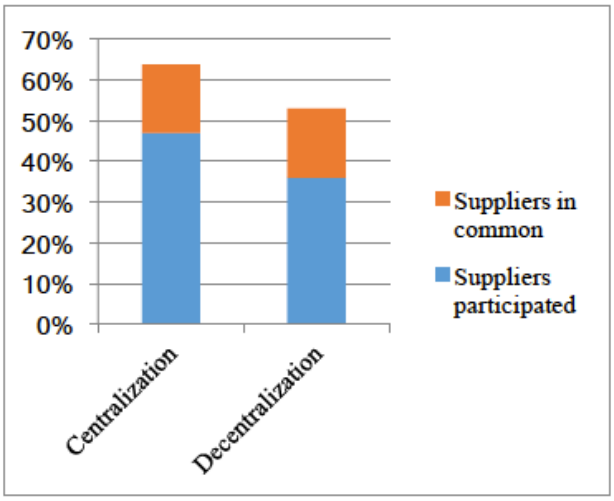


Figure 6. Suppliers participated in the period 2001-2003

6.2. Evaluation and estimation

6.2.1. Principle criteria for supplier's hidden quality costs

For the purpose of estimating supplier's hidden quality costs in the case of this hospital, we select three principle criteria: the delivery, the validity period as a single quantitative criterion, and quality as a multiple qualitative criterion, which is subdivided into five sub attributes. The objective is to estimate the supplier's hidden quality costs in seven cases as presented in the table 2.

6.2.2. Estimation of supplier's hidden quality costs

To apply the existing Taguchi loss function for quantitative quality characteristic, and the developed method for qualitative quality characteristic, we develop a purchasing model, in which we assume that the purchaser only knows the prices that might have different impacts on product quality and delivery. We estimate the loss associated to 7 suppliers according to these seven criteria.

- Quantitative quality characteristic

To calculate the loss associated to the quantitative criteria, we first need to specify the upper specification limit and lower specification limit for each criterion. For the case of this hospital, they are determined as follow:

- The maximum acceptable delivery and the minimum acceptable delivery are 7 days.
- The minimum period of validity is 18 months.

The average price of the market is 80.

As shown in table 3, we assume that the effect of price on delivery and validity period follows an exponential function, where delta price term shows the difference between the supplier price and the minimum price showed in the market. Then the estimation of the loss occurred in both structures, is determined in table 4 and 5.

- Qualitative quality characteristic

To estimate the loss associated to the quality criterion for each supplier, we suppose that: The maximum deviation from the ideal profile is 40%

- For the purchaser's orientation, the weight for each sub-criterion in the quality criterion is 0.2

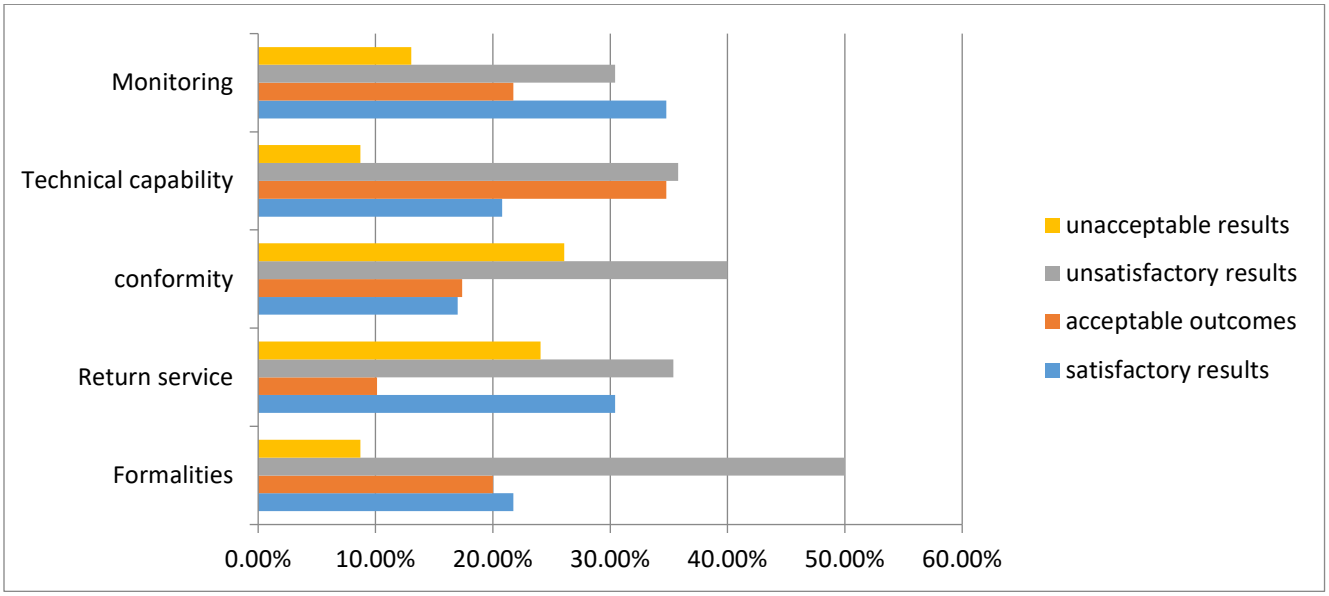


Figure 7. Supplier's evaluation criteria

Table 2. Criteria for supplier's hidden quality costs and their determination

Quantitative quality characteristic	1. Delivery		Costs to catch up the purchase unavailability as consequence of the supplier's delay or to pay extra fees to store the command for later use as a result of an order delivered in advance.
	2. Period of validity		Costs generated by expired medications due to suppliers that do not respect the validity period.
Qualitative quality characteristic	3.Quality	3.1. Monitoring	Costs to unlock the logistics problem in the hospital's upstream supply chain, generated by supplier's bad manufacturing practices, which include criteria for personnel, facilities, equipment, materials, manufacturing operations, labeling, packaging, quality control and, stability testing. These costs imply a supplier's bad conformity.
		3.2. Technical capability	Costs to maximize the return potential of the hospital's pharmacy generated by suppliers who have a bad experience in pharmaceutical returns service, and also to generate the minimum return credit for the hospital's outdated pharmaceuticals.
		3.3. Conformity	Costs to avoid unavailability of medicines in the pharmacy, and to monitor and check the supplier ongoing process of reviewing the degree to which programmed activities are completed and objectives are being met.
		3.4. Return service	Costs related to small errors of medications structure which can cause patients death. Hospital pharmacy policy requires providers to establish a continuous, systematic, and criteria-based evaluation system, such as the form and dose of drug use, that will help ensure the appropriate use of drugs.
		3.5.Formalities	Costs to assure continuity of hospital's supply and reliability of medicines quality, due to a supplier's bad technical capability. Hospital's pharmacy analysis confirms the reasonableness of the type and amount of resources proposed by the supplier. This analysis covers the proposed types and quantities of materials, labor, processes, special tooling, facilities, the reasonableness of scrap and spoilage, and other factors set forth in the supplier's proposal.

Table 3. Functions used to describe the effects of price on delivery and loss coefficients values.

	Delivery in advance	Late delivery	Period of validity in advance
Impact of price	High	Medium	Medium
Price function	$e^{\Delta price/-300}$	$e^{\Delta price/-200}$	$e^{\Delta price/-100}$
K	333.3	222.22	111.1

First, we have to determine for each supplier the independent variable y_i in equation 2. For this purpose, we rate each attribute of each alternative in the form of a matrix. To study the impact of these matrices values on final results, we assume that we have 2 experts; each one of them is assigned with the task of evaluating each supplier against each sub-criterion in each strategy. Then we determine the ideal solution, the ideal negative solution, and consequently the relative closeness to the ideal solution according to each expert (See table 6).

To get the values of k, we first need to specify the upper specification limit for each measure. As mentioned earlier, a loss of 100% will take place if a supplier image is 40% far away from the ideal profile. The other numerical measure that we need to model is how price affect quality characteristics. The expert might have different priorities regarding these criteria. For instance, one expert might be interested in high quality, while another in on-time delivery. For this reason, we assume that price effect is considered as medium or high in both centralized and decentralized structure. We use an exponential function to model the price effect for all experts.

Substituting these values in equation 2, we obtain the following loss coefficient values k, as shown in Table 7.

Table 4. Loss function results for delivery criterion

	Suppliers	y (days)	Loss L(y)
Centralization	S2	-8	21331.2
	S4	-5	8332.5
	S3	-9	26997.3
	S1	0	0
	S5	10	22220
	S6	7	10887.8
	S7	14	43551.2
Decentralization	S6	-9	26997.3
	S4	-4	5332.8
	S2	-3	2999.7
	S3	9	17998.2
	S5	10	22220
	S1	12	31996.8
	S7	14	43551.2

Table5. Loss function results for period of validity criterion

	Centralization		Decentralization	
	y (Month)	L(y)	y (Month)	L(y)
S1	2,8	871.024	2	444.4
S2	4,5	2249.775	3,4	1284.316
S3	2,3	587.719	5,2	3004.144
S4	3,6	1439.856	3,6	1439.856
S6	4,7	2454.199	3,1	1067.671
S5	5,9	3867.391	6,5	4693.975
S7	6,9	5289.471	6,7	4987.279

Table 6. Particular value specification results for each structure

	Centralization		Decentralization	
	Expert 1	Expert 2	Expert 1	Expert 2
S1	0.95299848	0.98239346	0.7435345	0.794534
S2	0.79439686	0.69798985	0.5469879	0.537699
S3	0.59252325	0.540469002	0.2435345	0.265476
S4	0.80358862	0.79417179	0.6967687	0.7176987
S5	0.99745556	0.96301608	0.2276879	0.19756789
S6	0.87174095	0.85333792	0.1632234	0.1624908
S7	0.52251934	0.55605416	0.2987698	0.283423

Table7. Functions used to describe the effects of price on quality and loss coefficients values

	Price function	k
Medium	$e^{\Delta price/-200}$	2222,22
High	$e^{\Delta price/-300}$	3333,3

After determining the y values for each supplier in each strategy, and the loss coefficients values k, we calculate the loss occurred in each case as presented in table 8.

Table8. The supplier's hidden quality loss values according to quality criterion

				Medium	High
			y	L(y)	L(y)
Centralization	Expert 1	S1	0,95299848	2018,2156	3027,3234
		S2	0,79439686	1402,35569	2103,53354
		S3	0,59252325	780,178424	1170,26764
		S4	0,80358862	1434,99603	2152,49404
		S5	0,99745556	2210,90588	3316,35882
		S6	0,87174095	1688,72152	2533,08228
		S7	0,52251934	606,719401	910,079101
	Expert 2			Medium	High
			y	L(y)	L(y)
		S1	0,98239346	2144,63835	3216,95753
		S2	0,69798985	1082,63324	1623,94986
		S3	0,60469002	812,547475	1218,82121
		S4	0,79417179	1401,56117	2102,34175
		S5	0,96301608	2060,86821	3091,30232
		S6	0,85333792	1618,17405	2427,26108
		S7	0,55605416	687,09586	1030,64379
Decentralization	Expert 1			Medium	High
			y	L1(y)	L2(y)
		S1	0,7435345	1228,52894	1842,79341
		S2	0,5469879	664,872824	997,309236
		S3	0,2435345	131,796577	197,694865
		S4	0,6967687	1078,84837	1618,27255
		S5	0,3876879	334,0009	501,001349
		S6	0,1632234	59,203582	88,805373
		S7	0,2987698	198,361113	297,541669
	Expert 2			Medium	High
			y	L1(y)	L2(y)
		S1	0,794534	1402,83992	2104,25988
		S2	0,537699	642,482941	963,724411
		S3	0,265476	156,615115	234,922673
		S4	0,7176987	1144,63616	1716,95424
		S5	0,356789	282,882483	424,323725
		S6	0,198769	87,7971562	131,695734
		S7	0,283423	178,506208	267,759312

6.3. Analysis and decision

We assume that the effect of price on delivery and validity

period is the same in both centralization and decentralization strategies. However, there is a big change at the level of the supplier's hidden costs. The figure 8 indicates that loss results for the delivery criterion are much greater in the case of decentralization. The purpose of having several suppliers allows to benefit from the specific superiorities of each supplier in his specialty and to spread their risks. However, a purchase in small quantity leads to less favorable price and costs relationship. This decreases the chance of negotiating effectively the delivery deadlines, and does not help to put pressure on suppliers to meet the delivery date. The figure 9 shows that for most of suppliers, the cost relative to period of validity is high in the case of centralization. In fact, most of suppliers have medicines on stock whose period of validity have been already started, which constitutes a loss to the organization as long as they are stored. In one hand, the supplier tries to get the maximum profit, by lowering the unit purchase price to encourage the buyer. In other hand, the buyer is mainly based on maximizing the cost criterion, and does not pay enough attention to the period of validity. As a result, the buyer procures to optimize purchasing costs; but, in return the stocks in the hospital pharmacy hides a huge loss.

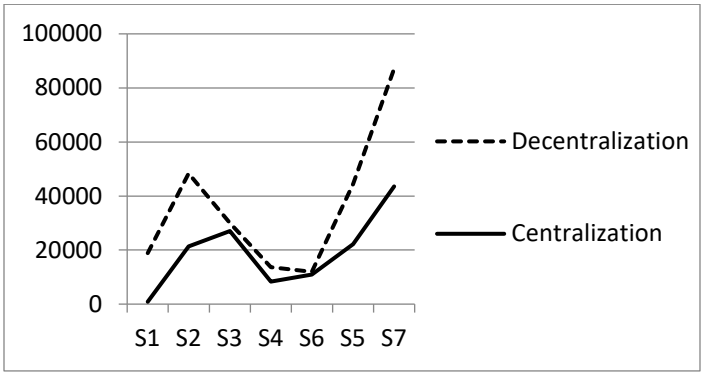


Figure 8. Loss results for delivery criterion in both structures

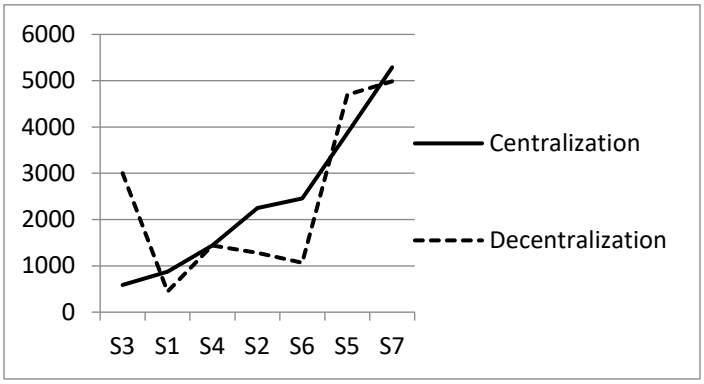


Figure 9. Loss results for period of validity criterion in both structures

In figure 10, the results determining the closest supplier to the best profile differ from one expert to another. In centralization strategy, the first expert indicates that S1 is the closest one, the second expert marks S3, while both experts' opinion answers for the decentralization strategy coincide at the same level for S6, which explains that experts' experiences bring different points of view on suppliers' evaluation.

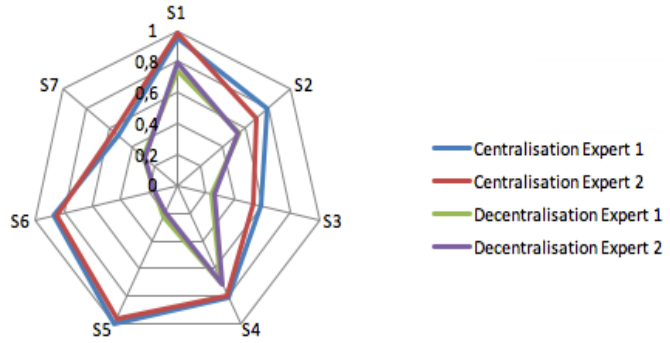


Figure 10. Expert's evaluation results

There are two parameters that might impact the supplier's loss function results: the value y, and the loss coefficient k. In one hand, the value y is influenced by the experts' evaluation (figure 11 shows that although both curves are almost symmetric for each strategy and follow the same trend, the maximum difference value is up to 5,6% in centralization strategy). In other hand, the loss coefficient is influenced by the determination of the impact of price on quality criterion, which is different from an organizational structure to another. As indicated in Figure 12, an increase in the loss coefficient results in an increase of the loss function, whose variation presented in an offset curve with a huge difference that comes to a maximum of 40%. Relatively speaking, the difference between experts' evaluations in centralization and decentralization strategy doesn't have a big impact on the variation of supplier's loss function in both structures which shows the effectiveness of our method.

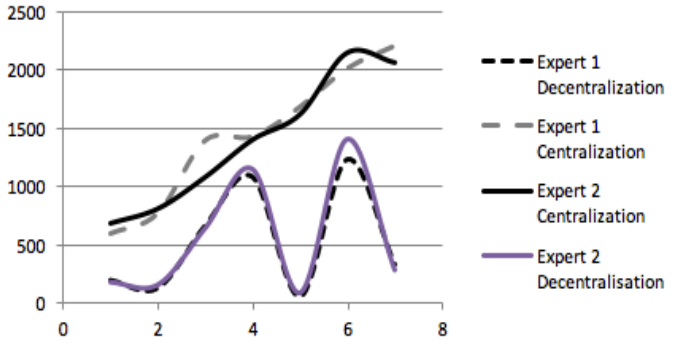


Figure 11. Variation of loss function in the case of medium impact of price on quality in both strategies

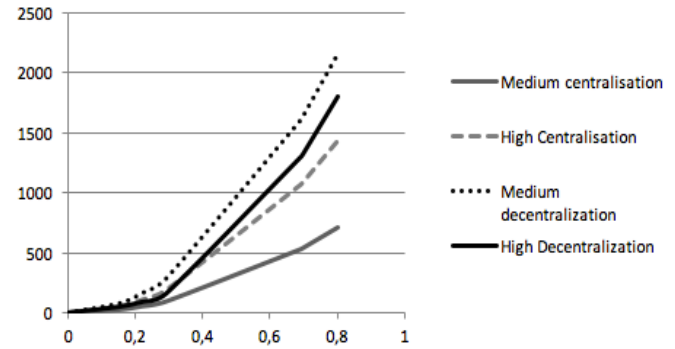


Figure 12. Variation of loss function in the case of first expert in both strategies

As shown in table 9, supplier's hidden quality costs are higher in the case of centralization, and this is due to:

- Centralization avoids sharing the responsibilities that can lead to coordination difficulties, and promotes decision-making consistent with the defined strategy.
- The centralized purchasing team may lack expertise on certain products, and their distance from user expectations may cause dissatisfaction.
- An error on a purchase will have multiple consequences, to the extent that the volumes purchased are more important, whether in terms of quality or delivery.
- Communication problems may occur between the various hospitals and the centralized purchasing site.
- Increased risk of dependence on suppliers as there is a risk of having too many purchases from few suppliers.

Table 9. Supplier's hidden quality costs in both organizational structures

	Centralization	Decentralization	Gain (C - D)/ C
S1	2889,2396	19226,7289	0,13190022
S2	22733,5557	27662,1728	0,03414751
S3	27777,4784	3131,49658	-0,1707576
S4	9767,49603	6411,64837	-0,0232507
S5	13098,7059	1401,6719	-0,0810419
S6	23908,7215	22279,2036	-0,0112899
S7	44157,9194	43749,56111	-0,00282927

7. Conclusion

The purchasing organizational structure is considerably changed according to many factors. However, this change doesn't take into account its impact on supplier's hidden quality costs. In this study, a deciding making approach is elaborated for the hospital sector. Regarding to this new approach, a new method of measuring supplier's hidden quality costs is developed. The idea is to use improved Taguchi quality loss function in order to measure the loss associated to supplier's bad quality and help the manager to take an appropriate decision regarding the purchasing structure. Based on this work, we conclude that an erroneous value of the loss coefficient can result in misallocation of the loss function value in centralization and decentralization strategy. Therefore, future research needs to address this coefficient, which represents the impact of price on quality, differs from a structure to another and which is considered as an important component of quality loss function.

References

[1] K. Jenoui, A. Abouabdellah, "Estimating supplier's hidden quality costs with Taguchi quality loss function and Topsis method". In 10th International Colloquium on Logistics and Supply Chain Management, Rabat, Morocco, 2017. https://doi.org/ 10.1109/LOGISTIQUA.2017.7962881

[2] A. Abouabdellah, A. Cherkaoui, "Decision Support System for Predicting the degree of a cancer patient's empowerment". Journal of Theoretical and Applied Information Technology, 60(3), 517-523, 2014.

[3] A. Marie, C. Giuliani, A. Abouabdellah, A. Cherkaoui, "The empowerment of patients factoring: Reporting to a holonic approach". In 9th Conference on service systems and service management, Troyes, France, 2006.

[4] D. Serrou, A. Abouabdellah, "Logistics in the hospital: Methodology for measuring performance". ARPN journal of engineering and applied sciences, 11(5), 250-256, 2016.

[5] D. Serrou, A. Abouabdellah, "Study grouping pharmacies impact on the performance of the hospital supply chain". In 6th International conference on industrial engineering and systems management, Seville, Spain, 2015.

[6] D. Serrou, A. Abouabdellah, "Study of Improved fiscal performance of hospital supply chain pharmacies by centralizing". In 45th International conference on computers & industrial engineering, Metz, France, 2015.

[7] WN. Pi, C. Low, Int J AdvManuf, "Supplier Evaluation and Selection Using Taguchi Loss Functions". The International Journal of Advanced Manufacturing Technology, 5 (1), 1123-1130, 2005. https://doi.org/10.1007/s00170-003-1975-5

[8] P. Trouiller, Technical support for improving the management and organization of hospital pharmacies, 2013.

[9] J. Dumoulin, Analyse du système d'approvisionnement pharmaceutique au Maroc : l'expérience de regroupement des achats depuis 2001, 2004.

[10] Saharidis, G.K.D., Kouikoglou, V.S., Dallery, Y. "Centralized and decentralized control polices for a two-stage stochastic supply chain with subcontracting". International Journal of Production Economics. 11(6), 117-126, 2009. https://doi.org/10.1016/j.ijpe.2008.10.001

[11] Chen, J.M, Chen, T.-H. "The multi-item replenishment problem in a two-echelon supply chain: the effect of centralization versus decentralization". Computers & Operations Research. 32(10), 3191-3207, 2005. https://doi.org/10.1016/j.cor.2004.05.007

[12] Chang, M.H, Harrington, J.E., Jr. "Centralization vs. Decentralization in a Multi-Unit Organization: A Computational Model of a Retail Chain as a Multi-Agent Adaptive System". Management Science. 46(10), 1427-1440, 2000. https://doi.org/10.1287/mnsc.46.11.1427.12085

[13] Behdani, B.Lukszo, Z.Adhitya, A.Srinivasan, "Decentralized vs. centralized management of abnormal situations in a multi-plant enterprise using an agent-based approach, in Computer Aided Chemical Engineering; 28(C), 1219-1224, 2010. https://doi.org/10.1016/S1570-7946(10)28204-4

[14] Hall, Ma, Tomkins, "A cost of quality analysis of a building project: Towards a complete methodology for design and build". Construction Management and Economics Journal. 19 (7), 727-740, 2001.

[15] A. Kazaz, M. Birgonul, "The evidence of poor quality in high rise and medium rise housing units: A case study of mass housing projects in Turkey". Building and environment journal. 40 (11), 1548-1556, 2005. https://doi.org/10.1016/j.buildenv.2004.11.023

[16] P. Love, "Auditing the indirect consequences of rework in construction: A case based approach". Management and audit journal. 17 (3), 138-146, 2002. https://doi.org/10.1108/02686900210419921

[17] P. Love, H. Li, "Knocking down walls and building bridges overcoming the problems associated with quality certification". Construction management and economics journal. 18(3), 321–332, 1999.

[18] P. Love, D. Edwards, "Determinants of rework in building construction projects". Engineering Construction & Architectural Management. 11(4), 259–274, 2004.

[19] R. Kethley, T. A. Waller, "Improving Customer Service in the Real Estate Industry: A Property Selection Model Using Taguchi Loss Functions". Total Quality Management. 13(6), 739-748, 2002.

[20] C. Quigley, C. McNamara, "Evaluating Product Quality: An Application of the Taguchi Quality Loss Concept". International Journal of Purchasing Materials. 28(3), 619-628, 1992.

[21] R. Kethley, T. A. Waller, "Improving Customer Service in the Real Estate Industry: A Property Selection Model Using Taguchi Loss Functions". Total Quality Management. 13(6), 739-748, 2002. https://doi.org/10.1080/0954412022000010109.

[22] H. C. Li, "Quality Loss Functions for the Measurement of Service Quality". The International Journal of Advanced Manufacturing Technology, 2(5), 29-37, 2003. https://doi.org/ 10.1007/s001700300004.

[23] R. Kethley, B. Waller, T. Festervand, "Improving customer service in the real estate industry: a property selection model using Taguchi loss functions". Total Quality Management, 13(6), 739–748, 2002. https://doi.org/ 10.1080/0954412022000010109

[24] M. W. Kim, W. Liao, "Estimating hidden quality costs with quality loss functions". Accounting Horizons. 8(1), 8–18, 1994.

[25] P. Kotler, "Strategies for marketing service excellence". The 1st Global

Conf, Management Centre, Europe, Brussels, 1991.
[26] S. K. Krishnan, "Increasing the visibility of hidden failure cost". Measuring Business Excellence, 10(4), 77–101, 2006.
[27] E. Timmerman, "An approach to vendor performance evaluation". Journal of Purchasing and Supply Management, 1(7), 27-32, 1986.
[28] R. E. Gregory, "Source selection: a matrix approach". Journal of Purchasing and Materials Management. 22(2), 24–29, 1986.
[29] W. R. Soukup, "Supplier selection strategies". Journal of Purchasing and Materials Management. 28(2), 7-12, 1987.
[30] K. Thompson, "Vendor prole analysis". Journal of Purchasing and Materials Management. 26 (1), 11-18, 1990.
[31] G. Barbarosoglu, T. Yazgac, "An application of the analytic hierarchy process to the supplier selection problem". Production and Inventory Management Journal, 38(1), 14–21, 1997.
[32] R. Nydick, R. P. Hill, "Using the Analytic Hierarchy Process to structure the supplier selection procedure". International Journal of Purchasing and Materials Management, 25(2), 31–36, 1992.
[33] J. Sarkis, S. Talluri, "A model for strategic supplier selection". In 9th International IPSERA Conference, London, UK, 2000.
[34] S. S. Chaudhry, F. G. Forst, J. L. Zydiak, "Vendor selection with price breaks". European Journal of Operational Research, 16(6), 52–66, 1993.
[35] E. C. Rosenthal, J. L. Zydiak, S. S. Chaudhry, "Vendor selection with bundling". Decision Sciences, 26 (1), 35- 48, 1995.
[36] F. P. Bua, W. M. Jackson, "A goal programming model for purchase planning". Journal of Purchasing and Materials Management, 19(3), 27–34, 1983.
[37] C. A.Weber, L. M. Ellram, "Supplier selection using multi objective programming: a decision support system approach". International Journal of Physical Distribution and Logistics Management. 23(2), 3-14, 1992.
[38] K. Jenoui, A. Abouabdellah, "System of multisourcing supplier's selection and evaluation in the hospital sector integrating the criteria: Total cost, Gap time, Risk performance". ARPN Journal of Engineering and Applied Sciences. 11(17), 10433-37, 2015.
[39] K. Jenoui, A. Abouabdellah, "Implementation of a decision support system heuristic for selecting suppliers in the hospital sector". In 6th International conference on industrial engineering and systems management, Seville, Spain, 2015.
[40] K. Jenoui, A. Abouabdellah, " Single or multiple sourcing strategy: a mathematical model for decision making in the hospital sector". In 11th International conference on intelligent systems: theories and applications, Mohamedia, Morocco, 2016.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 203-216(2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Novel Analysis of Synchronous and Induction Generators in Parallel Operation Mode in an Isolated Electric System

Vinicius Zimmermann Silva*, Ângelo Jose Junqueira Rezek, Rafael Di Lorenzo Corrêa

Federal University of Itajubá UNIFEI, Energy and Electric System Institute ISEE, 37500903, Brazil

ARTICLEINFO

Article history:
Received: 29 October, 2017
Accepted: 01 December, 2017
Online: 15 December 2017

Keywords:
Induction Generator
Isolated Electric System
Generators in Parallel
Synchronous Generator

ABSTRACT

This paper presents an analysis of a parallel connection of one synchronous generator and one self-excited induction generator that feeds a resistive load and an Induction Motor. The system voltage and frequency are controlled by a voltage control loop and a speed control loop connected to synchronous generator. The induction generator speed is controlled by its primary machine that is fed by an autotransformer and a diodes bridge. Through by voltages applied by an adjustable tap autotransformer connected to induction generator's primary machine, it is possible widen the range of its shaft speed if compared with the shaft speed caused by only field flux variation method. Then, by the autotransformer method is possible to widen the speed and power limits from the induction generator what increases the induction generator contributions and relieves the power supply from synchronous generator. Analysis of generators power supply and its interactions in various operational scenarios are shown. The results enable comparisons of the two methods of induction generator speed control, either by auto-transformer method or by field flux variation method. The first results in larger range of speed and power from the induction generator. Therefore, it has more representative features of actual field conditions.

Nomenclature and Abbreviations

V_{SG}	Synchronous generator voltage
I_{SG}	Synchronous generator current
f_{SG}	Synchronous generator frequency
$I_{fieldSG}$	Synchronous generator field current
n_{SG}	Synchronous generator speed
DCM_{IG}	Direct current motor coupled with induction generator
DCM_{SG}	Direct current motor coupled with synchronous generator
DCG_{SG}	Direct current generator coupled with synchronous generator
I_{aDCMSG}	Armature current of direct current motor coupled with synchronous generator
I_{fDCMSG}	Field current of direct current motor coupled with synchronous generator
V_{IG}	Induction generator voltage
f_{IG}	Induction generator frequency
I_{IG}	Induction generator current
n_{IG}	Induction generator speed
I_{aDCMIG}	Current of direct current motor coupled with induction generator
I_{fDCMIG}	Field current of direct current motor coupled with induction generator
I_c	Capacitor bank current
I_{loadR}	Resistors banks current
Q	Reactive power
pF, cosϕ	Power factor
Xc	Capacitive reactance
SG	Synchronous generator
IG	Induction generator
PIG	Induction generator power
PSG	Synchronous generator power
P_{DCMSG}	DCM power input coupled with synchronous generator
P_{DCMIG}	DCM power input coupled with induction generator
I_{WSG}	SG active current
I_{WIG}	IG active current
Va_{DCMIG}	DCM armature voltage coupled with IG
Va_{DCMSG}	DCM armature voltage coupled with SG
Ra	Armature resistance

* Vinicius Zimmermann Silva, vinicius.zimmermann@yahoo.com.br

https://dx.doi.org/10.25046/aj020625

C	Conjugate
Xs	Synchronous reactance
E	Back electromotive force
V_{pn}	Phase neutral voltage
FPSO	Floating production storage offloading
IM	Induction Motor
IdwMT	Motor reactive current
IwMT	Motor active current

1. Introduction

This paper is an extension of work originally presented in PEDG conference [1].

The use of induction generators in generation systems is usual in wind plants [2-4] or eventually in small hydroelectric generation units [5]. Otherwise, from studies started in [1], it is shown a new knowledge border which consider the use of IG in parallel of SG in an isolated electric system. The first results and behavior of this kind of generation system were shown in [1] and are followed by this current article with original and deeper analysis.

Then, this work is composed of analysis over an expanded system that includes an induction motor, a resistive load, an autotransformer connected to a diodes bridge to feed the DCM_{IG} armature and then permits to increase manually the output electric power limits if compared to flux variation method (3) presented in [1]. Therefore, this article shows the IG supplying more active power and the induction motor inserted in experiment what enables a novel analysis of the system behavior.

This paper presents results that show additional remarkable characteristics by operation in parallel mode among one SG and one IG. Some characteristics, such as reduced weight and size, easier maintenance, and shorter manufacturing and delivery time are more associated with induction generators and are relevant to oil platforms or offshore installation concerns. Besides, it has absence of dc supply for excitation and better transient performances [5].

One of the potential motivation for this study is to develop a potential alternative capable of optimizing the main electric system currently adopted in oil platforms and floating production storage and offloading (FPSO) ships to become cheaper, simpler, lighter, and more efficient.

The potential application of this study is the replacement of one or two synchronous generators by one or two induction generators in oil offshore platforms or FPSO. A typical offshore electric system uses three or four turbo-generators in 13800 V, 60 Hz, driven by dual fuel (fuel gas or diesel oil) turbines. In typical operation, three main turbo-generators operate and can supply all unit consumers with the fourth in standby. During the peaks of load, the fourth main turbo-generator may be required to meet the demand. Therefore, the whole electric system shall be suitable for this operational condition.

Therefore, the study target is to analyze various aspects of topology and operating involving an induction generator in parallel with a synchronous generator for application in an isolated electric system, and to establish the operational viability, advantages and challenges.

2. Isolated Electric System

The isolated electric system was sized as shown in the following equations and data. The automatic controls of the system consist of a voltage control loop and a speed control loop. Both are inserted in the SG scheme as shown in Figure 5 and Figure 6. The dataplates of the principal equipment are shown in Tables 1 to 6.

The parameterization used in both electronic control boards is indicated in [1], except for the references of voltage and speed for each of the control loops which are respectively 65.5 and 65.2 as shown in Figure 9.

2.1. Dataplate

In tables 1 to 4 are shown the main equipment's dataplates and the tables 5 to 6 are shown the load's dataplates used in the experiments along this work.

Table 1: DMC_{IG} Dataplate

Direct Current Motor $_{IG}$				
220 V	7.72 A	1.7 kW	1500 rpm	600 $mA_{fieldmax}$

Table 2: DMC_{SG} Dataplate

Direct Current Motor $_{SG}$				
220 V	9.1 A	2.0 kW	1800 rpm	600 $mA_{fieldmax}$

Table 3: IG Dataplate

Induction Generator (IG)					
220 V	7.5 A	1.86 kW	1410 rpm	0.8 PF	50 Hz

Table 4: SG Dataplate

Synchronous Generator (SG)					
230 V	5.0 A	2.0 kVA	1800 rpm	0.8 PF	60 Hz
V_{field}: 220 V			$I_{fieldmax}$: 600 mA		

Table 5: Load Dataplate

Resistive Load		
Load 1 (kW)	Load 2 (kW)	Load 3 (kW)
2/3	2/3	2/3

Table 6: Load, Induction Motor Dataplate

Induction Motor			
0.37 kW	1715 rpm	Cos ϕ: 0.71	60 Hz

2.2. Equations - Part I

Follow the main equations for analysis comprehension and to establish a studies base:

- Direct Current Motor

$$n = \frac{E}{k * \emptyset} \quad (1)$$

$$n = \frac{Ua - (\sum Ra) * Ia}{k * \emptyset} \quad (2)$$

$$C_{conjugate} = k * \emptyset * Ia \quad (3)$$

- Synchronous Generator

$$\dot{E} = \dot{U}a + (Ra + jX_s) * \dot{I}_a \quad (4)$$

- Induction Generator

$$S_{slip}\% = \frac{n_s - n_{IG}}{n_s} * 100 \quad (5)$$

$$n_s = \frac{120 * f_{SG}}{P_{number\ of\ poles}} \quad (6)$$

2.3. Capacitor Bank Sizing

As informed at the IG dataplate

$$\cos\phi = 0.8, \text{then } \sin\phi = 0.6$$

The reactive power is calculated to attend the reactive demand of the Induction Machine:

$$Q = \sqrt{3} * V * I * sen\phi \quad (7)$$

$$Q = \sqrt{3} * 220 * 7.5 * 0.6 => 1714.7 \text{ VAr}$$

$$Q_{generated} = Fc * Q \quad (8)$$

Fc =1.2 [6]

For the machine coupled at a resistive load which requires 7.5A, it is necessary for the reactive power generation to be approximately 2057.6 VAr as demonstrated below.

$$Q_{generated} = 1.2 * 1714.7 => 2057.6 \text{ VAr} \quad (9)$$

$$Q = \frac{3*V^2}{X_c} \quad (10)$$

$$Q = \frac{3*220^2}{X_c} => X_c = \frac{3*220^2}{2057.6} = 70.6\ \Omega \quad (11)$$

$$X_c = 70.6\ \Omega$$

$$X_c = \frac{1}{2*\Pi*f*C} = \frac{1}{2*\Pi*50*C} => \quad (12)$$

$$C = \frac{1}{2*\Pi*50*70.6} \cong 40\mu F \text{ per phase} \quad (13)$$

2.4. Resistive Divider Sizing:

- Field Control Loop Resistive Divider as in Figure 5 and Figure 6.

$$(1\ k\Omega + r_1) \rightarrow 300\ V_{(brigge\ output\ voltage)} \quad (14)$$

$$r_2(1\ k\Omega) \rightarrow 4V_{(MP\ 410T\ voltage\ limit)} \quad (15)$$

$$r_1 = 74\ k\Omega\ and\ r_2 = 1k\Omega \quad (16)$$

- Resistor Power Sizing:

$$P_{power} = {300^2}/{(75 * 10^3)} = 1.2\ W \quad (17)$$

- The resistors that were selected based on the sized resistor, were:

$$r_1 = 79.1\ k\Omega\ and\ r_2 = 947\Omega \quad (18)$$

- Speed Control Loop Resistive Divisor as in Figure 5 and Figure 6.

$$(1\ k\Omega + r_1) \rightarrow 36\ V_{(tacho\ generator\ output\ voltage)} \quad (19)$$

$$r_2(1\ k\Omega) \rightarrow 4V_{(MP\ 410T\ voltage\ limit)} \quad (20)$$

$$r_1 = 8\ k\Omega\ and\ r_2 = 1k\Omega \quad (21)$$

- Resistor Power sizing:

$$P_{power} = {36^2}/{(9 * 10^3)} = 1.14\ W \quad (22)$$

- The resistors chosen, based on the sized resistor, were

$$r_1 = 8{,}3\ k\Omega\ and\ r_2 = 947\Omega$$

2.5. Equations - Port II

Follow the system equations base to calculate the power and efficiencies shown in Table 7 and Table 8 for each generator and entire group of machines.

$$I_{SG}{}^2 = Iw_{SG}{}^2 + Ic_1{}^2 \quad (23)$$

$$I_{IG}{}^2 = Iw_{IG}{}^2 + Ic_2{}^2 \quad (24)$$

$$Ic - IdwMT = Ic_1 + Ic_2 \quad (25)$$

$$I_{wloadR} + IwMT = Iw_{SG} + Iw_{IG} \quad (26)$$

As I_{SG}, I_{IG}, Ic and I_{wloadR} are measured values, shown in Table 7 and Table 8, IdwMT(Motor's reactive current) and IwMT(Motor's active current) are values calculated in appendix 1. Then, the system has 4 four variables and 4-four equations. Then, for each scenario, the four variables Iw_{SG}, Ic_1, Iw_{IG} and Ic_2 were calculated by Matlab software and are shown in Table 7 and Table 8.

The entire group efficiency and the efficiency of each subgroup were calculated based on PIG, PSG, PDCMSG, and PDCMIG, as follows:

$$P_{SG} = \sqrt{3} * V_{SG} * Iw_{SG} \quad (27)$$

$$P_{IG} = \sqrt{3} * V_{IG} * Iw_{IG} \quad (28)$$

$$P_{DCMSG} = Va_{DCMSG} * Ia_{DCMSG} \quad (29)$$

$$P_{DCMIG} = Va_{DCMIG} * Ia_{DCMIG} \quad (30)$$

$$\eta_{group}\% = \frac{P_{SG}+P_{IG}}{P_{DCMSG}+P_{DCMIG}} * 100 \quad (31)$$

$$\eta_{SG}\% = \frac{P_{SG}}{P_{DCMSG}} * 100, \quad (32)$$

$$\eta_{IG}\% = \frac{P_{IG}}{P_{DCMIG}} * 100 \quad (33)$$

2.6. Equations - Part III

The equations e calculus used to determined the IM parameters and IM equivalent circuit as primary, secondary

impedances and currents, including IM currents showed in Table 8 are calculated in item 5 of appendix 1.

2.7. The Experiment and Schemes

The experiment was mounted in the laboratory as shown in Figure 1 and Figure 2. The detailed circuit is shown in Figure 6. Figure 2 shows another experiment view, including the taco generator and its connections. The detailed circuit is shown in Figure 6. Figure 3 shows the typical connections of both circuit board MP 410T which are implemented in this experiment for the voltage loop and speed loop. They are also indicated in Figure 5 and Figure 6.

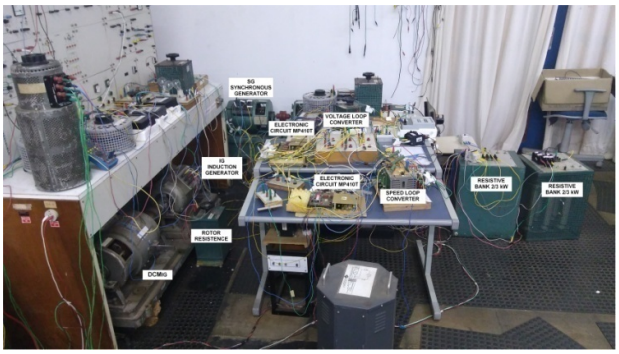


Figure 1 Laboratory assembly

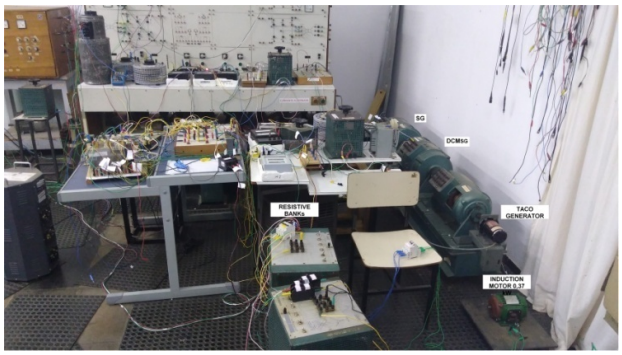


Figure 2 Another view of laboratory assembly

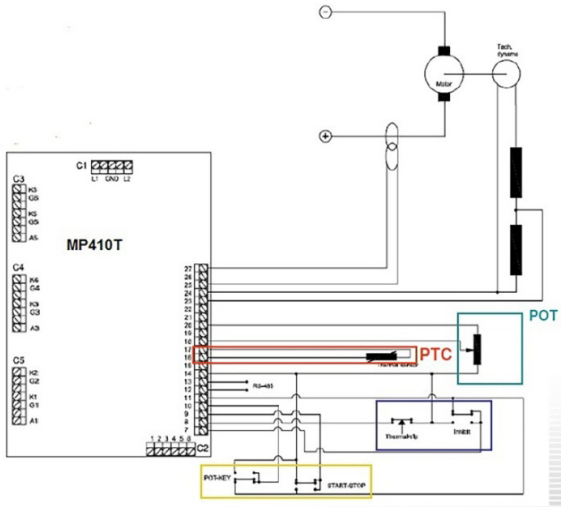


Figure 3. Connections arrangement of MP410T and devices

Figure 4 shows the 1B6C configuration used for the two thyristor converter bridges which were used in the control loops.

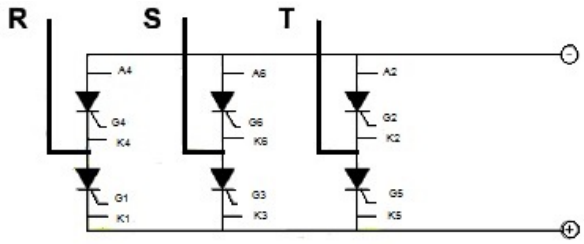


Figure 4. Converters configuration used in voltage and speed loops

As seen in [1], the power from IG, PIG, was limited because the I_{fDCMIG} reached the maximum viable value in accordance with DCM_{IG} current specifications. This limited speed from DCM_{IG} was a challenge because it was not showing a representative contribution from PIG. It was necessary to elevate the DCM_{IG} speed, n_{IG}.

Then, the methodology consists of implementing the comparative analysis of the power and efficiencies, starting with an analysis between the PIG from the scheme in [1] and PIG from the scheme in Figure 5 and Figure 6, covering the scenarios shown in Table 7 related to Figure 5 and Table 8 related to Figure 6. Finally, a comparative efficiency analysis was conducted for two subgroups, one composed of IG and DCM_{IG} and the other composed of SG and DCM_{SG} in scenarios related to the scheme in [1] and the schemes in Figure 5 related to Table 7 and Figure 6 related to Table 8.

Thus, the increase of DCM_{IG} speed, increase of PIG contributions and improvement of IG subgroup efficiencies are shown in results.

Figure 6 shows the scheme implemented in the laboratory to overcome the challenge related to limitation of IG speed [1] and bring the experiment closer to actual offshore conditions, taking into advantages of IG power capacity. It was used as an autotranformer connect to a diodes bridge to vary the voltage applied on the DCM_{IG} armature circuit and obtain a higher speed and PIG.

2.8. Voltage and Speed Control Loops

The Figure 7 and Figure 8 show the closed control loop used to control the SG speed and voltage via circuit boards.

Figure 9 (a) and Figure 9 (b) show the circuit boards with the respective working point or reference points defined during parameterization for the speed control loop, 65.2 and the voltage control loop, 65.5.

As [1], the master and slave behavior were verified by the generators, considering that SG is the master and IG is the slave. The frequency and voltage are determined by SG, while active power consumed at load is controlled by IG.

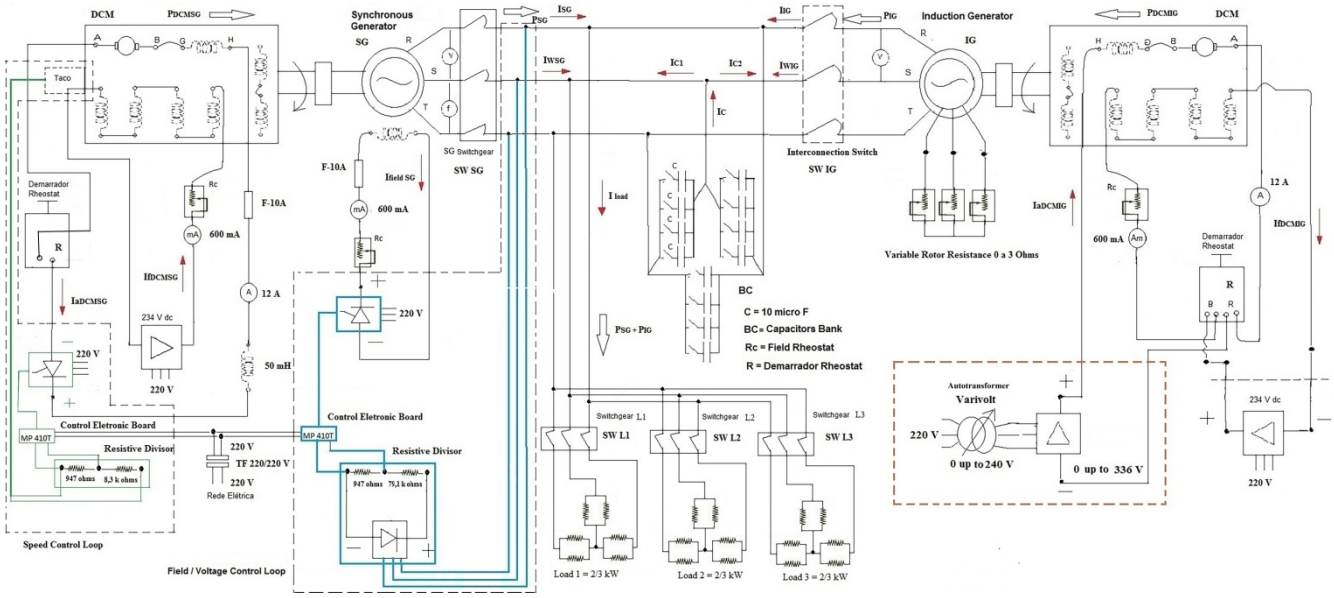


Figure 5 Synchronous generator in parallel with induction generator and its dc motor_ 234 Vdc steady source for DCM_{IG}

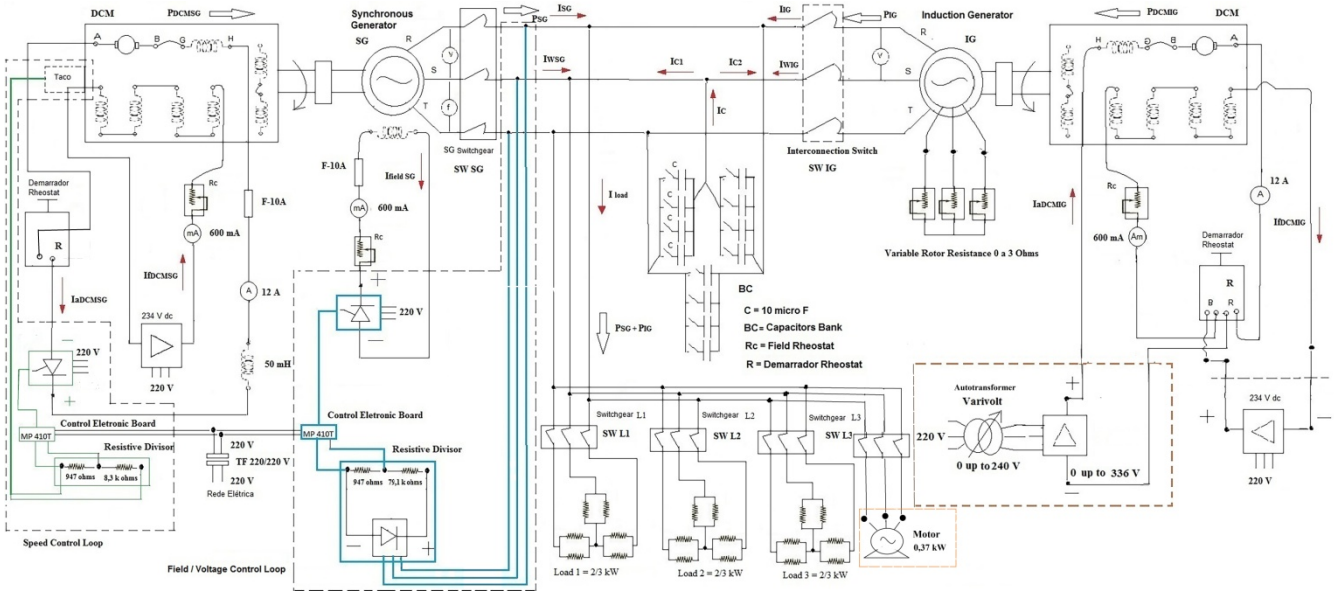


Figure 6: Synchronous generator in parallel with induction generator and its dc Motors _ Vdc variable source for DCM_{IG}

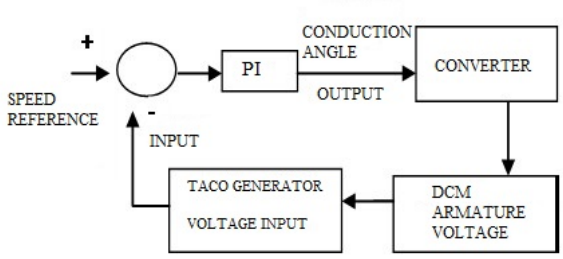


Figure 7. Speed Control Loop

Note: For both regulators, the proportional gain P was experimentally set at 0,010 and, in the same manner, the time constant was set at 0,04 s as in [1].

3. Experiment Data and Results

3.1. Experiment Data

The experimental data obtained in the laboratory are shown in Table 7 and Table 8. The Table 7 is referred to scheme in Figure 5 and Table 8 is referred to scheme in Figure 6. The unique difference is the presence of IM as additional load presented in Figure 6 scheme. Then, the Table 7 and Table 8 shows the data obtained from the scheme shown respectively in Figure 5 and Figure 6 which are similar to scheme shown in [1], except for the presence of an autotransformer in the DCM_{IG} field circuit for Figure 5 and Figure 6 and an IM as additional load in Figure 6.

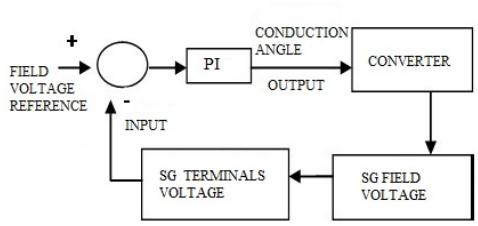
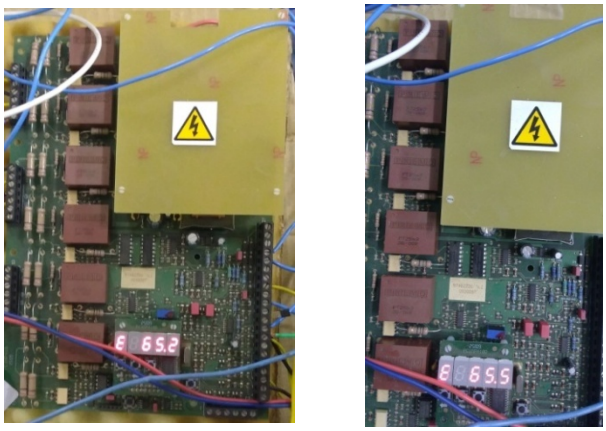


Figure 8. Field voltage control loop



(a) Speed Control Loop (b) Voltage Control Loop

Figure 9. Circuit boards MP410T used to implement the control loops

The autotransformer is responsible for applying a voltage range directly on the DCM_{IG} armature to obtain a larger speed range of the DCM_{IG} and to push more power to the IG, keeping the DCM_{IG} parameters into the rated values. Hence, the elevation of IG speed and power depend on just V_{aDCMIG}, it means this method is different from the scheme presented in [1] that depends on the DCM_{IG} field flux. Both V_{aDCMIG} and the DCM_{IG} field flux are adjusted manually. In summary, the [1] shows the data obtained from the scheme that has the 234Vdc steady source, Table 7 and Table 8 shows the data obtained from the schemes shown in Figure 5 and Figure 6, that has an autotransformer and a diodes bridge as substitute of the 234 Vdc steady source.

The data presented in Table 8 related to Figure 6 system performance results are compared to data obtained from Table 7 related to scheme in Figure 5 and also compared to data obtained from [1]. The scheme in Figure 5 and Table 7 shows the performance of the isolated electric system that considers the voltage application range on the DCM_{IG} armature by an autotransformer and an diodes bridge. The scheme in Figure 6 and Table 8 shows the performance of the isolated electric system that considers the same scheme presented in Figure 5 added by a induction motor, IM, as an additional load.

4. Results

The results and comparative analysis from experiments shown in Figure 5, Figure 6 and the results obtained in [1] are presented in the following.

The results show graphs and analysis of system configurations presented in Figure 5 and Figure 6, data obtained in [1] and analysis of contribution from each generator for each arrange of load, that, in this case, is ether just resistive banks or resistive banks added with an induction motor IM. It is noted that for all graphs obtained from [1] is related to a set system frequency of 55 Hz.

Table 7: Operational Scenarios in Controlled Mode and load being resistors bank

	SG and IG in parallel mode and no load. 40 µF / phase	SG and IG in parallel mode and load in. 40 µF /phase	SG and IG in parallel mode and load in. 40 µF /phase	SG and IG in parallel mode and load in. 40 µF /phase	SG and IG in parallel mode and load in . 40 µF /phase	SG and IG in parallel mode and load in. 40 µF /phase	SG and IG in parallel mode and load in. 40 µF /phase	SG alone and load in. 40 µF /phase	SG alone and no load. 30 µF /phase
Results	**Scenario 1B**	**Scenario 2B**	**Scenario 3B**	**Scenario 4B**	**Scenario 5B**	**Scenario 6B**	**Scenario 7B**	**Scenario 8B**	**Scenario 9B**
V_{SG} (V)	220.0	220.0	220.0	220.0	220.0	220.0	220.0	220.0	220.0
f_{SG} (Hz)	60	60	60	60	60	60	60,0	60,0	60,0
$I_{fieldSG}$ (mA)	190	320	300	300	320	350	370	340	100
V_{aDCMSG} (V)	272.9	275.0	276.2	276.8	277.5	278.0	280.7	278.5	270.0
I_{aDCMSG} (A)	2.0	2.5	3.8	4.7	6.2	7.4	9.5	10.6	2.0
I_{fDCMSG} (mA)	510.0	530.0	520.0	520.0	515.0	510.0	515.0	515.0	515.0
V_{IG} (V)	220.0	220.0	220.0	220.0	220.0	220.0	220.0	n.a.	n.a.
f_{IG} (Hz)	60	60	60	60	60	60	60	n.a.	n.a.
n_{IG} (rpm)	1793	1896	1878	1857	1848	1821	1794	n.a.	n.a.
V_{aDCMIG} (V)	291.3	323.6	316.7	308.3	301.3	293.0	284.1	n.a.	n.a.
I_{aDCMIG} (A)	0.5	8.0	7.0	6.0	4.7	3.5	2.0	n.a.	n.a.
I_{fDCMIG} (mA)	330.0	330.0	330.0	330.0	330.0	330.0	330.0	n.a.	n.a.
I_{wloadR} (A)	0.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	0.0
I_{IG} (A)	2.4	5.8	5.1	4.4	3.7	2.9	2.4	0.0	0.0
I_{SG} (A)	2.9	1.5	2.0	2.8	3.6	4.5	5.6	7.4	4.0
I_C (A)	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4	4.0
I_{WSG} (A)	**0.0**	***1.05***	***1.48***	***1.97***	***2.47***	***3.32***	***4.65***	***5.03***	**0.0**
I_{C1} (A)	***2.94***	***1.13***	***1.60***	***2.13***	***2.66***	***3.03***	***3.13***	***5.40***	**4.0**
I_{C2} (A)	***2.45***	***4.26***	***3.80***	***3.27***	***2.74***	***2.37***	***2.27***	***0.0***	**0.0**
I_{WIG} (A)	***0.0***	***3.95***	***3.51***	***3.03***	***2.53***	***1.68***	***0.35***	***0.0***	**0.0**
PSG (W)	***0.0***	***400.5***	***565.5***	***750.0***	***939.8***	***1266.6***	***1770.1***	***1915.8***	**0.0**
PIG (W)	***0.0***	***1504.8***	***1335.8***	***1155.3***	***965.5***	***638.67***	***135.17***	***0.0***	**0.0**
P_{DCMSG} (W)	***545.8***	***687.5***	***1049.6***	***1301.0***	***1665.0***	***2057.2***	***2666.7***	***2962.7***	**543.4**
P_{DCMIG} (W)	***145.7***	***2588.8***	***2216.9***	***1849.8***	***1416.1***	***1025.5***	***568.2***	***0***	**0.0**
η_{GROUP} (%)	***0.0***	***58.15***	***58.33***	***60.47***	***61.84***	***61.80***	***58.89***	***64.66***	**0.0**
η_{SG} (%)	***0.0***	***58.25***	***53.88***	***57.65***	***56.44***	***61.57***	***66.38***	***64.66***	**0.0**

η_{IG} (%)	***0.0***	***58.12***	***60.43***	***62.45***	***68.18***	***62.28***	***23.79***	***n.a.***	**n.a.**

Obs.: 1- Italic and bold text: calculated values by equations 23 to 33
2- Frequency was chosen based on the average between the SG and IG rated frequency
3- n.a: not applicable

Table 8: Operational Scenarios in Controlled Mode and load being resistors bank and induction motor

	SG and IG in parallel mode and no load. 40 µF / phase	SG and IG in parallel mode and load in. 40 µF /phase	SG and IG in parallel mode and load in. 40 µF /phase	SG and IG in parallel mode and load in. 40 µF /phase	SG and IG in parallel mode and load in. 40 µF /phase	SG and IG in parallel mode and load in. 40 µF /phase	SG and IG in parallel mode and load in. 40 µF /phase	SG alone and load in. 40 µF /phase	SG alone and no load. 30 µF /phase
Results	**Scenario 1C**	**Scenario 2C**	**Scenario 3C**	**Scenario 4C**	**Scenario 5C**	**Scenario 6C**	**Scenario 7C**	**Scenario 8C**	**Scenario 9C**
V_{SG} (V)	220.0	220.0	220.0	220.0	220.0	220.0	220.0	220	220.0
f_{SG} (Hz)	60	60	60	60	60	60	60	60	60
$I_{fieldSG}$ (mA)	170	420	410	400	400	430	475	390	100
V_{aDCMSG} (V)	278.9	277.0	277.5	280.5	281.0	281.1	282.9	279.2	271.7
I_{aDCMSG} (A)	2.0	2.9	4.0	5.0	6,3	7,6	9,6	10,5	2,0
I_{fDCMSG} (mA)	530.0	530.0	530.0	530.0	530.0	520.0	520.0	520.0	515.0
V_{IG} (V)	220.0	220.0	220.0	220.0	220.0	220.0	220.0	0.0	n.a.
f_{IG} (Hz)	60	60	60	60	60	60	60	0.0	n.a.
n_{IG} (rpm)	1803	1893	1868	1857	1843	1821	1793	0.0	n.a.
V_{aDCMIG} (V)	288.9	330.0	318.0	310.0	302.3	295.2	285.9	0.0	n.a.
I_{aDCMIG} (A)	1.0	8.0	7.0	6.0	4.7	3.5	2.0	0.0	n.a.
I_{fDCMIG} (mA)	330.0	330.0	330.0	330.0	330.0	330.0	330.0	0.0	n.a.
Iw_{loadR} (A)	0.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	0.0
I_{wMT} (A)	0.0	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.0
I_{dwMT} (A)	0.0	1.23	1.23	1.23	1.23	1.23	1.23	1.23	0.0
I_{IG} (A)	2,3	6,0	5,0	4,4	3,6	2,9	2,3	0.0	0.0
I_{SG} (A)	2.8	0.7	1.6	2.3	3.3	4.1	5.3	6.7	4.0
I_{C} (A)	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4	4.0
I_{WSG} (A)	***0.0***	***0.65***	***1.29***	***1.99***	***3.04***	***3.74***	***4.95***	**5.23**	**0.0**
I_{C1} (A)	***2.94***	***0.28***	***1.03***	***1.17***	***1.29***	***1.66***	***1.88***	**4.2**	**4.0**
I_{C2} (A)	***2.46***	***3.91***	***3.14***	***3.00***	***2.88***	***2.51***	***2.29***	**0.0**	**0.0**
I_{WIG} (A)	***0.0***	***4.55***	***3.91***	***3.22***	***2.16***	***1.45***	***0.25***	**0.0**	**0.0**
PSG (W)	***0.0***	***247.97***	***490.34***	***753.80***	***1156.9***	***1428.7***	***1887.8***	**1991.7**	**0.0**
PIG (W)	***0.0***	***1733.5***	***1491.1***	***1227.7***	***824.52***	***552.74***	***93.68***	**0.0**	**0.0**
P_{DCMSG} (W)	***557.8***	***803.30***	***1110.0***	***1402.5***	***1770.3***	***2136.4***	***2715.8***	**2931.6**	**624.5**
P_{DCMIG} (W)	***288.9***	***2640.0***	***2226.0***	***1860.0***	***1420.8***	***1033.2***	***571.8***	**0.0**	**0.0**
η_{GROUP} (%)	***0.0***	***57.55***	***59.40***	***60.73***	***62.09***	***62.52***	***60.27***	**67.59**	**0.0**
η_{SG} (%)	***0.0***	***30.87***	***44.17***	***53.75***	***65.35***	***66.88***	***69.51***	**67.59**	**0.0**
η_{IG} (%)	***0.0***	***65.66***	***66.99***	***66.00***	***58.03***	***53.50***	***16.38***	**0.0**	**n.a.**

Obs.: 1- Italic and bold text: calculated values by equations 23 to 33
2- Frequency was chosen based on the average between the SG and IG rated frequency
3- n.a: not applicable

For all experiments shown in this paper, the system frequency is 60 Hz. Then, the synchronous speed for the experiments in [1] is 1650 rpm (6) and 1800 rpm for the experiments shown in this paper.

To highlight the contributions from autotransformer and diodes bridge in the experiments in relation to the experiments without these devices, it will be presented some analysis based on graphs and experiment results.

Figure 10 shows that the PIG was limited to 1164 W [1] in scenario 2 [1] because the over current of I_{aDCMIG} 9.8 A [1] is a value greater than the rated I_{aDCMIG} of 7.72 A as Table 1. To overcome this barrier, an autotransformer and a diodes bridge were installed to widen the voltage range applied over the DCM armature, as shown in Figure 5 and Figure 6. The armature voltage range with the autotransformer and diodes bridge can vary from 0 to 336 Vdc. This range is greater than the previous V_{aDCMIG} of 234 Vdc. As the DCM_{IG} speed, n_{IG}, is directly proportional to V_{aDCMIG}, as shown in (2), the n_{IG} is elevated proportionally with the V_{aDCMIG} as shown in scenarios 2 to 8, in Table 7 and Table 8. Then, in this latter method, I_{aDCMIG} variation depends on the load conjugate variation only; as long as I_{aDCMIG} decreases, I_{aDCMSG} rises because of the load distribution to the generators, whereas the DCM_{IG} flux, ϕ, remains constant as (3).

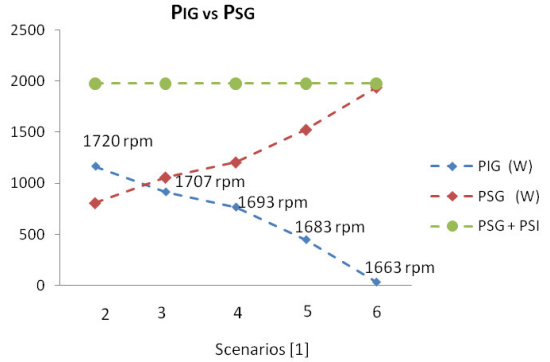


Figure 10. PIG vs PSG with DCM_{IG} field flux variation and V_{aDCM} kept constant

Figure 11 and Figure 12 shows that the power range supplied from IG, PIG, through the autotransformer method, is greater than PIG of 234 Vdc through the steady source method. Consequently, the power complement from SG is

lower than the PSG from 234 Vdc steady source method. Both methods feed the entire resistive load in Figure 10 and Figure 11, and resistive and motor load as shown in Figure 12.

In Figure 13, Figure 14 and Figure 15 shows the comparative results of IG and SG primary machine output power, P_{DCMIG} vs P_{DCMSG}. These performances are similar to PIG and PSG performances shown in Figure 10, Figure 11 and Figure 12. The P_{DCMIG} in Figure 13 elevates because the flux decreases and V_{aDCMIG} is kept constant [1] The P_{DCMIG} in Figure 14 and Figure 15 elevates because of the V_{aDCMIG} increases, which is manually adjusted by the autotransformer shown in Figure 5 and Figure 6.

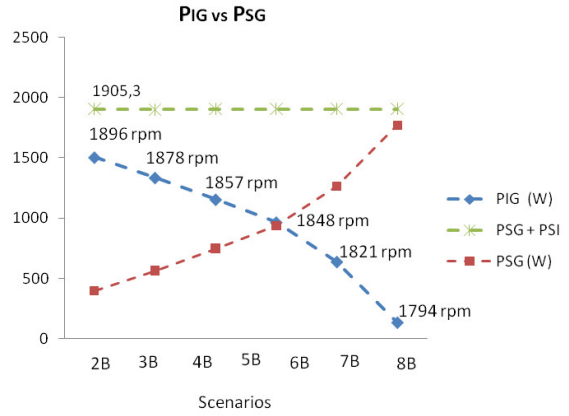


Figure 11. PIG vs PSG with V_{aDCMIG} variation and field flux kept constant (Only resistive load)

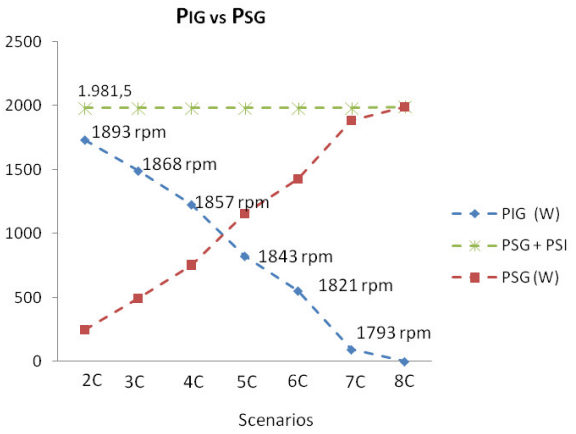


Figure 12. PIG vs PSG with V_{aDCMIG} variation and field flux kept constant (Resistive load plus IM)

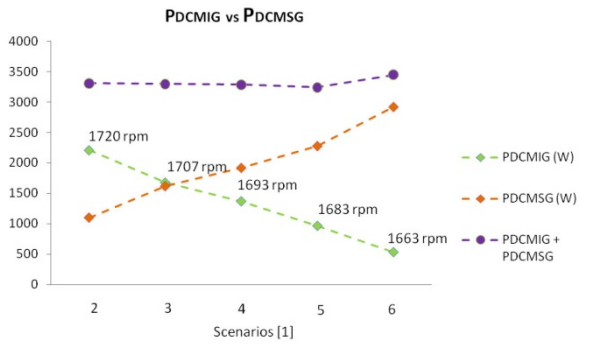


Figure 13. P_{DCMIG} vs P_{DCMSG} with DCM_{IG} field flux variation and V_{aDCMIG} kept constant

The efficiency results shown in Figure 16, Figure 17 and Figure 18 present each subgroup efficiency for each scenario and load conditions. Each subgroup efficiency is resulted from relation between a generator and a DCM as shown in (32) and (33). In summary, the target scenarios are: 2 to 7 (without autotransformer) as in [1], scenarios 2B to 8B (with autotransformer and just resistive load) and scenarios 2C to 8C (with autotransformer, resistive load and induction motor).

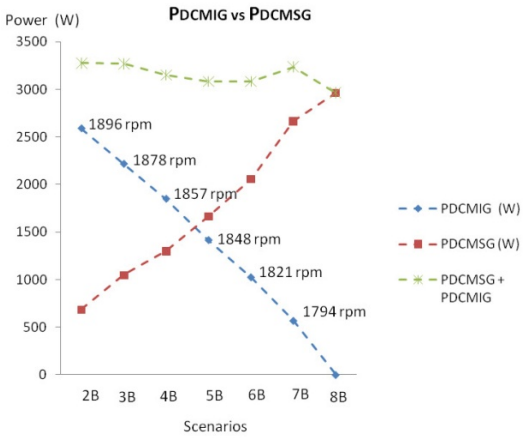


Figure 14. P_{DCMIG} vs P_{DCMSG} with V_{aDCMIG} variation and field flux kept constant (Only resistive load)

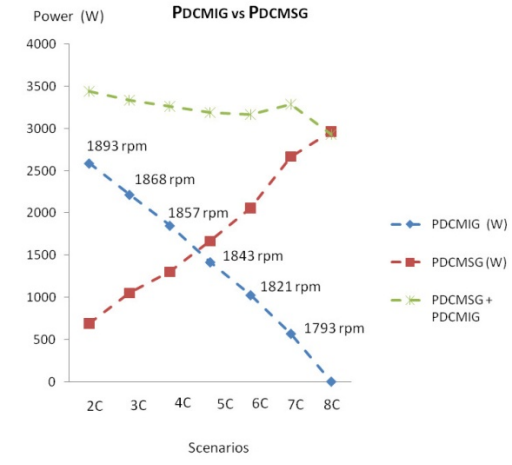


Figure 15. P_{DCMIG} vs P_{DCMSG} with V_{aDCMIG} variation and field flux kept constant (Resistive load plus IM)

The IG subgroup efficiencies shown in Figure 17 and Figure 18 are bigger than IG subgroup efficiencies from similar scenarios in Figure 16. Moreover, the elevation of power from IG due to the increase of DCM_{IG} speed, n_{IG}, is more representative of actual offshore conditions. The turbines can assume whatever speed required from generators in offshore platforms.

The reduction of losses contributed to increased IG subgroup efficiency as shown in Figure 17 and Figure 18. Fig. 16 shows the machine's efficiencies when the I_{fDCMIG} decreases to elevate the speed. The efficiency improves when the V_{aDCMIG} elevation method is applied as shown in Figure 17 and Figure 18.

The Figure 17 and Figure 18 are higher than the efficiency presented in Figure 16 considering that the Figure 16 indicates a DCM_{IG} speed range is more restricted as already clarified in [1].

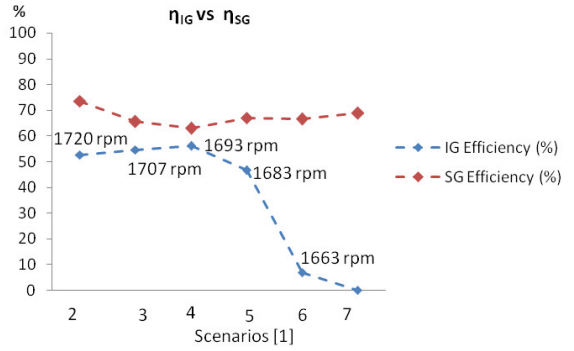


Figure 16. η_{IG} vs η_{SG} with DCM_{IG} field flux variation and V_{aDCMIG} kept constant

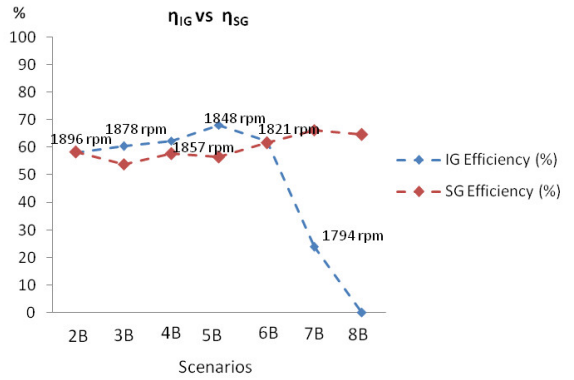


Figure 17. η_{IG} vs η_{SG} with V_{aDCMIG} variation and field flux kept constant (Only resistive load)

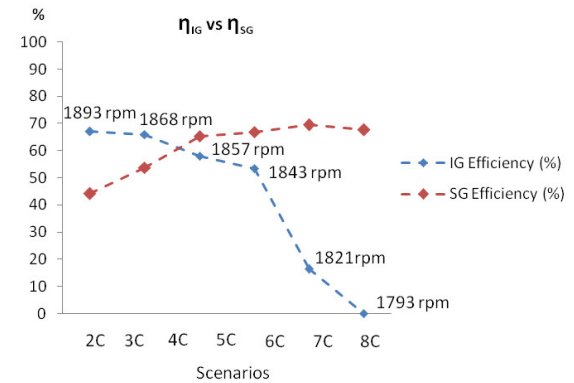


Figure 18. η_{IG} vs η_{SG} with V_{aDCMIG} variation and field flux kept constant (Resistive load plus IM)

5. Conclusion

The cited electric system shows voltage and frequency regulation for each scenario transition as demonstrated by results and respective analysis.

As it happened in [1], it was realized the master and slave behaviors to respectively SG and IG. SG controls the system voltage and frequency and IG follows the voltage and frequency defined by SG. The IG establishes the active power supplied to the system and the SG complements the rest of active power and reactive power required. The PIG depends on the IG shaft speed imposed by DCM_{IG}. The PSG and the system voltage and frequency are controlled by voltage and speed control loops presented in Figure 5 and Figure 6.

It was realized a greater PIG range in the scenarios in Table 7 and in Table 8, than the scenarios from [1] due to V_{aDCMIG} elevation method. The PSG complemented the PIG manually set to attend the load.

The V_{aDCMIG} elevation method resulted in increase of P_{DCMIG} and PIG and higher IG subgroup efficiency than the I_{fDCMIG} reduction method[1]. The increase of P_{DCMIG} and PIG were results of IG speed increase, $n_{IG,}$ that was obtained by elevation of V_{aDCMIG} as shown in Figure 11, Figure 12 and Figure 14. The use of V_{aDCMIG} range instead of Vdc steady source enables to reach the I_{aDCMIG} limit of 8,0 A and 1893 rpm shaft speed, it is 5,16% higher than synchronous speed. It is higher than Vdc steady source method limit that reached just 4,24% higher than synchronous speed, Figure 16 and [1].

The efficiencies from the V_{aDCMIG} elevation method were greater than efficiencies from the scheme with Vdc steady source [1], as shown in Figure 17 and Figure 18. V_{aDCMIG} elevation method does not have the additional current losses of Vdc steady source method. When the field flux ϕ, I_{fDCMIG}, is decreased, I_{aDCMIG} is increased by (3) and IG speed, $n_{IG,}$ is increased by (2). The main losses for the Vdc steady source method are related to higher I_{aDCMIG} elevation, $P(w)=Ra* I_{aDCMIG}^2$. Then, by the Vdc steady source method, besides I_{aDCMIG} be elevated due to the rise of PIG, I_{aDCMIG} is also elevated due to reduction of I_{fDCMIG}. Higher I_{aDCMIG} result in higher losses and lower efficiencies.

As shown in power performance graphs related to the V_{aDCMIG} elevation method, Figure 11, Figure 12, Figure 14 and Figure 15, the PIG and P_{DCMIG} are more representative of actual operation conditions in offshore platforms.

In continuation of this content, a study of load and generation transients has been done and submitted to publication in COBEP, Brazilian Power Electronic Conference that took place in November, 2017.

In the next stage of this work, it will be studied energy regeneration and ballast load as a additional way to the system regulation.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank the UNIFEI and PETROBRAS for the general support along the study.

References

[1] V. Z. Silva, A. J. J. Rezek, and R. L. Corrêa, "Analysis of synchronous and induction generators in parallel operation mode in an isolated electric system", Brazil, 2017, https://doi.org/10.1109/PEDG.2017.7972459

[2] S.P. Singh., B. Singh, M.P. Jain, "Comparative study on the performance of a commercially designed induction generator with induction motors operating as self-excited induction generators", IEE Proc. C, vol. 140, no.5, pp: 374 - 380, 1993, https://doi.org/ 10.1049/ip-c.1993.0055

[3] M. H. Haque, "Self-excited single-phase and three-phase induction generators in remote areas", IEEE conference, pages 38-42, 2008, https://doi.org/ 10.1109/ICECE.2008.4769169

[4] E. Touti, R. Pusca, J. F. Brudny, A. Châari, "Asynchronous generator model for autonomous operating mode", IEEE conference, pages 1-3, 2013, https://doi.org/ 10.1109/ECAI.2013.6636208

[5] D. M. Macedo, "The use of pumps operating as turbines and induction generators in electric power generation" (in Portuguese), Master in Science Dissertation, Federal University of Itajubá(UNIFEI), Itajubá, 2004.

[6] J. M. Chapallaz, P. Eichenberger, and G. Fischer, *Manual on Pumps used as Turbines.* MHPG Series, v. 11, Friedr. Vieweg&SohnVerlagsgesellschaftmb, Germany, 1992.

[7] J. R. Cogo, J.C. Oliveira, and J. P. G. Abreu, "Tests in induction machines", editor EFEI, Itajubá-MG, Brazil, 1983.

[8] A. P. Rizzi, " Electrical Measurements: power, energy, power factor, demand", in Portuguese, Livros Técnicos e Científicos Editora S.A., Rio de Janeiro-RJ, 1980.

Appendix 1_ Induction Motor Parameters

The follow calculus were developed on base in [7].

1. No Load and Locked Rotor Tests

The Figure 1 shows the circuit implemented in laboratory to no load test and locked rotor test.

The temperature during the test= 26ºC.

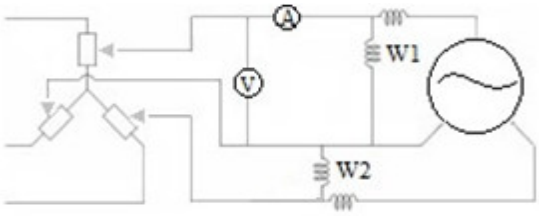
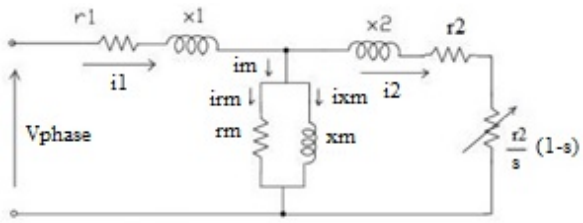


Figure 1: No load test and locked rotor test circuit

Figure 2: Induction Motor Equivalent Circuit



The Figure 2 shows the induction motor IM equivalent circuit.

5.1. No Load Test

In this section, it will be presented the no load test and locked rotor test to calculus the IM parameters such as impedances, currents, losses and efficiency

The Table 1 presented the no load test measurements.

Table 1: No load Test

Vapply (V)	W1 (W)	W2 (W)	V (V)	A (A)	W1+W2(W)
240	-120.0	+220.0	240.2	1.50	100.0
220	-100.0	+170.0	220.8	1.35	70.0
200	-70.0	+130.0	200.3	1.15	60.0
180	-60.0	+110.0	180.3	1.00	50.0
160	-40.0	+80.0	160.0	0.90	40.0
140	-30.0	+60.0	140.0	0.75	30.0
120	-20.0	+45.0	120.0	0.65	25.0
100	-15.0	+30.0	100.1	0.55	15.0
80	-10.0	+20.0	80.0	0.50	10.0
60	-4.0	+12.5	60.0	0.25	8.5
40	0.0	+5.0	40.0	0.15	5.0
20	0.0	+2.0	20.1	0.05	2.0
0	0.0	0.0	1.99	0.0	0.0

Figure 3 shows the curve of voltage applied and power measured by wattmeter (w1) and (w2). Note that the cross of curve and vertical axe result in estimated attrition and ventilation losses (*Pav*). In this case *Pav* is 1 W as demonstrated.

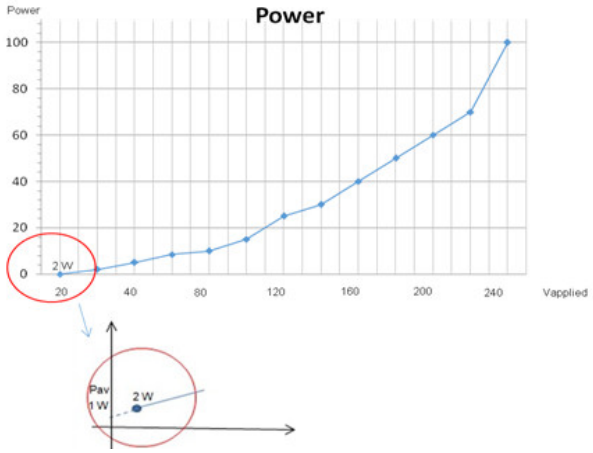


Figure 3: Results of no load test: W1+W2 vs Voltage applied

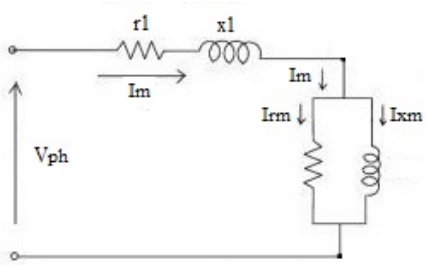


Figure 4: Equivalent circuit - No load test

- Magnetization Current

Im = Magnetization Current

Im= 1.35 A as shown in Table 1.

5.2. Locked Rotor Test

The Table 2 shows the locked rotor test measurements.

Table 2
Locked Rotor Test

W1 (W)	W2 (W)	V (V)	In (A)	W1+W2 (W)
20	65	39.34	1.80	85

- Rated Current

In = Rated Current

$$in = {}^{P}/_{\sqrt{3} * Vpp * cos\emptyset * \eta} \quad (1)$$

$$in = {}^{370}/_{\sqrt{3} * 220 * 0.71 * 0.76}$$

$$in = 1.80\ A \quad (2)$$

The Figure 3 shows the induction motor IM equivalent circuit during locked rotor test

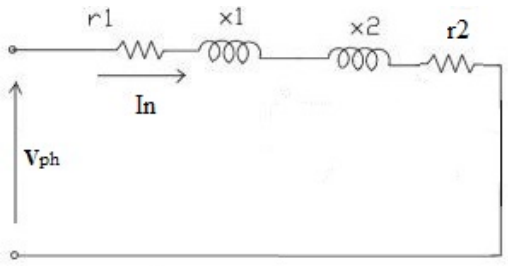


Figure 5: Equivalent circuit _ Locked rotor test

5.3. Magnetization Branch Parameters

Rm= Motor Average Measured Resistance
ry= Phase resistance in star connection
rd= Phase resistance in delta connection

Table 3:
Average Resistance

$R_{AA'}$	$R_{BB'}$	$R_{CC'}$
10.1	10.3	10.16

$$R_m = {(10.1 + 10.3 + 10.16)}/{3} = 10.16\ \Omega \quad (3)$$

- r1 Calculus

Follow the calculus memory of stator winding resistor r1.

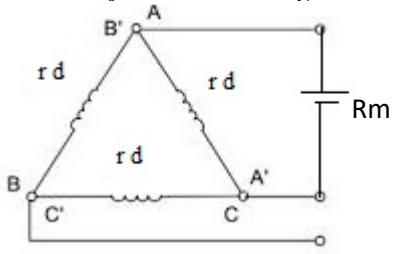


Figure 6: IM winding connection

$$R_m = {rd * 2rd}/{3rd} \quad (4)$$

$$R_m = {2}/{3}\,rd \quad (5)$$

$$rd = {3}/{2}\,rm \quad (6)$$

$$ry = {1}/{3}\,rd \quad (7)$$

(5) in (6)

$$ry = {Rm}/{2} = {10.16}/{2}$$

$$r1 = 5.08\ \Omega \quad (8)$$

- r2 Calculus

From Table XI, $W1 + W2 = 85\ W$

$$W1 + W2 = 3 * (\mathrm{r1} + \mathrm{r2}) * \mathrm{in}^2 \quad (9)$$

$${85}/{(3 * 1.8^2)} = 5.08 + r2$$

$$r2 = 3.66\ \Omega \quad (10)$$

- Reactive Power

$$Q = \sqrt{3} * Vpp * IdwMT * sen\emptyset = \sqrt{3} * (W2 - W1) \quad (11)$$

From Table XI, W2-W1=45

$$Q = \sqrt{3} * (65 - 20) = 77.94\ Var \quad (12)$$

$$Q = 3 * (x1 + x2) * In^2 \quad (13)$$

(13)and (2) in (14)

$$77{,}94 = 3 * (x1 + x2) * 1.8^2 \quad (14)$$

$$(x1 + x2) = 8.01\ \Omega \quad (15)$$

$$\begin{cases} x1 \rightarrow x2 \\ r1 \rightarrow r2 \end{cases} \quad [8] \quad (16)$$

(8) and (10) in (16)

$$\begin{cases} x1 \rightarrow x2 \\ 5{,}08 \rightarrow 3.66 \end{cases}$$

$$x1 = {r1 * x2}/{r2} \quad (17)$$

$$x1 = {5{,}08 * x2}/{3.66} \quad (18)$$

From (15) and (18), result in

$$5{,}08 * x2 + x2 * 3.66 = 29.32 \quad (19)$$

$$x2 * 8{,}74 = 29.32 \quad (20)$$

$$x2 = 3.35\ \Omega \quad (21)$$

(21) in (18), result in

$$x1 = {5{,}08 * 3.35}/{3.66} \quad (22)$$

$$x1 = 4.65\ \Omega \quad (23)$$

2. Power and Losses Calculus

The Pav was defined in Figure 3 and consist in 1 W and Pnl (no load losses) is no load test for 220 V, it resulted in 70 W (W1 +W2).

- Stator Joules Losses

$$\text{Pjstator26º= Stator Joule Losses} = 3 * r1 * im^2 \quad (24)$$

$$\text{Pjstator26º} = 3 * 5.08 * 1.35^2 \quad (25)$$

$$\text{Pjstator26º} = 27.77\ \text{W} \quad (26)$$

- Phf_ histerese and foucault losses

$$P_{hf} = (W1 + W2) - 3 * r1 * im^2 - Pav \quad (27)$$

$$P_{hf} = 70 - 3 * 5.08 * 1.35^2 - 1 \quad (28)$$

$$P_{hf} = 41.23\ W \quad (29)$$

5.4. No Load Reactive Power

$$Q_o = \sqrt{3} * (W2 - W1) \quad (30)$$

From Table 1, $W2 - W1 = 270W$

$$Q_o = \sqrt{3} * 270\ \text{W}$$

$$Q_o = 467.65\ \text{Var} \quad (31)$$

5.5. Magnetization Branch

- No load Reactive Power in Magnetization Branch

$$Q_m = Qo - 3 * X1 * io^{\,2} \quad (32)$$

$$Q_m = 467.65 - 3 * 4.66 * 1.35^{\,2} \quad (33)$$

$$Q_m = 442.23\ Var \quad (34)$$

- Magnetization Branch Power Factor Angle

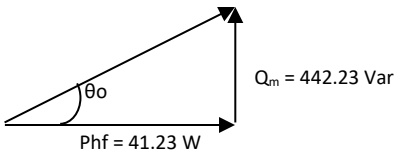


Figure 7: Magnetization branch power factor

$$\theta o = arctg\ {}^{442.23}/_{41.23} \quad (35)$$

$$\theta o = 84.67^{o} \quad (36)$$

- Magnetization Branch Currents

$$\text{Irm} = \text{Im} * \cos\theta o \quad (37)$$

$$\text{Irm} = 1.35 * \cos 84.67^{o} \quad (38)$$

$$\text{Irm} = 0.1254\ A \quad (39)$$

$$\text{Ixm} = \text{Im} * \text{sen}\theta o \quad (42)$$

$$\text{Ixm} = 1.35 * \text{sen}84.67^{o} \quad (40)$$

$$\text{Ixm} = 1.34\ \text{A} \quad (41)$$

- Magnetization Branch Resistance

$$\text{rm} = {}^{\text{Phf}}/_{(3 * \text{Irm}^2)} \quad (42)$$

$$\text{rm} = {}^{41.23}/_{(3 * 0.1254^2)} \quad (43)$$

$$\text{rm} = 873.97\ \Omega \quad (44)$$

- Magnetization Branch Reactance

$$\text{xm} = {}^{\text{Qm}}/_{(3 * \text{Ixm}^2)} \quad (45)$$

$$\text{xm} = {}^{442.23}/_{(3 * 1.34^2)} \quad (46)$$

$$\text{xm} = 82.10\ \Omega \quad (47)$$

3. Parameters of Induction Motor

Figure 8 shows all IM parameters calculated as resistors, reactance and currents as data obtained during experiment, without temperature correction because the motor operated no load.

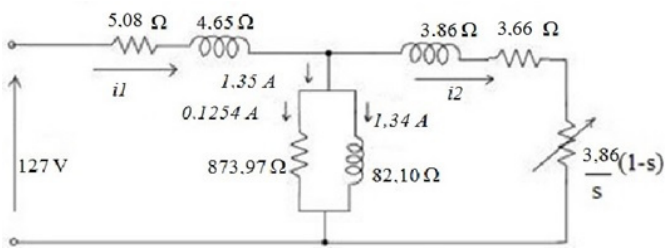


Figure 8: Parameters of Induction Motor _IM

4. Induction Motor Efficiency

It will be calculated the temperature correction factor, corrected resistences to 40ºC, corrected winding losses, Pwind40ºC, and total losses, Ptl, before induction motor efficiency calculus.

5.6. Temperature Correction Factor and Resistence Correction

- Temperature Correction Factor Kt 40ºC

$$\text{Kt } 40^{o} = {}^{234.5 + 40}/_{(234.5 + 26)} \quad (48)$$

$$\text{Kt } 40^{o} = 1.054 \text{ to r1 and r2} \quad (49)$$

- Resistence Correction

$$\text{r1 } 40^{o} = \text{r1 } 26^{o} * 1.054 \quad (50)$$

$$r1\ 40^{o} = 5.08 * 1.054 \quad (51)$$

$$r1\ 40^{o} = 5.35\ \Omega \quad (52)$$

$$r2\ 40^{o} = r2'26^{o} * 1.054 \quad (53)$$

$$r2\ 40^{o} = 3.66 * 1.054 \quad (54)$$

$$r2\ 40^{o} = 3.86\ \Omega \quad (55)$$

5.7. Corrected Winding Losses and Total Losses

- Winding Losses

The winding losses, Pwind, measured in locked rotor test, Table 2, need to be corrected to 40ºC. Follow the corrected winding losses.

$$Pwind40^{o} = kt\ 40^{o} * Pwind26^{o}\ (\text{locked rotor test}) \quad (56)$$

$$Pwind40^{o} = 1.054 * 85W \quad (57)$$

$$Pwind40^{o} = 89.59\ W \quad (58)$$

- Total Losses

The total losses, Ptl, calculus:

$$Ptl = Phf + Pav + Pwind40^{o} \quad (59)$$

$$Ptl = 41.23W + 1W + 89.59W \quad (60)$$

$$Ptl = 131.82\ W \quad (61)$$

5.8. Rated Motor Efficiency Estimate

Pout = Motor output power (Motor dataplate_Table 6)
Pin=Motor Input power

$$\eta = {Pout}/{Pin} \quad (62)$$

$$Pin = Pout + losses \quad (63)$$

$$Pin = 370\ W + 131.82\ W \quad (64)$$

$$Pin = 501.82\ W \quad (65)$$

$$\eta = {370}/{501.82} \quad (66)$$

$$\eta = 73.73\ \% \quad (67)$$

5. Induction Motor Currents

Follow the calculus of IM reactive and active currents (IdwMT and IwMT). This values are shown in Table 8 for all scenarios tested.

- Reactive Current

$$Q = \sqrt{3} * 220 * IdwMT = \sqrt{3} * (W2 - W1) \quad (68)$$

$$\sqrt{3} * 220 * IdwMT = 467.65 \quad (69)$$

$$IdwMT = 1.23\ A\ ** \quad (70)$$

** The difference between IdwMT and Ixm is perfectly acceptable. The difference close to 0,1 A between expected and found values is due to small measurements errors and instruments precision and accuracy errors. It does not impact the presented modeling. Therefore, the results and modeling continue valid.

- Active Current

$$P = (W1 + W2) = 70.0\ W \quad (71)$$

$$Q = \sqrt{3}(W2 - W1) = 467.6\ W \quad (72)$$

$$S = \sqrt{70.0^2 + 467.6^2} = 472.86\ VA \quad (73)$$

$$\cos\emptyset = {P}/{S} = {70}/{472.86} \quad (74)$$

$$\cos\emptyset = 0.148 \quad (75)$$

$$IwMT = Io * \cos\emptyset = 1.35 * 0.148 \quad (76)$$

$$IwMT = 0.20\ A \quad (77)$$

In Figure 9 are shown resistor banks current and no load IM current.

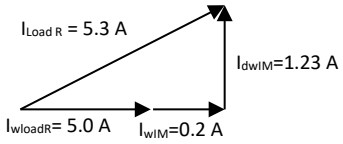


Figure 9: Resistors bank and IM currents in scenarios C

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 205-212 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Design of an Automatic Forward and Back Collision Avoidance System for Automobiles

Tasneem Sanjana*, Ferdus Wahid, Mehrab Masayeed Habib, Ahmed Amin Rumel

Faculty of Electrical and Electronic Engineering, American International University-Bangladesh (AIUB), 1213, Bangladesh

ARTICLE INFO

Article history:
Received: 14 November, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
Forward collision avoidance system
Rear collision avoidance system
ATMEGA32
Sharp Distance Sensor
Motor driver
Actuator

ABSTRACT

This paper is the extended reflection of work originally presented in conference of Electrical, Computer and Communication Engineering (ECCE)-CUET 2017, entitled "Automated Anti Collision System for Automobiles". Automated collision avoidance system is a trending technology of science in automobile engineering. The aim of this paper is to design a system which will prevent collision from the front as well as the back for automobiles. This paper gives an overview of secure and smooth journey of car (vehicles) as well as the certainty of human life. This system is controlled by microcontroller ATMEGA32. Two Sharp distance sensors are used to detect object within the danger range where one is for front detection and other is for back detection. A crystal oscillator is used to produce the oscillation and generates the clock pulse of the microcontroller. An LCD and a GLCD are used to give information about the safe distance for front and rear respectively, and a buzzer is used as alarm. An actuator is used as automatic brake and inside the actuator there is a motor driver that runs the actuator. For coding "microC PRO for PIC" is used and "Proteus Design Suite Version 8 Software" is used for simulation.

1. Introduction

All of the greatest achievement of the science, automobile is the most probably the one, which significantly changed human life. A collision avoidance system is a current emerging technology in the field of automobile. It is also known as pre-crash system, forward collision warning system or collision mitigating or car anti-collision system. An automated car anti-collision system is an automobile safety system design to reduce the accidents as road traffic accidents are the largest cause of injury-related deaths worldwide .When driver himself is not concentrating on driving or any other parameters, that time it may cause damage to vehicle (car) as well as a life, that time automated ant-car collision device can play an effective role regarding the danger ahead. An automated anti-collision system device is placed within a car to warn its driver of any dangers that may lie ahead on the road. Once the detection is done this system either provide a warning to the driver when there is an imminent collision or take action autonomously without any driver command (by braking or steering both). This system runs by microcontroller that detects obstacle with speed sensor and stops the vehicle by giving instruction to the actuator. [1].

*Tasneem Sanjana, American International University, Bangladesh
Email: tsanjana909@gmail.com

2. Literature Review

In 2003, Honda introduced the first pre-crash system with autonomous braking. In 2003 Toyota added an "automatic partial pre-crash braking system" to the Celsior36. In 2006 Audi introduced "Braking guard" radar-assisted forward collision warning on AudiQ7. But these could not put a satisfactory performance. These all were not fully automated. In 2009, the U.S. National Highway Traffic Safety Administration (NHTSA) had begun studying to make frontal collision warning systems and lane departure warning systems. In 2012 a research by the Insurance Institute for Highway Safety examined how particular features of crash-avoidance systems affected the number of claims under various forms of insurance coverage. This indicates that two crash-avoidance features provide the biggest benefits; these are (a) autonomous braking that would brake on its own when the driver would not alert to avoid a forward collision and (b) adaptive headlights that would shift the headlights in the direction to the driver steers. At the circa 2012 stage of development, it has found from this research that lane departure systems is not helpful and perhaps harmful. Periodically anti-car collision features are rapidly making their way into the new vehicle fleet. Many car companies like BMW, Audi, Mercedes- Benz , Ford, Toyota, Volvo developed vehicle collision avoidance system from 2003-2009, but not necessarily they were all time effective.[1] Few years back,

some research were carried out on car anti-collision system device using ad hoc wireless network, V2V communication, GPS and radar implementation but all this effort were informatory in type which would give signal to the driver or produced some buzzes or sounds only but finally action would be taken by the driver own self in which there were quite chances of the collision. [2] Since the emergence of mechanical vehicles, road safety has been a major concern, the automated anti-car collision system has gained increasing attention in the last three years. As a result, various automated anti car collision system have originated for assuring safety than the existing system. Research of the Queensland department of Transport and Main Roads has revealed that installation of anti-collision technology could reduce 20 to 40 percent in terms of number and severity of fatal crashes and 30-50 percent of all injuries. Thus it is the main aim to ensure the safety of human life during the collision, the U.S National High Way Traffic Safety Administration has decided to make the pre-crash system mandatory for vehicles. In similar way all new vehicles in Europe will have advanced emergency automated braking system installed by November,2015. A group of automakers and suppliers companies including General Motor(GM), Ford Motor, Nissan Motor are doing research of this automated vehicles collision avoidance system at the University of Michigan. General Motor company has a plan to install this automated vehicles avoidance system in their cars by next two years. Toyota has already given an official announcement that they are going to give this automated vehicles avoidance system in their Toyota and Lexus models by 2017. On the other hand, BMW has already introduced satisfactory and effective anti-car collision system on BMW 7series in 2012. Volvo company has launched this system in Volvo FH by February 2013. Also car company Renault introduced automated active braking system on Renault EspaceV, in October 2014 as well as Volkswagen Passat introduced too on Volkswagen Passat/B8 in 2014. [2] [3] Today, automated anti car collision system generally has camera sensor, distance sensor, actuator, buzzer to detect and warn the drivers of any danger lie ahead on the road, it could be a car, human, pedestrian or any stationary object just as tree or pole etc.

3. System Overview

The block diagram of proposed system is given in Fig.1.Advantage of proposed block diagram over existing technologies is, this system will provide forward collision protection as well as helps to avoid rear collision by displaying messages to the vehicles coming from the back regarding to slow down their vehicle. Here microcontroller does the prime work and it controls the entire task. Two distance sensors are connected to the microcontroller. A LCD, A GLCD and an alarming device are also connected to the microcontroller. A Motor Driver is connected with micro controller to operate an actuator, as actuator needs high current to turn on and microcontroller can't provide this amount of current so motor driver will provide high current to actuator to turn on when it's needed. Distance sensors will be placed in front and back side of a car. They will sense the distance of vehicle in front and rear of the car and send this information to microcontroller. Microcontroller will compare this distance with some preset distances. If the measured distance from the front distance sensor crosses the preset safe distance value, microcontroller will send command to LCD to display close distance and also to write value of that exact distance, At the same time microcontroller will send a pulsating signal to buzzer and buzzer will be buzzed according to that signal. If the measured distance by the front distance sensor crosses the preset value of close distance in microcontroller, microcontroller will tell LCD to show critical distance and the value of that corresponding distance measured by the sensor.

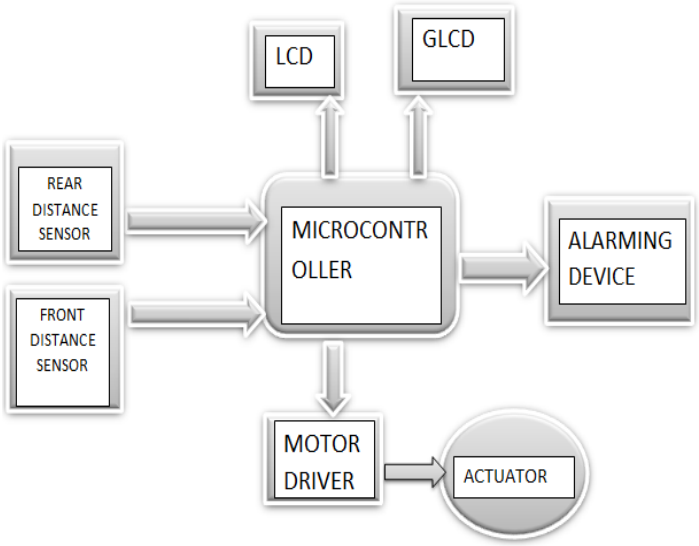


FIG.1. Block diagram of forward and back collision avoidance system for automobiles

At the same time a continuous signal will be send to buzzer and it will buzz continuously to notify the driver that vehicle is exceeding close distance and it's forwarding towards critical distance. Whenever vehicle crosses critical distance and entered into the distance at which collision is must if the car isn't slow down, that moment microcontroller will send signal to motor driver to turn on the actuator and it will stop the car automatically, at the same time LCD will "show actuator on" with distance information. Here, actuator itself works as a brake. For rear distance sensor it will sense the distance of the incoming vehicle and it will sent information to micro controller, if the distance exceed the safe distance value microcontroller will instruct GLCD to display slow down along with distance information, and if the incoming vehicle crosses critical distance microcontroller will instruct GLCD to display "Emergency! Stop" along with distance information to the driver of the incoming vehicle from the back. GLCD could be placed at the rear glass or just above the registration plate of the car containing this proposed safety technology.

4. Preparation to Proteus model

In this section components that were used for simulation and connection diagram has been explained.

4.1. ATMEGA32

ATMEGA32 is the high-performance, low-power Microchip 8-bit AVR RISC-based microcontroller combines 32KB ISP flash memory with read-while-write capabilities, 1KB EEPROM, 2KB SRAM, 54/69 general purpose I/O lines, 32 general purpose working registers, a JTAG interface for boundary-scan and on-chip debugging/programming, three flexible timer/counters with

compare modes, internal and external interrupts, serial programmable USART, a universal serial interface (USI) with start condition detector, an 8-channel 10-bit A/D converter, programmable watchdog timer with internal oscillator, SPI serial port, and five software selectable power saving modes. The device operates between 1.8-5.5 volts. By executing powerful instructions in a single clock cycle, the device achieves output through approaching 1 MIPS per MHz, balancing power consumption and processing speed. [4]

4.2. Sharp Distance Sensor: Model (GP2Y0A21):

A sensor is an electrical device which is used to measure physical properties and gives corresponding electrical output which is a transmitting impulse as for measurement or control for operation. In this model for measuring the distance of a vehicle Sharp Distance Sensor (GP2Y0A21) is used. Sharp Distance sensor follows triangulation measuring method for measuring distance of an object in front of it, It transmits an IR (Infrared Ray), if any obstacle is present in the path of that IR ray it will bounces back to the sensor and that's how distance is measured. Output of this sensor is analog and this output varies a range from 3.1V at 10cm to 0.4V at 80cm.

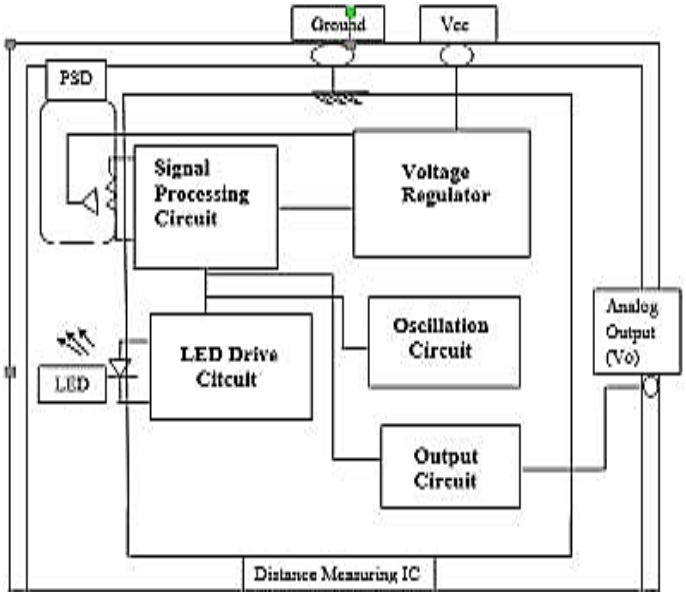


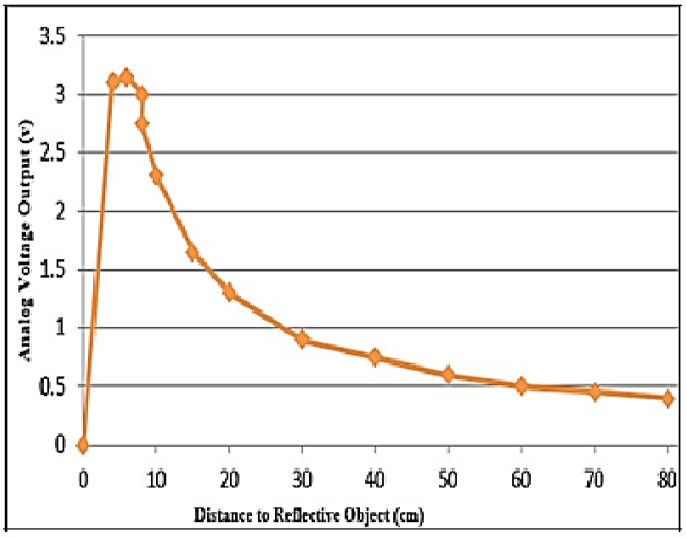
FIG.2.Internal Block Diagram of Sharp Distance Sensor [1]

FIG.2. indicates that this sensor contains IR LED armed with lens which discharges narrow light beam, after reflecting from the object, the beam will be directed through the second lens on a position-sensible photo detector (PSD) and the conductivity of PSD depends on the point where the beam falls, finally conductivity converts to voltage and voltage digitalizes by Analog to digital converter (ADC) [5]. Sharp Distance sensor utilize following equations to convert Distance into digitalized voltage with the help of ADC, and the relation between distance and the Voltage is

$$1 / (d + k) = a * ADC + b \qquad (1)$$

Where, distance is in centimeters.
k is corrective constant (fund using tial-and-error method)
ADC is digitalized value of voltage.
A is linear member (value is determined by the trend line equation)
b is free member (value is determined by the trend line equation).
So, distance d can be expressed from the formula [6]:

$$d = (1 / (a * ADC + B)) – k \qquad (2)$$



If a graph is plotted using equation (2) it will look like FIG.3. After analyzing this graph it can be concluded that the output voltage of the Sharp distance sensor is decreasing with the increased distance, which indicates an inversely proportional relation between distance and output voltage of sharp distance sensor.

4.3. Linear Actuator:

Linear actuator is a device which converts circular motion produced by a conventional electric motor into linear motion. Linear actuators are used in machine tools and industrial machinery, in computer peripherals such as disk drives and printers, in valves and dampers, and in many other places where linear motion is required. Hydraulic or pneumatic cylinders inherently produce linear motion. Many other mechanisms are used to generate linear motion from a rotating motor. In the majority of linear actuator designs, the elementary principle of operation is an inclined plane. The threads of a lead screw act as a continuous ramp that consents a small rotational force to be used over a long distance to complete the movement of a large load over a short distance. The system converts rotary motion (in the form of an electric motor) to linear motion. When an electrical source is applied to the motor, the thread shaft is driven and the nut rides up and down the shaft in the corresponding direction. This action delivers an extension and retraction capability for tasks requiring a linear displacement. Note that in this example the nut and red tube do not rotate, merely "ride" the threads on the spinning shaft up and down. [7]

4.4. Crystal Oscillator (16 MHz):

A crystal oscillator is an electronic oscillator circuit which uses in the mechanical resonance of a vibrating crystal of piezoelectric material to create an electrical signal with a very precise frequency and this frequency is usually used to keep track of time (as in quartz wristwatches), to provide a stable clock signal for digital integrated circuits, and to alleviate frequencies for radio transmitters and receivers, The most common type of piezoelectric resonator used is the quartz crystal, so oscillator circuits incorporating them became known as crystal oscillators but other piezoelectric materials including polycrystalline ceramics are used in similar circuits. [8]. A crystal is a solid in which the constituent atoms, molecules, or ions are packed in a repeatedly ordered, repeating pattern extending in all three spatial dimensions. Almost any object made of an elastic material could be used like a crystal, with appropriate transducers, since all objects have natural resonant frequencies of vibration. For example, steel is very elastic and has a high speed of sound. It was often used in mechanical filters before quartz. The resonant frequency depends on size, shape, elasticity, and the speed of sound in the material. High-frequency crystals are typically cut in the shape of a simple, rectangular plate. Low-frequency crystals, such as those used in digital watches, are typically cut in the shape of a tuning fork. For applications not needing very precise timing, a low-cost ceramic resonator is often used in place of a quartz crystal. When a crystal of quartz is suitably cut and mounted, it can be made to distort in an electric field by applying a voltage to an electrode near or on the crystal. This property is accepted as electrostriction or inverse piezoelectricity. When the field is removed, the quartz will generate an electric field as it returns to its previous shape, and this can generate a voltage. The result is that a quartz crystal behaves like a circuit composed of an inductor, capacitor and resistor, with a precise resonant frequency. Quartz has the further advantage that its elastic constants and its size change in such a way that the frequency dependence on temperature can be very low. The specific characteristics will depend on the mode of vibration and the angle at which the quartz is cut (relative to its crystallographic axes. Therefore, the resonant frequency of the plate, which depends on its size, will not change much, either. This means that a quartz clock, filter or oscillator will remain accurate. For critical applications the quartz oscillator is mounted in a temperature-controlled container, called a crystal oven, and can also be mounted on shock absorbers to prevent perturbation by external mechanical vibrations and oscillation.[8] In this design crystal oscillator is for to activate micro controller internal clock operating circuit.

4.5. Motor Driver (L293D):

L293D is a dual H-bridge motor driver integrated circuit (IC). Motor drivers act as current amplifiers since they take a low-current control signal and provide a higher-current signal. This higher current signal is used to drive the motors.L293D contains two inbuilt H-bridge driver circuits. In its common mode of operation, two DC motors can be driven simultaneously, both in forward and reverse direction. The motor operations of two motors can be controlled by input logic at pins 2 and 7 and 10 and 15. Input logic 00 or 11 will stop the corresponding motor. Logic 01 and 10 will rotate it in clockwise and anticlockwise. Enable pins 1 and 9 (corresponding to the two motors) must be high for motors to start operating. When an enable input is high, the associated driver gets enabled. As a result, the outputs become active and work in phase with their inputs. Similarly, when the enable input is low, that driver is disabled, and their outputs are off and in the high-impedance state. [9]

4.6. LCD Display (16x2):

LCD (Liquid Crystal Display) screen is an electronic display module and find a wide range of applications. A 16x2 LCD display is very elementary module and is very usually used in various devices and circuits. These modules are preferred over seven segments and other multi segment LEDs. The reasons is LCDs are economical, easily programmable, have no limitation of displaying special and even custom characters (unlike in seven segments), animations and so on. A 16x2 LCD means it can display 16 characters per line and there are 2 such lines. In this LCD each character is displayed in 5x7 pixel matrix. This LCD has two registers, namely, Command and Data. The command register stores the command instructions given to the LCD. A command is an instruction given to LCD to do a predefined task like initializing it, dissipating its screen, setting the cursor position, controlling display etc. The data register stores the data to be displayed on the LCD. The data is the ASCII value of the character to be displayed on the LCD. [10]

4.7. GLCD (AMPIRE128x64):

GLCD is a graphic liquid crystal display. AMPIRE (128*X*64) graphic LCD contains 128 coulombs and 64 rows. It can display data in (128x64) matrix. Graphical LCDs are different from the ordinary alphanumeric LCDS, like (16x1), (16x2), (16x4), (20x1), (20x2) etc. They (ordinary) can print only characters or custom made characters. They have a fixed size for displaying a character normally (5x7) or (5x8) matrix. Where as in graphical LCD we have 128x64=8192 dots each dot can be lit up as per coding or can make pixels with 8 dots that is 8192/8=1024 pixels, graphical LCD is controlled by two KS0108 controllers. A single KS0108 controller is capable of controlling 40 characters so for controlling a graphical LCD we need two KS0108 controllers and the 128x64 LCD is divided into two equal halves with each half being controlled by a separate KS0108 controller. [11] Special types of shapes or picture can be drawn in GLCD by exciting it DOTS.

4.8. MOSFET:

MOSFET is a three terminal semi-conductor device it consists of gate drain and source. It works in three region such as cutoff active and saturation. To use MOSFET as a switch it need to operate in active region and for amplification any input signal it will work in saturation region. Whenever a signal will be given in its gate it will activate. In this model MOSFET is used as a switch for alarming circuit (buzzer circuit). It's used in a way that voltage across its drain to source is close to zero to make it work as a switch.

4.9. Proteus Model:

FIG.4 is the proteus model of an automatic forward and backward collision avoidance system. This microcontroller works within 20nsec. Its operating voltage range is: 2.0V to 5.5V. Two

sharp distance sensors are used to measure the distance of the obstacle for front and rear respectively, crystal oscillator is also connected to the microcontroller, it produces the oscillation and generates the clock pulse of the microcontroller. There is a buzzer used as alarm and it is controlled by a MOSFET, microcontroller turns on the MOSFET which work as a switch for buzzer. A motor driver L293D is used for controlling the actuator. Actuator will be connected to the car braking system.

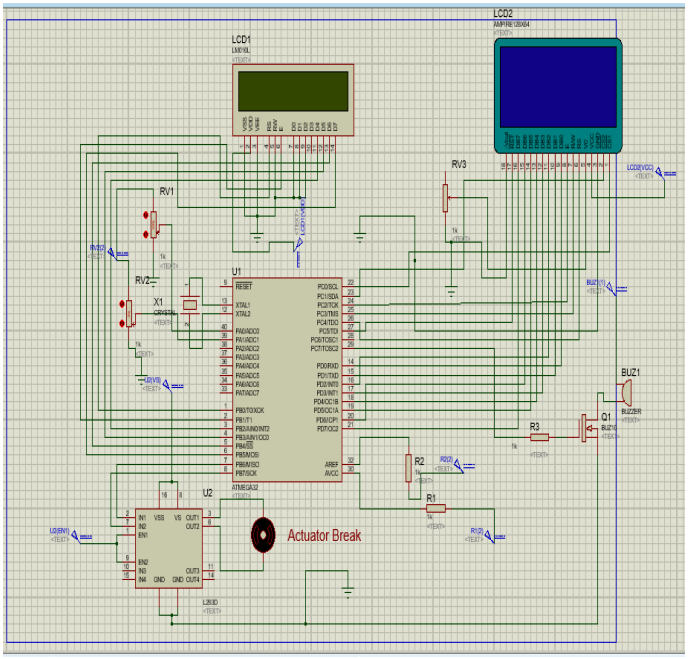


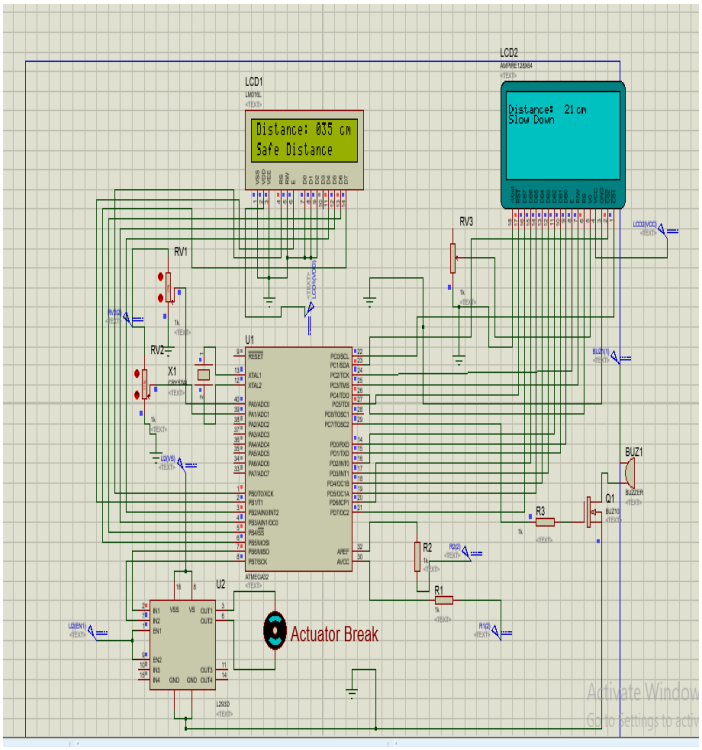
FIG.4. Proteus model of the system

From FIG.4 it is found that microcontroller pin7 (PB6/MISO) and pin8 (PB7/SCK) is connected to motor driver pin2 (IN1) and pin7 (IN2). From this connection of the configuration, motor driver gets command from microcontroller. 1k ohm resistors are connected to microcontroller pin30 and 32 with constant 5V DC for operating of microcontroller and ADC comparison respectively. ADC comparison is used in microcontroller because it will help microcontroller to process the measured distance by distance sensor in digital format. Microcontroller pin30 and 40 are connected to sharp distance sensors but there is no distance sensor device in Proteus so potentiometers are used instead of it. The highest readings of these potentiometers are 5k ohm and operating voltage +2.5V. Two port of actuator is connected to motor driver pin3 (OUT1) and pin6 (OUT2) A DC gear motor is used as an actuator in simulation for analyzing the motor driver behavior to actuator due to absence of an actuator in proteus model. Two ports of crystal oscillator are connected to microcontroller pin13 and pin12 is used for generating clock pulse of microcontroller. Here is a LCD display which pin11 (D4), pin12 (D5), pin13 (D6), pin14 (D7) are connected to microcontroller pin3 (PB2), pin4 (PB3/PGM), pin5 (PB4), pin6 (PBS) respectively, so that, LCD display can get signal from microcontroller and gives information about the obstacle.

Microcontroller pin29 (PC7) is connected to a MOSFET by a 1kohm resistor R2. Microcontroller itself cannot turn the buzzer on for this reason a MOSFET has been used. Microcontroller gives instruction to the MOSFET to turns the buzzer on and it operates at +12V. GLCD pin9-16 (DB0-DB7) is connected with microcontroller pin14-21 (PD0-PD7) which will give information regarding rear obstacle distance. GLCD control pins, 1(CS1), 2(CS2), 6(RS), 7(RW), 8(E), 17(RST) is connected to microcontroller pin22 (PC0),pin23 (PC1), pin26 (PC4),pin25 (PC3), pin24 (PC2),pin27 (PC5) respectively, so that GLCD can get signal from microcontroller and gives information about the distance which should be kept to avoid collision, or to stop to avoid collision towards the car approaching from the back. A program is written in MIKROC software, which is the instruction for microcontroller action and HEX file of that program is used for simulation.

5. Simulation and Result

From FIG.5.we can conclude that if there is a vehicle in front of a car containing this collision avoidance system at a distance of 35 cm microcontroller will command LCD to Display “safe distance “to the driver of the car, and at the same time if there is a vehicle approaching from rear and if it’s distance is 21 cm micro controller will instruct GLCD to display the distance information along with the message “Slow Down’’ to the incoming car from the back.



From FIG.6. we can conclude that if there is a vehicle in front of a car containing this collision avoidance system at a distance of less than 26 cm microcontroller will command LCD to Display “Close Distance” to the driver of the car and microcontroller will send activation signal towards buzzer circuit, and at the same time if there is a vehicle approaching from rear and if it’s distance is less than 15 cm micro controller will instruct GLCD to display the distance information along with the message “Emergency! Stop’’ to the incoming car from the back.

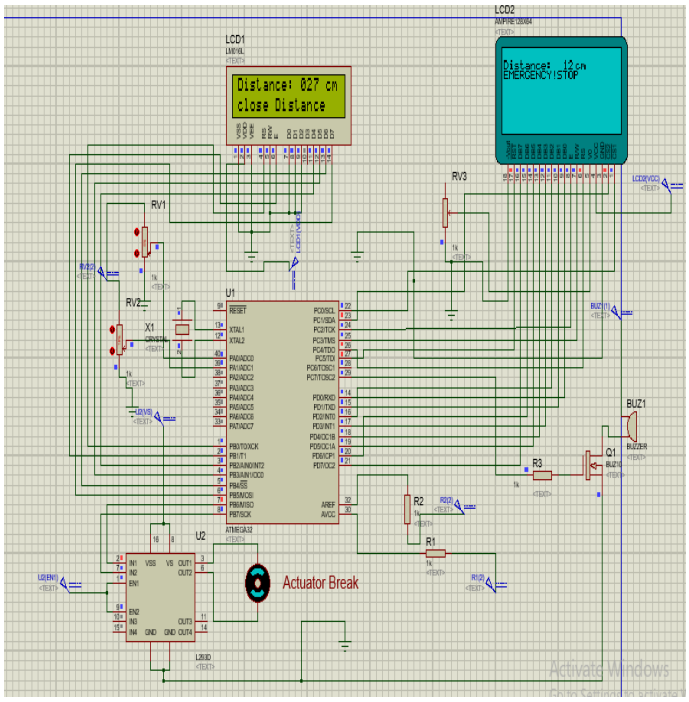


FIG.6. Result for close distance for front and emergency stop distance for rear

From FIG.7. shows that if there is a vehicle in front of a car containing this collision avoidance system at a distance of 18 cm microcontroller will command LCD to Display "Critical Distance" to the driver of the car and microcontroller will send activation signal towards buzzer circuit, and at the same time if there is a vehicle approaching from rear and if it's distance is27 cm micro controller will instruct GLCD to display the distance information along with the message "Maintain This Distance'' to the incoming car from the back.

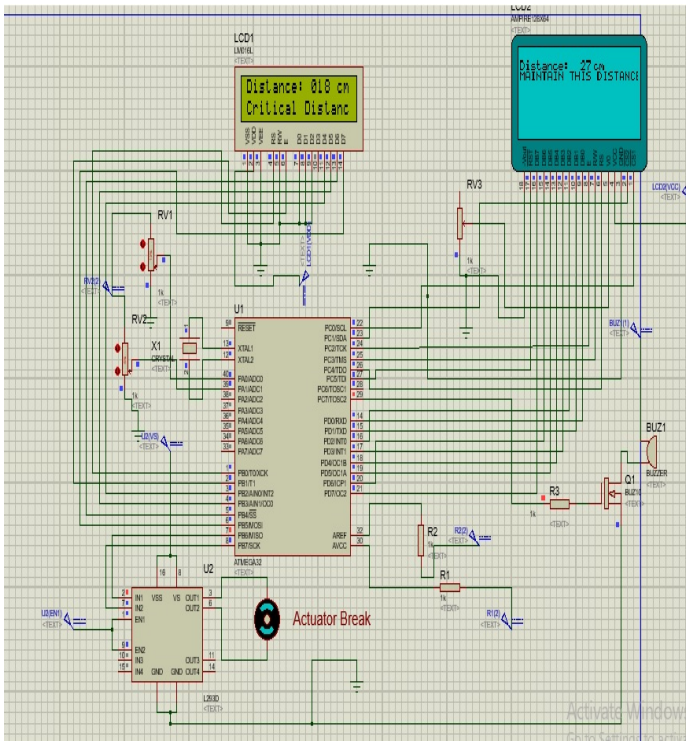


FIG.7. Result for critical distance for front and safe distance for rear

From FIG.8. we can conclude that if there is a vehicle in front of a car containing this collision avoidance system at a distance of 13 cm microcontroller will command LCD to Display "Actuator On" to the driver of the car and microcontroller will instruct motor driver to activate actuator, motor driver will send high current towards actuator which will cause actuator to turn on and this actuator will automatically press the brake of the car to stop. If there is a vehicle approaching from rear and if it's distance is less than 19 cm micro controller will instruct GLCD to display the distance information along with the message of "TENT TO COLLISION'' to the incoming car from the back.

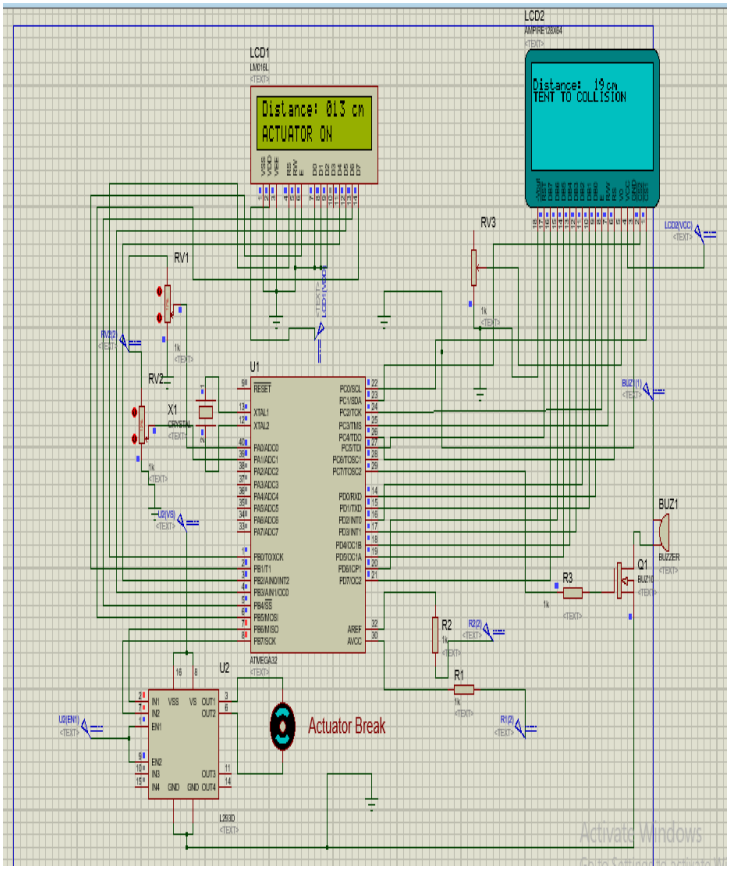


FIG.8. Result for Actuator On distance for front and tent to collision distance for rear

6. Result analysis:

During simulation distance was taken in small scale as in cm. For forward collision prevention system instruction for microcontroller was set in a way if measured distance by distance sensor is above 30 cm it will consider this distance to safe distance and it will instruct LCD to display to show the message safe distance along with that distance value. If the value of distance varies from 20 cm to 29 cm microcontroller will instruct LCD to show Close distance along with the value of the distance measured by the sharp distance sensor, further more microcontroller will send a pulsating signal to buzzer and buzzer will be turn on, this pulsating signal will cause change in the sound frequency of that buzzer. And if distance ranges from 15cm to 19 cm micro controller will instruct LCD to display the message "critical distance along with the measured distance "at the same time microcontroller will send a continuous signal to the buzzer and it

will buzzed, this buzzing frequency will be different from the previous one. And any distance less than 15 cm microcontroller will instruct motor driver to turn on the actuation system to stop the car. At the same time LCD will display "Actuator on along with distance information". From the simulated result, which were shown in the previous section; it is seen that when sharp distance sensors measure a distance of 35 cm LCD shows the message of safe distance which is larger than 30 cm. When measured distance by sensor was found 27 cm LCD shows the messages that this distance is close distance and at the same time a pulsating signal was sent to buzzer circuit. This measured distance was less than 29 cm. When distance was found about 18 cm LCD display message that this distance is critical distance, which was less than 20 cm. At the same time buzzer circuit will be turn on as micro controller sends a continuous signal to buzzer. From FIG.7 it can be seen that a red dot is appearing in front of the R3 resistor and which was connected with the microcontroller PC7 port and same red dot is appearing in that particular pin which indicates that microcontroller sending continuous signal to buzzer to buzz. Finally when distance was measured 13 cm which is less than 15 cm LCD displays the message that actuator on along with the measured value of distance in cm. From FIG.8 it is seen that there were two red dots in pin 2 and 7 of the motor driver which were connected to microcontroller pin 7 and pin 8 and same dots appear across those pins, which indicates microcontroller is sending high signal to motor driver to turn on the actuator or the motor of the actuator system. As motor driver needs high or low logic in its both pin to stop the motor of actuating signal and microcontroller is sending high signal to that motor driver is evidence that microcontroller is actually control the braking system of the vehicle and it will cause the car to stop.

For back collision avoidance system, the set of instruction for microcontroller was if measured distance is above 25cm it will be considered as a safe distance and GLCD will show "Maintain this distance" along with measured distance. Any distance less than 25 cm and greater 20 cm microcontroller will instruct GLCD to show "slowdown" along with measured distance value. Any distance ranges 19 cm to 15 cm microcontroller will instruct GLCD to display "tent to collision" along with the distance that was measured by sharp distance sensor. Finally any distance measured by the sharp distance sensor less than 15 cm microcontroller will instruct GLCD to display "Emergency! Stop". FIG.5 to FIG.8 in the simulation and result section it's seen that the measured distance by rear sharp distance sensor was 21cm, 12c m, 27 cm, 19 cm respectively, this distances follow the ranges that were mentioned above . And GLCD is displaying same sort of the message for that particular range of distances. From the above discussion it can be concluded that microcontroller is executing its instruction perfectly as it was given to it by a program which was written in MIKROC PRO for AVR, which results in that proposed design system is perfectly working during simulation. But when this system will be practically implemented for vehicle this distance range wouldn't be applicable, virtually all vehicles' or cars' road braking performance test indicate stopping distances for 60 mph are typically 36m to43m [12]. So when this system will be implemented in reality only the set of instructions will need to be changed which will be loaded in microcontroller.

7. Implemented forward collision system

The design of the research work in this paper is the extended design of the implementation of *"Automated Anti Collision System for Automobiles"* [1] which has been implemented successfully in practical by prototype prior, where a Sharp distance sensor, microcontroller PIC16f876a, motor driver, linear actuator alarming device and LCD display were used. From following figures we will see the complete view of previously implemented prototype device along with its distance ranges.

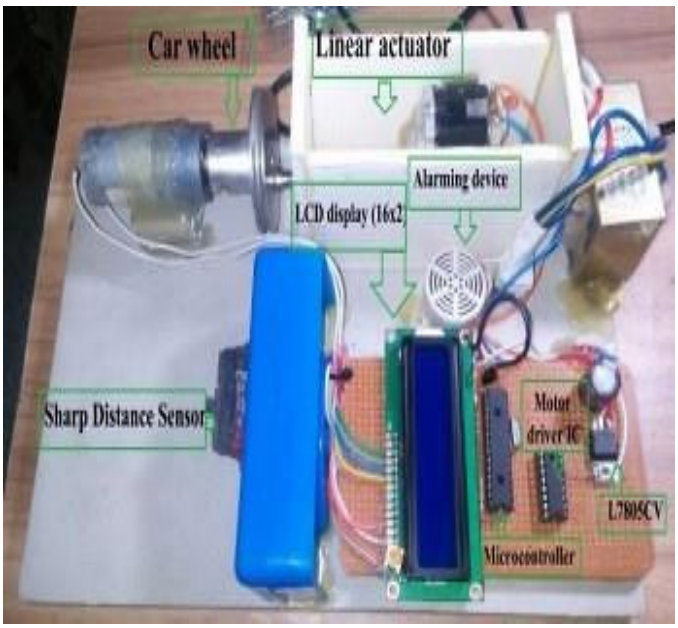


FIG.9: Implemented Prototype Device

From FIG.9.we can see the complete view of previously implemented prototype device.



FIG.10. Safe distance of the prototype device





From FIG.10.we can see the distance is 30 cm which was considered as safe distance and the actuator did not work. In FIG.11.and FIG.12.we can see distances as follow 24 cm and 09 cm whereas, 21 cm gave alarm to the actuator and in 09 cm, and actuator started working. In addition, the research of this paper has been done with the back collision avoidance system along with this mentioned forward collision avoidance system. Also more progress has been done about the forward obstacles detection distance ranges. Due to shortage of time this recent extended research work prototype was not possible to implement.

8. Application and Improvement plan of the system

Proposed system which is described in this paper will be implemented by prototype practically very soon. Near future, during manufacture it can be implanted in motor bikes taxi-cab, car, and truck even in aircraft system. In winter when roads might be rarely seen due to fogs this system will notify the person who will be in control of that vehicle about other vehicles or any other obstacles ahead of his vehicle. Which will help him drive safely his vehicle, and it will make him feel comfortable while driving and it will also extinguish his fear. This will reduce terrifying collision of the vehicles and not only the winter, in other season it will notify driver to maintain a safe distance by alarming sound or showing distance information in LCD. If driver begin to avoid the alarming sound and start to drive recklessly system will forcefully stop the car to avoid accident. Researchers of this paper has some future improvement plan for making this system more convenient and add other features to make this system more accurate. For improving this system sonar system will be implemented due to its long range detection, Emergency light or special kind of alarm will be used to notify the vehicle to maintain safe distance, graphical symbols will be added along with messages in GLCD. Wireless Fidelity (Wi-Fi) technology can be used for transferring information between successive vehicle about the distance and what speed should the kept for safety, furthermore, Dual Infra-Red Obstacle Detector (DROD40) can be used in this system to determine obstacle in every side such as left, right and ahead distinctly. In future, researchers of this paper have plan to design a system which will monitor driver health condition to judge his fitness for driving if he isn't fit, vehicle will not start. So, this will reduce almost 60% of accident.

Conclusion

Technologies, new inventions and transport system play an important role in human lives. Mankind is becoming habituated to these new inventions. Especially transport system without this mankind even think about their daily activities and even can step outside from their home. As every other invention this transport system has adverse effect too, and as usual mankind is responsible for this kind of adverse effect. People are always in rush to reach their destination as early as possible to save their time or to reach their destination on time without thinking about their lives value, what would happen if they met up with an accident which means colliding with other vehicle for their tendency to go somewhere within a flash. This results in injury or loses of their lives. The prime motive of the design is to reduce these accidents. In other words it can be concluded that this design is used to save human lives from the loss due to accident and also to ensure safe and comfortable journey.

Acknowledgment

First and foremost, we want to express our gratitude towards Almighty creator for his blessings without which it would not be possible to complete this work successfully afterwards want to acknowledge the constant support received from our family. It is an honor to express our heartiest gratitude to MS. Muhit, Assistant Professor, Faculty of Electrical and Electronic Engineering, AIUB, for his immense support and encouragement to improvise this designed system. And finally especial gratitude to honorable Mr. Chowdhury Akram Hossain, Assistant Professor, Faculty of Electrical and Electronic Engineering, AIUB. Without his motivation this work wouldn't be possible to advance so far. Finally, any errors are our own and should not tarnish the reputations of these esteemed persons.

References

[1] Tasneem Sanjana, Kazi Ahmed AsifFuad, Mehrab Masayeed Habib, Ahmed Amin Rumel "Automated Anti-Collision System for Automobiles" in International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar Bangladesh February 16-18, 2017 http://**DOI:** 10.1109/ECACE.2017.7913024

[2] http://www-nrd.nhtsa.dot.gov/pdf/esv/esv19/05-0322-O.pdf [Cited: 14 May, 2015]. Available:

[3] N., Mechanisms and Mechanical Devices Source book 4th Edition (2007), 25, McGraw- Hill [Cited: 17 May 2015].

[4] http://www.microchip.com/wwwproducts/en/ATmega32 [Cited: 13 May, 2015]. Available:

[5] http://www.sharpsma.com/webfm_send/1208 [Cited: 17 May, 2015]. Available:

[6] Roboticlab website abailable at, http:// home. roboticlab.Eu/en/examples/sensor/ir_distance(Cited : 17, May 2016).

[7] https://www.firgelliauto.com/blogs/news/18090539-linear-actuator-how-to[Cited: 17 May, 2015]. Available:

[8] http://www.electronics-tutorials.ws/oscillator/crystal.html[Cited: 17 May, 2015]. Available

[9] http://www.engineersgarage.com/sites/default/files/L293D.pdfCited: 7 May 2015]. Available:

[10] http://www.engineersgarage.com/sites/default/files/LCD%2016x2.pdf [Cited: 17 May , 2015]. Available:

[11] https://www.electronify.org/basic-component-introduction/graphics-lcd [Cited: 17 May, 2015]. Available:

[12] Divya P.&MurugesanA.,"Intelligent Car Braking System with collisinAvoidance and ABS",International Journal of Computer Applications(0975– 8887),National Conference on Information and Communication Technologies (NCICT , 2015).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 206-210 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

The impact of Big Data on the Android Mobile Platform for Natural Disaster Situations

Zijadin Krasniqi[*,1], Adriana Gjonaj[2]

[1]*Department of Information Technology, Tax Administration of Kosovo.*

[2]*Department of Informatics, Mathematics, and Statistics, European University of Tirana, Albania.*

ARTICLE INFO

Article history:
Received: 10 January ,2018
Accepted: 06 March ,2018
Online: 31 March ,2018

Keywords:
Android
OCEMA
Big Data
DWH
Natural Disasters

ABSTRACT

This application developing for the project OCEMA comes as result of architecture building in Android, then the development of a professional modeling in Talend Open Studio for Big Data, which enabled the integration of data from many data sources. One of its uses is the quick identification of people found in areas affected by natural disasters. The application identifies the persons who do not have an ID card, or another identification document, by using identification through fingerprint or personal number. OCEMA Application has access to all the agencies involved in natural disasters managing, such as ISK, HMIK, MIK, FSK and CRA. This application is developed to connect to web applications as well, such as applications, which gather real time information on earthquakes and weather in the world.
The study of literature and the actual work with these systems has shown some important components of success for DWH systems, DataMart and mobile application development.

1. Introduction

In the year 2002 alone, disasters affected a staggering 680 million people worldwide. Natural disasters include bring earthquakes, flooding, droughts, heat waves, cold spells and events that destroy countries, businesses and individual property and well-being.
The goal of this study is to investigate the adoption of cell phones as preparedness efforts during the natural disaster seasons in a developing country. Since Kosovo is a country vulnerable to natural disasters, the present study aims to investigate the adoption of cell phones for natural disaster preparedness [1]

OCEMA applications initially are secured by a username and password, and then divided in 5 buttons, such as: ISK, CRA, HMIK, MIK and FSK. This application enables access to real time information, such as in the case of the agency of civil register, offering two possibilities to identify the victims of natural disasters.

It uses the identification through personal number, but in cases where there is no ID card, the identification can be done using the fingerprints. Part of this application is access to institutions such as the Institution of Seismology, the Medical Institution, the Hydro meteorological Institution and the Firefighter's Service.

These institutions have updated information. Beside this, the OCEMA application has the possibility to access important web pages in order to monitor climate and seismic conditions in world level.

2. Mobile Platforms

The four most common smart phone operating systems, by market share in the third quarter of 2013 (IDC 2013) are:
•Android by Google, with a distinct dominance at 81 %,
•iOS, on Apple's iPhone, at 12.9 %,
•Windows Phone by Microsoft at 3.6 %
•and BlackBerry at 1.7 %.

There are many significant differences between these platforms. Some are visible to the users such as availability of specific features or design of the user interface (UI), and some are invisible to the users but affect the developers of mobile applications.

2.1. Android architecture

Starting from the bottom we have Linux Kernel, Android is built up on the Linux Kernel. Linux is already being used extensively from so many years and its kernel had received so many security patches. Linux Kernel provides basic system functionality like

[*]Zijadin Krasniqi, Tax Administration of Kosovo, zijadinkrasniqi@hotmail.com

process management, memory management, device management like camera, keypad, display etc.

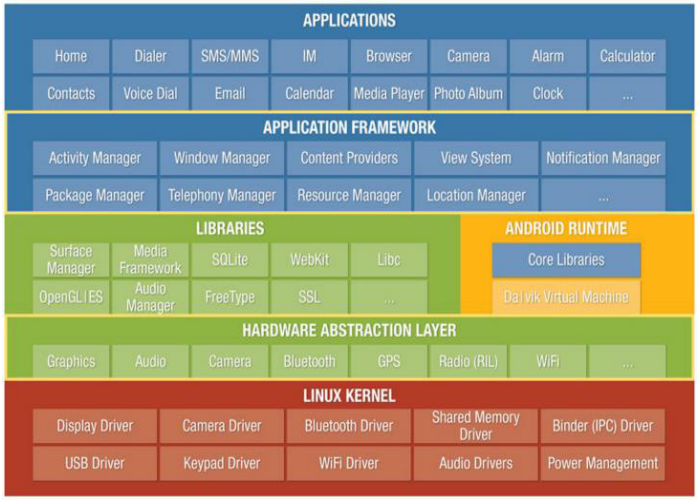


Figure 1: Android Architecture [2]

As the base for a mobile computing environment, the Linux kernel provides Android with several key security features, including:

- A user-based permissions model
- Process isolation
- Extensible mechanism for secure IPC
- The ability to remove unnecessary and potentially insecure parts of the kernel.

As a multiuser operating system, a fundamental security objective of the Linux kernel is to isolate user resources from one another. The Linux security philosophy is to protect user resources from one another [2]

Taking into consideration the Android as a target platform for application comes as the result of reviewing various works using Android technology.

The Android SDK allows application development with great ease. There are many inbuilt features and tools in an Android device, which can be integrated and programmed to be used as and when required from within the application [3]

3. Case study: SQLite_OCEMA database integration can be implemented in Android mobile devices

The operational Centre of the Emergency Management Agency requires analysis in daily, weekly and monthly bases. What is most important is the capability of real time data analysis. Therefore, this increases the need for data collection and data integration in a single database. The scope is to be able to analyze possible problems faced during emergency situations and natural hazards and then determine the highest levels of entities from which we can collect data.

The following data form the bases of the conceptual model of OCEMA database, which we are going to build.

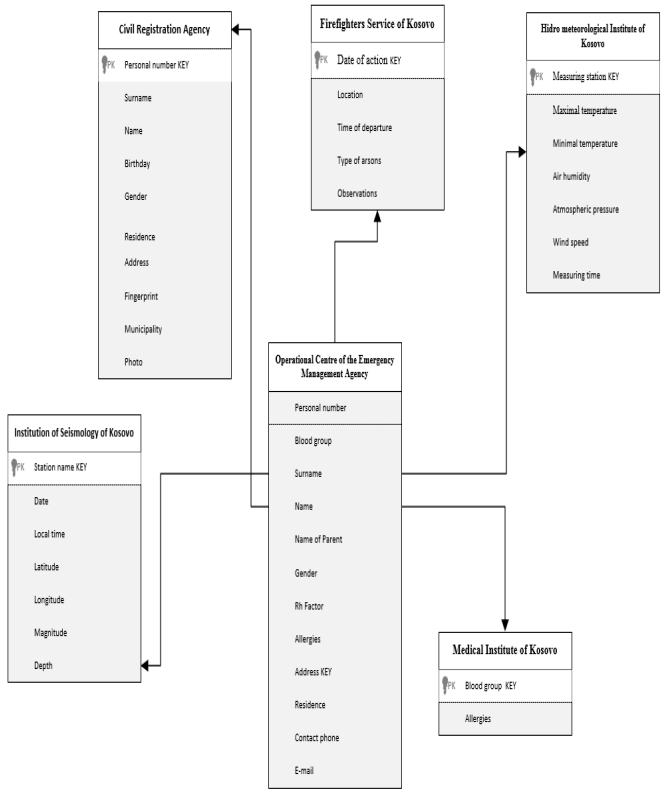


Figure 2: The Entity Relationship diagram of OCEMA database

We start with the description of the entities used in the OCEMA database.

The entity named Medical Institute of Kosovo is composed of the following data: Blood group for the patients and the eventual allergies present in the population. Entry data from MIK are an important information source at national level as far as the health of the population is concerned.

The entity named Civil Registry Agency includes several fields involving records on entry data regarding the civil registration at a national level, among them including the following characteristics: personal number, surname, name, date of birth, residence, address, fingerprints, district and photo.

All these attributes of the CRA entity include the population's records. In our study, the principal searching attributes on individuals will be the personal number and the fingerprints.
The entity named Institute of Seismology of Kosovo includes the following attributes: name of the weather station from where we get the real time data on seismic waves and information from the instruments of online monitoring of seismic waves.

Then in this entity we include: the date of seismic waves, local time, latitude, longitude, magnitude and seismic depth.
The data entity generated by the Hydro meteorological Institute of Kosovo includes a very significant dynamics of available data aiming the comprehensive real time monitoring in order to prevent flooding from rainfall, snow and also extreme draught caused by extreme maximal temperatures.

This entity includes the following attributes: the weather measuring station, maximal temperature, minimal temperature, air humidity, atmospheric pressure, wind speed, measuring time. It is very important to keep under control the continuous monitoring and observation of many factors, which affect arsons, such as the case of arsons involving homes, forests and pastures.

The entity of Firefighters Service is composed of the following attributes: date of action, location, time of departure, type of arsons, observations. All of these offer a very good opportunity to get information from this very important institution.

4. Data integration in SQLite_OCEMA through Talend Open Studio for Big Data

We have chosen SQLite as the main database, where we will integrate the tables from Oracle SQL, such as: CRA, HMIK, MIK, ISK and FSK. Most of the enterprise companies use Oracle in their applications in order to load data. In the case of data loading in SQLite_OCEMA, we have built the following job in Talend Open Studio for Big Data, as shown in the Figure below:

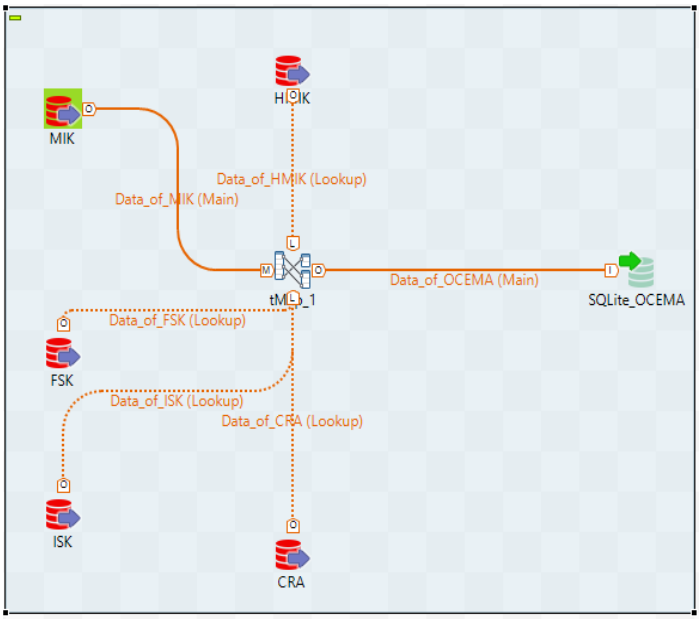


Figure 3: Talend's Job for data loading from the ORACLE SQL tables in SQLite_OCEMA

These heterogeneous data sources, HMIK, MIK, FSK, ISK and CRA, then will be stored in SQLite_OCEMA databases, as shown in the TOS schema.

An important step in ensuring the job developing with ETL is the successful connection between TOS and SQLite_OCEMA database.

4.1. Projecting the SQLite _OCEMA database

Unlike most RDBMS products, SQLite does not have client/server architecture. Most large-scale database systems have a large server package that makes up the database engine.

The database server often consists of multiple processes that work in concert to manage client connection, I/O files, caches, query optimization and query processing. A database instance typically consists of a large number of files organized into one or more directory trees on the server's file system.

In order to access the database, all the files must be present and correct. This can make it somewhat difficult to move or reliably back up a database instance. To access the database, client software libraries are typically provided by the database vendor. These libraries must be integrated into any client application that wishes to access the database server.

These client libraries provide APIs to find and connect to the database server, as well as set up and execute database queries and commands. The Figure 4 below shows how everything fits together in a typical client/server RDBMS.

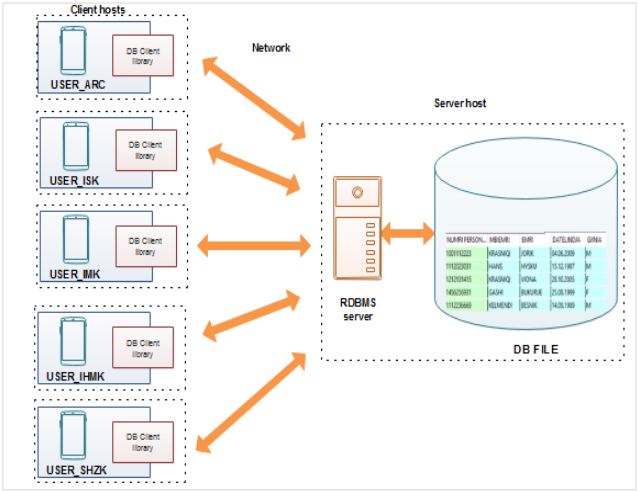


Figure 4: Client/server RDBMS architecture [4]

In contrast, SQLite does not have a separate server. The entire database engine is integrated into whatever application needs to access a database.

The only shared resource among applications is the single database file as it sits on disk. If we need to move or back up the database, we can simply copy the file. Below, in the Figure 5, we will see the SQLite architecture.

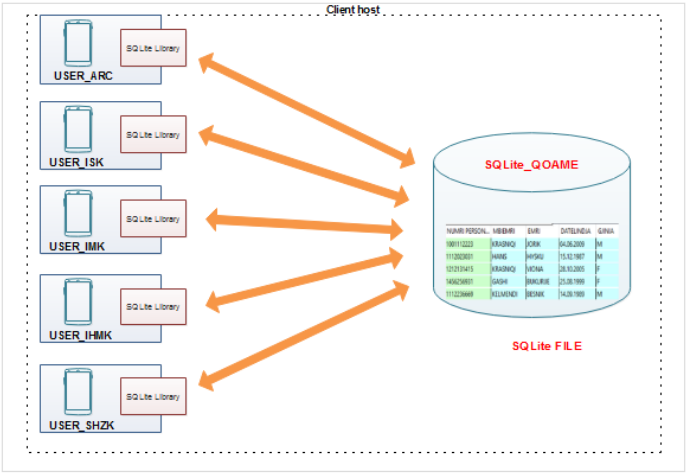


Figure 5: The SQLite server architecture [4]

By eliminating the server, a significant amount of complexity is removed. This simplifies the software components and nearly eliminates the need for advanced operating support. Unlike a traditional RDBMS server that requires advanced multitasking and high performance inter-process communication, SQLite requires little more than the ability to read and write to some type of storage.SQLite is designed to be integrated directly into an executable. This eliminated the need for an external library and simplifies distribution and installation. Removing external dependencies also removes most versioning issues.

If the SQLite code is built right into the application, we never need to worry about linking to the correct version of a client library or that the client library is version-compatible with the database server. Eliminating the server imposes some restrictions. SQLite is designed to address localize storage needs, such as a web server accessing the local database. This means it isn't well suited for situations where multiple client machines need to access a centralized database. That situation is more representative of client/server architecture and is better serviced by a database system that uses the same architecture. [4]

4.2. SQLite_OCEMA database-the case of android application for CRA

Android uses the SQLite database system, which is an open-source, stand-alone SQL database, widely used by many popular applications.

SQLite is a lightweight transactional database engine that occupies a small amount of disk storage and memory, thus, it is a perfect choice for creating databases on many mobile operating systems such as Android and iOS [2]

The Android platform offers full support for SQLite databases. The created database will be accessible by name from each of the classes inside the application, not those from outside. In our case, when we are searching the data from the Civil Registration Agency in case of emergencies and natural disasters, we have the possibility to search for missing people during lifesaving operations and also in the warning period of natural hazards.

Thus, we use the personal number identification method to find people struck there where the natural disaster happened. If we take the Civil Registration Agency table of the SQLite_OCEMA database as in the following Table 1:

Table 1: Civil Registration Agency data table in the SQLite_OCEMA database.

Personal number	Surname	Name	Birthday	Gender	Municipality	Address
1001112223	KRASNIQI	JORIK	04.06.2009	M	PRISHTINE	"Bajram Bahtiri"
1112023031	HANS	HYSKU	15.12.1987	M	PEJE	"Lidhja e Prizrenit"
1212131415	KRASNIQI	VIONA	28.10.2005	F	PODUJEVE	"Ismail Dumoshi"
1456256931	GASHI	BUKURIJE	25.08.1999	F	MITROVICE	"Bajram Bahtiri"
1112236669	KELMENDI	BESNIK	14.09.1989	M	FERIZAJ	"Lidhja e Prizrenit"
1488523691	HALILI	LUM	25.03.2011	M	PRIZREN	"Bajram Bahtiri"
2581473691	KADRIU	LIRI	18.02.2000	F	PEJE	"Bajram Bahtiri"
3698521471	GASHI	ISA	23.04.1988	M	PEJE	"Lidhja e Prizrenit"
1234567893	KODRA	GEZIM	01.01.1975	M	PRISHTINE	"Ismail Dumoshi"
1593574562	TELAKU	TRIM	02.02.2012	M	DEÇAN	"Bajram Bahtiri"
1235661114	RUGOVA	JORIK	14.09.2014	M	PRISHTINE	"Lidhja e Prizrenit"
4569871233	HALILI	RRON	14.05.2007	M	PEJE	"Ismail Dumoshi"
1488523691	HALILI	LUM	25.03.2011	M	PRIZREN	"Bajram Bahtiri"

5. Searching the SQLite_OCEMA data using the personal number

If we set as criteria the searching of data using the personal number method from the SQLite _OCEMA database and we use the recommended methods to create a new SQLite database.

When a database has been successfully opened, the SQLite Open Helper will cache it, so you can (and should) use these methods each time you query or perform a transaction on the database, rather than caching the open database within your application.

A call to getWritableDatabase can fail due to disk space or permission issues, so it's good practice to fall back to getReadableDatabase method for database queries if necessary. In most cases this method will provide the same, cached writeable database instance as getWritableDatabase [5]. An example is shown in the Figure 6.

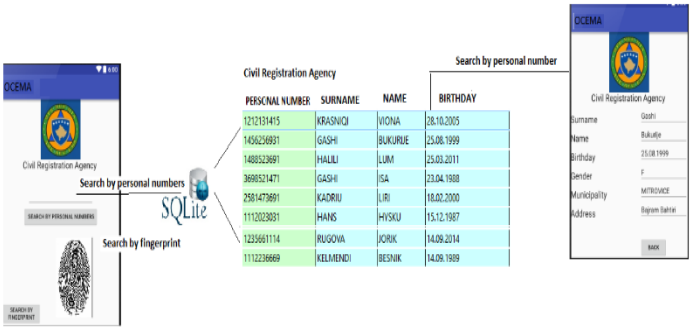


Figure 6: Searching by personal number in SQLite_OCEMA database [6]

This is a part of the application development. In this part, we search for people struck by natural disasters and emergency situations by using the fingerprint method, as shown in the Figure.

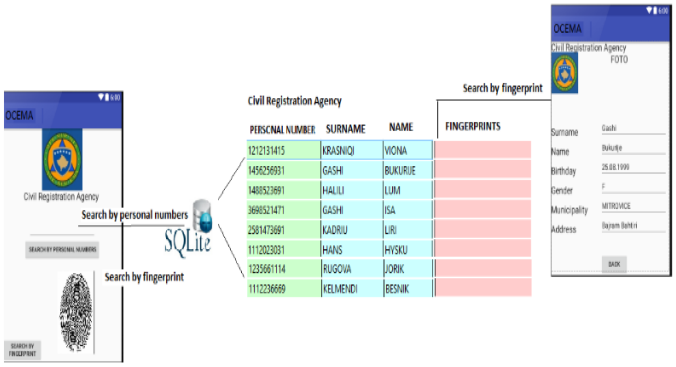


Figure 7: Searching by using the fingerprints in SQLite_OCEMA database.[6]

This result is displayed every time the user launches the matching activity from our proposed application, by comparing the initial fingerprint with the ones stored in database.

One utility of the proposed application consists in quick identification of the people struck in areas hit by natural hazards or emergency operations, in the quick identification of those people who do not have an ID card or another document of identification, by using the fingerprints identification method.

Also, the application could be deployed in the access systems of the public institutions like schools, libraries, hospitals and so on[7]

6. SQLite_OCEMA database- the case of android application for ISK

The Institution of Seismology has a large number of weather measuring stations across the territory of Kosovo. These stations transmit their data to the institute on daily, weekly and monthly bases. Upon creating and building the SQLite_OCEMA database, these data can be reported in a shorter time, better to say in real time across all the weather measuring stations.

By clicking on the ISK (Institute of Seismology of Kosovo) button in the OCEMA basic application, we gain access to search the latest updated ISK data, as shown below in the Figure 8.

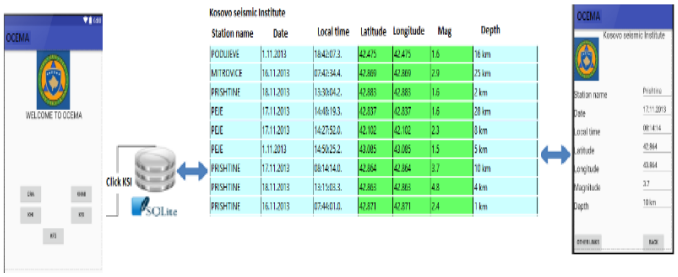


Figure 8: Access to the data of Institute of Seismology of Kosovo. [6]

After the creation of the Institution of Seismology of Kosovo table in the SQLite_OCEMA database, we can click on the ISK button to search the latest updated data of this institution, including magnitude of seismic waves, seismic depth, geographical position, latitude-longitude, date and time of the seismic wave.

7. SQLite_OCEMA database, the case of Android application for HMIK

The Hydro meteorological Institute of Kosovo has a large number of weather measuring stations across the territory of Kosovo. These stations transmit their data to the institute on daily, weekly and monthly bases. Upon creating and building the SQLite_OCEMA database, these data can be reported in a shorter time, better to say in real time across all the stations, including data from the river flows, data from sensors placed in locations where there is more snow.

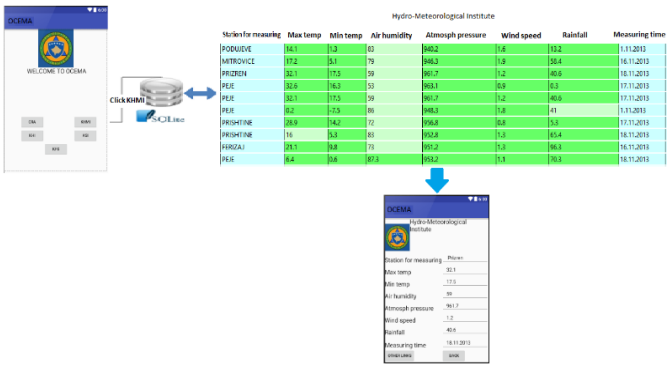


Figure 9: Data from SQLite database for the HMIK case [6]

Thanks to this application we have been able to get detailed information from all the weather measuring stations in the territory of Kosovo, which provide a representation of warnings in case of flooding from atmospheric precipitations, wind speed, air humidity and the maximum and minimum temperatures.

8. Conclusion

Mobile technology has shown a considerable impact on society over the past few years. Gradually, this technology has become part of every person's life and is the easiest mode to deliver information. The development of application OCEMA for the doctoral project comes as a result of building architecture in Android Platform, then a professional model in Talend Open Studio for Big Data, which enabled the data integration from multiple data sources. Then we have developed the applications in Android Platform using the Java programming language. We have made it possible to reach near real time representation of the sources.

It is widely known that during natural disasters and emergencies, we do not have all the information available, but in those moments, everything is important, above all the information related to the human factor. EMA should be aware of these kinds of situations and precede the capacity building in systems communication and developing applications for the management of emergencies and natural disasters.

References

[1] A. Lawson, L.Willoughby, "Adoption of cell phones as preparedness efforts for natural disasters", Issues in Information Systems, International Association for Computer Information Systems, USA, 2012, Vol. 13, pp 11-20 http://iacis.org/iis/2012/79_iis_2012_11-20.pdf

[2] A.Agrawal "Android Application Security Part 2-Understanding Android Operating System", 2015, http://manifestsecurity.com/android-application-security-part-2

[3] T. Tejassvi, "An Android application to support flash flood disaster response management in India" M.sc Thesis, University of Twente -Netherlands, 2014.

[4] J. Kreibich, Using SQLite, Published by O'Reilly Media, 2010, USA.

[5] R. Meier, Professional Android™ 4 Application Development, John Wiley & Sons, Inc., 2012, Canada.

[6] EMERGENCY MANAGEMENT AGENCY-LOGO (2015) https://ame.rks-gov.net/en/Home/PgrID/494/PageID/4.

[7] O.Dospinescu,I. Lîsîi ,"The Recognition of Fingerprints on Mobile Applications – an Android Case Study" , IBIMA Publishing , Romania,2016 http://ibimapublishing.com/journals/JEERBE/2016/813264/813264.html

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 213-217 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A High Efficiency Ultra Thin (1.8 μm) CdS/CdTe p-i-n Solar Cell with CdTe and Si as BSF layer

Nahid A. Jahan, Md. Minhaz Ul Karim and M. Mofazzal Hossain*

Department of Electronics and Communications Engineering, East West University, Dhaka 1212, Bangladesh

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
CdTe/CdS
BSF layer
Cell efficiency
Intrinsic layer

ABSTRACT

CdTe-based photovoltaic (PV) cells provide the lowest EBPT (energy payback time) and emit less amount of GHG (green house gases) among different types of PV cells. Thus, it is very essential to enhance the efficiency of CdTe-based solar cells. A high efficiency CdTe/CdS p-i-n heterostructure solar cell is designed and the performance of the cell is investigated. All the simulations have been done using AMPS 1D (Analysis of Microelectronic and Photonic Structures) simulator. The cell consists of a transparent conducting oxide (TCO) layer (ZnO), a window layer (n-type CdS), an intrinsic layer (CdTe), an absorber layer (p-type CdTe) and a highly doped CdTe and Si as back surface field (BSF) layer. After the optimization of layer thicknesses and doping densities of different layers, we obtained a conversion efficiency of 26.01% for a total cell thickness of 1.8 μm. It is also found that the conversion efficiency can be increased by simply increasing the thickness of intrinsic layer. At 1.5 AM solar irradiance, the proposed cell structure attained a V_{oc} of 1.09 V, a J_{sc} of 26.78 mA/cm^2, and a fill factor of 89%, reaching an overall conversion efficiency of 26.01% with CdTe as BSF layer.

1. Introduction

This paper is an extension of our previous work presented in *ICAEESE 2016* [1]. The rapidly rising demand of world energy supply and the depletion of fossil fuel impose crucial challenges concerning today's energy requirement context. It is also worth mentioning that, the surfeit burning up of fossil fuels renders threatening global warming emissions (GWE) ascribing alarming rise of terrestrial temperature while imposing serious environmental pollutions by particulate materials [2]. These aforementioned issues necessitate the blooming of alternatively dependable, affordable-cost consistent energy sources. To meet these requirements various plausible renewable energy technologies have been thriven so far. Among these alternative technologies solar photovoltaic (PV) is thought to be as one of the most potential and reliable green energy concepts that assures the direct conversion of incident solar irradiation into electrical energy. The power extracted from PV architectures is popularly being adopted and deployed in the remote areas while ensuring the minimization of electrical power transmission losses. PV power can also be potentially applied in vehicular transportation that can ensure the diminishing of global warming emissions in near future. From purely economic point of view, the flourishing market of PV is generating a shortfall in the storage of highly cost silicon wafers which takes almost 50% of the total manufacturing cost. Therefore, it fosters an opportunity for thin film solar technology to create credentials in the market of PV modules because of its moderate cost. Among different types of thin film based modules, CdS/CdTe thin-film solar cells reveals significant suitability owing to its distinctive characteristics encompassing the following aspects. Firstly, the energy bandgap of CdTe is 1.45 eV, which is favorably well-matched to the spectrum of solar radiation. It also lies in the category of direct bandgap semiconductor guiding to fairly strong light absorption with reduced generation of heat. Besides, the doping of CdTe as p-type semiconductor film as well as the formation of p-n heterojunction with the inclusion of CdS is found to be much easier. Moreover, CdS exhibits wide energy-gap of 2.4 eV and can be fabricated as n-type adopting undemanding usual techniques of film deposition [3-5]. In the meanwhile, various less-costly and manageable atomic congregation techniques have been developed for the growth of CdS/CdTe solar cells [6-7]. Also it should be mentioned that, a thickness of around 4 microns absorber layer of CdTe is almost appropriate for capturing the entire solar spectrum due to acceptable range of absorption coefficient [8-9] of CdTe (~10^5 cm^{-1}) in the visible solar spectrum. All these properties

*M. Mofazzal Hossain, Aftabnagar, Dhaka-1212, Bangladesh, +8801796587888 & Email: dmmh@ewubd.edu

of CdTe attribute to quite less material cost of the cell if compared with cells based on Si-wafers. In addition, from various respective studies, it is revealed that the energy-payback-time (average) of thin film CdS/CdTe solar cell is around 1.4 years [10] that is found to be as the rock-bottom among all PV technologies. Moreover, CdTe solar cell generates the lowest amount of GWE for per kWh electrical energy compared to other PV technologies [10].

In CdS/CdTe thin film solar cell, irrespective of superstrate or substrate configurations, light enters into the cell through the TCO and CdS (n-type) window layers. From the earlier attempts, the experimental efficiency of single junction CdS/CdTe solar cell has been reported so far around 17% [11-13]. Approximately 18.6% conversion efficiency from a CdS/CdTe solar cell incorporating complex BSF and TCO layers has been reported by Amin et al [14].

In this work, we explore the possibilities and chances of enhancing the efficiency of CdS/CdTe heterojunction solar cell with the insertion of an intrinsic layer of CdTe. Additionally, a back surface field (BSF) layer of highly doped CdTe (p^+ layer) and Si (p^+ layer) are sandwiched between back contact and the absorber layer of p-CdTe in order to diminish the recombination loss and thereby, magnifying the conversion efficiency. In this proposed structure as depicted in Figure 1, a TCO layer of ZnO is inserted to enhance the level of V_{oc} by minimizing the series resistance of the cell.

2. Methodology

In this work the proposed cell is simulated in AMPS 1D simulator to optimize the layer thicknesses of TCO, window, absorber, intrinsic and BSF layers, and the doping density of BSF layer to achieve the maximum possible conversion efficiency under 1.5 AM solar irradiance. The properties of n-CdS, p-CdTe, a-Si and n-ZnO, those are given in Table 1, are taken from [13-15].

Table 1 Material Properties

Parameters	a-Si	CdTe	CdS	ZnO
ε_r	11.9	9.4	9.0	9.0
μ_n ($cm^2.V^{-1}.s^{-1}$)	20	500	350	100
μ_p ($cm^2.V^{-1}.s^{-1}$)	2	60	50	25
E_g (eV)	1.72	1.45	2.42	3.0
N_C (cm^{-3})	2.5×10^{20}	8×10^{17}	2.4×10^{18}	2.2×10^{18}
N_V (cm^{-3})	2.5×10^{20}	1.8×10^{19}	1.79×10^{19}	1.8×10^{19}
χ (eV)	3.8	4.28	4.5	4.35

3. Function of Intrinsic Layer

To comprehensively explain the function of intrinsic layer the simplified energy band diagram for thermal equilibrium condition is shown in Figure 2. It is worth mentioning that while depicting the energy band diagram, gradient of the quasi-Fermi level has not been shown for simplicity. Literally, when light falls and photo-voltage is generated it biases the p-n junction in forward bias mode and therefore the Fermi level should not be continuous throughout the junction of cell. However, it is evident from this diagram that, the inclusion of intrinsic layer increases the width of depletion layer where mainly the photons are absorbed and electron-hole pairs are generated. From Figure 2, it is clear that the generated electrons can move easily towards the TCO layer (left to n-CdS) due to the slope of electron energy and the effect of tunneling. Similarly photo-generated holes can move towards the back contact layer (right to BSF layer) due the slope of holes energy.

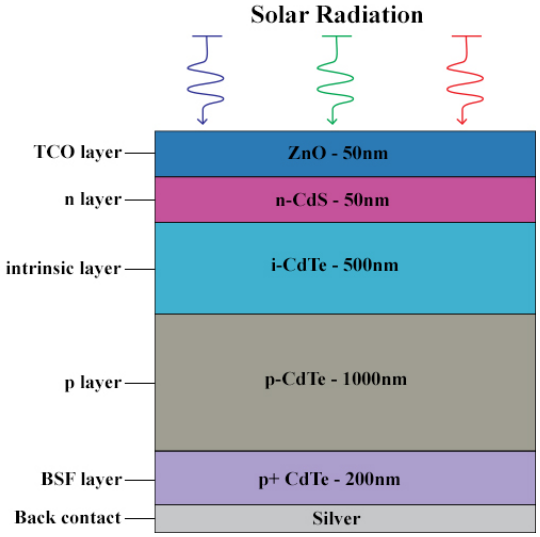


Figure 1The structure of the proposed p-i-n CdS/CdTe solar cell

4. Results and Discussion

In this work, using AMPS 1D simulator we have optimized the thickness of different layers of the cell in order to improve the efficiency. Throughout the simulation, the thicknesses of window layer (n-CdS) and absorber layer (p-CdTe) are kept constant at 50 nm and 1000 nm respectively (as shown Figure 1). We only explored the possibilities of enhancing the conversion efficiency of the cell by optimizing the thicknesses of BSF and intrinsic layers. The proposed structure is simulated in AMPS 1D environment and the I-V characteristics of the cell under 1 sun illumination is extracted from AMPS 1D simulator which is exemplarily shown in Figure 3. After simulation under different doping densities and thicknesses we obtained several figures and results from which we extracted the cell efficiency for desired doping density and thickness.

We explored the effects of doping density of p-CdTe BSF layer on the conversion efficiency for a BSF layer thickness of 50 nm and the results are shown in Figure 4. From Figure 4 it is found that beyond doping density of 2×10^{18} cm^{-3}, the conversion efficiency increases exponentially. The effects of the insertion of amorphous Si (a-Si) and CdTe as BSF layer on the efficiency are also investigated.

Figure 5 and Figure 6 depict the effects of BSF layer thickness (highly doped p-Si and p-CdTe) on the conversion efficiency without the inclusion of intrinsic layer. For this analysis, the doping density of a-Si and CdTe are considered as 6.0×10^{17} cm^{-3} and 1.0×10^{19} cm^{-3} respectively . At a BSF layer thickness of 200 nm, the maximum efficiencies for a-Si and CdTe are attained to be 23.0% and 24.46% respectively. It is observed that beyond 200 nm thickness of BSF layer, in case of a-Si the efficiency becomes

saturated whereas in case of CdTe the efficiency increases almost linearly. From Figure 5 and 6 it is also noticed that in the proposed CdS/CdTe solar cell, CdTe outperforms a-Si as a BSF layer.

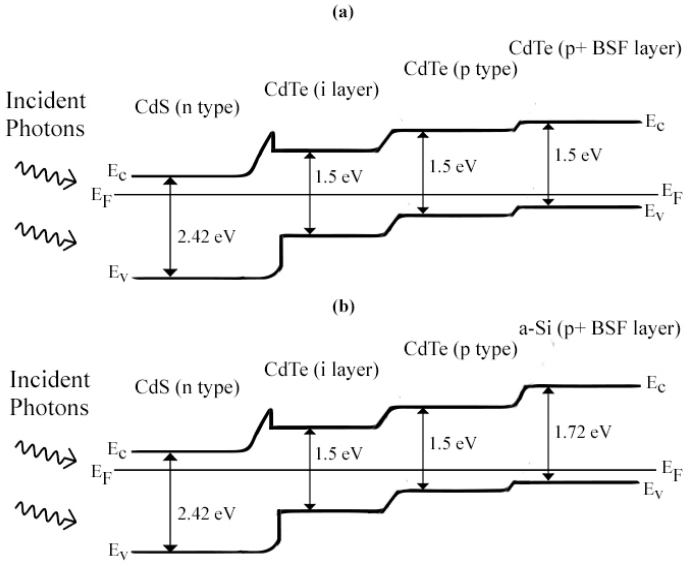


Figure 2 Simplified energy band diagram of proposed (a) p-i-n CdS/CdTe (BSF layer: CdTe) (b) p-i-n CdS/CdTe (BSF layer: a-Si) solar cells

Figure 7 and 8 show the effects of intrinsic layer (CdTe) thickness on the cell's conversion efficiency considering a-Si and CdTe as BSF layers separately. It is observed that in both cases the conversion efficiency increases linearly with the increase of intrinsic layer thickness. However, the maximum conversion efficiency at thickness 500 nm of intrinsic layer for CdTe and a-Si (as BSF layer) are found to be 26.01% and 24.51% respectively. The dependence of short circuit currents on the variation of intrinsic layer thickness for both CdTe and a-Si as BSF layers are also studied as depicted in Figure 9. From this comparison it is realized that CdTe as BSF layer provides larger short circuit current than that of a-Si.

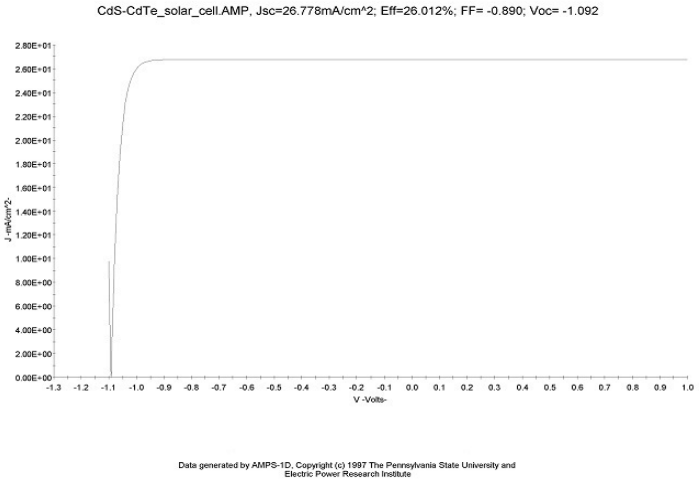


Figure 3 I-V characteristics of the proposed p-i-n CdS/CdTe solar cell.

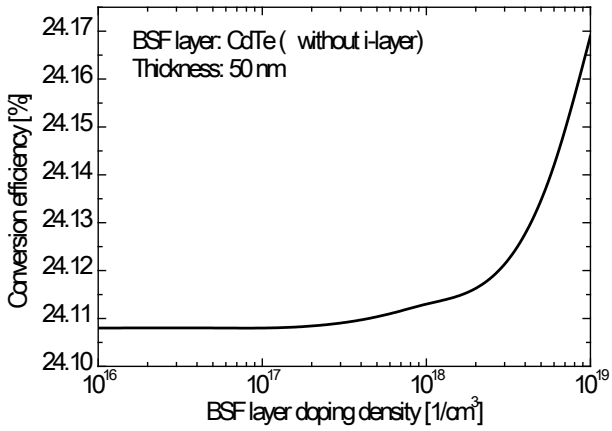


Figure 4 Effects of BSF layer doping density on the conversion efficiency

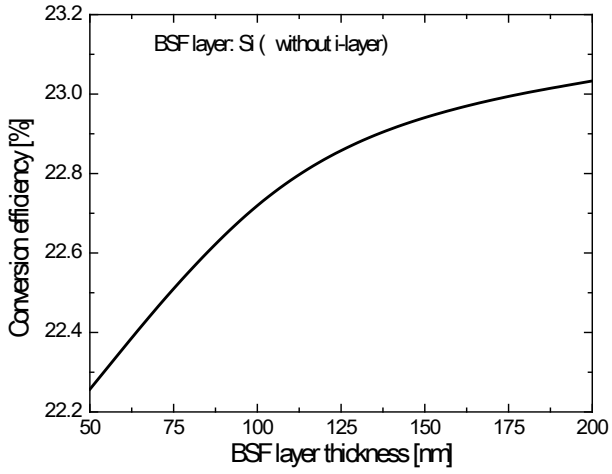
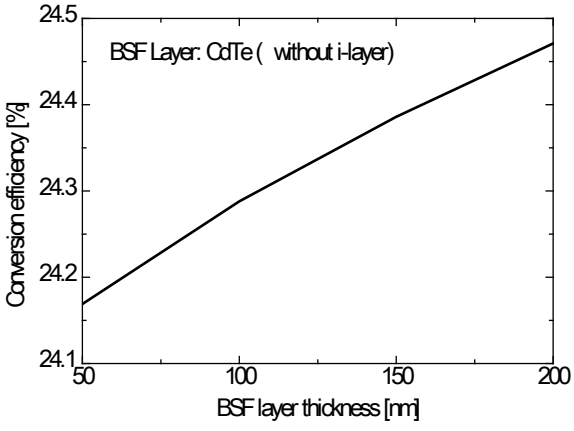


Figure 5 Effects of BSF (Silicon) layer thickness on the conversion efficiency



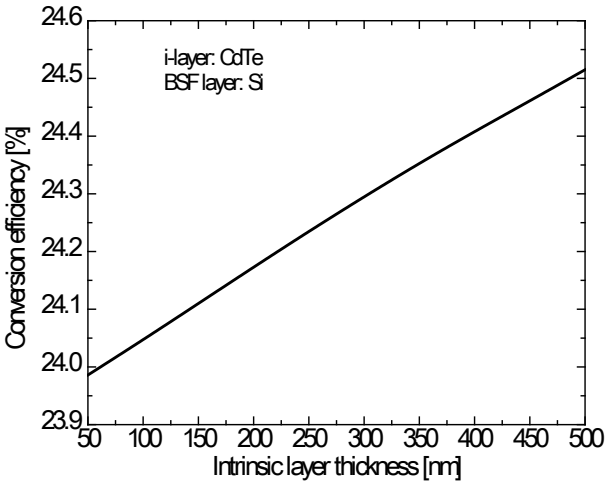


Figure 7 Effects of thickness of intrinsic layer on the conversion efficiency

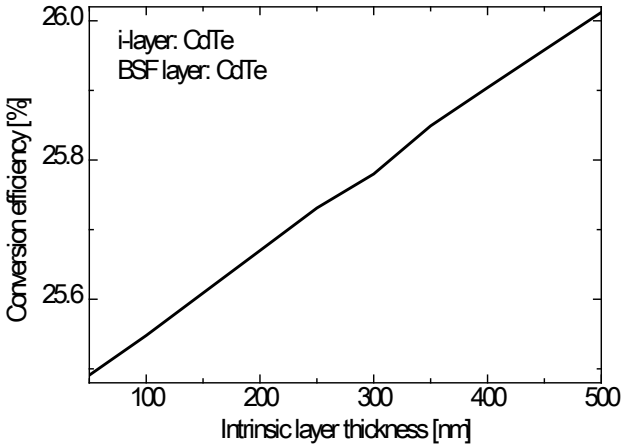


Figure 8 Effects of thickness of intrinsic layer on the conversion efficiency

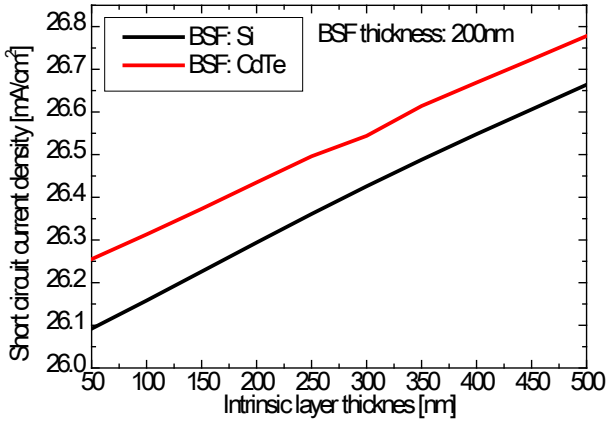


Figure 9 Comparison of current density for Si and CdTe as BSF layer

5. Conclusions

An ultrathin (1.8 μm) p-i-n CdS/CdTe PV cell is designed and simulated using 1D AMPS simulator. The thicknesses of intrinsic layer and BSF layer and the doping density of BSF layer are optimized. It is found that CdTe outperforms Si as a BSF layer. The maximum conversion efficiency of the proposed solar cell is found to be 26.01% taking CdTe as BSF layer. Simulation results revealed that a highly doped p^+ CdTe layer (BSF layer) has significant contribution in achieving such a high efficiency. At 1.5 AM solar irradiance, the proposed cell structure with CdTe as BSF layer achieved an open-circuit voltage of 1.09 V, a short-circuit current density of 26.78 mA/cm^2, and a fill factor of 89% and the corresponding overall conversion efficiency of 26.01%. The same structure with Si as BSF layer achieved an open-circuit voltage of 1.06 V, a short-circuit current density of 26.66 mA/cm^2, and a fill factor of 87%, and the corresponding overall conversion efficiency of 24.51%. Compared to our previous work [1] the cell thickness is reduced by 30% with a sacrifice of 0.73% cell efficiency.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors are thankful to East West University, Dhaka, Bangladesh for financial supports.

References

[1] M. Mofazzal Hossain, Md. Minhaz Ul Karim, S. Banik, Nahid A. Jahan and M. A. Matin, "Design of a high efficiency ultrathin CdTe/CdS p-i-n solar cell with optimized thickness and doping density of different layers," *ICAEESE*, Putrajaya, Malaysia 2016, 14-16 Nov.
DOI: 10.1109/ICAEES.2016.7888058

[2] A. Nishimura, Y. Hayashi, K. Tanaka, M. Hirota, S. Kato, M. Ito, K. Araki and E.J. Hu "Life cycle assessment and evaluation of energy payback time on high-concentration photovoltaic power generation system," Applied Energy **87**, pp. 2797-2807, 2010.
https://doi.org/10.1016/j.apenergy.2009.08.011

[3] G. C. Morris and S. K. Das, "Some fabrication procedures for electrodeposited CdTe solar cells," International Journal of Solar Energy, **12**(1-4), pp. 95-108, 1992.
http://dx.doi.org/10.1080/01425919208909753

[4] T. L Chu, "Thin film cadmium telluride solar cells by two chemical vapor deposition techniques," Solar Cells, **23**, pp. 31-48, 1988.
https://doi.org/10.1016/0379-6787(88)90005-1

[5] S. Ikegami, "CdS/CdTe solar cells by screen-printing-sintering technique: Fabrication, photovoltaic properties and applications," Solar Cells, **23**(1-2), pp. 89-105, 1988.
https://doi.org/10.1016/0379-6787(88)90009-9

[6] T. L Chu and S. S. Chu, "High efficiency thin film CdS/CdTe solar cells," International Journal of Solar Energy, **12**(1-4), pp. 121-132, 1992.
http://dx.doi.org/10.1080/01425919208909755

[7] V. M. Nikale, S. S. Shinde, C. H. Bhosale, and K. Y. Rajpure, "Physical properties of spray deposited CdTe thin films: PEC performance," Journal of Semiconductors, **32**(3), pp. 033001-7. 2011.
http://doi.org/10.1088/1674-4926/32/3/033001

[8] Marple, D. T. F., "Optical Absorption Edge in CdTe: Experimental," Phys. Rev. **150**(2), pp. 728-734, 1966.
https://doi.org/10.1103/PhysRev.150.728

[9] Kim Mitchell1, Alan L. Fahrenbruch and Richard H. Bube, "Photovoltaic determination of optical-absorption coefficient in CdTe," J. Appl. Phys. **48**, pp. 829-830, 1977.
https://doi.org/10.1063/1.323636

[10] Jinqing Peng, Lin Lu, and Hongxing Yang, "Review of life cycle assessment of energy payback and greenhouse gas emission of solar photovoltaic systems," Renewable and Sustainable Energy Reviews **19,** pp.255-274, 2013.
https://doi.org/10.1016/j.rser.2012.11.035

[11] First Solar Inc (2014), http://investor.firstsolar.com/releasedetail.cfm?ReleaseID=828273

[12] NREL, 2001. Available in www.nrel.gov/news/press/2001/1501_record.html

[13] J. Britt and C. Ferekides, "Thin film CdS/CdTe solar cell with 15.8% efficiency," Applied Physics Letters, **62(**22), p. 2851-2852, 1993.
https://doi.org/10.1063/1.109629

[14] Nowshad Amin, M A Matin, M M Aliyu, M A Alghoul, M R Karim and K Sopain, "Prospects of Back Surface Field Effect in Ultra-Thin High-Efficiency CdS/CdTe Solar Cells from Numerical Modeling," Int. J. Photoenergy, **2010**, pp. 1-8, 2010.
http://dx.doi.org/10.1155/2010/578580

[15] Minami, Tadatsugu, "Transparent conducting oxide semiconductors for transparent electrodes," Semiconductor Sci. and Technol. **20**(4), pp. S35-S44, 2005.
https://doi.org/10.1088/0268-1242/20/4/004

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 217-224 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Measurement of the Bridge Surface Deflections Using Near-Field Amplitude of Secondary Radiators System

Oleksandr Poliarus[*,1], Yevhen Poliakov[1], Andrii Lebedynskyi[1], Viktor Ivanov[2], Ruslan Pashchenko[2]

[1]*Department of metrology and life safety, Kharkiv National Automobile and Highway University, 61002, Ukraine*

[2]*O. Ya. Usikov Institute for Radiophysics and Electronics of the National Academy of Sciences of Ukraine, 61085, Ukraine*

ARTICLE INFO

Article history:
Received: 01 November, 2017
Accepted: 17 November, 2017
Online: 15 December 2017

Keywords :
Remote measurement
Radiating element
Antenna directivity pattern
Field amplitude measurement
Surface deflection

ABSTRACT

In remote measurements on large-sized objects main focus is on measuring deformations of bridge surfaces and determination of the dynamic amplification factor in particular. The proposed method estimates deflections of entire lower surface of a bridge due to placement of secondary radiators on this surface. Antenna array directivity pattern (ADP) is distorted during deflection of the bridge surface. ADP distortions contain the information about deflections of the individual points of the surface. Minimization of the distance in a functional space with a quadratic metric (the functional) between a distorted measured ADP and theoretical one allows us to determine these deflections. The main problem is measurement of the secondary radiators system ADP in far zone which often cannot be achieved for real bridges in practical situations. Therefore, a new method for determining surface deflections based only on field amplitude measurements in near-field zone of a secondary radiators system is considered in the article. This amplitude of field strength depends on the bridge surface deflection. The modulus of difference between the normalized amplitude distribution of the measured field at the outputs of the receiving array elements, which is created by the radiators system of an unloaded bridge, and the similar amplitude field distribution of a loaded bridge is minimized and deflection of the bridge at the points of the radiating elements placement is calculated.

1. Introduction

This paper is an extension of the work originally presented in the International Conference on Antenna Theory and Techniques [1].

In engineering practice, an important place is taken by measurements of surface deflections of large-sized objects. The dynamics of these deflections can serve as a basis for diagnosing the objects' state. Deflection measurements for lower part of the surface are one of the most important actions in maintenance of bridges. For this purpose, contact sensors are usually mounted directly on the lower part of the bridge. The installation of contact sensors and transducers interrupts traffic on the bridges and can be time-consuming. Non-contact measurements on bridges which are carried out remotely have recently become an alternative to the contact ones. The analysis of remote method for estimating bridge surface deflections was carried out in [1]. It is based on measurement of the antenna directivity pattern (ADP) distortion of the radiators which are installed on the lower surface of a bridge.

2. Modern Methods of Bridges Surface Deflections Measurement

Let us consider the methods of deflection measurements on bridge structures. The number of deflection measurement methods has increased essentially in the last decade. Bridge safety monitoring is based, in a great extent, on measurements of deflections that become indispensable for this aim. New technologies include intelligent robots that feature a high speed of measurement and precision, laser image deflection measurement, GPS satellite-surveying systems, electro-optical imaging deflection, wireless inclination sensor systems [2, 3] and techniques of digital image correlation between images of undeformed and deformed configurations [4]. In [5], for measuring deflection and vibration of bridges a ground-based interferometric radar is suggested. The maximum measurement range is equal to 1000 m and the errors of displacement radial component can reach

[*]Oleksandr Poliarus, Email : poliarus.kharkov@ukr.net

0.1 mm. Most of the proposed methods are expensive or very expensive.

The main characteristics of any bridge are dynamic amplification factor (DAF), bridge natural frequency and bridge acceleration [6]. The dynamic amplification factor can be evaluated as the maximum total load effect divided by the maximum static effect [7]

$$DAF = \frac{\varepsilon_{dyn}}{\varepsilon_{stat}}, \quad (1)$$

where ε_{dyn} – maximum dynamic response which is usually determined as vertical deflection of the bridge surface due to dynamic loads (for example, moving vehicles), ε_{stat} – maximum deflection under static loads (for example, immovable vehicles).

In [8], an optimal scheme for measuring DAF of a bridge using radar is proposed. This radar measures the phase of reflected electromagnetic wave in noisy environments. In its turn, the phase depends on deflections of the bridge surface.

Increased measurement accuracy can be provided by using measuring devices which operate on the different physical principles. An optimal scheme for determining bridge surface deflection by measuring it simultaneously with a mechanical sensor and radar has been developed in [9]. Practically, operation quality of such scheme does not depend on noise or interference. Since radar-based methods use phase measurements that are sensitive to moving both bridge and earth surfaces, seismic fluctuations of the earth surface reduce the accuracy of the deflections measurement due to traffic, for example. In addition, DAF estimation (1) involves measuring static and dynamic deflections that are separated in time due to technological reasons. This fact imposes strict requirements to stability of radar frequency that increases the cost of a measuring system.

Consequently, all modern measuring technologies for bridge structures can be divided into three groups. The first group includes the use of radar systems of millimeter and optical ranges of electromagnetic waves [10-15] for estimation of deformations and deflections of bridge elements, including stay-cables which support the buildings. For this purpose, in particular, phase interferometers with a high-resolution range are used. These systems measure dynamic characteristics of bridge structures at local points. The second group includes the use of opto-mechanical systems of technical vision [16-18] followed by image processing based on developed algorithms, which allows to analyze the dynamics of bridge structures in distributed areas. These methods do not determine the DAF. The third group is based on application of the algorithms for processing GPS data with Kalman filtration to obtain the value of the bridge DAF [19].

DAF measurement is performed at one point of the bridge structure. To measure it at another point, it is necessary to change the direction of the main maximum of the transmitting and receiving antennas directivity pattern. The existing measuring methods are not able to estimate the dynamics of deflection changes in the entire lower surface of the bridge. In order to fix that, a system for deflection measurements on the lower surface of a bridge by creating, on the surface, an equidistant linear antenna array that re-radiates the incoming electromagnetic wave was proposed in [1] (Figure 1). When the surface is bent due to the process of bridge loading, the radiators also move together with it and the ADP is distorted.

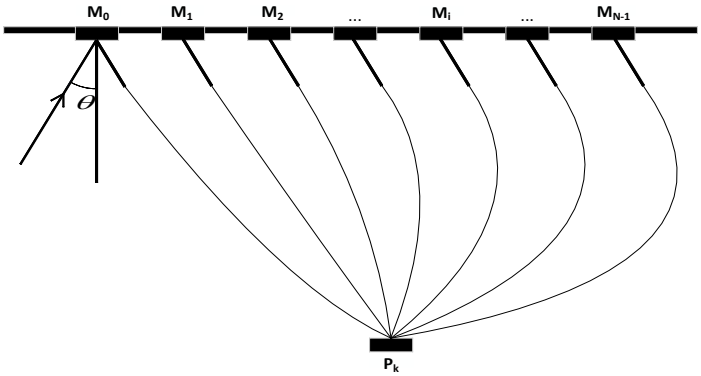


Figure 1: The system of secondary radiators on the lower part of a bridge surface

The antenna directivity pattern is measured in far zone of the antenna and then it is compared with the undisturbed one which was previously measured in the absence of deflections on the unloaded bridge. The integral of square of the difference between the two ADPs creates a functional which depends on the displacements of the secondary radiators at the deformed surface. Minimization of this functional by the method of global random search (using a genetic algorithm) leads to estimation of all the displacements of radiators and, consequently, the deviation of the points on the bridge surface. A similar method is used in [20] for antennas synthesis and even earlier in solving the inverse problem of measurement in linear inertial systems or systems with memory [21].

Let us consider this method. In [1], the authors have developed the functional,

$$J(s_0, s_1, ..s_{N-1}) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(f_m(\theta) - \sum_{i=0}^{N-1} A_i \cdot e^{j\left[\phi_i + \frac{2\pi}{\lambda} l_i \cos\theta - \frac{2\pi}{\lambda} s_i \sin\theta\right]} \right)^2 d\theta, \quad (2)$$

in which s_i, $i = 1, 2, ... N-1$ is a deflection of the surface at the point of the secondary radiator placing, $f_m(\theta)$ – the measured ADP of the radiators system from which the theoretical ADP of the same system, that is described in (2), is subtracted.

For an unloaded bridge $s_i = 0$

$$J(s_0, s_1, ..s_{N-1}) \approx 0$$

$s_i \neq 0$, there may be a difference between the two ADPs of the radiators system. Minimization of the functional (2) using a genetic algorithm results in the determination of unknown deflections s_i. It should be known, however, that formula (2) is the only one and has N unknown deflections. Since the dependence of the functional (2) on the deflections in the N

space is complicated, it is possible to find a non-global, but a local minimum of (2) which leads to incorrect ("phantom") solutions of the deflections determination. One or more of them may have an abnormal meaning. Since the deflections were fixed during the simulation, the "phantom" solutions could easily be distinguished from the correct ones. As shown in [1], when the ratio of the distance d between the radiators to the wavelength λ exceeds 0.5, a certain probability of "phantom" solutions begins to appear. It may reach a value of 0.6 with an increase in the ratio $\frac{d}{\lambda}$ to one and more. The reason for this phenomenon is the appearance of array diffraction maxima and increased number of the local extrema in the search space of genetic algorithm. Consequently, "phantom" solutions can be almost completely eliminated by rational choice of the ratio $\frac{d}{\lambda}$ of the radiating antenna array. In addition, the use of a priori information about surface deflections reduces their probability significantly, even when the ratio $\frac{d}{\lambda}$ is large. This information is often known from previous bridge research or can be obtained by other methods, for example, using the measurements made by mechanical sensors.

It should be noted that instead of passive secondary reflectors, one can use the antenna array of active radiators located at the same points. However, the cost of such system is probably higher.

3. Results of Experimental Research

For experimental verification of this effect, a "Kredo" coherent-pulsed radar with an electromagnetic wave length of 1.8 cm and a pulse duration of 0.4 μs (Figure 2) was used. The experimental setup also included a digital oscilloscope DSO-2250, USB (analogue-digital converter (ADC) with a sampling frequency of 100 MHz) and a laptop. In Figure 2a, a photo of a radar sounding path is shown, in Figure 2b – the placement of the angle reflectors against the background of a reinforced concrete wall, and in Figure 2c – the experimental installation. At a distance of 0.7 m from a reinforced concrete wall and at a height of 1.5 m, the array of five secondary radiators in the form of triangular horns (Figure 2b) was constructed; the distance between horns was equal to 1.5 m. The length of the horn side was 30 cm, and its depth was 15 cm.

The analog radar signals reflected from the horns were received by a mirror antenna, converted to the intermediate frequency and arrived at the input of the amplitude detector. The signal from the output of this amplitude detector entered the ADC of the oscilloscope, was converted into a digital code, and then into a text file (file with extension .txt), which is available in various computer mathematical packages.

To estimate the amplitude of reflected signals, an informative signal (Figure 3) consisting of 200 points was cut out from a long signal (5000 points). It corresponded to the distances at which the angle reflectors were located. In Figure 3, a blanking signal (long, up to a distance of approximately 80 m) which corresponds to the duration of the sounding signal is displayed initially and then a signal (of a shorter duration), that describes the amplitude of the signal reflected from the horns, appears.

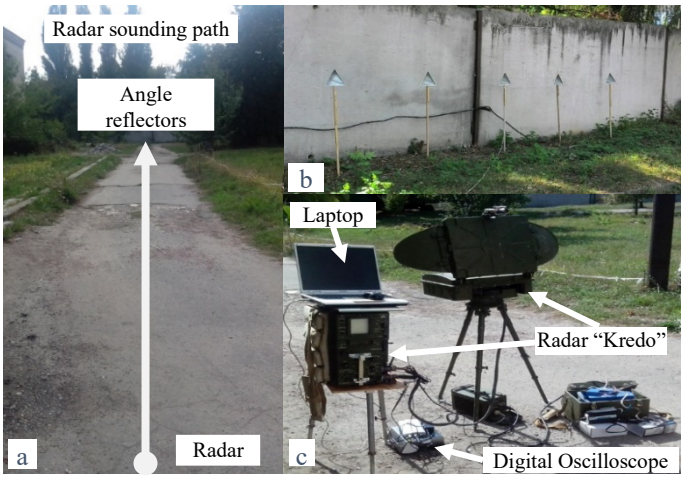


Figure 2: Photographic scheme of the experiment: radar sounding path (a); placement of the angle reflectors (b); general view of the experimental installation (c)

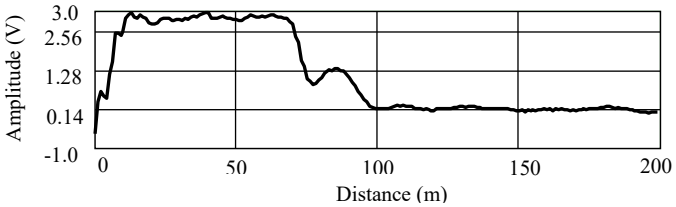


Figure 3: Distance distribution of amplitude of the signal received from angle reflectors

During the experiment the following positions of the angle reflectors were considered: a reinforced concrete wall without reflectors (variant 1); five reflectors on the same line (Figure 2b) (variant 2); one central reflector that was shifted forward by 20 cm (variant 3); two reflectors (central and next to the right) were shifted forward by 20 cm (variant 4); three reflectors (central and two next to the right) were shifted forward by 20 cm (variant 5); one reflector (rightmost) has been moved forward by 10 cm, two reflectors (central and next to it to the right) were shifted by 20 cm forward (variant 6); two rightmost reflectors were shifted forward by 10 cm, the central reflector was shifted by 20 cm forward (variant 7); three reflectors (central and two next to it to the right) were shifted forward by 10 cm (variant 8).

The reflected signals for the above-mentioned variants of the angle reflectors positions were recorded 10 times and then their values were averaged. The obtained experimental data were used to estimate the amplitude of the signals that were reflected from the angle reflectors with different placements relative to each other. The total field amplitude of secondary radiation of the horns at a range of 85 m was measured using the radar. The field amplitude at the radar receiver output was changed if any radiator was shifted in the radial direction (toward the radar location), that shows its sensitivity to movements of the radiators. For example, in the above-described variants, the experimental values in volt of the field amplitude were as follows: 1.16, 1.30, 1.39, 1.51, 1.39, 1.48, 1.40, 1.34. The shape of ADP of the secondary radiators system in experiments was not measured.

4. New Method for Estimating Bridges Surface Deflections

The problem of determination of the bridge surface deflections, described above, is technically complicated. The practical measurement of ADP for the secondary radiators system causes the greatest difficulty. First, we need to measure the field amplitude dependence on the angular coordinate, and it is not easy to do near a real bridge. Secondly, the measurements should be made in the far-zone of antenna at distance r_{fz}, calculated as two squares of antenna size L divided by the wavelength λ. For example, if $L = 30$ m and $\lambda = 30$ cm, then $r_{fz} \geq 6$ km and this value of far-zone distance is impossible for practice. Of course, with increasing wavelength and surface dimensions, whose deformation is measured, there are other problems associated with, for example, the efficiency of radiation of the passive elements of array. Consequently, practical application of the method presented in [1] is limited by the conditions of bridge sounding.

This problem can be solved on the basis of the amplitude-phase $(A_i,\ \phi_i)$ distribution measurement in the near-field zone of the antenna and obtaining the ADP of the radiators system on the basis of these measurements. Such problems were solved by many authors, for example, [22]. In practice, high-precision measurements of the amplitude and phase of the field in real weather and other conditions are complicated and expensive which reduces the value of this approach. Consequently, simpler methods of remote measurements of bridge surface deflections are required. This forces us to give up high-precision measurements of the amplitude and phase in the near-field zone of antenna.

The main idea of this article is to use only amplitude measurements in the near-field zone of the radiators system. The obtained data do not allow to determine the ADP but our task is to estimate only the bridge surface deflections. In this case, there is no need for complex operations of ADP measuring. Instead of these operations, we use a linear receiving antenna consisting of Q receiving elements, for example, horns located at points $P_q, (q = 1, 2, .., Q)$ (Figure 4) at the same distance δ from each other.

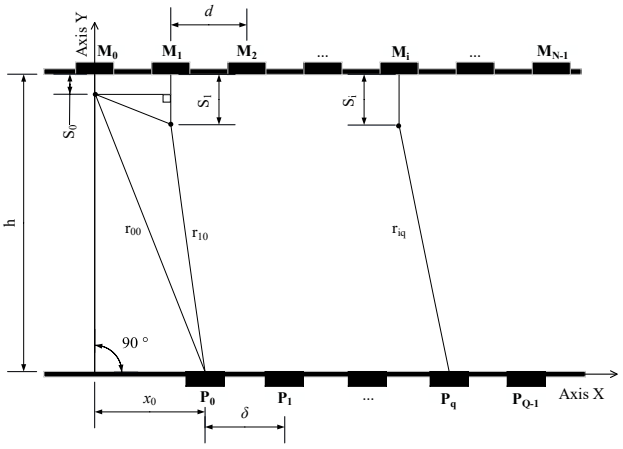


Figure 4: The layout of the receiving antenna array elements (points P_q) relative to the radiating elements position (points M_i)

The receiving element P_q is located from the transmitting element M_i at a distance which is easily determined from the geometry of Figure 4:

$$r_{iq} = \sqrt{(h - S_0)^2 + (x_0 + \delta \cdot q)^2 + (d \cdot i)^2 + (S_i - S_0)^2 - 2\Lambda_{iq} \cdot \Upsilon_{iq}}, \quad (3)$$

where $\Lambda_{iq} = \sqrt{(h - S_0)^2 + (x_0 + q \cdot \delta)^2}\sqrt{(d \cdot i)^2 + (S_i - S_0)^2}$,

$$\Upsilon_{iq} = \cos\left(\frac{\pi}{2} - \operatorname{arctg}\left(\frac{x_0 + q \cdot \delta}{h - S_0}\right) - \operatorname{arctg}\left(\frac{S_i - S_0}{d \cdot i}\right)\right),$$

h – distance between the lower surface of a bridge and array of receiving elements placed in parallel to the transmitting array;

S_i – amplitude of a bridge surface deflection at the points M_i, i. e. in locations of radiators;

x_0 – distance between the origin of the coordinates and the first in order of the receiving element P_0.

The complex amplitude of field strength at the output of each q receiving element is the sum of the fields created by all transmitting elements and is determined by the known formula [23] which for this case has the following form

$$\dot{E}_q = \sum_{i=0}^{N-1} \frac{\dot{A}_i}{r_{iq}} e^{-j\frac{2\pi}{\lambda} r_{iq}}, \quad (4)$$

where $\dot{A}_i = A_i e^{j\phi_i}$; A_i, ϕ_i – amplitude and phase of the field, which is created by the i radiator (further, for simplicity, initial phase field is taken equal to zero, i. e. $\phi_i = 0$).

We believe that all secondary radiators ADPs are equal and wide so that the extreme elements of the transmitting array irradiate all elements of the receiving array by energetically equivalent waves. The amplitude A_i of the field in the simulation was the same. Thus, the distance (3) makes the largest contribution to the total amplitude distribution of field strength. The amplitude of the field created by the secondary radiators at an unloaded bridge is determined similarly (4)

$$\dot{E}^0_q = \sum_{i=0}^{N-1} \frac{\dot{A}_i}{r^0_{iq}} e^{-j\frac{2\pi}{\lambda} r^0_{iq}}, \quad (5)$$

where the upper index "0" indicates variables describing the position of radiators on an unloaded bridge, the surface shape of which is not necessarily a plane. These variables are easy determined before carrying out measurements. Due to the specificity of the problem geometry (Figure 4), the amplitudes of field strength will be different at the output of every receiving element. After the bridge loading the deflections s_i

λ

elements E_q depend only on the distances r_{iq} containing the deflections. The simulation was carried out in several stages:

- Determination of the field amplitudes E_q^0 at the outputs of the receiving elements of the array for an unloaded bridge, provided that the distances r_{iq} are given;
- Measurement of deflections s_i by mechanical sensors. In this article the typical values of bridge deflections were set by authors;
- Determination of the field amplitudes E_q at the outputs of the array receiving elements for a loaded bridge, provided that the distances r_{iq} and the deflections s_i are specified.

Thus, amplitudes of the field at the outputs of receiving elements of the array for an unloaded bridge and a loaded one were obtained in the simulation process for different conditions. It was further assumed that the deflections S_i, which were previously given, are now unknown. To determine these deflections and compare them with the given ones or for this case with reference deflections, a functional was composed:

$$J(s_0, s_1, ..s_{N-1}) = \left| E_q^0 - E_q \right| \Rightarrow \min \qquad (6)$$

It is minimized with the help of a genetic algorithm by variations on S_i. As a result, the "unknown" values of the bridge surface deflections were determined.

5. The Simulation Results

In Figure 5, 6 normalized field amplitudes at the outputs of the receiving antenna array elements located at distances $h = 10$ m (Figure 5) and $h = 50$ m (Figure 6) from the lower surface of a bridge are presented. The length of the electromagnetic wave was $\lambda = 20$ cm. Similar figures for the wavelength $\lambda = 5$ cm are shown in Figures 7, 8.

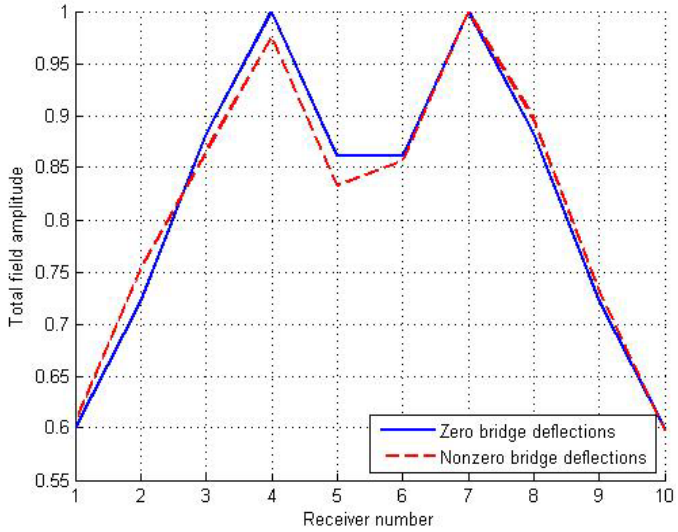


Figure 5: Normalized field amplitude at the outputs of the array receiving elements for an even (solid line) and deformed (dotted line) bridge surfaces for $\lambda = 20$ cm, h=10 m

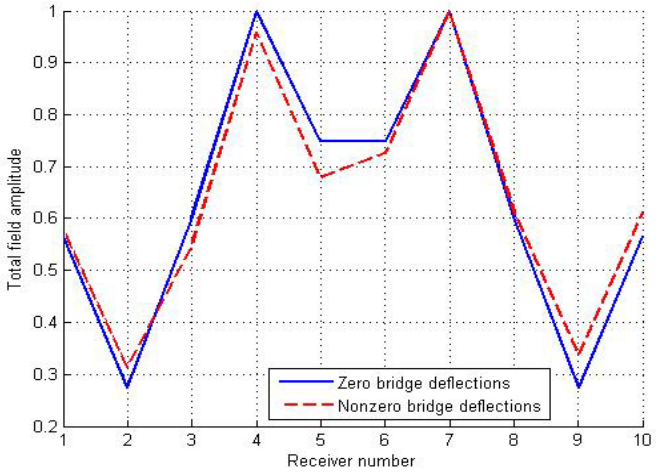


Figure 6: Normalized field amplitude at the outputs of the array receiving elements for an even (solid line) and deformed (dotted line) bridge surfaces for $\lambda = 20$ cm, h=50 m

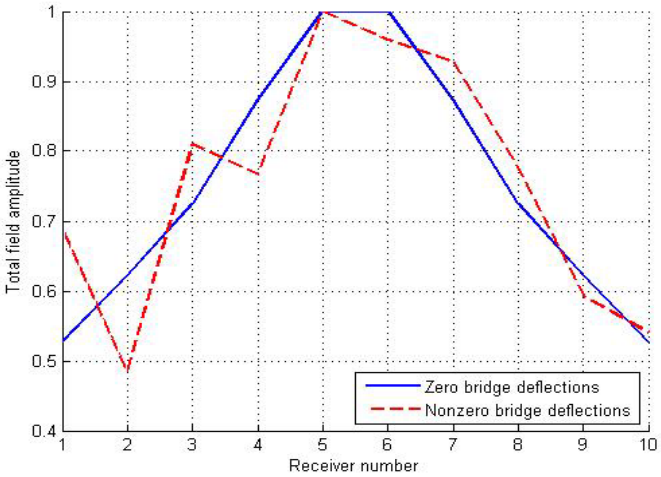


Figure 7: Normalized field amplitude at the outputs of the array receiving elements for an even (solid line) and deformed (dotted line) bridge surfaces for $\lambda = 5$ cm, h=10 m

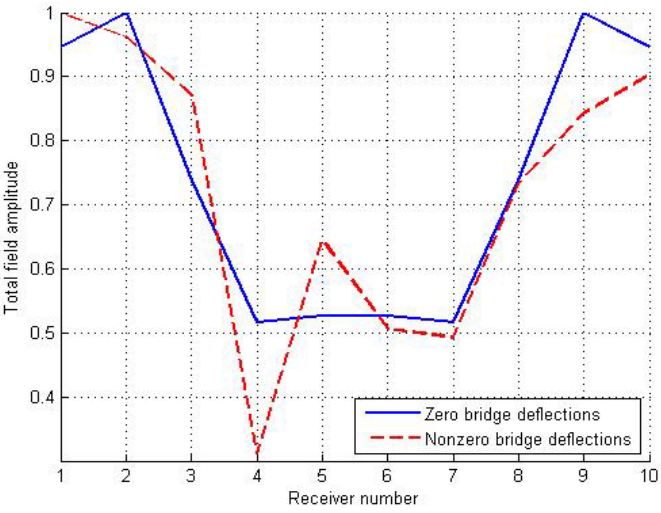


Figure 8: Normalized field amplitude at the outputs of the array receiving elements for an even (solid line) and deformed (dotted line) bridge surfaces for $\lambda = 5$ cm, h=50 m

The genetic algorithm in the process of its work "deforms" the dotted line in such a way that it takes a place of the continuous line. Since the latter is obtained for an unloaded bridge, the algorithm actually reduces existing deflections s_i almost to zero and at the same time determines their absolute values. In this case, the errors of these deflections determination arise. For modeling, the bridge surface deflections in millimeters were selected by authors as follows: $s_0 = 0$, $s_1 = 5$, $s_2 = 4$, $s_3 = 7$, $s_4 = 6$, $s_5 = 4$, $s_6 = 2$, $s_7 = 3$, $s_8 = 4$, $s_9 = 2$. The profiles of the deformed bridge surfaces and the profiles of the reconstructed surfaces that were determined by the genetic algorithm, are shown in Figure 9. Between them, there is some discrepancy which was interpreted as an error of deformed surface restoration.

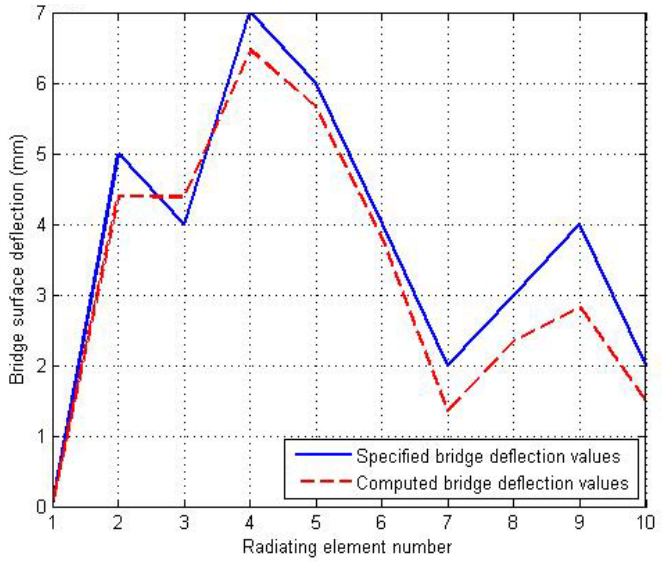


Figure 9: The profile of the deformed bridge surface (solid line) and the profile of the surface that was determined by genetic algorithm (dotted line)

The dependence of relative errors of surface profile restoration on the number of receiving elements in array for different wavelengths is shown in Figure 10.

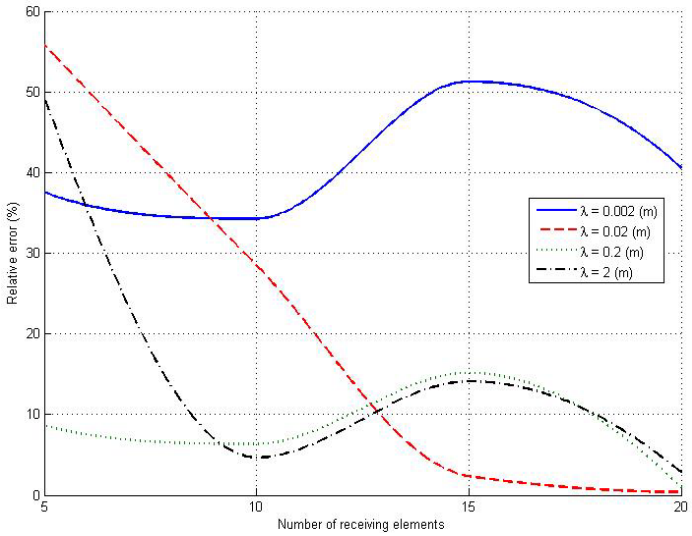


Figure 10: Dependences of the relative errors of the bridge deformed surface profile restoration on the number of receiving elements in array for different wavelengths

In the centimeter and decimeter range of wavelengths the relative errors of bridge surface profile determination are decreased with the increase in array receiving elements number and when the number 10 of receiving elements is exceeded, these errors are less than 10%. This quantity is limited in practice by reception conditions, because the method requires that all elements must receive the radiation from all transmitting elements. Relative error also depends on the wavelength which is used for radiation (Figure 11). The number of receiving elements is equal to 10.

For each bridge the dependence (Figure 11) will be different but there always is a wavelength range which provides the minimum error of the bridge surface deflection determination.

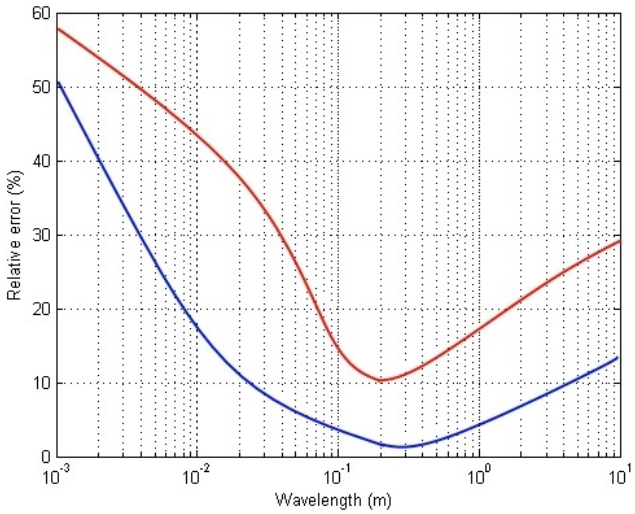


Figure 11: Dependences of the relative errors of the bridge deformed surface profile restoration on the wavelength: blue line – $h = 50$ m, $\delta = 3$ m; red line – $h = 100$ m, $\delta = 2$ m

Dependence of the relative errors of the bridge surface profile restoration on signal / noise ratio is represented in Figure 12. A deterministic signal describing the field amplitude and a white noise of different intensity were added in simulation. In this case the work of the genetic algorithm for finding the global minimum became more complicated.

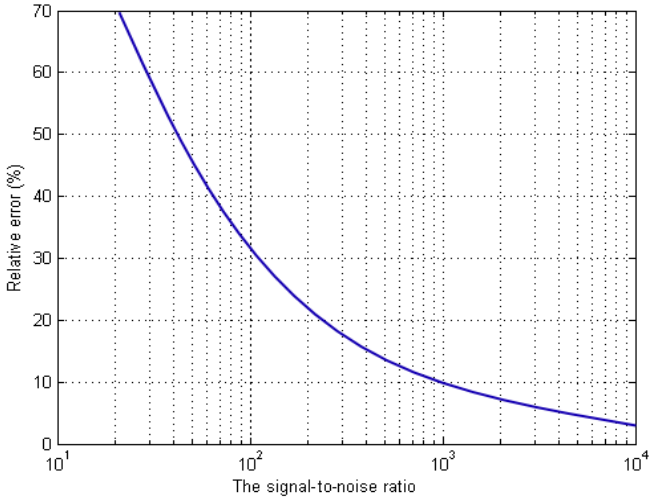


Figure 12: Dependence of the relative error of the bridge deformed surface restoration on the signal-to-noise ratio

The dependence of the "phantom" solutions probability on the ratio $\frac{d}{\lambda}$, that is, on the electrical distance between the radiators, is shown in Figure 13. Figure 14 demonstrates similar dependence on the number of receiving elements. The wavelength was 20 cm, and the distance between the lower surface of the bridge and the receivers was 50 m.

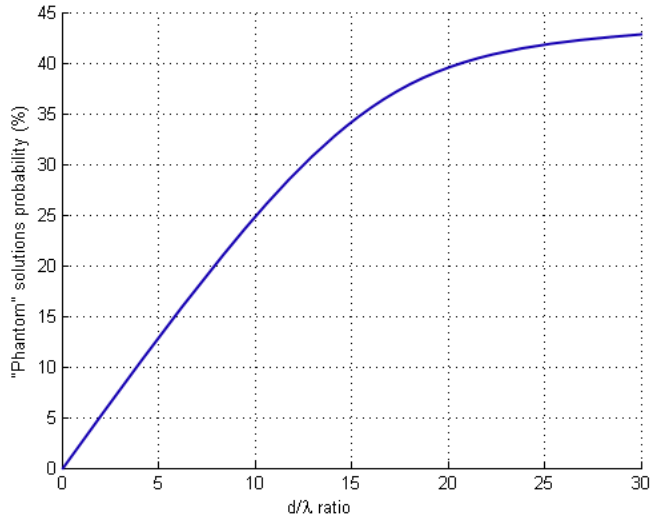


Figure 13: Dependence of "phantom solutions" probability on the ratio $\frac{d}{\lambda}$

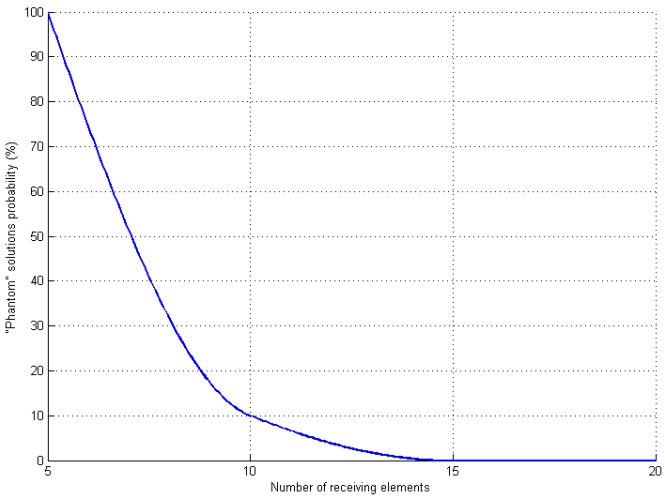


Figure 14: Dependence of "phantom solutions" probability on the number of receivers

The solution was considered as "phantom" one if the error of the bridge deformed surface restoration exceeded 10%. Figures 13 and 14 impose requirements for the number of radiators and receiving elements of the appropriate antenna arrays. Increased number of secondary radiators not only improves the quality of the curved surface profile measurement, but also significantly reduces the probability of the "phantom" solutions.

In the article, the problem of bridge surface deflections determination was solved for simplified conditions in the absence of radio waves reflection from the concrete part of the bridge surface and other parts, neglecting the influence of the steel elements of the reinforced concrete structure, etc. However, one should expect that the approach proposed in the article will not change. Indeed, in the process of the field amplitude measuring at the outputs of receiving elements of the array at an unloaded bridge, all the influencing factors which remain at a loaded bridge, will be taken into account. The difference between these fields appears only in the presence of bridge deflections in the second case. Hence the functional (6) will probably remain the same, that is, the method is expedient to implement in real conditions, too. However, the errors estimation for bridge surface deflections restoration in such conditions remains a challenging problem.

6. Conclusions

New methods for measuring surface deflection of bridges and other large objects were discussed in the paper. It is based on the processing of measurement information about the electromagnetic field amplitude which can be obtained from systems of secondary (passive) radiators (reradiators) or active radiators installed on the lower part of the bridge. The measuring information is obtained at the output of the receiving elements of a linear antenna array which can be installed at any convenient distance from the lower part of the bridge. Minimization of the functional developed in the study allows us to obtain numerical data on the movement of radiators, as well as the deflection of the entire surface of the bridge respectively. The paper defines the conditions under which the errors of deflections restoration in the process of the functional minimizing do not exceed the permissible ones. They depend on the electromagnetic wave length, the geometry of placement of the radiators and receiving elements, as well as on the level of noise in the measurement location. The application of a global random search of the functional extremum (using genetic algorithms) for the global minimum determination in certain situations leads to appearance of wrong ("phantom") solutions when the genetic algorithm finds the minimums that are not global. Recommendations to limit the probability of obtaining "phantom" solutions to a value of zero were also discussed. A mathematical model of electromagnetic waves scattering was made according to a simplified scheme, which did not take into account the reflection from the concrete surface of the bridge and other elements. The consideration of these factors will not make any significant changes to the principle of surface deflection determination, since in any case surface deflections certainly lead to a change in the field amplitude in the observation point, and these changes are the basis for determining deflections. The sensitivity of the field amplitude to the displacement of secondary radiators is proved in the paper experimentally. The proposed method can be used for determination of deflections of other large objects.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Poliarus O. V., Poliakov E. O., Brovko Ja. S. "Measurement method of the object surface deflections using antennas theory". - 2017 XI International Conference on Antenna Theory and Techniques ICATT'2017, May 24-27, 2017. - Kyiv, Ukraine, pp. 416-418.

[2] Yan Yu, Hang Liu, Dongsheng Li, Xingquan Mao, Jinping Ou. "Bridge Deflection Measurement using Wireless MEMS Inclination Sensor Systems". – International Journal on Smart Sensing and Intelligent Systems, 2013, vol. 6, №1, pp. 38…57.

[3] Xuefeng Zhao, Hao Liu, Yan Yu, Xiaodong Xu, Weitong Hu, Mingchu Li. "Bridge Displacement Monitoring Method Based on Laser Projection". – Sensing Technology. – Sensors, 2015, 15, pp. 8444…8463.

[4] Satoru Yoneyama and Hiroki Ueda. "Bridge Deflection Measurements Using Digital Image Correlation with Camera Movement Correction". – Material Transactions, 2012, vol. 53, №2, pp. 285…290.

[5] P. Kuras, T. Owerko, Ł. Ortyl, R. Kocierz, P. Kohut, K. Holak, K. Krupiński. Comparison of methods for measuring deflection and vibration of bridges. - Proceeding of the Joint international symposium on Deformation monitoring: 2–4 November 2011, Hong Kong, China, 2011, pp. 1..8.

[6] Pacglite, A. Pacglitis. "The Dynamic Amplification Factor of the Bridges in Latwia". – Journal Procedia Engineering, 2013, vol. 10, pp. 851…858.

[7] Brady S. P., O'Brien J. O., Znidaris A. "Effect of vehicle velocity on the dynamic amplification of a vehicle crossing a simply supported bridge". – Journal of Bridge Engineering, 11 (2), pp. 241…249.

[8] Poliarus O. V., Barchan V. V. "Radio engineering system for estimating of the dynamic amplification factor of bridge structure". – Kharkiv: Applied electronics, 2011, vol. 10, №3, pp. 368…371. (in original: Полярус О. В., Барчан В. В. Радіотехнічна система оцінки коефіцієнта динамічності мостової конструкції. – Харьков: Прикладная радиоэлектроника, 2011, т. 10, №3.)

[9] Poliarus O. V., Barchan V. V., Poliakov E. O. "Method of accuracy increasing of vibrational displacements measurements for a machine on the basis of measuring system integration". – Kharkiv: Technology of instrument making, 2012, №1, c. 25…29. (in original: Полярус О. В., Барчан В. В., Поляков Є. О. Методика підвищення точності вимірювань вібропереміщень машини на основі комплексування вимірювачів. – Харьков: Технология приборостроения, 2012, №1, с. 25…29)

[10] Carmelo Gentile. "Application of Microwave Remote Sensing to Dynamic Testing of Stay-Cables". - Remote Sensing, 2010, vol. 2, pp. 36-51.

[11] Patrick Paultre, Omar Chaallal, Jean Proulx. "Bridge dynamics and dynamic amplification factors – a review of analytical and experimental findings". – Can. J. Civ. Eng., 1992, №19, pp. 260-278.

[12] Cheol-Hwan Kim, Byung-Wan and Jin-Taek. "Application of Laser Vibrometer to the Measurement and Control of Cable Tensile Forces in Cable-Stayed Bridges". - International Journal of Distributed Sensor Networks, 2012, vol. ID 810682, 7 pages.

[13] Charles R. Farrar, Timothy W. Darling, Albert Migliori and William E. Baker. Microwave Interferometers for non-contact vibration measurements on large structures. - Los Alamos National Laboratory, 2000. – 14 p.

[14] Larry D. Olson, P. E. Olson. Recent advances in NDE and SHM of bridge superstructure with sonic and radar methods. – FHWA 2010, Nondestructive Evaluations of Bridges, 2010, pp. 1-12.

[15] Mayer, L., Yanev, B., Olson, L.D. and Smyth, A.W. "Monitoring of the Manhattan Bridge for Vertical and Torsional Performance with GPS and Interferometric Radar Systems", Proceedings CD, Transportation Research Board 89th Annual Meeting, 2010.

[16] Piotr Olaszek. "Investigation of the dynamic characteristic of bridge structures using a computer vision method".- Measurement, 1999, vol. 25, pp. 227–236.

[17] Shien Ri, Tatsuro Numayama, Masumi Saka, Kenichi Nanbara and Daisuke Kobayashi. "Noncontact Deflection Distribution Measurement for Large-Scale Structures by Advanced Image Processing Technique" - Materials Transactions, Vol. 53, No. 2, 2012, pp. 323…329.

[18] "Displacement Measurement of Burr Arch-Truss Under Dynamic Loading Based on Image Processing Technology". - 6th International Conference on Advances in Experimental Structural Engineering 11th International Workshop on Advanced Smart Materials and Smart Structures Technology August 1-2, 2015, University of Illinois, Urbana-Champaign, United States.

[19] R. Kaloop Mosbeh, Li Hui. "Analysis of bridge movements and damage based on GPS monitoring technique". – Интерэкспо Гео-Сибирь, 2010, вып. 1, т. 1, с.1-5.

[20] Poliarus O. V., Lebedynskiy A. V. "Methods of studying the antennas synthesis problem in practical lessons". – X Anniversary International Conference on Antenna Theory and Techniques ICATT'2015, April 21-24, 2015.-Kharkiv, Ukraine. – pp. 87-89.

[21] Полярус О. В., Поляков Є. О. "Наближене розв'язання оберненої задачі вимірювань та його метрологічне забезпечення": монографія . - Х. : Видавництво «Лідер», 2014. – 120 с. (in Ukrainian)

[22] Hirokazu Kobayashi, Yoshio Yamaguchi, Yi Cu. "Simple Near-field to Far-field Transformation Method Using Antenna Array-factor". - Journal of Wireless Networking and Communications, 2012, №2(4), pp. 43-48.

[23] Harish A. R., Sachidananda M. "Antennas and Wave Propagation". – Oxford University Press, 2007. – 411 p.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 218-224 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Domain Independent Feature Extraction using Rule Based Approach

Sint Sint Aung[*,1], Myat Su Wai[2]

[1]*Department of Academic Affairs, University of Computer Studies, Mandalay, +95, Myanmar*

[2]*Web Mining Lab, University of Computer Studies, Mandalay, +95, Myanmar*

A R T I C L E I N F O

Article history:
Received: 16 November, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords:
Sentiment Analysis
Features Extraction
Opinions

A B S T R A C T

Sentiment analysis is one of the most popular information extraction tasks both from business and research prospective. From the standpoint of research, sentiment analysis relies on the methods developed for natural language processing and information extraction. One of the key aspects of it is the opinion word lexicon. Product's feature from online reviews is an important and challenging task in opinion mining. Opinion Mining or Sentiment Analysis is a Natural Language Processing and Information Extraction task that identifies the user's views or opinions. In this paper, we developed an approach to extract domain independent product features and opinions without using training examples i.e, lexicon-based approach. Noun phrases are extracted using not only dependency rules but also textblob noun phrase extraction tool. Dependency rules are predefined according to dependency patterns of words in the sentences. StandfordCoreNlp Dependency parser is used to identify the relations between words. The orientation of words is classified by using lexicon-based approach. According to the experimental results the system gets good performance in six different domains.

1. Introduction

Sentiment analysis is one of the most popular information extraction tasks both from business and research prospective. It has numerous business applications, such as evaluation of a product or company perception in social media [1]. From the standpoint of research, sentiment analysis relies on the methods developed for natural language processing and information extraction. One of the key aspects of it is the opinion word lexicon. Opinion words are such words that carry opinion. Positive words refer to some desired state, while negative words to some undesired one. For example, "good" and "beautiful" are positive opinion words, "bad" and "evil" are negative [2].

Opinion phrases and idioms exist as well. Many opinion words depend on context, like the word "large". Some opinion phrases are comparative rather than opinionated, for example "better than". Auxiliary words like negation can change sentiment orientation of a word[3].

[*]Corresponding Author: Sint Sint Aung, Email: ssaung@gmail.com

Opinion words are used in a number of sentiment analysis tasks. They include document and sentence sentiment classification, product features extraction, subjectivity detection etc. [4]. Opinion words are used as features in sentiment classification. Sentiment orientation of a product feature is usually computed based on the sentiment orientation of opinion words nearby [5]. Product features can be extracted with the help of phrase or dependency patterns that include opinion words and placeholders for product features themselves. Subjectivity detection highly relies on opinion word lists as well, because many opinionated phrases are subjective in [6]. Thus, opinion lexicon generation is an important sentiment analysis task. Detection of opinion word sentiment orientation is an accompanying task.

Opinion lexicon generation task can be solved in several ways. The authors of [7] point out three approaches: manual, dictionary-based and corpus-based. The manual approach is precise but time-consuming. The dictionary based approach relies on dictionaries such as WordNet. One starts from a small collection of opinion words and looks for their synonyms and antonyms in a dictionary [8]. The drawback of this approach is that the dictionary coverage

is limited and it is hard to create a domain-specific opinion word list. Corpus-based approaches rely on mining a review corpus and use methods employed in information extraction. The approach proposed in [9] is based on a seed list of opinion words. These words are used together with some linguistic constraints like "AND" or "OR" to mine additional opinion words.

Clustering is performed to label the mined words in the list as positive and negative. Part of speech patterns are used to populate the opinion word dictionary in and Internet search statistics is used to detect semantic orientation of a word. In [10], the authors extend the mentioned approaches and introduces a method for extraction of context-based opinion words together with their orientation. Classification techniques are used in [11] to filter out opinion words from text. The approaches described were applied in English. There are some works that deal with Russian. For example, in [12], the author proposed to use classification. Various features, such as word frequency, weirdness, and TF-IDF are used there.

Most of the research done in the field of sentiment analysis relies on the presence of annotated resources for a given language. However, there are methods which automatically generate resources for a target language, given that there are tools and resources available in the source language. Different approaches to multilingual subjectivity analysis are studied and are summarized in [13].

In one of them, subjectivity lexicon in the source language is translated with the use of a dictionary and employed for subjectivity classification. This approach delivers mediocre precision due to the use of the first translation option and due to word lemmatization. Another approach suggests translating the corpus. This can be done in three different ways: translating an annotated corpus in the source language and projecting its labels; automatic annotation of the corpus, translating it and projecting the labels; translating the corpus in the target language, automatic annotation of it and projecting the labels. Language Weaver 1 machine translation was used on English-Roman and English-Spanish data. Classification experiments with the produced corpora showed similar results. They are close to the case when test data is translated and annotated automatically. This shows that machine translation systems are good enough for translating opinionated datasets. In [14], the authors also confirmed when they used Google Translate 2, Microsoft Bing Translator 3 and Moses 4.

Multilingual opinion lexicon generation is considered in [that presents a semi-automatic approach with the use of triangulation. The authors use high-quality lexicons in two different languages and then translate them automatically into a third language with Google Translate. The words that are found in both translations are supposed to have good precision. It was proven for several languages including Russian with the manual check of the resulting lists. The same authors collect and examine entity-centered sentiment annotated parallel corpora [15].

The process of automatic extraction of knowledge by means of opinion of others about some particular product, topic or problem. Opinion mining is also called sentiment analysis due to large volume of opinion which is rich in web resources available online. Analyzing customer review is most important, it tend to rate the product and provide opinions for it which is been a challenging problem today. Opinion feature extraction is a sub problem of opinion mining, with the vast majority of existing work done in the product review domain. Main fields of research in sentiment analysis are Subjectivity Detection, Sentiment Prediction, Aspect based Sentiment Summarization, Text summarization for opinions, Contractive viewpoint, Summarization, Product Feature Extraction, Detecting opinion spam [16].

2. Related Works

Many researchers have addressed the problem of constructing subjective lexicon for different languages in recent years. In [17] to compile a subjective lexicon, the author investigated three main approaches and they are outlined in this section.

knowledge to extract the domain-specific sentiment lexicon based on constrained label propagation. According to [18], the authors had divided the whole strategy into six steps. Firstly, detected and extracted domain- specific sentiment terms by combining the chunk dependency parsing knowledge and prior generic sentiment lexicon. To refine the sentiment terms some filtering and pruning operations were carried out by others. Then they selected domain-independent sentiment seeds from the semi-structured domain reviews which had been designated manually or directly borrowed from other domains. As the third step, calculated the semantic associations between sentiment terms based on their distribution contexts in the domain corpus. For this calculation, the point-wise mutual information (PMI) was utilized which is commonly used in semantic linkage in information theory. Then, they defined and extracted some pair wise contextual and morphological constraints between sentiment terms to enhance the associations. The conjunctions like "and" and "as well as" were considered as the direct contextual constraints whereas "but" was referred to as a reverse contextual constraint. The above constraints propagated though out the entire collection of candidate sentiment terms. Finally, the propagated constraints were incorporated into label propagation for the construction of domain-specific sentiment lexicon. In [19], the authors proposed approach showed an accuracy increment of approximately 3% over the baseline methods. Opinion analysis has been studied by many researchers in recent years. Two main research directions are sentiment classification and feature-based opinion mining. Sentiment classification investigates ways to classify each review document as positive, negative, or neutral. Representative works on classification at the document level include. These works are different from ours as we are interested in opinions expressed on each product feature rather than the whole review.

In [20], ssentence level subjectivity classification is studied, which determines whether a sentence is a subjective sentence (but may not express a positive or negative opinion) or a factual one. Sentence level sentiment or opinion classification is studied in. Our work is different from the sentence level analysis as we identify opinions on each feature. A review sentence can contain multiple features, and the orientations of opinions expressed on the features can also be different, e.g., "the voice quality of this phone is great and so is the reception, but the battery life is short." "voice quality", "reception" and "battery life" are features. The opinion on "voice quality", "reception" are positive, and the opinion on

"battery life" is negative. Other related works at both the document and sentence levels include those in [21].

Most sentence level and even document level classification methods are based on identification of opinion words or phrases. There are basically two types of approaches: (1) corpus-based approaches, and (2) dictionary-based. approaches. Corpus-based approaches find co-occurrence patterns of words to determine the sentiments of words or phrases, e.g., the works in [22].

In [23], the authors proposed the idea of opinion mining and summarization. It uses a lexicon-based method to determine whether the opinion expressed on a product feature is positive or negative. In [24] and [25]. these methods are improved by a more sophisticated method based on relaxation labeling. We will show in Section 5 that the proposed technique performs much better than both these methods. In [26], a system is reported for analyzing movie reviews in the same framework. However, the system is domain specific. Other recent work related to sentiment analysis includes in. In [27], the authors studied the extraction of comparative sentences and relations, which is different from this work as we do not deal with comparative sentences in this research.

Our holistic lexicon-based approach to identifying the orientations of context dependent opinion words is closely related to works that identify domain opinion words . In [28], the authors used conjunction rules to find such words from large domain corpora. In [29], the conjunction rule basically states that when two opinion words are linked by "and" in a sentence, their opinion orientations are the same. For example, in the sentence, "this room is beautiful and spacious", both "beautiful" and "spacious" are positive opinion words. Based on this rule or language convention, if we do not know whether "spacious" is positive or negative, but know that "beautiful" is positive, we can infer that "spacious" is also positive. Although our approach will also use this linguistic rule or convention, our method is different in two aspects. First, we argue that finding domain opinion words is still problematic because in the same domain the same word may indicate different opinions depending on what features it is applied to. For example, in the following review sentences in the camera domain, "the battery life is very long" and "it takes a long time to focus", "long" is positive in the first sentence, but negative in the second. Thus, we need to consider both the feature and the opinion word rather than only the opinion word as in [30].

Opinion target and opinion word extraction are not new tasks in opinion mining. There is significant effort focused on these tasks. They can be divided into two categories: sentence-level extraction and corpus-level extraction according to their extraction aims. In sentence-level extraction, the task of opinion target/word extraction is to identify the opinion target mentions or opinion expressions in sentences. Thus, these tasks are usually regarded as sequence-labeling problems. Intuitively, contextual words are selected as the features to indicate opinion targets/words in sentences. Additionally, classical sequence labeling models are used to build the extractor, such as CRFs and HMM. In [31], the authors proposed a lexicalized HMM model to perform opinion mining. In [32], the authors used CRFs to extract opinion targets from reviews. However, these methods always need the labeled data to train the model. If the labeled training data are insufficient or come from the different domains than the current texts, they would have unsatisfied extraction performance. Although in [33], the authors proposed a method based on transfer learning to facilitate cross- domain extraction of opinion targets/words, their method still needed the labeled data from out-domains and the extraction performance heavily depended on the relevance between in-domain and out-domain.

Although many target extraction methods exist, we are not aware of any attempt to solve the proposed problem. According to [34], although in supervised target extraction, one can annotate entities and aspects with different labels, supervised methods need manually labeled training data, which is time-consuming and labor-intensive to produce. Note that relaxation labeling was used for sentiment classification in, but not for target classification.

3. Proposed method

This section presents the detailed of step by step process about the system.

3.1. Preprocessing the Input sentences

Input sentences with xml file are prepared before parsing to the StanfordCoreNLP parser. Xml tag are removed. ASCII code characters are replaced with Unicode characters because StanfordCoreNLP cannot process non-Unicode characters.

3.2. Rules for Features and Opinions Extraction

In this section, we describe how to extract opinion and product features using extraction rules. They are the most important tasks for text sentiment analysis, which has attracted much attention from many researchers. Based on the relations between features and opinions, there are four main rules in the double propagation;

1. extracting features using opinion words
2. extracting features using the extracted features
3. extracting opinion words using the extracted features
4. extracting opinion words using both the given and the extracted opinion words

In the following extraction rules, O is opinion word, H is the third word, {O} is a set of seed opinion lexicon, F is product feature, and O-Dep is part-of-speech information and dependency relations. {JJ}, {VB} and {NN} are sets of POS tags of potential opinion words and features, respectively. And {DR} contains dependency relations between features and opinions such as mod, pnmod, subj, s, obj, obj2, conj. We used rule 1 and 2 to extract features, and rule 3 and 4 use to extract opinion words. Moreover, we also used some additional patterns to extract features and opinions.

R11: If a word F whose POS is NN is directly depended by an opinion word O through one of the dependency relations mod, pnmod, subj, s, obj and obj2, then F is a feature. It can be defined as follows;

$$O \rightarrow O\text{-}Dep \rightarrow F$$

$$F \rightarrow F\text{-}Dep \rightarrow O$$

such that O ∈ {O}, O-Dep and F-Dep ∈ {DR}, where {DR} = {mod, pnmod, subj, s, obj, obj2, desc} and P OS (T) ∈ {N N}. For eg. Overall a sweet machine.

R12: If an opinion word O and a word F, whose POS is NN, directly depend on a third word H through dependency relations except conj, then F is a feature. It can be expressed as follows;

$$O \rightarrow O\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$$

such that O ∈ {O}, O-Dep and F-Dep ∈ {DR}, POS(F) ∈ {NN}. For eg. Canon is the great product.

R13: If a word F whose POS is NN is indirectly depended by an opinion word O through another word H through two dependency relations dobj and amod or nmod:poss, then F is a feature. It can also be expressed as follows;

$$O \rightarrow O\text{-}Dep \rightarrow H \rightarrow F\text{-}Dep \rightarrow F$$

$$O \leftarrow O\text{-}Dep \leftarrow H \leftarrow F\text{-}Dep \leftarrow F$$

such that O ∈ {O}, O-Dep ∈ {DR}, F-Dep ∈ {DR}, POS(F) ∈ {NN}. For eg. I like the computer’ s battery.

R21: If a word Fj, whose POS is NN, directly depends on a feature Fi through conj, then Fj is a feature. It can also be expressed as follows;

$$Fi \rightarrow Fi - Dep \rightarrow Fj$$

such that Fi ∈ {F}, Fj-Dep ∈ {CONJ}, POS(Fj) ∈ {NN}. For eg. Overall, I like the system features and performance.

R22: If a word Fj, whose POS is NN, and a feature Fi, directly depend on a third word H through the same dependency relation, then Fj is a feature. It can also be expressed as follows;

$$Fi \rightarrow Fi\text{-}Dep \rightarrow H \leftarrow Fj\text{-}Dep \leftarrow Fj$$

such that Fi ∈ {F}, Fi-Dep ∈ {DR}, Fj-Dep ∈ {DR}, POS(Fj) ∈ {NN}. For eg. Canon has done an excellent job.

R31: If a word O whose POS is JJ or VB directly depends on a feature F through one of the dependency relations mod, pnmod, subj, s obj, obj2 and desc, then O is an opinion word. It can also be expressed as follows;

$$O \rightarrow O\text{-}Dep \rightarrow F$$

$$F \rightarrow F\text{-}Dep \rightarrow O$$

such that F ∈ {F}, O-Dep and F-Dep ∈ {DR}, POS(O) ∈ {JJ, VB}. For eg. Overall a sweet machine.

R32: If a word O whose POS is JJ or VB and a feature F directly depend on a third word H through dependency relations except conj, then O is an opinion word. It can also be expressed as follows;

$$O \rightarrow O\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$$

such that F ∈ {F}, O-Dep and F-Dep ∈ {DR}, POS(O) ∈ {JJ, VB}. For eg. Canon is the great product.

R33: If a word O whose POS is JJ or VB indirectly depends on a feature F through another word H through dependency relations dobj, amod and nmod:poss, then O is an opinion word. It can also be expressed as follows;

$$O \rightarrow O\text{-}Dep \rightarrow H \rightarrow F\text{-}Dep \rightarrow F$$

$$O \leftarrow O\text{-}Dep \leftarrow H \leftarrow F\text{-}Dep \leftarrow F$$

such that F ∈ {F}, O-Dep and F-Dep ∈ {DR}, POS(O) ∈ {JJ}. For eg. I like the computer’ s battery.

R41: If a word Oj, whose POS is JJ or VB, directly depends on an opinion word Oi through dependency relation conj, then Oj is an opinion word. It can also be expressed as follows;

$$Oi \rightarrow Oi\text{-}Dep \rightarrow Oj$$

such that Oi ∈ {O}, Oi-Dep ∈ {CONJ}, POS(Oj) ∈ {JJ, VB}. For eg. Nice and compact.

R42: If a word Oj, whose POS is JJ or VB, and an opinion word Oi, directly depend on a third word H through the same dependency relation, then Oj is an opinion word. It can also be expressed as follows;

$$Oi \rightarrow Oi\text{-}Dep \rightarrow H \leftarrow Oj\text{-}Dep \leftarrow Oj$$

such that Oi ∈ {O}, Oi-Dep == Oj-Dep, POS(Oj) ∈ {JJ, VB}. For eg. The screen size and screen quality is amazing.

3.3. Features and Opinion Extraction

In this section This system takes raw data as input and xml tag and ascii code characters are removed. After that, word tokenization, part-of speech tagging and dependency identification between words are done by using StandfordfordCoreNLP dependency parser. We used the algorithm also from [17]. Table 1 shows some examples of English stop word list.

Table 1. Some English Stopwords List

to	Of	I	Me	My
Mine	You	At	They	In
Which	With	On	Under	Below
Above	Thing	Things	Some	Someone
Sometime	Something	Somebody	No one	nobody

To start the extraction process, a seed opinion lexicon, a list of general words, review data and extraction rules are input to the proposed algorithm. The extraction process uses a rule-based approach using the relations defined in above. The system assumed opinion words to be adjectives, adverbs and verbs in some cases. And product features are nouns or noun phrases and also verbs in some cases.

Its primary idea is that opinion words are usually associated with product features in some ways. Thus, opinion words can be recognized by identified features, and features can be identified by known opinion words. So, the extracted opinion words and product features are used to identify new opinion words and new product features. The extraction process ends when no more opinion words or product features can be found.

Table 2. Some Unigram Features

Price	System	Computer	Laptop	Reviews
Customer	Battery	Camera	Power	Screen
Staff	Food	Wine	Window	Notebook
Work	Use	Download	Connection	Cost
D-link	Device	Feature	Model	quality

Moreover, textblob is used to extract ngram noun phrase words from the sentences that are not covered with extraction rules. It can increase the performance of the system in term of precision, recall, and f1-score. In order to increase the accuracy, stop words and some general words are removed during the extraction time. Table 2, 3 and 4 describe some extracted of unigrams, bigrams, trigrams and n-gram words from the system.

Table 3. Some Bigram Features

Picture quality	Battery life	Tech support	Screen quality
Power supply	Printer sharing	Return policy	Set up
Security setting	Service tech	Setup software	Mac support
Guest feature	8GB RAM	web cam	charger unit
connect quality	cooling system	cordless mouse	data rate
Window 7	Cd drive	Hard drive	Service tech

The system constructs n-gram dictionary to refine the noun phrase extraction. If the phrase contains in this dictionary, the system extracts it as a feature. Otherwise, remove it.

Table 4. Sme N-gram Features

D-Link support crew	customer service agents
cover for the DVD drive	Dell's customer disservice
D-Link support crew	extended life battery
Garmin GPS software	sound quality via USB
customer service center	design based programs
direct Electrical connectivity	built it web cam
fingerprint reader driver	technical service for dell

4. Classifying Polarity Orientation

Opinions are classified by using Vader lexicon and qualitative analysis techniques developed by C.J. Hutto and Eric Gilbert, 2014. This deep qualitative analysis resulted in isolating five generalizable heuristics based on grammatical and syntactical cues to convey changes to sentiment intensity. They incorporate word-order sensitive relationships between terms:

Punctuation, namely the exclamation point (!), increases the magnitude of the intensity without modifying the semantic orientation.

Capitalization, specifically using ALL-CAPS to emphasize a sentiment relevant word in the presence of other non-capitalized words, increases the magnitude of the sentiment intensity without affecting the semantic orientation

Degree modifiers (also called intensifiers, booster words, or degree adverbs) impact sentiment intensity by either increasing or decreasing the intensity.

The **contrastive conjunction** "but" signals a shift in sentiment polarity with the sentiment.

By examining the **tri-gram** to deeply analyze the intensity of sentiment orientation.

Table 5. shows some example of polarity classification. Polarity score included negative sign (-) indicates negative opinion. And, score with positive sign (+) indicates positive opinion.

Table 5. Polarity Classification of the System

Words	Polarity Score	Polarity Label
Set up	0.4404	positive
excellent	0.5719	positive
work	0.7264	positive
installation disk	-0.296	negative
problem	-0.4019	negative

5. Experimental Results

For experiment, we use core i7 processor, 4GB RAM and 64-bit Ubuntu OS, And, we implement the proposed system with python programming language (PyCharm 2016.3 IDE for python).

Table 6. Dataset Used in the System

Dataset	no of sentences	no of features
ABSA15 restaurant	1083	1193
ABSA15 Hotel	266	212
Router	245	304
Speaker	291	435
Computer	239	346
iPod	161	293
Linksys Router	192	375
Nokia 6000	363	633
Norton	210	302
Diaper Champ	212	239

In this paper, 10 product review datasets are collected and evaluated in term of precision, recall and f1-score. Table 6 shows the domains according to their names, the number of sentences

and the number of features. For performance evaluation on product feature extraction, the comparative results between the proposed approach and Qiu's approach are also analyzed. Precision, recall and f1-score of both are described in figure 1, 2 and 3.

From figure 1, we can see that the proposed approach has higher precision in 6 datasets, dropped in 3 datasets and nearly the same in 1 dataset over Qiu's approach. The proposed approach relies only on the review data itself and no external information is needed.

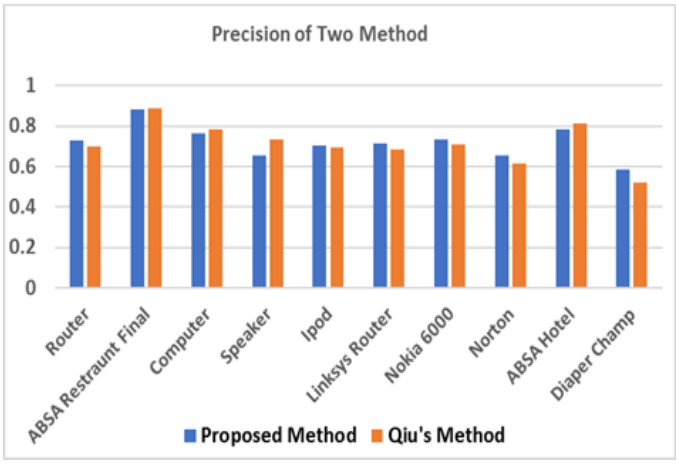


Figure 1: Comparison of Precision on Feature Extraction between Two Approaches

According to the experimental results, the highest precision 0.8819 (88%) are achieved in ABSA Restaurant dataset. In this dataset, the precision, recall and f1-score are nearly the same because there are no verb product features in this dataset.

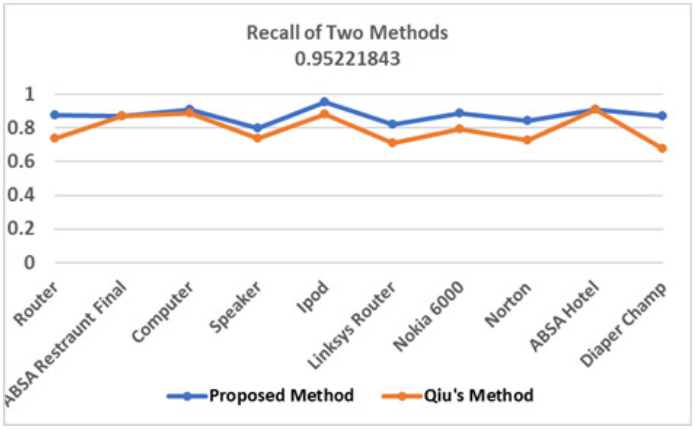


Figure 2: Comparison of Recall on Feature Extraction between Two Approaches

Figure 2 shows that proposed approach outperforms all the other approaches in recall except ABSA Restaurant dataset and ABSA Hotel dataset. The proposed approach has about 14% improvement in Router dataset, about 3% in Computer dataset, about 6% in Speaker dataset, about 7% in iPod dataset, about 10% in Linksys Router dataset, about 9% in Nokia 6000 dataset, about 12% in Norton dataset, about 20% in Diaper Champ dataset. In ABSA Restaurant and ABSA Hotel datasets, the recall of the two approaches are the same. So, to sum up, the proposed approach outperforms over the Qiu's approach according to the recall.

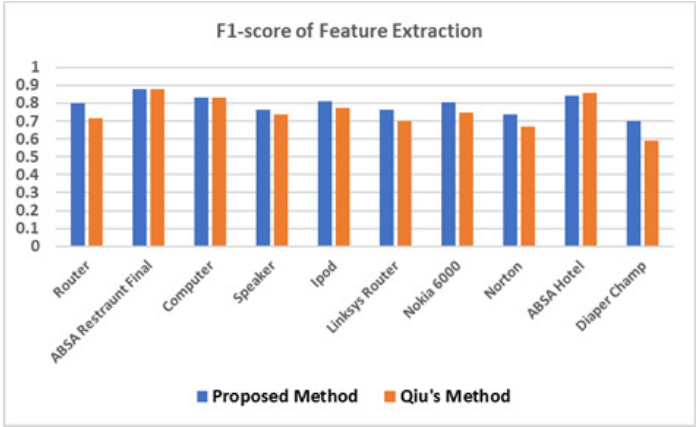


Figure 3: Comparison of F1-score on Feature Extraction between Two Approaches

The comparative results of f1-score in feature extraction are shown in figure 3. According to the experimental results, the proposed approach has higher f1-score (about 7%) in 7 datasets and about 0.003% drop in ABSA Restaurant and Computer datasets and 1% drop in ABSA Hotel dataset

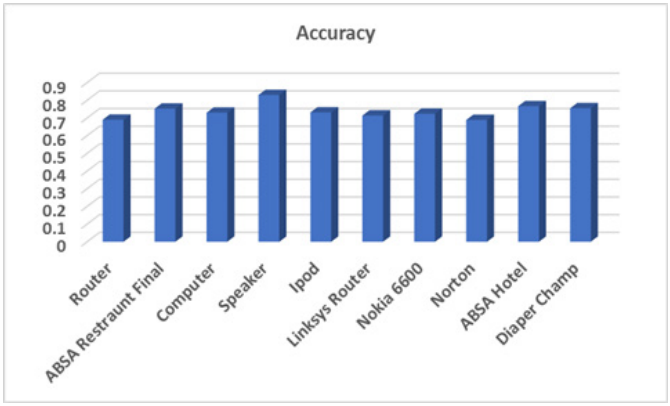


Figure 4 Experimental results of Polarity Classification

The performance analysis of polarity classification of the proposed system is evaluated in all datasets. Figure 4 shows the experimental results of polarity classification in all datasets. The system achieves Highest accuracy 83% in speaker dataset.

6. Conclusion

Opinion mining and sentiment classification are not only technically challenging because of the need for natural language processing. In this work, an effective opinion lexicon expansion and feature extraction approach is proposed. Features and opinions words are extracted simultaneously by using proposed algorithm based on double propagation. So, unlike the existing approach, context dependent opinion words are extracted and domain independence. According to experimental results, the proposed system works well in all datasets and get domain independency without using training examples. As the future extension, we will analyze the performance of the proposed system with more different datasets from SemEval research group.

And we will apply more dependency relations in extraction process.

References

[1] Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., and Hussain, A. "Concept-level sentiment analysis with dependency-based semantic parsing: A novel approach". Cognitive Computation, pages 1–13, 2015.

[2] Asghar M, Khan A, Ahmad S, Kundi F, "A Review of Feature Extraction in Sentiment Analysis", Journal of Basic and Applied Scientific Research, 4(3): 181–186, 2014.

[3] Broß, J. "Aspect-Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques". PhD thesis, Freie Universität Berlin, 2013.

[4] Chen, L., Wang, W., Nagarajan, M., Wang, S., Sheth, A. P. "Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter", In the Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media(ICWSM), 50–57, 2012.

[5] Chinsha, T. and Joseph, S. "A syntactic approach for aspect-based opinion mining". In Semantic Computing (ICSC), 2015 IEEE International Conference on, pages 24–31. IEEE.

[6] De Marneffe, M.-C. and Manning, C. D. "Stanford typed dependencies manual. Report", Technical report, Stanford University, 2016.

[7] Fabbrizio G, Aker A, Gaizauskas R012) "Summarizing Online Reviews Using Aspect Rating Distributions and Language Modeling", IEEE Intelligent Systems, 28(3): 28–37, 2012.

[8] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," Computational Linguistics 37, no.1, 2011, pp. 9–27.

[9] Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," In Proc. of the 49th Ann. Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 151-160. Association for Computational Linguistics, 2011.

[10] Khan, K., Baharudin, B., Khan, A., and Ullah, A. "Mining opinion components from unstructured reviews: A review". Journal of King Saud University-Computer and Information Sciences, 26(3):258–275, 2014.

[11] Kundi F, Ahmad S, Khan A, Asghar, "Detection and Scoring of Internet Slangs for Sentiment Analysis Using SentiWordNet", Life Science Journal, 11(9):66–72, 2014.

[12] Liu, B. "Web data mining: exploring hyperlinks, contents, and usage data". Springer Science & Business Media, 2011.

[13] Liu, B. "Sentiment analysis and opinion mining". Synthesis Lectures on Human Language Technologies, 5(1):1–167, 2015.

[14] Liu, B. "Sentiment analysis mining opinions, sentiments, and emotions". 1:1–386, 2015.

[15] Liu, L. Xu and J. Zhao, "Opinion target extraction using word-based translation model", Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1346-1356, 2012.

[16] Moghaddam, A. S., "Aspect-based opinion mining in online reviews". PhD thesis, Applied Sciences: School of Computing Science, 2014.

[17] Myat Su Wai and Sint Sint Aung, "Simultaneous Opinion Lexicon Expansion and Product Feature Extraction", Proceeding of 16th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2017), ISBN: 978–1–5090–5506–7.

[18] Neviarouskaya A, Aono M, "Sentiment Word Relations with Affect, Judgment, and Appreciation", IEEE Transactions on Affective Computing, 4(4): 425–438, 2014.

[19] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I., "Semeval-2015 Task 12: Aspect Based Sentiment Analysis". In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 486–495, Denver, Colorado. Association for Computational Linguistics.

[20] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S.. "Semeval-2014 Task 4: Aspect Based Sentiment Analysis". In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

[21] S. Poria, B. Agarwal, Gelbukh, A. Hussain, N. Howard, "Dependency-based semantic parsing for concept-level text analysis", In Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Proceeding, Part I. LNCS, vol. 8403, pp. 113–127. Springer, Heidelberg, 2014.

[22] S. Poria, E. Cambria, G. Winterstein, G. B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis,". Knowledge-Based Systems, vol.69, pp. 45-63, 2014.

[23] S. Poria, E. Cambria, L.W. Ku, C. Gui, A. Gelbukh, "A rule-based approach to aspect extraction from product reviews," In Workshop Proc. of the 25th International Conference on Computational Linguistics, COLING'14, pp. 28-37, 2014.

[24] Shariaty, S. and Moghaddam, S.. "Fine-grained opinion mining using conditional random fields". In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pages 109–114. IEEE.

[25] Souza M, Vieira R, Busetti D, Chishman R, Alves I, "Construction of a Portuguese Opinion Lexicon from Multiple Resources", in the proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL' 2011), 59–66. Springer, pp.411-448, 2007.

[26] Xu, T. J. Zhao, D. Q. Zheng, S. Y. Wang, "Product features mining based on Conditional Random Fields model", Proceedings of the 2010 International Conference on Machine Learning and Cybernetics, pp. 3353-3357, 2010.

[27] Xu, X., Cheng, X., Tan, S., Liu, Y., and Shen, H. "Aspect-level opinion mining of online reviews". China Communications, 10(3):25–41.

[28] Y. Wu, Q. Zhang, X. Huang, and L. Wu, "Phrase dependency parsing for opinion mining,". In EMNLP'09, Volume 3-Volume 3, pp. 1533- 1541. Association for Computational Linguistics, 2009.

[29] Y. Zheng, L. Ye, G. Wu and X. Li, "Extracting product features from Chinese customer reviews," Proceedings of 3rd International Conference on Intelligent System and Knowledge Engineering, pp. 285-290, 2008.

[30] Z. Hai, K. Chang, G. Cong, "One seed to find them all: mining opinion features via association", Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 255-264, 2012

[31] Zhai, Z., Liu, B., Xu, H., and Jia, P., "Clustering product features for opinion mining". In Proceedings of the fourth ACM international conference on Web search and data mining, pages 347–354. ACM.

[32] Zhang D, Dong H, Yi J, Song L, "Opinion summarization of customer reviews", In proceedings of the International Conference on Automatic Control and Artificial Intelligence (ACAI 2012), 1476–1479.

[33] Zhang, L. and Liu, B. "Aspect and entity extraction for opinion mining". In Data mining and knowledge discovery for big data, pages 1–40. Springer.

[34] Zhang, Y. and Zhu, W. "Extracting implicit features in online customer reviews", In Proceedings of the 15th ACM international conference on Information and knowledge management, pages 43–50. ACM.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 267-271 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

2x1 Microstrip Patch Array Antenna with Harmonic Suppression Capability for Rectenna

Nur Aisyah Amir*, Shipun Anuar Hamzah, Khairun Nidzam Ramli

Faculty of Electrical and Electronic, Universiti Tun Hussein Onn Malaysia, 86400, Malaysia

ARTICLE INFO

Article history:
Received: 30 October, 2017
Accepted: 02 December, 2017
Online: 23 December, 2017

Keywords :
Key 1 Rectangular inset feed antenna
Key 2 PBG structure
Key 3 Band pass

ABSTRACT

This paper is an extension of work originally presented in 2016 IEEE Asia-Pacific Conference on Applied Electromagnetics (APACE). A 2×1 microstrip patch array antenna integrated with photonic bandgap (PBG) and stubs is designed and analyzed. The performance of the PBG and stubs structure are explained and analyzed in terms of the elimination of the resonance at the harmonic frequencies of the antenna. The proposed antenna is designed on FR-4 substrate with thickness of 1.6 mm and operated at 2.45 GHz frequency suitable for rectenna design application. From the simulated result, the first harmonic frequency (5.4 GHz), the second harmonic frequency (6.6 GHz) and the third harmonic frequency (7.8 GHz) are successfully suppressed. For instance, the radiation to the forward of the stubs-PBG antenna is suppressed at more than 15 dB at the second and third harmonic frequencies.

1. Introduction

The term rectenna is commonly denoted as rectifying antenna at microwave power transmission system. It is a combination of the antenna and high efficient rectifier circuit. The ability of the wireless power transmission system is very important and it depends on the ability of each process; the conversion of signal DC-to-RF at the transmitter, at the transmission and at the conversion of signal RF-to-DC at the receiver. Hence, the receiving antenna is one of the main elements in the rectenna system development.

One of the most popular types of receiving antenna that is used for rectenna application is the microstrip patch antenna [1 - 4]. Some of the advantages of employing this type of antenna is low profile, low cost, light-weight and suitable if integrated with RF devices. However, it also exhibits disadvantages such as the excitation of surface waves that exist in the substrate layer. Surface waves are exceptionable because when a patch antenna radiates, a part of the total available radiated power becomes trapped along the surface of the substrate. It will diminish the total available power for radiation to space wave, and the harmonic frequency is produced. In addition, for arrays antenna, surface waves have a significant impact on the mutual coupling between array elements [5]. Moreover, in rectifying circuit, diode that used to convert the RF signal to DC power supply also generates the unwanted radiation of harmonic frequencies. Thus, the rectenna cannot operate well and will disturb the overall performance of the system.

Several design methods have been proposed in the past in order to overcome these unwanted generated harmonic problems. One of the most popular methods is called the photonic band gap (PBG) or electromagnetic band gap (EBG) [6 - 8] and defected ground structure (DGS) [9 - 11]. Several other techniques for instance antenna with slit and stub structure [12], circular sector patch antenna [13], slot antenna and notch antenna [14], Low Pass Filter (LPF) [15 - 17] also have been discussed earlier as a one of the method to suppress the harmonic frequencies.

According to that, a harmonic suppression microstrip array patch antenna is proposed in order to improve the system performance. In this paper, the 2x1 microstrip patch array antenna integrated with stubs and two-dimensional (2-D) PBG pattern in the ground plane beneath the square patch, is proved experimentally and the capability of the stubs and PBG structure for the suppression of the harmonics are discussed [18].

2. Antenna design procedure

2.1. Numerical method

The structure of the microstrip patch array antenna with PBG substrate (PBG antenna) is presented in Figure 2. A 2x1 array antenna is fabricated on the FR-4 substrate with dimensions of 143 x 52 x 1.6 mm^3 and the relative permittivity of $\varepsilon r = 4.4$. The source signal is fed directly to the antenna by the 50Ω microstrip line with

*Nur Aisyah Amir: Universiti Tun Hussein Onn Malaysia, 01127865561, aisyahamir91@gmail.com

width length (Wd2) = 4.13 mm. The rectangular inset feed antenna dimensions are calculated prior to designing the array antenna by using [19]:

$$w=\frac{c}{2f_0\sqrt{\frac{(\varepsilon_r+1)}{2}}} \quad (1)$$

$$\varepsilon_{eff}=\frac{\varepsilon_r+1}{2}+\frac{\varepsilon_r-1}{2}\left(1+12\frac{h}{w}\right)^{-\frac{1}{2}} \quad (2)$$

$$L_{eff}=\frac{c}{2f_o\sqrt{\varepsilon_{eff}}} \quad (3)$$

$$\Delta L=0.412h\frac{(\varepsilon_{eff}+0.3)\left(\frac{w}{h}+0.264\right)}{(\varepsilon_{eff}-0.258)\left(\frac{w}{h}+0.8\right)} \quad (4)$$

$$L=L_{eff}-2\Delta L \quad (5)$$

Where:

W = Width of the patch
L = Length of the patch
Ɛr = Dielectric constant of substrate
Ɛreff = Effective dielectric constant of substrate
c = Speed of light
f_0 = Frequency reference
h = Height of dielectric substrate

There are advantages of PBG structure which are increasing the antenna's performance in terms of the gain and radiation pattern and also react as a filter to suppress the harmonic frequency. The PBG structure is designed based on the formula [20]:

$$\lambda=\frac{c}{f_o\sqrt{\varepsilon_r}} \quad (6)$$

$$r=\lambda/8 \quad (7)$$

$$a=\lambda/2 \quad (8)$$

After calculation, the 9 x 5 squares with the dimensions 7 x 7 mm^2 are etched in the ground plane at the period (g_b) of 15 mm as shown in Figure 1.

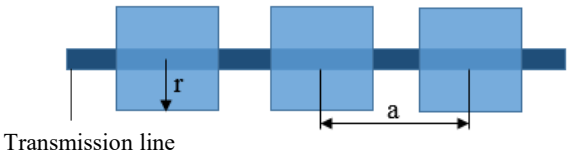
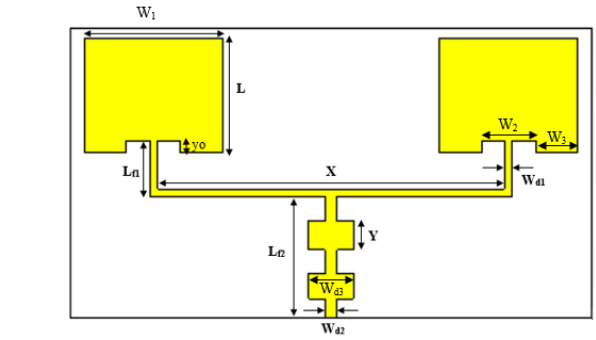


Figure 1: PBG structure at ground plane

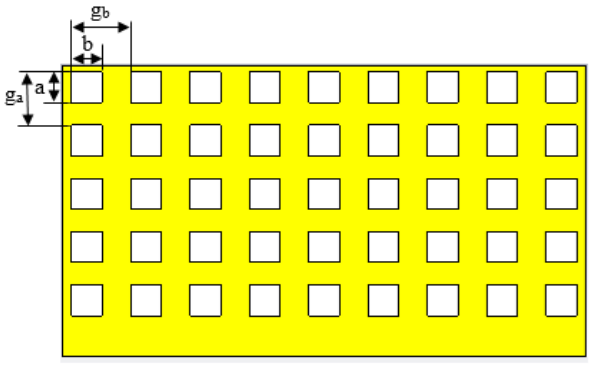
2.2. Design simulation

By using CST microwave software, the rectangular inset feed antenna is designed and simulated to get the operating frequency at 2.45 GHz. Some enhancement are made to obtain the best performance of the antenna. After completing the single element geometry, a 2x1 array antenna is designed. The parameters of the feed line network can be chosen by setting the feed line impedance to 50 Ω (Z1 = 50 Ω), which splits into two 100 Ω (Z2 = 100 Ω) [21]. Similarly, some optimization is done in order to get the greatest achievement of the array antenna.

After done with the conventional microstrip patch array antenna design, the 9 x 5 PBG square are etched at the ground plane. However, some size optimization of PBG square size is done in order to match with the 2x1 array patch antenna design.



(a)



(b)

Figure 2: 2x1 microstrip patch array antenna with stubs and PBG structure (a) view from the front (b) view from the back.

Table 1: Dimension of antenna (unit: mm)

Parameter	Dimension	Parameter	Dimension
L	29.16	**W_{d3}**	11.3
W_1	37.61	**X**	88.32
W_2	10	**Y**	5
W_3	13.805	**a**	7
L_{f1}	5.5	**b**	7
L_{f2}	15.7	**g_a**	15
W_{d1}	0.71	**g**	15
W_{d2}	4.13		

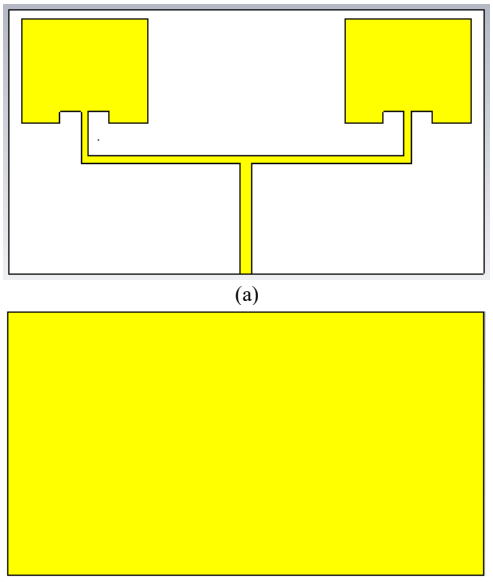
3. Result and discussion

This proposed antenna is divided into three parts which are normal array patch antenna, PBG array antenna and array patch antenna with PBG and stub structure. Each part is discussed and analyzed according to their characteristics.

3.1. Normal array patch antenna

The normal patch antenna without PBG structure shows in Figure 3 is simulated and measured for the comparison of the basic

characteristics. The resonant frequency of the fundamental is adjusted to be 2.45 GHz for the antenna. The simulated and measured S-parameter (return loss) result of the array antenna without stubs and PBG structures are shown in Figure 4. As can see, the return loss is at -13 dB and has a few harmonic frequencies which are not good for an antenna.



(a)

(b)

Figure 3: 2x1 rectangular inset feed array antenna (a) front view (b) back view

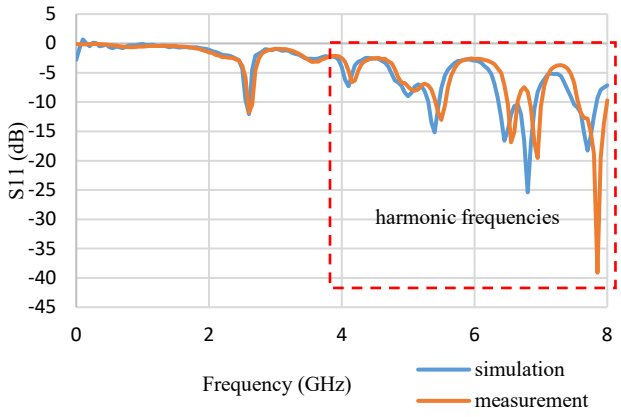


Figure 4: Simulated and measured result for S-parameter (return loss) of 2x1 rectangular inset feed array antenna

3.2. PBG structure

As can see in Figure 4, there are a lots of harmonic frequencies produces above 4 GHz. Hence, PBG structure is beneath at the ground in order to suppress the harmonic frequencies. At the same time, the gain of the antenna also can be improved. Figure 5 shows the PBG structure at ground plane. This arrangement of the PBG lattice produces the band pass characteristics of the transmission parameter at more than -20 dB from 4.6 GHz to 8 GHz shown in Figure 6 (a) when the uniform microstrip line is fabricated on this substrate instead of the patch antenna. The effect of the PBG can be seen by observing the S-parameter (return loss) in Figure 6 (b). It shows that the harmonic frequencies above 4.6 GHz is successfully suppressed when the S-parameter of PBG structure is compare with return loss of 2x1 rectangular inset feed array antenna with and without PBG structure. While Figure 7 shows the comparison of return loss of 2x1 rectangular inset feed array antenna with PBG structure only between simulation and measurement.

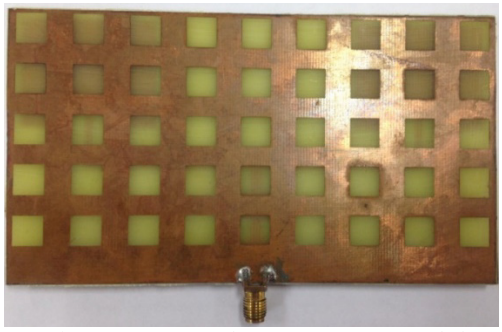
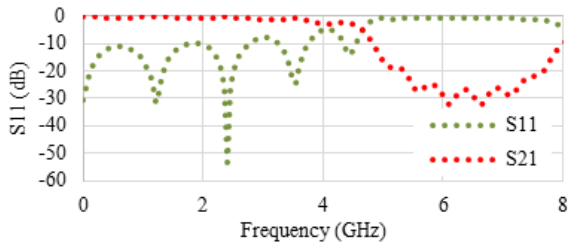
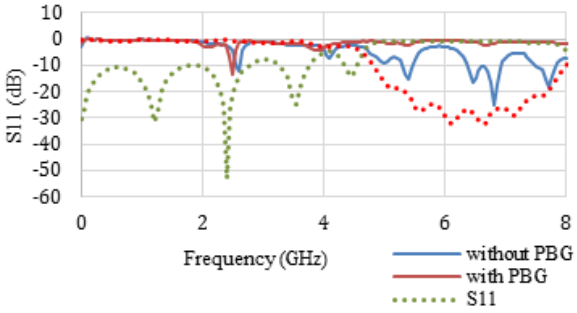


Figure 5: Prototype of PBG structure at ground plane



(a)



(b)

Figure 6: The effect of PBG structure (a) S-parameter of PBG square structure (b) comparison S-parameter of PBG square with return loss of 2x1 rectangular inset feed array antenna with and without PBG structure

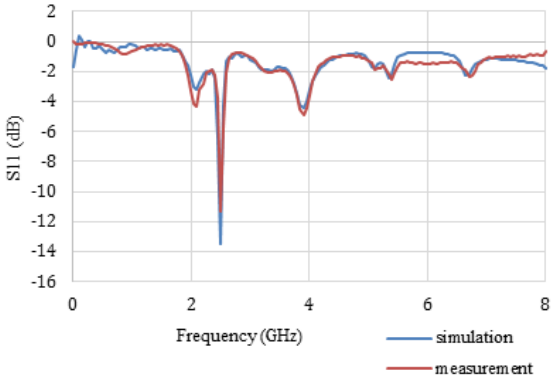


Figure 7: Simulated and measured result for return loss of 2x1 rectangular inset feed array antenna with PBG structure only

3.3. Array patch antenna with PBG and stub structure

The final design is combination of normal array patch antenna with PBG and stubs structure as shown is Figure 8. The S-parameter (return loss) of the array antenna with PBG and stubs structure, and the array antenna without PBG and stubs structure are shown in Figure 9.

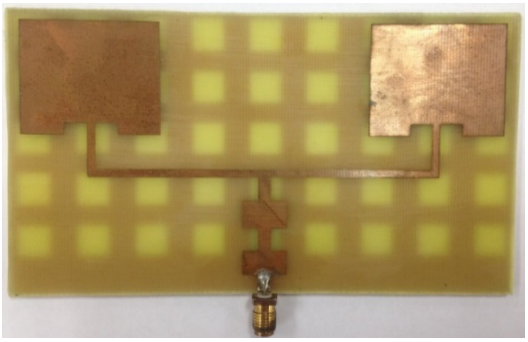


Figure 8: Prototype of 2x1 rectangular inset feed array antenna with PBG structure and stub

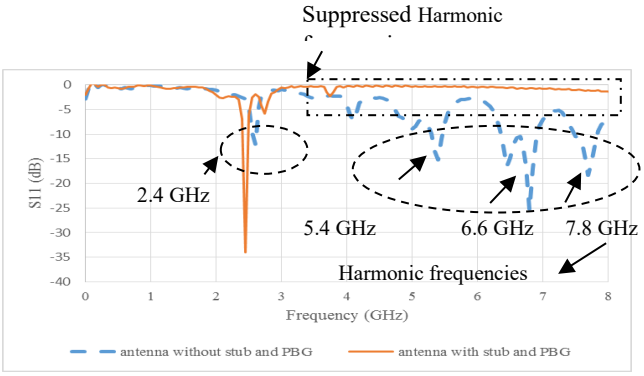


Figure 9: Measured S-Parameter (return loss) of the proposed array antenna. Graph of the comparison for antenna with and without stubs and PBG

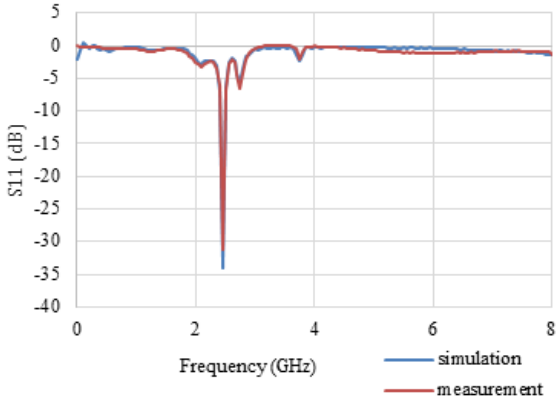


Figure 10: Comparison of return loss between simulation and measurement for 2x1 array patch antenna with PBG and stub structure

Figure 9 shows, a comparison between conventional array antenna with array antenna with PBG and stubs structure indicates that the harmonic frequencies are indeed suppressed. According to the theoretical design, the return loss is observed at 2.45 GHz. However from measurement result, the S-Parameter (return loss) is observed at 2.6 GHz for conventional antenna while for stubs-PBG antenna is at 2.45 GHz. Through observation, the antenna performances are increased. The return loss is revised from -12.05 dB to -34.99 dB after the stub and PBG is added. Besides, based on to the characteristic of PBG, it also can suppress the surface wave in the band-gap. A periodic structure surrounding the patches in addition to one underneath the patches is essential for effective suppression of the harmonic and surface wave. Figure 10 depicts the comparison between simulated and measured return loss of the proposed array antenna.

3.4. Gain

As mentioned before, besides suppressing the harmonic frequency, PBG structure also can increase the performance of the antenna in terms of gain. As can be seen in Figure 11 (a) and (b), there are comparison antenna's gain between simulation and measurement for normal array patch antenna and PBG array patch antenna.

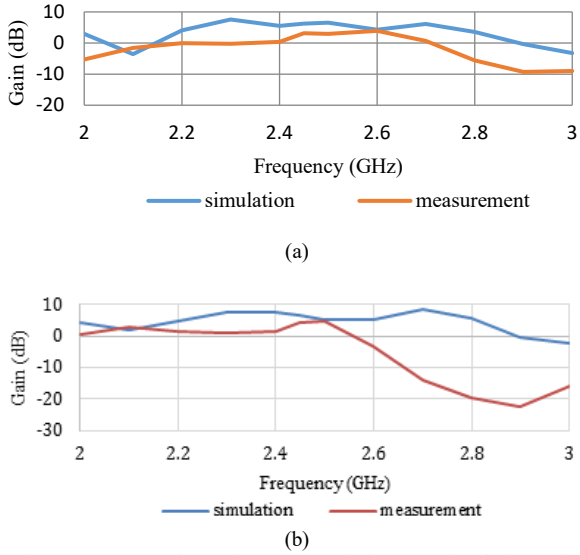


Figure 11: Comparison of antenna's gain between simulation and measurement (a) normal array patch antenna (b) PBG array patch antenna

Figure 11 (a) shows the gain comparison between simulation and measurement for normal array patch antenna. From the graph, the gain for 2.45 GHz is 6.263 dB for simulation while 3.15 dB for measurement. The gain increases to 6.436 dB when the PBG structure is added as shows in Figure 11 (b). While for measurement, the gain is 4.31 dB.

In Table 2, due to the effective suppression of the harmonic and the surface wave, and changes in the current distribution, the antenna array with stub and PBG has higher gain compared to the normal array antenna. The gain increase about 82.41% which is from 3.156 dB to 5.757 dB. While for normal array antenna and array antenna with PBG structure, the result shows a minimal gain increase by 36.66%. Thus, it shows that with the addition of stub structure at the feed line, the gain's performance will improve more. Moreover, the mixed current distribution is useful for diminishing the mutual coupling between the patches and decreasing the maximum side-lobe level.

Table 2: Measured antenna gain base on structure added

Antenna structure	Gain (dB)
Array antenna without PBG and stub	3.156
Array antenna with PBG	4.313
Array antenna with PBG and stub	5.757

4. Conclusion

2x1 microstrip patch array antenna has been presented in this effort. The proposed structure is integrated with two-dimensional PBG pattern in the ground plane under the square patch and stubs. The effectiveness of the PBG and stubs structure for harmonic suppression are analyzed. The simulation and measurement works verified that the combination of the array antenna with PBG and stubs is efficient in eliminating unwanted frequencies. Moreover, the performances of the antenna in term of gain also increase.

Acknowledgment

The authors would like to thank the Centre Graduate Study UTHM for supporting this study. Much appreciation also goes to the Electromagnetic Compability Lab (EMC Centre), Universiti Tun Hussein Onn Malaysia for providing the measurement facilities.

References

[1] G. A. Vera, A. Georgiadis, A. Collado, and S. Via, "Design of a 2.45 GHz rectenna for electromagnetic (EM) energy scavenging," Radio and Wireless Symposium (RWS), 2010 IEEE (pp. 61-64). IEEE, 2010.

[2] U. Olgun, C. C. Chen and J. L. Volakis, "Wireless power harvesting with planar rectennas for 2.45 GHz RFIDs,"Electromagnetic Theory (EMTS), 2010 URSI International Symposium on (pp. 329-331). IEEE, 2010

[3] H. Takhedmit, B. Merabet, L. Cirio, B. Allard, F. Costa, C. Vollaire and O. Picon, "A 2.45-GHz low cost and efficient rectenna," Antennas and Propagation (EuCAP), 2010 Proceedings of the Fourth European Conference on (pp. 1-5). IEEE, 2010

[4] J. Zhang, Y. Huang and P. Cao, "Harvesting RF energy with rectenna arrays," Antennas and Propagation (EUCAP), 2012 6th European Conference on (pp. 365-367). IEEE, 2012

[5] D. M. Nashaat Elsheakh, M. F. Iskander, E. A.-F. Abdallah, H. A. Elsadek, and H. Elhenawy, "Microstrip Array Antenna with New 2D-Electromagnetic Band Gap Structure Shapes to Reduce Harmonics and Mutual Coupling," Prog. Electromagn. Res. C, vol. 12, pp. 203–213, 2010. http://onlinewww.jpier.org/PIERC/pierc12/17.09112008.pdf

[6] M. N. Aktar, M. S. Uddin, M. R. Amin, and M. A. Ali, "Enhanced gain and bandwidth of patch antenna using ebg substrates," Int. J. Wirel. Mob. Networks, vol. 3, 2011. https://www.researchgate.net/profile/Mst_Nargis_Aktar/publication/49965985_Enhanced_Gain_and_Bandwidth_of_Patch_Antenna_Using_EBG_Substrates/links/0912f50973187d409b000000.pdf

[7] Y. Horii and M. Tsutsumi, "Harmonic control by photonic bandgap on microstrip patch antenna," Microwave and Guided Wave Letters, IEEE, 9(1), 13-15, 1999

[8] W. Zhang, J. Mao, X. Sun, R. Qian, and D. Zhang, "Microstrip patch antenna array on ground with circular PBG," Microwave and optical technology letters, 41(2), 127-130, 2004.

[9] V. S. Kushwah and G. S. Tomar, "Size reduction of microstrip patch antenna using Defected Microstrip Structures," Proc. - 2011 Int. Conf. Commun. Syst. Netw. Technol. CSNT 2011, pp. 203–207, 2011.

[10] R. Dua, H. Singh, and N. Gambhir, "2.45 GHz Microstrip Patch Antenna with Defected Ground Structure for Bluetooth," … J. Soft Comput. Eng. …, no. 6, pp. 262–265, 2012. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.2510&rep=rep1&type=pdf

[11] M. S. Ghaffarian, G. Moradi, and R. Zaker, "Harmonic suppression slot antenna using rectangular defected ground structure," Electrical Engineering (ICEE), 2011 19th Iranian Conference on (pp. 1-4). IEEE, 2011.

[12] R. A. Rahim, F. Malek, S. F. W. Anwar, S. L. S. Hassan, M. N. Junita, and H. F. Hassan, "A harmonic suppression circularly polarized patch antenna for an RF ambient energy harvesting system," CEAT 2013 - 2013 IEEE Conf. Clean Energy Technol., pp. 33–37, 2013.

[13] F. J. Huang, T. C. Yo, C. M. Lee, and C. H. Luo, "Design of circular polarization antenna with harmonic suppression for rectenna application," IEEE Antennas Wirel. Propag. Lett., vol. 11, pp. 592–595, 2012.

[14]R. A. Rahim, M. N. Junita, S. I. S. Hassan, and H. F. Hassan, "Harmonics suppression circular polarization elliptical shape microstrip patch antenna," Proc. 2014 2nd Int. Conf. Technol. Informatics, Manag. Eng. Environ. TIMEE 2014, pp. 147–150, 2015.

[15] M. Han, S. Jung and H. Sohn, "High efficient rectenna using a harmonic rejection low pass filter for RF based wireless power transmission," Wireless Communications Systems (ISWCS), 2014 11th International Symposium on (pp. 423-426). IEEE, 2014.

[16] D. Gangwar and R. L. Yadava, "Design and analysis of a pentagonal rectenna," Signal Processing and Integrated Networks (SPIN), 2014 International Conference on (pp. 654-658). IEEE, 2014.

[17] J. Zhang, Y. Huang and P. Cao, "A wideband cross dipole rectenna for RF wireless harvesting," Antennas and Propagation (EuCAP), 2013 7th European Conference on (pp. 3063-3067). IEEE, 2013.

[18] N. A. Amir, S. A. Hamzah, K. N. Ramli, and M. Esa, "2× 1 microstrip patch array antenna with harmonic suppression capability" Applied Electromagnetics (APACE), 2016 IEEE Asia-Pacific Conference on(pp. 272- 276). IEEE, 2016.

[19] Joshua, M.S, Mayank, M, Prafull, and S, "Design and Optimization of Microstrip Patch Antenna," Int. J. Emerg. Trends Technol. Comput. Sci., 2013.

[20]I. Garde, M. J. Yábar, and C. Del Río, "Simple modeling of DGS to design 1D-PBG low-pass filters," Microw. Opt. Technol. Lett., vol. 37, no. 3, pp. 228–232, 2003. http://onlinelibrary.wiley.com/doi/10.1002/mop.10878/full

[21] Y. S. H. Khraisat, "Design of 4 elements rectangular microstrip patch antenna with high gain for 2.4 GHz applications," Mod. Appl. Sci., vol. 6, no. 1, pp. 68–74,2012.http://ccsenet.org/journal/index.php/mas/article/view/12540/9541

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 225-235 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Taxonomy for Enhancing Usability, Flexibility, and Security of User Authentication

Susan Gottschlich*

Raytheon Co., 1001 Boston Post Road East, Marlborough, MA, 01752, 508-490-2339

ARTICLE INFO

Article history:
Received: 29 August, 2017
Accepted: 03 December, 2017
Online: 15 December, 2017

Keywords:
Continuous authentication
Multifactor authentication
Internet of Things
Software Defined
Wearable devices
Cybersecurity
Biometrics

ABSTRACT

Two technology trends - a move toward software defined capabilities and toward networked devices – support both unprecedented innovations and requirements for security. A fundamental aspect of security is user authentication, which allows devices and software applications to establish their user's identity and identity is in turn used to establish which of its capabilities the user is authorized to access. While multiple authentication steps, known as multifactor authentication, are being used more widely throughout the military, government, businesses, and consumer sectors, the selection and implementation of which authentication factors to require is typically defined by security policy. Security policy is in turn typically established by a security organization that may have no formal metrics or means to guide its selection of authentication factors. This paper will present a taxonomy for describing authentication factors including important attributes that characterize authentication robustness to aid in the selection of factors that are consistent with the user's mission. One particular authentication factor that I have developed will be discussed in the context of this taxonomy to motivate the need to broaden current definitions and security policies. The ultimate goal of this paper is to inspire the development of standards for authentication technologies to both support mission aware authentication innovation and to inform decision making about security policies concerning user authentication and authorization. Further, this paper aims to demonstrate that such an approach will fundamentally enhance both security and usability of increasingly networked, software-defined devices, equipment and software applications.

1. Introduction

Mobile devices, such as cell phones and tablets, can be extremely useful for military personnel, security personnel, employees, and consumers in performing an entire gambit of tasks from routine day to day tasks to time critical emergency response. Current cybersecurity best practices encourage the use of user authentication in order for these devices to be accessed. Increasingly, equipment that was once thought of as hardware, is becoming essentially software defined. This includes military radios, medical devices, and autonomous cars. This trend supports the possibility of requiring users to authenticate with their cars, medical devices, and radios, for example, before gaining access to their functionality, thereby enhancing security and discouraging theft and/or hacking. Further, Internet of Things (IoT) devices are becoming pervasive throughout homes and industries, and security policies for IoT device access have not consistently been established and implemented.

This paper is an extension of work originally presented in [1], where a method for continuous secondary factor authentication for military or security personnel required to perform missions while in harm's way was motivated and described. Specifically, this paper will use the Concept of Operation (ConOps) and method discussed in [1] and related ConOps to motivate the need for mission aware authentication factors and in turn to describe and validate authentication factors so that appropriate ones can be selected. This paper will further argue that standards and/or specifications and a means of validating compliance is crucial to support innovation and transition to productization.

In [2], the National Institute of Standards and Technology (NIST) defines multifactor authentication as "Authentication

* Corresponding Author: Susan Gottschlich, susan.gottschlich@raytheon.com

using two or more different factors to achieve authentication. Factors include: (i) something you know (e.g., password/PIN); (ii) something you have (e.g., cryptographic identification device, token); or (iii) something you are (e.g., biometric)." When only two factors are used, this is referred to two factor or dual factor authentication. Dual factor authentication is currently a requirement for NIST's Risk Management Framework (RMF). RMF compliance is a requirement being levied on all military information systems.

Continuous authentication is an emerging authentication factor methodology that implements authentication as a process rather than a simple event such as entering a user name and password. The idea is that once a user establishes his or her identity through another authentication factor or as an initialization step in a continuous authentication factor, the user is continuously monitored by some means so that their identity is continuously established.

Continuous authentication can enhance usability by alleviating the need for a user to re-authenticate, for instance, if they allow their authenticated session to become idle for longer than some inactivity period prescribed by security policy. Continuous authentication can also enhance security by potentially detecting instantaneously when a user may have lost control of his or her device or software application session, eliminating the need for security policies to establish arbitrarily the length of inactivity periods allowed before re-authentication is required.

This paper argues that in order to enhance both security and usability, user authentication should be pervasively implemented on software-defined networked or stand-alone devices or software applications, but that the factors used in authentication should be selected to be consistent with the user's underlying mission and/or need for using the device or application. Thus, the factors employed may be far different than current approaches that typically involve entering a user name and password and/or pin along with a secondary factor (something the user is or has).

Section II will present background on current work on authentication factors and on biometrics which are increasingly being used to support 'something you are' authentication factors.

2. Background

Until recently, user authentication has pervasively been implemented via the use of a password or pin as a primary factor (something you know). If a secondary factor is warranted, several options for 'something you have' have been implemented including an RSA key, a CAC card, an email account, or a token. Alternatively biometric factors (something you are) such as finger prints, iris scans or voice spectrograms have been used. Section 4 will review ConOps examples where such approaches might not be practical given the user's mission.

Various implementations of an emerging authentication approach – continuous authentication – are beginning to emerge. Some of these implementations are described in industry publications [3, 4], and in conference proceedings [5]. Continuous authentication is the idea that a user is monitored in such a way that they are being continuously authenticated. An approach presented in [5], for instance, uses a combination of color information of users' clothing and face information in order to robustly monitor a user who may not always be facing the computer being used. Other continuous authentication approaches presented in [3, 4] use behavior recognition, such as keystroke patterns (e.g. typing rhythm, mouse movement) potentially used in combination with other biometric factors such as iris patterns.

Continuous authentication can bring about more secure authentication that is also more usable than current pervasive security policies implement. For instance, if a user who is not being continuously authenticated walks away from his or her computer briefly, a nefarious agent can assume this user's computer session. Locking computer sessions during inactivity periods is meant to counter this possibility but if the inactivity period is too short, it can become extremely counterproductive for a user who must frequently re-authenticate.

Alternatively, allowing for a lengthy inactivity period negatively impacts the security posture of the computer session. Continuous authentication is theoretically more secure because it tracks the user continuously, thus no timeout or inactivity period need be implemented. It may also be considered more user friendly because users do not need to re-authenticate.

An important thing to note about continuous authentication implementations is that they are generally tied to the type of device being used. For instance, key stroke and mouse behavior monitoring applies to computers but not necessarily cell phones or IoT devices. Further, they are often dependent on a user's mission or role. Tracking clothing, irises, and keyboard behavior are certainly appropriate for office and home environments. They may be less appropriate in environments where users are wearing uniforms (clothing is always the same color across users and days) and protective eyewear, using their devices occasionally while they perform other challenging tasks (e.g. driving, patrolling), or put into high stress situations that may alter their behavior or voice substantially.

In [1] I presented an approach that I developed that uses a fitness tracking device paired with a mobile device to continuously monitor a user performing potentially hazardous missions. This approach locks the mobile device if conditions suggest that the user is under severe duress (possible captured), dangerously wounded or killed. In other words, this approach monitors the user and when it detects that the user may have lost control of his or her device, it fails the user's authentication with the device. Otherwise, it allows a user persistent uninhibited access to the mobile device.

The novelty of the approach is based on the recognition that all of the conditions that suggest a user has lost control of his or her device will bring about a sudden change in the user's biometrics as monitored by the fitness tracking device. Because this is the case, the various biometric sensors can be monitored

using Kalman filters which can detect jump (or sudden) changes in a biometric variable.

In [1] I describe experiments conducted with this continuous monitoring approach. The experimental results, which were derived from live experiments with 'normal' conditions but simulated experiments with dangerous conditions, confirm the approach's ability to provide strong authentication under the assumed user conditions.

As far as I am aware, this is only continuous authentication approach published that considers how to detect and lock down a device when conditions suggest that the user has lost control of his or her mobile device.

While this is generally not a concern for office workers or personal device users, it may be an important concern for military and security users. This differentiation further suggests the need for selecting mission-aware authentication factors.

When contrasted against NIST's definition of authentication factors, the fitness tracking device might be considered both something you have and a means to identifying something you are. In addition to tracking biometric readings, the fitness tracker's API also allows the authentication software to monitor band contact, so it is possible to detect when the user is no longer wearing the device or the tracker is no longer communicating. This paper will argue that combination factors and other innovations will be encouraged if security organizations are given the tools to evaluation their effectiveness.

Further, I submit that the emergence and commercialization of many new technologies including biometric feature sensing and recognition, wearable technologies, and machine learning (ML) will support more innovative and specialized approaches to user authentication.

There is substantial related literature on the biometric readings that might be drawn from in order to help assess the strength and risks associated with a biometric factor. In his Ph.D. dissertation, Gari D. Clifford discusses heart rate variability (HRV) and its causes [6]. Clifford shows that a person's heart rate is variably by nature. For instance, Circadian rhythms will drive variability throughout the day and factors such as general stress level and caffeine intake will cause day to day variations.

Several publications, including [7, 8, 9] consider heart rate variability, Galvanic skin response (which can be measured with some currently available fitness trackers), and other factors in analyzing what happens during exercise, duress, and the differences between the two. Based upon the research presented in [6-9], I developed rules presented in [1] to differentiate conditions that result in an increased heart rate.

Note that all of the biometric factors used in the algorithm in [1] can be measured by the Microsoft® Band 2, which was used in the experiments that were described. Skin conductance delta (also known as Galvanic skin response), for instance, supports the estimate of duress, but other factors can be used in differentiation. My research indicates that conditions resulting in a sudden decrease in heart rate (e.g. extreme cold, cardiac arrest) generally indicate the user is severely stressed by health or environmental issues, rather than an explicit threat.

Research presented in [10] performed a case study on students under stress due to a university examination. In [11] the use of wearable sensors is compared to electrocardiograms and concluded that the former is sufficient for detecting stress conditions.

Much of the research on heart rate measurements are focused on detecting an abnormal condition that may occur during a medical procedure such as surgery. The approaches discussed in such research generally involve preprocessing the measurement signal and performing machine learning (ML) techniques.

The algorithmic approach presented in [1] rejected the application of a pure ML techniques because measurable biometrics, such as heart rate, will vary from day to day due to many factors such as time of day, intake of caffeine or other drugs, and the user's current activity, and it is not practical to control for these factors.

Rather, the algorithm focuses on detecting jump changes in biometric measurements so that normal variances in measurements are implicitly handled.

The work in [12] describes the use of Electroencephalogram (EEG) for authentication. When mature, this technology may be trusted to uniquely authenticate an individual as a single factor (e.g. as opposed to dual factor or multifactor authentication). An unanswered question is how well these approaches work under high stress and maybe duress situation.

The work in [13] addresses biometric recognition techniques, including facial recognition, voice recognition, finger print recognition and iris recognition, being applied to authentication. It further describes how to assess biometric characteristics in terms of five qualities: robustness, distinctiveness, availability, accessibility and acceptability. The work further developed a taxonomy of uses. The focus of this work, however, is in 'snapshot' authentication – authenticating a user at a particular point in time using a snapshot reading (finger print, facial image, etc.) as opposed to the continuous streams of readings I proposed using in [1].

Nevertheless the discussion in [13] can easily be extended into the continuous monitoring domain if it can be assumed that there are mechanisms to continuously take snapshots of the discussed biometric factors.

In the following I will first propose a taxonomy for authentication factors that expands on the work presented in [13] in Section III. Specifically, I decompose the qualities presented in [13] into measurable and verifiable attributes. Further, these attributes are intended to support a broad array of potential authentication methodologies.

In Section IV, I present three ConOps to consider with regards to the taxonomy. I will conclude with a discussion of next steps.

3. Authentication Factor Taxonomy

The purpose of this section is to present a draft taxonomy for authentication factors and the user's that they are designed to identify and authorize. The goal is to motivate the development of industry and government standards organization that develops and maintains taxonomies and related standards that can be utilized by manufacturers of authentication devices and/or algorithms as well as security organizations or home users who must ultimately determine authentication policy. This draft has been developed informally as a means to motivate a formal taxonomy following a research methodology, for instance, as described in [14, 15].

The draft ontology is decomposed into three related groups of attribute/value pairs for describing authentication factors, user missions, and security policy.

Table I provides a list of key attribute/value pairs that can be used to provide specifications for an authentication factor. It is anticipated to be used by suppliers of authentication factors to provide an unambiguous description of the level of security a given authentication factor offering is able to support. It is also anticipated to be used by security organization to specify policy for the conditions under which a given methodology can be used.

This taxonomy can be seen as an extension of the characterization described in [13] because it encompasses the full range of authentication factors and it including attributes that might be used in a risk analysis (e.g. some connection protocols may be considered to be less secure than others). It further attempts to decompose the qualities into attributes that are measurable and verifiable.

This taxonomy strives to address current and emerging authentication methodologies. It is understood that future innovations may bring about methodologies that will require a further extension in the taxonomy. Nevertheless it is quite possible that future innovations will not change the attributes in the taxonomy but instead extend the range of potential values associated with each attribute.

It is the only taxonomy that I am aware of that is meant to model emerging authentication factor technology including continuous authentication and combination factors (e.g. something you have and something you are). I believe that research aimed at characterizing authentication factors is essential to promote innovation in authentication factor offerings. Without such a mechanism, innovators may have a hard time determining whether or not a proposed authentication concept will, if developed, be acceptable by security organizations and users, and thus less likely to expend resources developing innovative authentication approaches.

The first attribute, category, provides a high level simple characterization. Security policies may use this characterization to structure the definition of acceptable authentication factors. In turn, users may use it to determine the best factor(s) to implement their ConOps.

Table 1: Authentication factor definition

ATTRIBUTE	VALUE
Category (structure which provides high level distinctions to support policy definition)	-Discrete, continuous, or continuous with initial discrete authentication factor -Something you know, something you are, something you have, or a combination factor.
Strength (structure that varies depending on the underlying factor)	-Something you know: number of characters, special characters (for single entry), number of questions (for security question knowledge based authentication) -Something you have: restrictions on how the device is obtained, if/when it expires, policies for how lost, stolen, or damaged devices are replaced - Something you are: frequency of interrogation
False Acceptance Rate	Expressed as a percentage. May be calculable based on strength settings.
False Rejection Rate	Expressed as a percentage. May be calculable based on strength settings.
Validation Approach	The approach used to validate strength, false acceptance rate and false rejection rate established.
Connections required (something you have, something you are)	Wired interfaces including: USB, DS-101. Wireless interfaces including: WiFi, Bluetooth, ZigBee (IoT), Proximity card, Near-Field Communications (NFC)
Acceptable user conditions	A list of well-defined conditions including: Continuous use, continuous use with occasional interruptions, wearable-occasional use, pocketed-occasional use, protective face equipment, protective eye equipment, protective hand equipment, protective finger equipment, exertion-tolerant, stress-tolerant, voice use, covert use.
Authenticated devices and applications	A list of well-defined devices, including computer (keyboard, mouse, camera), tablet, phone, software defined equipment, IoT device, etc.as well as applications that the factor can support.
Acceptable devices for authentication factor (if required)	Varies depending on the factor (e.g. USB token, wearable device, software algorithm to run on device for use).
Additional events detected	A list of well-defined events including: duress, medical issue, fatigue, theft, tamper.

The next three attributes – strength, false positive rate and false negative rate – aim to support objective measurable specifications that address the 'robustness' and 'distinctiveness' qualities described in [13]. The strength attribute represents methods by which an authentication factor can be configured (from a supplier's point of view) or must be configured (from a security policy point of view). The strength attribute is particularly important for something that you know factors (e.g. passwords). For continuous factor authentication, strength may be described in terms of the sampling rate (how frequently classification/recognition is performed). For knowledge-based authentication (e.g. answering security questions), strength may be described in terms of the number of questions required to authenticate. The strength attribute is usually configurable.

False acceptance rate and false rejection rate are important for biometric and continuous authentication factors. They are specialized from the universally accepted receiver operating characteristic (ROC) curve mechanism for statistically characterizing the sensitivity and specificity of a binary classifier (for instance see [16]).

In the realm of user authentication, these measures are intended to capture statistically how likely a false user successfully authenticates or a true user fails to authenticate. This occurs because biometric recognition technologies (face recognition, voice recognition, etc.) are not perfect. Specifically, distinctiveness is addressed by false acceptance rate. A low false acceptance rate suggests that the underlying algorithm is able to distinguish a true user from false users.

These attributes are extremely important in helping to select which of perhaps many similar technologies is appropriate for a given application. They may not, however, be meaningful for more traditional authentication factors (e.g. something that you know) such as passwords as there is virtually no way for these factors to be misrecognized.

False acceptance and rejection rates are usually not directly configurable. However, it is frequently possible to tune them. For instance, lower false acceptance rates may be achievable by tuning algorithmic parameters in such a way that the false rejection rate increases as a result. Other factors, possible tied to strength, might also be tuned to decrease these rates. For instance, acceptance rates may increase or decrease depending on the frequency of classification/recognition.

The validation approach attribute is important in specifying how strength and acceptance rate specifications are validated. False acceptance rates and false rejection rates quotations may be very low if, for instance, statistical validation is performed using a small homogeneous sample set. Further, the validation approach may help a security organization determine under which conditions an authentication factor may be useful. For instance, voice recognition technologies may be robust in quiet lab environments but less so in outdoor public environments. This attributes addresses, in part, the acceptability and accessibility qualities from [13].

The connections requirement is particularly important for security organizations that need to perform a vulnerability analysis. Any connection interface is potentially vulnerable to exploitation, spoofing and/or denial. Further, organization may have implemented mechanisms for minimizing vulnerability for given connection types and thus prefer some technologies over others. Note that this attribute may not be relevant to 'something you know' authentication. This attribute addresses, at some level, the availability, accessibility, and acceptability qualities from [13].

The acceptable user conditions attribute supports the acceptability and accessibility qualities. Emerging authentication factors will likely be specialized for a given user condition or situation in order to support appropriate robustness while also maximizing convenience. For instance, continuous authentication based on facial recognition is only appropriate in situations where the user is expected to always be facing the face image capture component of the authentication factor. This may be appropriate for continuous use of a computer in an office environment, but may be less appropriate for occasional use devices that are frequently pocketed.

The next two attributes Authenticated devices and applications and Acceptable devices for authentication factor are simply characterizations of the equipment/software that a factor is intended to authenticate (e.g. computer, bank account session) and equipment used to perform the authentication. These attributes anticipate that suppliers may want to provide software-based solutions that can run on a variety of equipment. For instance, a factor that involves the use of a fitness tracker could potentially run on a variety of commercially available trackers. Thus, this attribute addresses the 'accessibility' quality discussed in [13].

Finally, the additional events detected attribute formalizes the concept that many potential authentication factors may have applications well beyond that of simple authentication. In particular, continuous biometric authentication factors may be able to detect events not directly related to authentication or security such as health events or fatigue. This attribute addresses the 'acceptability' quality from [13]. In particular, an organization may be more willing to invest in an authentication factor approach if it may bring additional benefits beyond robust authentication.

It should be noted that Table I is used to describe 'opt in' technologies where the user has opted to use an authentication factor in order to identify themselves. In biometric recognition and related areas, 'opt out' technologies are used to identify individuals without their direct compliance. While opt out technologies define an important research area, I do not believe that they are appropriate for authentication factors as most current uses for authentication implicitly involve users opting in in order to utilize the equipment or service being secured.

To achieve security policies that are appropriately flexible while maintaining a strong security posture, it is important to take into account a user's mission. In many situations today, security policies only allow a single combination of authentication factors. The result is that users may not use equipment or services that is too onerous to access, may seek

mechanisms to subvert the security policies decreasing the security posture, or may put up with the inconveniences reducing their overall effectiveness and safety. Future security policies might allow a user community to select their own factors and instead provide guidance on how the selected factors must be configured and used in order to be allowable.

Toward this end, Table II provides a draft taxonomy for users in the context of the mission they are performing and for which they need devices and applications authenticated. It is anticipated to be used by a mission leader. The first attribute, mission class, is intended to be a high level characterization to simplify selection and guidance of authentication factors.

Table 2: User (Mission) Definition

ATTRIBUTE	VALUE
Mission Classes	A list of well-defined mission classes including: Desk worker, Office worker not at desk, Factory worker, Home indoor user, home outdoor user, tactical user in hazardous situations, tactical user not likely in hazardous situations, tactical user in extreme environments, etc.
Anticipated session duration	Varies but used to help determine how often a user might be required to re-authenticate
False Acceptance Rate tolerance	Taken together these might be used to select appropriate factors and/or their configurations. For instance, a factor may be configured for extremely low acceptance rate while allowing higher rejections rates (possibly forcing user re-authentication)
False Rejection Rate tolerance	
Acceptable user anomalies	A list of well-defined conditions that are acceptable for a user to exhibit during a valid authentication session including: Stress, strenuous activity, covert use, extreme environmental conditions, etc.
Mission restrictions and conditions	A list of well-defined restrictions and conditions including: outdoor use, wired and wireless restrictions, etc.
Devices and/or applications requiring authentication	A list of devices or applications including computer, tablet, phone, software defined equipment, IoT devices and web site account sessions.
Authentication factor limitations	A list of well-defined restrictions including: no-keyboard requirement, no carry requirement

The second attribute, expected duration, is important in selecting technologies and methods. For instance, more rigorous initialization and/or calibration steps may be more acceptable for long duration uses. If the duration is anticipated to span multiple days then authentication approaches must deal with users sleeping and bathing. For some technologies, expected battery life on equipment is an important consideration.

False acceptance and rejection rate tolerances are important in helping mission leaders specify acceptable risks. For instance, a mission leader may need to consider the risk of an authenticated device falling into an adversary's hands. To the extent that recognition algorithms are tunable, a mission leader may insist on a tuning that is either very conservative for false acceptance rate or very conservative for false rejection rate.

Mission restricts and conditions are important for helping users consider what authentication technologies may be acceptable. Outdoor usage suggests weatherproofing and often battery operation. Wireless transmissions are ultimately observable and/or disruptable by an adversary, but some are less so than others.

Listing all of the equipment or services that a user might authenticate with in order to carry out their mission may spur the usage of authentication factors that can be applied to multiple devices simultaneously. For instance, a wearable device may potentially interface with all of the equipment a user may need to use during a given mission so that they are less encumbered by individual specific authentication processes.

Authentication device restrictions may be used to reject authentication factors. Something you have factors such as tokens or cards are problematic if a user damages, loses or forgets to bring them, particularly where there may be no viable back up options. This case may suggest a 'no carry' restriction. Users wearing protective hand gear may reject any approach requiring a keyboard or keypad.

Table III is intended to allow a mission leader to work with a security organization to develop appropriate security policies and to allow the mission leader to determine how best to perform his or her mission given acceptable risks. A security policy may allow a variety of primary factors to be used as long as they meet requirements spelled out in the security policy. The same should be possible for secondary factors. A supplier offering an authentication factor that is intended as a secondary factor might provide appropriate interfaces to a variety of primary factors in order to facilitate integration.

As wearable devices and biometric recognition algorithms become more capable, I believe that it should be possible to develop an authentication factor that can be used to 'unlock' a number of devices or services. In particular, software defined equipment could interface with other equipment to obtain authenticated user credentials providing a stronger security posture without placing additional requirements on the user.

Thus, future security policies should define to what degree they will allow multiple simultaneous devices and/or applications to be authenticated by a given authentication factor or combination of factors. Today, a similar concept is in common use, referred to as single sign on, where once a user authenticates with their computer, for instances, their credentials

are made available to software applications (avoiding the need to re-authenticate).

Table 3: Authentication Policy definition

ATTRIBUTE	VALUE
Primary factor requirements	Restrictions on the primary authentication factor. These requirements should specify restrictions on strength and mechanisms for re-authentication if 'something you know' is forgotten or 'something you have' is misplaced, forgotten, or damaged. It may further provide restrictions on secondary factors paired with it.
Secondary factor(s) requirements	Restrictions on the secondary factor(s) (can be more than one). Includes re-authentication mechanisms, restrictions on strength
Acceptable devices or services for use	A list of devices or services that can be simultaneously authenticated via a single (multifactor) authentication.
Anomaly policy	A list of well-defined actions to be taken including logging, notification, etc.
Rejection policy	A description of what happens when user authentication is rejected including logging, notification, locking session, etc.
Recovery policy	A description of how a valid user 'recovers' from authentication factor issue including forgetting information, losing, damaging or forgetting tokens, biometric measurement issues, etc.

Anomaly policy provides guidance on what actions are taken if anomalies are detected. For instance, an anomaly on a user's tablet might trigger a notification to the mission leader that might help the leader decide whether or not to take action.

Finally, rejection policies in the future may be more flexible than they are today. If a user loses control of their device, it may make sense to lock or even zeroize the device. However, if the possibility exists for them to regain control, a policy that might allow partial access to emergency features (e.g. ability to send emergency requests to a mission leader) or re-authentication features.

Note that Table I-III are not meant to be completely exhaustive. Rather their intent is to trigger discussions about how to completely define authentication technologies vis a vis user missions and security policies so that standards and/or specifications can be developed that are inclusive of a wide range of authentication technologies.

While the scope of situations for which this taxonomy is intended to address is extremely wide, Table IV attempts to outline a potential approach to using the three part taxonomy presented in this section to select an appropriate authentication methodology for a given mission.

4. ConOps and Authentication Limitations

To better motivate the need for innovation and flexibility in authentication factors and the security policies used to provide restrictions and requirements for them, this section will describe 2 ConOps for which currently acceptable assumptions made about authentication implementations may not hold and 1 for which they do.

The first ConOps was originally presented in [1]. Here a tactical user, such as a soldier, Marine, law enforcement officer, guard or emergency responder, uses a network-enabled cell phone or tablet to support their operations. The user's device might be of critical importance in helping them identify and locate persons of interests or threats, send out requests for back up or exfiltration, or report back relevant intelligence and observations.

The challenge here is that user authentication is extremely important to insure that if the user's device or they themselves are captured by a red force actor (enemy combatant, criminal, or other threat) that the red force actor is not able to exploit the device to gain access to critical information and/or intelligence. On the other hand, the user authentication should allow the tactical user uninhibited access to the device up until the moment that the tactical user loses control of it.

A fundamental assumption that we made in pursuing this work is that the target user community can be characterized by the following constraints:

- The user needs to access their device even when they are in harm's way and there is a possibility that the device may fall into the hands of adversaries including insider threats.
- The user will need to access their device while performing other tasks that will take precedence over device interaction.
- These may include driving, surveilling or monitoring, or physically interacting with civilians, criminals or other red force actors.
- Thus, authentication factors tied to inactivity periods are not practical as the tactical user may be inactive, vis a vis the device, throughout the majority of their tactical mission.
- The user may want to be able to access sensitive data and applications covertly (for instance, if they are undercover).
- This implies the use of commonly available devices, not specialized restricted equipment.

Table 4: Selecting Authentication Methodology for a Given Mission

CONSIDERATION	TAXONOMY RELEATIONSHIP	COMMENTS
What devices and/or applications are required for the mission and do they need to be protected?	Table II – Mission class, expected duration, devices and/or applications requiring authentication	Include the following: • Software defined equipment (e .g. radios, networks, sensors) • Applications running on computing devices such as computers, tablets, phones Consider if it is possible to implement authentication directly into the device or application: • Can authentication be added to or credentials shared with software defined equipment? • Can authentication be added to applications? Consider if alternative mechanism to protecting devices is possible: • Physical device not require protection but cloud based applications running on them do need to be protected. Is there an effective mechanism for authentication on the cloud? • Software defined equipment does not need to be protected as long as the infrastructure that it integrates with is able to provide protections
Risk of false acceptance	Table II - False Acceptance Rate tolerance	If a red force actor gains access to mission devices and/or applications, what are the risks (critical information access, ability to reengineer underlying technology, ability to perform malicious actions with capture devices and/or applications)? Are there other mechanisms to mitigate the risks?
Risk of false rejection	Table II - False Rejection Rate tolerance	If a user can no longer access their devices and/or applications, what are the issues (inability to communicate with teammates, inability to navigate mission area, loss of knowledge access critical to perform mission, inability to perform actions required for mission). Are there other mechanisms to mitigate the loss of access?
Restrictions on authentication factors	Table II - Authentication factor limitations	Consider if a user can use their fingers at all times that device interaction is required. Consider if a user can wear a device and where (wrist, finger, around neck). Consider if a user may forget something they know. Consider if a user may loss, damage, or forget something they have. Consider if a biometric feature will be available to measure (iris scan, finger prints).
What category(s) of authentication methodology are acceptable	Table I – Categories	Work with Security organization to find categories for which policies exist or develop new policies for given category.
What is the best authentication implementation?	Table I – Remaining attributes	Work with Security organization to select authentication factors and configurations to meet mission requirements and provide appropriate protection.
How to implement security policy?	Table III all attributes	Work with Security organization to determine how best to implement policy to provide appropriate protection and mission flexibility.

- The user's organization does not wish to waive cybersecurity controls associated with identification and authentication because the cybersecurity risk of the waiver is considered higher than is acceptable.

To support this ConOps, we developed a continuous authentication algorithm for monitoring the availability of a fitness tracker, worn by the user, and connected to the device via a consented Bluetooth wireless connection. The algorithm further estimating a user's health status based on biometric sensor readings read from the fitness tracker. In particular, we implemented the algorithm using both a Microsoft® Band 2 and Microsoft® Band 1 fitness tracker connected to a Microsoft® Surface tablet.

While the ultimate use of the proposed algorithm must be determined by the security policies of the organization responsible for managing the mobile devices, the proposed use of the algorithm is to provide a continuous stream of user health status estimations to be used as a secondary authentication factor. Certain statuses, when they are identified, will cause the device to lock and/or wipe sensitive data depending on organizationally defined procedures. In particular, the following situations may indicate authentication failure or partial failure limiting device access, especially to sensitive data:

- The fitness tracker loses contact with the wearer or the connection with the device implying that the user may no longer have his or her device and/or fitness tracker.
- The tactical user is under extreme duress implying that the user may no longer be in full control of the device or may be being manipulated by the threat.
- The tactical user may be dead or their health may be extremely compromised indicating that the user is unable to maintain control of the device.

In [1] we presented a jump Kalman filter (JKF) that continuously monitors the fitness tracker sensor readings looking for 'jump' discontinuities that might indicate one of the above three situations. Jump discontinuities are required to differentiate between, for example, duress and extreme physical exertion, the latter being an acceptable and even expected situation for the tactical user.

The benefit of using JKFs over explicit ML techniques is that JKFs are always using recent history to predict a user's next biometric reading. As long as the biometric readings do not change suddenly, the JKF's estimates will be very close to the actual readings. At times when the variability of the readings is high, the estimated covariance is high, so the allowable difference between the estimates and actual readings is increased. It is only when a true discontinuity in readings is observed that a jump is identified. This means that regardless of what a user's health state is after the user is initially authenticated by whatever means, as long as it does not change suddenly during the continuous monitoring, the JKF does not recognize a duress or other disqualifying condition. See [17,18] for other work related to JKF algorithms.

ML techniques can be very powerful, but many will require an initial training period. This may require retraining during every authentication session when using sensor readings or outputs that change a lot over time, e.g. from day to day. Of course, if the degree of difficulty involved in retraining is low, ML techniques might be very useful.

The implementation of the JKF algorithm and its configuration can be used to illustrate important considerations in user authentication. Ultimately, for a security organization to trust such an algorithm, they should require objective metrics on its performance in all relevant conditions. These include probability of false acceptance and probability of false rejection. Ideally such metrics might be computed by a supplier of this kind of authentication factor but then an important question is how these metrics are collected. Ultimately, it may be difficult for a commercial organization to thoroughly vet such an algorithm with actual users because it might require putting users into, at least convincingly, dangerous situations.

Further, this algorithm performs best when calibrated. Kalman filters normally are calibrated by simply collecting a small amount of history at the beginning of a session. In [1] we used 10 seconds of history to calibrate the JKFs. Jumps were recognized by the difference between actual and predicted readings exceeding the filter's covariance by greater than a predetermined constant factor:

$$|\boldsymbol{x_k} - \boldsymbol{z_k}| > C\ {\sigma_z}^2 \quad (1)$$

where $\boldsymbol{z_k}$ is the measurement, $\boldsymbol{x_k}$ the prediction, ${\sigma_z}^2$ the covariance, and C is a predetermined constant factor. One potential use of ML in conjunction with a JKF is in establishing the value of C, which may or may not be a constant. For instance, it is possible to imagine that the value of C may be best expressed as a function of other variables including the covariance. Ultimately, rigorous calibration mechanisms that are not unduly onerous for the user community are needed and security organizations need confidence in these mechanisms.

It should be noted that in addition to supporting user authentication, the approach presented in [1] supports a host of auxiliary requirements and features, including enhanced situation awareness of tactical users' status, early indicators of conflict or environmental hazards, and possibly early warning of health related issues (e.g. experiencing duress when not in a hazardous situation). It is likely that many of the other approaches being invented for user authentication might also be capable of supporting ancillary features.

A second ConOps is that of a tactical user, such as an Airman conducting an airborne mission. In such missions, physical access to mission aircraft is heavily restricted by armed guards. Because of this, in the past, authentication requirements may have been waived. In recent times, however, concern over insider threats, which may be inadvertent, accidental or deliberate, have forced tactical users to perform dual factor authentication. User authentication both limits access to the aircraft systems to trusted users and enforces nonrepudiation – the concept that an authenticated user cannot perform an action and later deny it.

An issue with requiring 'something you have' such as a token as a secondary authentication factor in such situations, is that if a user forgets, losses, or damages the card it may be necessary to scrub the mission.

Biometric factors may be complicated by the requirement to wear personal protective equipment. A further concern with biometric factors, such as finger prints, is that a red force actor can physically force a user (conscious or unconscious) to provide a reading to the authentication system.

One authentication approach for the tactical user that I believe will support strong security while alleviating concerns over authentication factor requirements involves a continuous authentication factor. Similar to the approach presented in [1], a continuous authentication factor tied to something that a tactical user may normally be wearing, such as a fitness tracker, may be used to establish and maintain the user's identity.

Further, this wearable device may be used to establish connections with all of the devices, such as radios and mission computers that the user needs to interact with in order to conduct their mission. The user could be automatically authenticated to all devices as long as the continuous authentication and connection to the devices remains viable. The continuous

authentication algorithm used to establish the user's identity and continually monitor the user may be changed to reflect different aspects of the mission than the one described in [1].

Wireless technologies are frequently disallowed in classified environments which may include mission aircraft. In order to use a continuous monitoring approach such as the one suggested, either a wired device would be required to monitor the user, or a wireless protocol that is not considered vulnerable must be utilized. Bluetooth, ZigBee (utilized in IoT devices), or Near Field communications (NFC) protocols are probably the most practical for such situations but they may require further encryption or other restrictions before they are allowed in a classified environment.

A third ConOps to consider is one that is implicitly addressed in much of the literature involving continuous authentication. It is that involving a knowledge worker who is usually using a computer or tablet. During the course of a normal day's work the worker would not be expected to undergo severe stress or duress. Any such situation would be considered abnormal and would warrant locking the worker out of their session until he or she reestablishes it via more traditional means such as requiring the entry of a password or PIN or calling a help desk. In this situation, video analytics using the computer or tablet camera, keystroke behavior, or a variety of other factors are likely practical to support strong and usable security.

5. Next Steps

This paper has sought to demonstrate that an increasingly networked software-defined technology footprint in our work and leisure environments is expanding the need and/or opportunity for user authentication before allowing access to computers, phones, IoT devices, services, and general equipment. At the same time, technology advances in wearable devices, biometric sensors, behavior analytics and a host of other fields is supporting the potential for more secure and less onerous authentication methodologies. Developing innovative new authentication factors to the point of productization and mass production, however, will be accelerated if inventors, investors, and manufacturers understand that security organizations and even home users will accept alternative authentication factor technologies to the relatively simplistic ones commonly utilized today.

Toward this end, this paper has attempted to motivate the formation or expansion of a standards body concerning user authentication technologies and policies. The taxonomy presented in Section III is primarily meant to provide early talking points in the formulation of a standards or specification body.

In one somewhat related example of how standards bodies can spur innovation and technology development, the U.S. National Security Agency (NSA) developed and published the Commercial Solutions for Classified Program (CSfC) specification [19]. Manufacturers and solution providers have used this specification to develop capability sets that they in turn have been able to have registered with NSA after providing evidence of compliance with the CSfC specification. This allows third parties to purchase and integrate compliant capability sets into their environments. In the end, the overall process of getting a classified environment certified by NSA has become far more flexible, affordable, and expedient because of the specification and associated registration process.

A similar approach applied to user authentication may ultimately support stronger and less onerous user authentication, and it may further support the expansion of user authentication and access restriction's use within an overall cybersecurity plan further enhancing security and usability at an enterprise level. For instance, if all equipment in a factory, office, or military platform required user authentication to read and/or modify equipment settings and/or to use the equipment, it may be much more difficult for the equipment to be 'hacked' because network penetration alone does not provide access to the equipment and software running on it.

While such an approach may seem unacceptable today, acceptance of emerging authentication methodologies will increase within security organizations as their strength and acceptance rates are proven and their risks are understood and mitigated. Further, acceptance within user communities will increase as authentication processes become simpler and less prone to error.

Similarly, user authentication techniques that are easy to understand and utilize might be used to enhance home security in a consistent and predictable fashion reducing deliberate and/or inadvertent actions causing harm to occupants and/or damage to or loss of property.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

This work was funded in part by Raytheon Internal Research and Development (IRAD) funds.

References

[1] S. Gottschlich, "Incorporating health monitoring and duress detection into mobile device authentication" in 2017 IEEE International Symposium on Technologies for Homeland Security

[2] Joint Task Force and Transformation Initiative, "Security and privacy controls for federal information systems and organizations" NIST Special Publication 800 (2013): 53

[3] B. Violino, "Continuous authentication: Why it's getting attention and what you need to know," in CSO, Mar 14, 2017

[4] R. Walters, "Continuous Authentication: The future of Identity and Access Management (IAM)" in Network World, Sep 16, 2016

[5] K. Niinuma and A. K. Jain. "Continuous user authentication using temporal information" in Proceedings SPIE, 7667(1), 2010.

[6] G. D. Clifford, "Signal processing methods for heart rate variability," Diss. Department of Engineering Science, University of Oxford, 2002.

[7] http://scienceline.org/2007/06/ask-hsu-fightorflight/.

[8] S. Boettger, V.K. Yeragani, L. Donath, H. J. Müller, H. H. Gabriel, K. J. Bär, "Heart rate variability, QT variability, and electrodermal activity during exercise" Med Science Sports Exercise 42(3), 443-448, 2010

[9] D. K. Spierer, E. Griffiths, T. Sterland, “Fight or flight,” The Tactical Edge, Summer 2009.

[10] P. Melillo, M. Bracale, L. Pecchia, “Nonlinear Heart Rate Variability features for real-life stress detection. Case study: students under stress due to university examination” BioMedical Engineering OnLine, 10(1), 96, 2011

[11] J. Choi, R. Gutierrez-Osuna, "Using heart rate monitors to detect mental stress" in IEEE Sixth International Workshop on Wearable and Implantable Body Sensor Networks, 219-223, 2009

[12] R. J. Rodriguez. “An Electroencephalogram (EEG) Based Biometrics Investigation for Authentication: A Human-Computer Interaction (HCI) Approach,” Doctoral dissertation, College of Engineering and Computing, Nova Southeastern University, 2015.

[13] J. Wayman, A. Jain, D. Maltoni , D. Maio, “An Introduction to Biometric Authentication Systems,” in J. Wayman, A. Jain, D. Maltoni, D. Maio, (eds), Biometric Systems. Springer, 2005.

[14] R. C.Nickerson, U.Varshney, J. Muntermann, “A method for taxonomy development and its application in information systems,” European Journal of Information Systems, 22(3), 336-359, 2013

[15] A. Hevner, S. Chatterjee, Design research in information systems: theory and practice (Vol. 22), Springer Science & Business Media, 2010

[16] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics" Journal of biomedical informatics, 38(5), 404-415, 2005

[17] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Estimation from lossy sensor data: Jump linear modeling and Kalman filtering" in Proceedings of the 3rd international symposium on Information processing in sensor networks, 251-258, 2004.

[18] L. Galleani, Lorenzo, P. Tavella, "Detection of atomic clock frequency jumps with the kalman filter" IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 59(3), 504-509, 2012

[19] National Security Agency, “Commercial Solutions for Classified,” available: https://www.nsa.gov/resources/everyone/csfc

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 225-233 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Velocity obstacles for car-like mobile robots: Determination of colliding velocity and curvature pairs

Emese Gincsainé Szádeczky-Kardoss[*,1], Zoltán Gyenes[2]

[1]*Associate professor at Department of Control Engineering and Information Technology, Budapest University of Technology and Economics, H-1117, Hungary*

[2]*Student at Department of Control Engineering and Information Technology, Budapest University of Technology and Economics, H-1117, Hungary*

ARTICLE INFO

Article history:
Received: 29 November, 2017
Accepted: 15 January, 2018
Online: 30 January, 2018

Keywords:
Mobile robots
Motion planning
Velocity obstacles

ABSTRACT

This paper addresses the motion planning problem of Reeds-Shepp-type car-like mobile robots moving among static and dynamic obstacles. If the positions and the velocity vectors of the obstacles are known or well estimated, the Velocity Obstacles (VO) method and its non-linear version (NLVO) can be used to plan a collision-free trajectory for a robot in the dynamic environment. VO and NLVO algorithms determine a velocity vector for the robot which corresponds not necessarily to the orientation of the robot, hence a nonholonomic car-like mobile robot cannot apply it exactly. Previously, the NLVO method was adopted for Dubins-like mobile robots, which are able to move forward only. In this paper, a method similar to NLVO is presented, but it results motions feasible for Reeds-Shepp-type robots, which are able to drive both forward and backward. Longitudinal velocities and curvatures of turning circles are calculated, which ensure collision-free motion if the arbitrary movement of the obstacles are known for some time-horizon.

1. Introduction

This paper is an extension of work originally presented in the 25th Mediterranean Conference on Control and Automation [1].

One of the main tasks of autonomous mobile robots is to execute a safe motion in their workspace from the actual position to a desired goal configuration. In some applications, the environment is fix, i.e. the positions and velocities of the obstacles are known. Otherwise the robot has to use some sensors to be able to estimate this information. Robots should be able to plan their motions such that they will not collide with static or moving obstacles. The literature presents several path planning methods for the avoidance of static obstacles (see e.g. [2,3]).

There are two possibilities for motion planning among moving obstacles: 1, The planning can be done in two steps: geometric path planning and then velocity planning. 2, The geometry of the path and the velocity profile along it are determined simultaneously.

[*]Corresponding Author: Emese Gincsainé Szádeczky-Kardoss,
Email: szadeczky@iit.bme.hu

In the first case, the geometry of the path and its time distribution are calculated separately (e.g. [4,5]). At the path planning phase, the moving obstacles are disregarded. The planner calculates a path which ensures collision-free motion to the goal among the static obstacles. The motions of the obstacles influence only the velocity profile of the path.

The second possibility is to calculate both the path geometry and the velocity profile in one step (e.g. [6]). In this second case the shape of the path depends also on the positions and movements of the dynamic obstacles.

If the information about the obstacles (positions, velocities) are known, a global path planner can be applied. If the environment is unknown, a local reactive obstacle avoidance algorithm has to be used based on the sensory information (e.g. [7]).

Several motion planning methods in static environment can be extended to solve the dynamic problem as well. Some methods use the configuration space of the robot to find a feasible path among static obstacles (e.g. [3]). If the workspace of the robot contains moving obstacles as well, the configuration space should be

extended by a temporal dimension, and the motion should be searched in this extended space. Using this solution, one has to modify the distance metric to deal with the temporal dimension [5].

The Artificial Potential Field (APF) algorithm is a quite simple method which can be used for path planning in static environment [8]. Applying special potential force functions, one can solve the motion planning problem in case of moving target and obstacles using APF [9,10].

The Dynamic Window (DW) method is a local obstacle avoidance algorithm used in static environment [11]. The planning is done in the velocity space of the robot. Reachable and admissible velocity values are selected. Reachable values satisfy the kinematic and dynamic constraints of the robot and admissible values guarantee that the robot can stop before hitting an obstacle. An adaptation of DW can be used with moving obstacles as well [7].

The inevitable collision states (ICS) for robots are presented in [12]. If the robot is in an IC state, it surely collides with an obstacle independently from the future trajectory of the robot. If a state is non-ICS, there exists at least one motion possibility for the robot to avoid collision. Using the ICS concept, the motion planning problem in dynamic environment can also be solved [13].

The concept of velocity obstacles (VO) was introduced in [6] for such environments where the velocity vectors of the obstacles are supposed to be unchanged for some time-interval. Using VOs, an avoidance maneuver can be determined in the velocity space of the robot, based on the current positions and velocities of the robot and obstacles. Velocity obstacles represent the set of robot velocities that would result in a collision with a static or moving obstacle. The basic VO method was extended for obstacles moving along arbitrary trajectories (non-linear velocity obstacles – NLVO [14]).

The inverse version of NLVO (INLVO) can be used to plan the velocity profile for a robot along a path with fix geometry [15]. A modified VO method can be used to plan autonomous navigation for unmanned surface vehicles as well [16]. The hybrid reciprocal VO (HRVO) is a method to plan the motion of multiple mobile robots without central coordination [17].

Our conference paper [1] presented a modified version of VO to plan the motion for Dubins-like mobile robots (VOD - Velocity Obstacles for Dubins-like robots). These robots go only forward. VOD defined pairs of colliding velocity and turning radius.

In this paper, the goal is to define velocity obstacles for Reeds-Shepp-type car-like mobile robots, which are able to drive both forward and backward. The determination of colliding velocity and curvature pairs is discussed. In this paper, the curvature is used instead of turning radius, hence the graphical representation of the colliding pairs is easier. The method and the equations of [1] were modified to deal with negative velocities and to use curvature instead of turning radius.

The paper is organized as follows. Section 2 gives a short review of velocity obstacle methods. Section 3 presents the properties of Reeds-Shepp-type mobile robots. Section 4 describes how the 'velocity' of a car-like robot is represented in this work. In Section 5, the velocity obstacles for car-like mobile robots (VOCL) are presented. Some simulation results are given in Section 6. Finally, a short section concludes the paper.

2. A Short Review of Velocity Obstacles

VO method can be used to find a feasible velocity vector for the robot such that the robot is able to avoid static and moving obstacles. It is assumed that the position and velocities of the obstacles are known, and the obstacles follow a straight-line path for some time-horizon. The VO method uses circular representation of the robot and the obstacles with known radii.

Given are a robot A and some moving obstacles B_i ($i = 1 \dots n < \infty$), where n denotes the number of obstacles. (According to this concept, a static obstacle is a moving obstacle with zero velocity.) The velocity obstacle VO_i contains all the robot velocity vectors $\mathbf{v}$ which would result a collision with obstacle B_i:

$$VO_i = \{\mathbf{v} | \exists t: A(t, \mathbf{v}) \cap B_i(t) \neq \emptyset\} \quad (1)$$

where $A(t, \mathbf{v})$ denotes the robot at time t if velocity $\mathbf{v}$ was applied. The shape of VO_i is a cone (see Figure *1*). The union of the individual VO_i reads

$$VO = \cup_{i=1}^{n} VO_i. \quad (2)$$

Selecting a velocity vector outside VO guarantees that no collision will occur between the robot and the obstacles.

An example for a point robot A and two moving (B_1, B_2) and a static obstacle (B_3) is presented in Figure *1*.

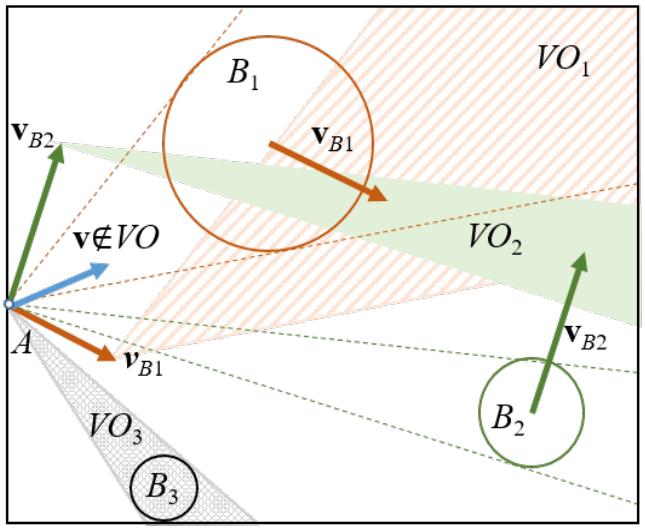


Figure 1. Velocity obstacles for a point robot A with two moving (B_1, B_2) and a static obstacle (B_3). $\boldsymbol{v}_{B1}, \boldsymbol{v}_{B2}$ are the velocities of the obstacles. VO_i denotes the velocity obstacle corresponding to obstacle B_i. A collision-free example is depicted for robot velocity: $\boldsymbol{v} \notin VO = \cup_{i=1}^{3} VO_i$.

2.1. Non-Linear Velocity Obstacles

The non-linear velocity obstacle (NLVO) defines the set of all linear robot velocities that would result a collision with obstacle B_i (t) moving along arbitrary known trajectory. At time instant t,

one can define robot velocity vectors which move the robot to a position during time $t - t_0$ such that a collision occurs with B_i:

$$NLVO_i(t) = \frac{c_i(t)+B_i}{t-t_0} \tag{3}$$

where $c_i(t)$ denotes the trajectory of the obstacle B_i. Considering all $t > t_0$, one gets

$$NLVO_i = \bigcup_{t>t_0} \frac{c_i(t)+B_i}{t-t_0}. \tag{4}$$

The shape of the non-linear velocity obstacle $NLVO_i$ is a warped cone with apex at A (see Figure 2). Similar to (2), the union of the individual $NLVO_i$ defines the set of all robot velocity vectors which result a collision for some $t > t_0$.

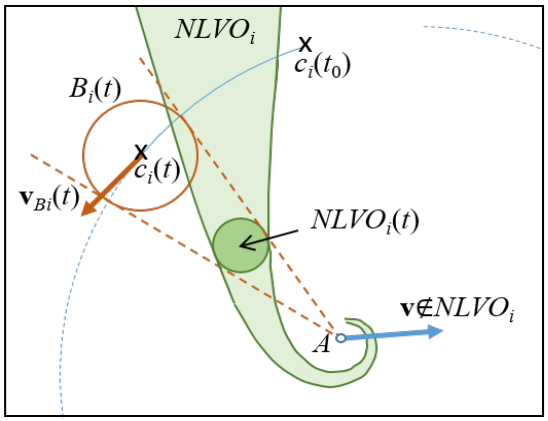


Figure 2. $NLVO$ for a point robot A with obstacle B_i moving along a circular path.

2.2. Generalized Velocity Obstacles

The concept of velocity obstacles was generalized to apply it for car-like robots [18]. The obstacle is defined in the control space of the robot:

$$GVO = \{u | \exists t: \|A(t,u) - B(t)\| < r_A + r_B\} \tag{5}$$

where u is the control input of the robot, $A(t,u)$ is the position of A if control u was applied up to t. r_A and r_B denote the radii of the circular robot and obstacle.

The controls are sampled, and for each control u, the minimum distance between A and B is determined numerically, if control u was applied to the robot. If the minimum distance is smaller than the sum of the radii, $u \in GVO$ and it means that a collision will occur between B and A for the control u.

2.3. Idea for Extension

In this paper, the velocity obstacles are applied to similar robots as presented in Subsection 2.2. This solution will not be restricted to sampled control inputs, as suggested by [18]. The presented method represents the velocity obstacles for car-like robots as a subset of a two-dimensional plane similar to the methods of VO and NLVO. In this work, this plane is determined by the velocity of the robot and the curvature of its path.

3. Reeds-Shepp-Type Car-Like Mobile Robots

A car-like mobile robot (see Figure 3) can move in a two-dimensional workspace. The state of the robot is defined by its position and orientation. The position is given by the (x, y) coordinates of the midpoint of its rear axle. The orientation of the robot is denoted by θ, and it is defined by the angle of the positive x axis of the coordinate system and the longitudinal axis of the robot. Ackermann-steering is supposed (see [19]), and the movement of the robot is described by the motion of a bicycle putting on the longitudinal axis of the robot. The inputs of the robot are the longitudinal velocity v and the angle of the front turning wheel (δ). Notice, that v is a scalar.

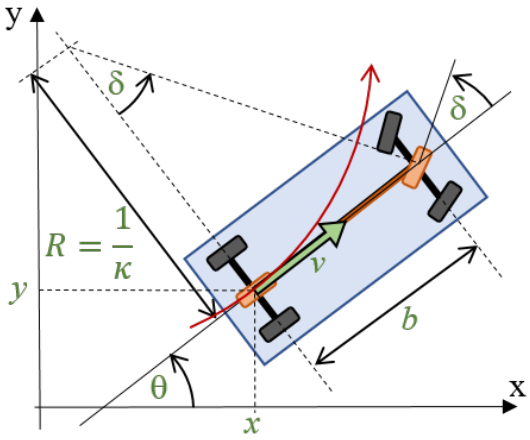


Figure 3. Car-like mobile robot.

The kinematics of the robot reads:

$$\begin{bmatrix}\dot{x}\\ \dot{y}\\ \dot{\theta}\end{bmatrix} = \begin{bmatrix}\cos\theta\\ \sin\theta\\ \frac{\tan\delta}{b}\end{bmatrix} v, \tag{6}$$

where b is distance between the front and the rear wheels. The turning angle δ of the front wheel determines the radius R of the turning circle of the robot: $R = \frac{b}{\tan\delta}$. The reciprocal of the turning radius is called curvature: $\kappa = \frac{1}{R} = \frac{\tan\delta}{b}$.

The turning radius R cannot be arbitrary small, since the turning angle δ of the front wheels is also limited. Depending on the limit of δ and the geometrical b parameter of the robot, a minimal turning radius can be determined: $R_{min} \leq |R|$. Similarly, the curvature is also limited: $|\kappa| \leq \kappa_{max}$.

The car-like mobile robot is a nonholonomic system since the direction of its velocity vector is constrained by the following equation:

$$\dot{x}\sin\theta - \dot{y}\cos\theta = 0. \tag{7}$$

There are different types of car-like mobile robots: Dubins-robots can only move forward (i.e. $v > 0$). It was proved that for given start and goal states the path with the minimal length consists of three segments: two or three circular segments with maximal

curvature (i.e. with minimal turning radius) and, if needed, a straight-line between two circles if no obstacles are presented in the workspace [20]. The Reeds-Shepp-type mobile robots are able to travel both forward and backward. It was also shown that the path with minimal length between given start and goal states contains (like Dubins-robots) circular segments with maximal curvature, and it may contain straight-lines and cusp points where the driving direction changes [21].

4. Velocity Representation for Car-Like Mobile Robots

A Reeds-Shepp-type mobile robot can move forward or backward on a circular path segment with constant curvature (κ), i.e. with constant radius ($R = \frac{1}{\kappa}$), or on a straight-line. The curvature may have negative values, which shows that the robot turns in the negative direction (i.e. clockwise). The straight-line motion can also be described by a zero curvature. The time-distribution along the path depends on the longitudinal velocity v of the robot. The sign of v shows the direction of motion. These imply that the actual motion of the robot can be described by the pair (v, κ).

If (v, κ) is constant for a time period Δt, one can calculate the displacement in position and in orientation for Δt, if the motion starts at $t_0 = 0$ from $[x_0, y_0, \theta_0] = [0,0,0]$ by integrating (6):

$$\begin{bmatrix} x(\Delta t) \\ y(\Delta t) \\ \theta(\Delta t) \end{bmatrix} = \begin{bmatrix} \frac{2}{\kappa} \sin\frac{\kappa v \Delta t}{2} \cos\frac{\kappa v \Delta t}{2} \\ \frac{2}{\kappa} \sin^2\frac{\kappa v \Delta t}{2} \\ \kappa v \Delta t \end{bmatrix}. \quad (8)$$

This works in the reverse direction as well: if a point (x, y) is given in the workspace of the robot, one can determine the constant curvature κ and the constant forward longitudinal velocity $v^f \geq 0$ and backward velocity $v^b < 0$, which can move the robot from $t_0 = 0$ and $[x_0, y_0, \theta_0] = [0,0,0]$ to this point during time Δt:

$$\kappa = \frac{2y}{x^2+y^2}, \quad (9)$$

$$v^f = \frac{x^2+y^2}{y\Delta t} \operatorname{atan}\frac{y}{x}, \quad (10)$$

$$v^b = \frac{x^2+y^2}{y\Delta t} \left(\operatorname{atan}\frac{y}{x} - sgn\left(\operatorname{atan}\frac{y}{x}\right)\pi\right). \quad (11)$$

Notice, that (10) (and (11)) has several solutions for given x and y. If the robot does not go more than ones around in a circle, (10) (or (11)) will have a single solution for v^f (or for v^b) with $-\pi < \operatorname{atan}\frac{y}{x} \leq \pi$.

If the robot goes straight, its motion is described by: $x = v\,\Delta t, y = 0, \kappa = 0, R = \infty, v = \frac{x}{\Delta t}$. In the sequel, the goal is to determine the (v, κ) (i.e. (v^f, κ) and (v^b, κ)) pairs which result a collision-free motion for the robot.

5. Velocity Obstacles for Car-Like Robots

Let the robot be at the $[x_0, y_0, \theta_0] = [0,0,0]$ initial state. First, such (v, κ) pairs are determined which result a collision with static obstacles. Then the collision with moving obstacles is also considered.

5.1. Collision with Static Obstacles

Suppose that a static obstacle is presented in the workspace of the robot. Let a circle represent the obstacle. Its position is given by the coordinates of its center: (x_o, y_o) and its radius is r_o. For the sake of simplicity, the robot is also represented by a circle with radius r_r. (A non-circular robot or obstacle can be approximated by one or more circles.) To examine the collision for a point-like robot, the radius of the obstacle should be enlarged by the radius of the robot: $r_o' = r_o + r_r$ similar to the method proposed in [6].

First, such circular motions (i.e. curvatures κ_g) are considered, where the robot grazes (touches) the obstacle. Geometrically, the problem is the following: given a point (position of the robot), a straight-line going through this point (line of the robot's orientation) and a circle (obstacle with enlarged radius r_o'), such a circle has to be found, which is tangential to the circle of the obstacle and grazes the line of orientation at the given point (robot's position).

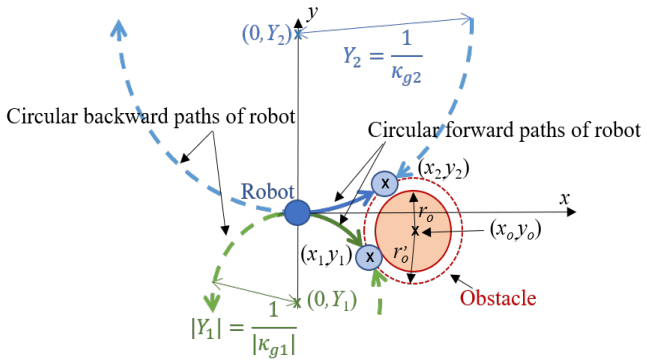


Figure 4. Grazing a static obstacle.

There are two possibilities (see Figure 4). In both cases the center of the turning circle lies on the y-axis (i.e. the x coordinate of the center of the turning circle is 0). The y coordinate of the center of the turning circle defines the radius of the circle. The two possible values for the curvatures (or for reciprocals of y coordinates of the center) are:

$$\kappa_{g1} = \frac{1}{Y_1} = \frac{2(y_o - r_o')}{x_o^2 + y_o^2 - r_o'^2}, \quad (12)$$

$$\kappa_{g2} = \frac{1}{Y_2} = \frac{2(y_o + r_o')}{x_o^2 + y_o^2 - r_o'^2}. \quad (13)$$

All curvatures between κ_{g1} and κ_{g2} will result in a collision with the static obstacle. More precisely, moving on a circle with curvature $\kappa_i \neq 0$ will cause a collision if

$$\min(\kappa_{g1}, \kappa_{g2}) < \kappa_i < \max(\kappa_{g1}, \kappa_{g2}). \quad (14)$$

$\kappa_i = 0$ can only cause a collision if

$$sgn(\kappa_{g1}) \neq sgn(\kappa_{g2}) \wedge sgn(x_o) = sgn(v). \quad (15)$$

The coordinates of the grazing points ((x_1, y_1) resp. (x_2, y_2)) can be calculated as well:

$$x_1 = \frac{x_o(x_o^2+y_o^2-r_o'^2)}{x_o^2+(y_o-r_o')^2}, \quad (16)$$

$$y_1 = \frac{(x_o^2+y_o^2-r_o'^2)(y_o-r_o')}{x_o^2+(y_o-r_o')^2}, \quad (17)$$

$$x_2 = \frac{x_o(x_o^2+y_o^2-r_o'^2)}{x_o^2+(y_o+r_o')^2}, \quad (18)$$

$$y_2 = \frac{(x_o^2+y_o^2-r_o'^2)(y_o+r_o')}{x_o^2+(y_o+r_o')^2}. \quad (19)$$

Consider now the time instant t. The (v, κ) input pairs can be determined, which cause collision with the static obstacle B_i at t. The velocity obstacle for a car-like robot (VOCL) is the union of these points (see Figure 5):

$$VOCL_i(t) = \{ (v, \kappa) | A(t, v, \kappa) \cap B_i \neq \emptyset \}, \quad (20)$$

where $A(t, v, \kappa)$ represents the robot at time-moment t moving from $t_0 = 0$ and from the origin on a circular path with radius $\frac{1}{\kappa}$ with velocity v. κ and v can be calculated from (x, y) position using (9)-(11).

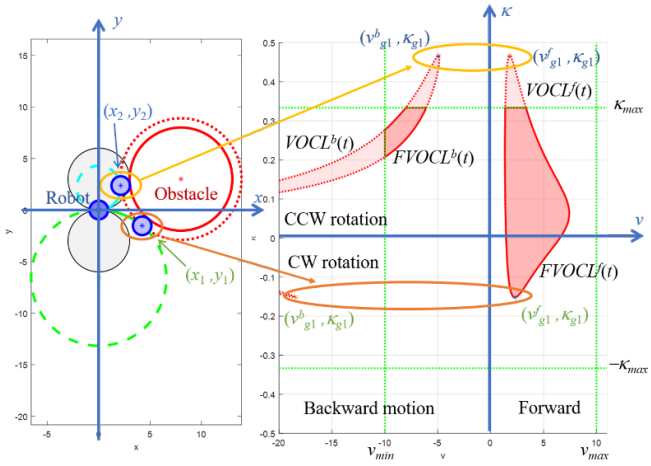


Figure 5. a. Robot and obstacle at a grazing case. b. Velocity obstacle $VOCL(t)$ and feasible VO $FVOCL(t)$ for time instant t

From the grazing points (x_1, y_1) and (x_2, y_2), one can determine the corresponding points in $VOCL_i(t)$. The curvatures κ_{g1} and κ_{g2} are used and the forward velocities v_{g1}^f, v_{g2}^f and backward velocities v_{g1}^b, v_{g2}^b can also be calculated if (x_1, y_1) or (x_2, y_2) should be reached on a circular path during time t.

(v_{g1}, κ_{g1}) and (v_{g2}, κ_{g2}) lie on the boundary of $VOCL_i(t)$. All other points on the boundary of $VOCL_i(t)$ correspond to a motion (i.e. (v, κ) values) which moves the robot during time t to a boundary point of the obstacle, but before or after the time instant t a collision occurs.

$VOCL_i(t)$ according to (20) may contain (v_j, κ_j) pairs, such that $|\kappa_j| > \kappa_{max}$ or $|v_j| > v_{max}$. These κ_j or v_j values are not feasible. Feasible VO ($FVOCL_i(t) \subseteq VOCL_i(t)$) contains only feasible curvature and velocity values:

$$FVOCL_i(t) = \{(v, \kappa) | A(t, v, \kappa) \cap B_i \neq \emptyset \wedge$$

$$|v| \leq v_{max} \wedge |\kappa| \leq \kappa_{max}\}. \quad (21)$$

In Figure 5 $VOCL_i(t)$ is delimited by dotted line and $FVOCL_i(t)$ is bounded by solid line.

Taking time-moments in a time-interval $t \in [t_0, t_h]$ one can get $FVOCL_i$ for a static obstacle B_i:

$$FVOCL_i = \bigcup_{t \in [t_0, t_h]} FVOCL_i(t). \quad (22)$$

If the robot moves according to $(v, \kappa) \notin FVOCL_i$, the robot will not collide with the static obstacle B_i in the given time interval $[t_0, t_h]$.

Notice, that the selection of t_h influences the effectiveness and the computational demand of the method. If a small value was selected for t_h, a mobile robot may collide with a moving obstacle. On the other hand, a large time-horizon can result that VO includes almost the complete velocity space [13].

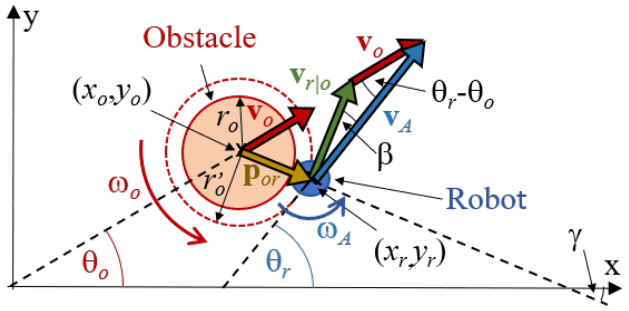
5.2. Calculating VOCL for Moving Obstacles

For a moving obstacle $B_i(t)$ the definition of $VOCL_i(t)$ for a given time-moment t is similar to (20):

$$VOCL_i(t) = \{ (v, \kappa) | A(t, v, \kappa) \cap B_i(t) \neq \emptyset \}. \quad (23)$$

$FVOCL_i(t)$ and $FVOCL_i$ can be determined as well, according to (21) and (22).

The main difference is the following: the grazing points in $VOCL_i(t)$ ((v_{g1}, κ_{g1}) and (v_{g2}, κ_{g2})) do not represent grazing cases any more if the obstacle is moving since the position of the grazing points depends on the motion of the obstacle.



To analyze VOCL and the grazing points, the following notations are used (see Figure 6):

- A time instant t_A is considered.
- The position of the moving obstacle at t_A is denoted by $(x_o(t_A), y_o(t_A))$. The angle between the positive x-axis and

the velocity vector $\mathbf{v}_o(t_A)$ of the obstacle is denoted by $\theta_o(t_A)$, the absolute value of the velocity is $v_o(t_A)$. $\omega_o(t_A)$ denotes the turning rate of the obstacle. The vector $\boldsymbol{\omega}_o(t_A)$ is parallel to the positive z-axis, if the direction of the rotation is positive, and to the negative z-axis in case of clockwise rotation.

- An input pair (v_A, κ_A) is selected.
- If a robot is moving with a velocity v_A on a circle with curvature κ_A for time t_A, its position $(x_r(t_A), y_r(t_A))$ and orientation $\theta_r(t_A)$ can be determined according to (8). The absolute value of the robot's velocity v_A remains unchanged during the circular motion, but the vector of the velocity $\mathbf{v}_A(t_A)$ changes, since the orientation of the robot is also modified. The turning rate of the robot is $\omega_A = \kappa_A\, v_A$. The direction of vector $\boldsymbol{\omega}_A$ is parallel to the positive z-axis, if $\omega_A > 0$, and to the negative z-axis if $\omega_A < 0$. The vector $\boldsymbol{\omega}_A$ is constant during the circular motion, since its direction and length (i.e. ω_A) are unchanged.
- The relative velocity of the robot and the obstacle at time moment t_A is denoted by

$$\mathbf{v}_{r|o}\,(t_A) = \mathbf{v}_A(t_A) - \mathbf{v}_o(t_A). \tag{24}$$

- The angle between the velocity vector $\mathbf{v}_A(t_A)$ of the robot and the relative velocity $\mathbf{v}_{r|o}(t_A)$ is:

$$\beta(t_A) = \operatorname{atan}\frac{v_o(t_A)\sin(\theta_r(t_A)-\theta_o(t_A))}{v_A - v_o(t_A)\cos(\theta_r(t_A)-\theta_o(t_A))}. \tag{25}$$

- $\mathbf{p}_{or}$ is a vector connecting the center of the obstacle to the center of the robot.
- The angle between the positive x-axis and $\mathbf{p}_{or}$ is:

$$\gamma(t_A) = \operatorname{atan}\frac{y_r(t_A)-y_o(t_A)}{x_r(t_A)-x_o(t_A)}. \tag{26}$$

The following two propositions follow directly from the definition of $VOCL_i(t)$.

Proposition 1:

$$\sqrt{\left(x_r(t_A)-x_o(t_A)\right)^2 + (y_r(t_A)-y_o(t_A))^2} \le r_o' \Leftrightarrow$$

$$(v_A, \kappa_A) \in VOCL_i(t_A). \tag{27}$$

If $|\kappa_A| \le \kappa_{max}$ and $|v_A| \le v_{max}$, then $(v_A, \kappa_A) \in FVOCL_i(t_A)$ is also true. Similar to [14], the boundary points of the set $VOCL(t)$ are denoted by $\delta VOCL(t)$.

Proposition 2:

$$\sqrt{\left(x_r(t_A)-x_o(t_A)\right)^2 + (y_r(t_A)-y_o(t_A))^2} = r_o' \Leftrightarrow$$

$$(v_A, \kappa_A) \in \delta VOCL_i(t_A). \tag{28}$$

For the grazing points (v_{g1}, κ_{g1}) and (v_{g2}, κ_{g2}) in case of a static obstacle still holds true that $(v_{g1}, \kappa_{g1}) \in \delta VOCL_i(t_A)$ and $(v_{g2}, \kappa_{g2}) \in \delta VOCL_i(t_A)$ where δ denotes the region boundary. But these points are not grazing points any more, if the obstacle is moving.

Theorem 3: (v_A, κ_A) is a grazing point in $VOCL_i(t_A)$ if

$$\theta_r(t_A) + \beta(t_A) - \gamma(t_A) = \frac{(2a-1)\pi}{2}, a \in \mathbb{Z}. \tag{29}$$

Proof: $\theta_r(t_A) + \beta(t_A)$ denotes the angle between the relative velocity $\mathbf{v}_{r|o}(t_A)$ and the positive x-axis. $\gamma(t_A)$ is the angle between the positive x-axis and $\mathbf{p}_{or}$ (the vector connecting the center of the robot and the obstacle). The condition for the grazing situation is that the relative velocity $\mathbf{v}_{r|o}(t_A)$ is perpendicular to $\mathbf{p}_{or}$ which is equivalent to the case where the relative velocity $\mathbf{v}_{r|o}(t_A)$ is tangent to the circle of the enlarged obstacle with radius r_o'.

Similar to [14], it can be proven that if $\mathbf{v}_{r|o}(t_A)$ is perpendicular to $\mathbf{p}_{or}$, the robot grazes the obstacle.

A point P is selected on the boundary of the robot. $\mathbf{p}_P$ is a vector which goes from the center of the robot to P. The velocity of the point P is:

$$\mathbf{v}_P(t_A) = \mathbf{v}_A(t_A) + \boldsymbol{\omega}_A \times \mathbf{p}_P. \tag{30}$$

Similarly, a point Q can be selected on the boundary of the obstacle. $\mathbf{p}_Q$ is a vector which goes from the center of the obstacle to Q. The velocity of Q reads:

$$\mathbf{v}_Q(t_A) = \mathbf{v}_o(t_A) + \boldsymbol{\omega}_o(t_A) \times \mathbf{p}_Q. \tag{31}$$

If the point P of the robot grazes point Q on the obstacle, vector $\mathbf{p}_{or}$ has to be parallel to $\mathbf{p}_Q$ and $-\mathbf{p}_P$. Moreover, grazing case can only occur, if the relative velocity of point P according to point Q is tangent to the circle of the obstacle, i.e. if it is perpendicular to $\mathbf{p}_{or}$ (and accordingly to $\mathbf{p}_Q$ and $\mathbf{p}_P$ as well). The relative velocity of P according to Q is:

$$\mathbf{v}_{P|Q}(t_A) = \mathbf{v}_P(t_A) - \mathbf{v}_Q(t_A) =$$

$$= \mathbf{v}_A(t_A) + \boldsymbol{\omega}_A \times \mathbf{p}_P - \mathbf{v}_o(t_A) - \boldsymbol{\omega}_o(t_A) \times \mathbf{p}_Q. \tag{32}$$

$\mathbf{v}_{P|Q}(t_A)$ is perpendicular to $\mathbf{p}_{or}$, if:

$$\mathbf{v}_{P|Q}(t_A) \cdot \mathbf{p}_{or} = 0$$

$$(\mathbf{v}_A + \boldsymbol{\omega}_A \times \mathbf{p}_P) \cdot \mathbf{p}_{or} - \left(\mathbf{v}_o + \boldsymbol{\omega}_o \times \mathbf{p}_Q\right) \cdot \mathbf{p}_{or} = 0. \tag{33}$$

Both $(\boldsymbol{\omega}_A \times \mathbf{p}_P) \cdot \mathbf{p}_{or}$ and $\left(\boldsymbol{\omega}_o(t_A) \times \mathbf{p}_Q\right) \cdot \mathbf{p}_{or}$ equal 0, hence one gets

$$\left(\mathbf{v}_A(t_A) - \mathbf{v}_o(t_A)\right) \cdot \mathbf{p}_{or} = 0, \tag{34}$$

which is equivalent to that $\mathbf{v}_{r|o}(t_A)$ is perpendicular to $\mathbf{p}_{or}$.■

Notice, that the obstacles can move on arbitrary paths with arbitrary velocity profiles. The algorithm only supposes that for every time moment $t_A \in [t_0, t_h]$ the positions and the velocity vectors of each obstacles are known.

5.3. Trajectories Avoiding Obstacles

If a goal position is given in the workspace of the robot, a motion should be planned which arrives to this point. During the motion, no collision should occur.

Fiorini and Shiller presented some heuristic search methods using Velocity Obstacles [6]. Applying these, one can easily select at each time moment a feasible velocity vector which moves the robot to the direction of the goal such that it avoids collisions with the obstacles. Using these heuristics alone does not guarantee to find an optimal solution. They designed an off-line global search method as well. A tree was generated for avoidance maneuvers using VO.

Shiller et al. presented avoidance maneuvers using NLVO [14]. These avoidance maneuvers can be used in a local or global motion planner as well.

The VOCL method can also be used in all above methods to plan the motion for car-like mobile robots to a given goal avoiding static and moving obstacles.

5.4. The Safest Solution

Another possibility for the selection of (v, κ) pair is to determine the safest motion. This method is called as the Safety Velocity Obstacles method (SVO).

The most important goal of the SVO method is to ensure the safest path for the robot during the motion.

This method checks, how far a possible $(v, \kappa) \notin FVOCL$ pair is from the nearest $FVOCL_i$ (name this distance VO_{dist}). If the distance is bigger than a predefined threshold d_{max}, then the value of VO_{dist} is set to this maximum distance.

Hence, a normalized distance can be defined:

$$VO_{norm} = \frac{VO_{dist}}{d_{max}} \tag{35}$$

where $VO_{norm} \in [0,1]$. After that the value of the cost can be defined as

$$VO_{cost} = 1 - VO_{norm}. \tag{36}$$

If the selected (v, κ) pair is far from the nearest $FVOCL_i$, then the value of VO_{norm} will be a big number (near to 1). So, such (v, κ) pair has to be chosen for the robot, where the value of VO_{cost} is minimal.

So, with the SVO method one can plan the safest motion in dynamic environment.

6. Simulation Results

An example for VOCL is presented here. A circular robot with radius $r_o = 0.9\text{m}$ is given. Its minimal turning radius is $R_{min} = 3\text{m}$ (i.e. $\kappa_{max} = \frac{1}{3}\text{m}^{-1}$) and its maximal velocity is $v_{max} = 7\frac{\text{m}}{s}$. Seven circular obstacles with different radii are presented in the workspace (see Figure 7). Four obstacles move (B_1, B_2, B_3, B_4) and there are three static obstacles (B_5, B_6, B_7). The velocity vectors and the paths of the dynamic obstacles are also depicted in Figure 7. The VOCL and FVOCL for $t_h = 10\text{s}$ is given in Figure 8. White areas represent (v, κ) pairs which correspond to collision-free motion for t_h.

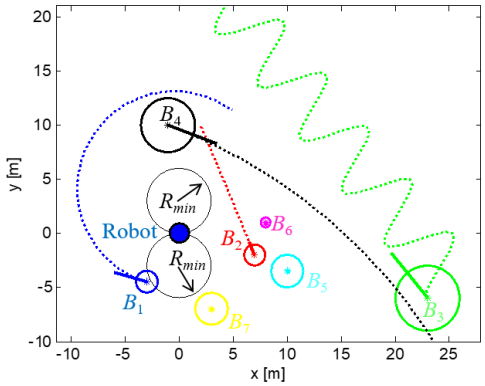
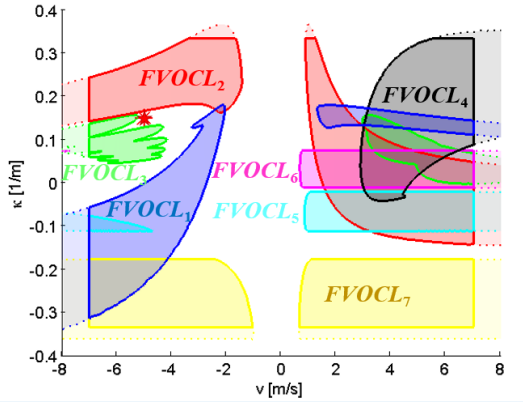


Figure 7. An example with seven obstacles at $t_0 = 0\text{s}$. The velocity vectors of dynamic obstacles are depicted by solid lines, the paths of the obstacles are denoted by dotted lines (with $t_h = 10s$).

A collision-free example is also presented here. $\kappa = 0.15\text{m}^{-1}$ ($R = 6.\dot{6}\text{m}$) and $v = -5\frac{\text{m}}{s}$ was selected. The corresponding (v, κ) point is depicted by a red star in Figure 8. The motion of the robot and the movement of the obstacles are depicted in Figure 9.



7. Conclusion

Velocity obstacles (VO) and non-linear velocity obstacles (NLVO) methods can be used to plan a collision-free motion for a planar robot moving among static and moving obstacles. VO supposes that the obstacles move on straight-lines with constant velocities. If the path of the obstacle is not a straight-line, NLVO can be applied. These methods determine a velocity vector for the robot, which results in a collision-free motion in a time-interval. Applying this velocity, the robot will move on a straight-line according to the direction of the selected velocity vector. Both methods assume that the position and the motion of the obstacles are known.

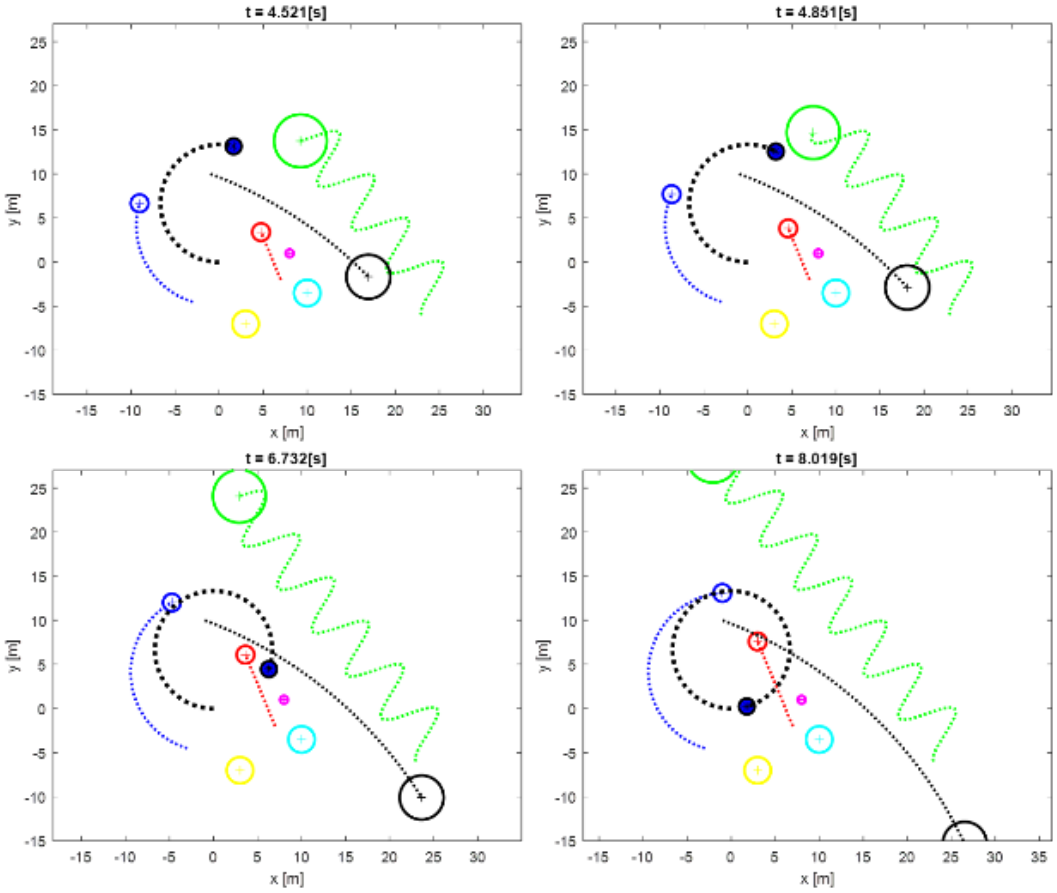


Figure 9. Motion of robot and obstacles in the example ($v = -5\frac{m}{s}$, $\kappa = 0.15m^{-1}$). The robot is depicted by a blue disc, and its path is shown by the thick black dotted line.

The method, presented in this paper is similar to NLVO but in this case the robot is a car-like mobile robot. This robot cannot move to arbitrary direction. The direction of its velocity vector is determined by its orientation. The robot can move on a straight-line according to its orientation or on a circular path. Hence, the presented VOCL method determines only the magnitude of the velocity vector and, additionally, the curvature of the circular path to follow to avoid collisions.

The next stage of this work is to implement the VOCL method for a car-like mobile robot, such that the collision-free motion is determined in real-time.

Our future goal is to take the non-circular shape of the robot (e.g. car-like rectangle) also into consideration during the construction of VOCL.

[22]. In this case a continuous curvature path could be determined. The drawback is that the computational complexity would be larger.

References

[1] E. Gincsainé Szádeczky-Kardoss, B. Kiss, "Velocity obstacles for Dubins-like mobile robots" in 25th Mediterranean Conference on Control and Automation (MED), Valletta, Malta, 2017. https://doi.org/10.1109/MED.2017.7984142

[2] J. C. Latombe, Robot motion planning, Kluwer, Boston, 1991.

[3] S. M. LaValle, Planning algorithms, Cambridge University Press, 2006.

[4] K. Kant, S. W. Zucker, "Toward efficient trajectory planning: The path-velocity decomposition" Int. J. Robot. Res., **5**(3), 72–89, 1986. https://doi.org/10.1177/027836498600500304

[5] E. Szádeczky-Kardoss, B. Kiss, "Motion planning in dynamic environments with the rapidly exploring random tree method" Int. Rev. of Automatic Control, **1**(1), 109–117, 2008.

[6] P. Fiorini, Z. Shiller, "Motion planning in dynamic environments using velocity obstacles" Int. J. of Robot. Res., **17**(7), 760–772, 1998. https://doi.org/10.1177/027836499801700706

[7] M. Seder, I. Petrovic, "Dynamic window based approach to mobile robot motion control in the presence of moving obstacles" in International Conference on Robotics and Automation, Roma, Italy, 1986–1991, 2007. https://doi.org/10.1109/ROBOT.2007.363613

[8] O. Khatib, "Real-Time obstacle avoidance for manipulators and mobile robots". Int. J. Robot. Res., **5**(1), 90–98, 1986. https://doi.org/10.1177/027836498600500106

[9] C. Qixin, H. Yanwen, Z. Jingliang, "An evolutionary artificial potential field algorithm for dynamic path planning of mobile robot" in IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 3331–3336, 2006. https://doi.org/10.1109/IROS.2006.282508

[10] P. Shi, J. N. Hua, "Mobile robot dynamic path planning based on artificial potential field approach" Adv. Mat. Res., **490–495**, 994–998, 2012. https://doi.org/10.4028/www.scientific.net/AMR.490-495.994

[11] D. Fox, W. Burgard, S. Thrun, "The dynamic window approach to collision avoidance" IEEE Robot. Autom. Mag., **4**(1), 23–33, 1997. https://doi.org/10.1109/100.580977

[12] T. Fraichard, H. Asama, "Inevitable collision states – a step towards safer robots?" Adv. Robotics, **18**(10), 1001–1024, 2004. https://doi.org/10.1163/1568553042674662

[13] L. Martinez-Gomez, T. Fraichard, "Collision avoidance in dynamic environments: an ICS-based solution and its comparative evaluation" in IEEE International Conference on Robotics and Automation, Kobe, Japan, 100–105, 2009. https://doi.org/10.1109/ROBOT.2009.5152536

[14] Z. Shiller, F. Large, S. Sekhavat, C. Laugier, "Motion planning in dynamic environments: Obstacles moving along arbitrary trajectories". in IEEE

International Conference on Robotics and Automation, Seoul, South Korea, 3716–3721, 2001. https://doi.org/10.1109/ROBOT.2001.933196

[15] C. Moon, W. Chung, "Trajectory time scaling of a mobile robot to avoid dynamic obstacles on the basis of the INLVO" Adv. Robotics, **27**(15), 1189–1198, 2013. https://doi.org/10.1080/01691864.2013.819604

[16] Y. Kuwata, M. T. Wolf, D. Zarzhitsky, T. L. Huntsberger, "Safe maritime autonomous navigation with COLREGS, using velocity obstacles" IEEE J. Oceanic Eng., **39**(1), 110–119, 2014. https://doi.org/10.1109/JOE.2013.2254214

[17] J. Snape, J. v. d. Berg, S. J. Guy, D. Manocha, "The hybrid reciprocal velocity obstacle". IEEE T. Robot., **27**(4), 696–706, 2011. https://doi.org/10.1109/TRO.2011.2120810

[18] D. Wilkie, J. v. d. Berg, D. Manocha, "Generalized velocity obstacles". in IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 5573–5578, 2009. https://doi.org/10.1109/IROS.2009.5354175

[19] W. A. Wolfe, "Analytical design of an Ackermann steering linkage" J. Eng. Ind.- T. ASME, 11, 11–14, 1959.

[20] L. E. Dubins, "On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents" Am. J. Math., **79**(3), 497–517, 1957. https://doi.org/10.2307/2372560

[21] J. A. Reeds, L. A. Shepp, "Optimal paths for a car that goes both forwards and backwards". Pac. J. Math., **145**(2), 367–393, 1990

[22] T. Fraichard, A. Scheuer, R. Desvigne, "From Reeds and Shepp's to continuous-curvature paths" in International Conference on Advanced Robotics, Tokyo, Japan, 585–590, 1999.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 231-237 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

An Aggregation Model for Energy Resources Management and Market Negotiations

Omid Abrishambaf, Pedro Faria*, João Spínola, Zita Vale

GECAD – Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development, Institute of Engineering – Polytechnic of Porto (ISEP/IPP), Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal

ARTICLE INFO

Article history:
Received: 04 December,2017
Accepted: 27 January,2018
Online: 06 April,2018

Keywords:
Demand Response
Aggregator
Distributed Generation
Smart grid
K-Mean Clustering

ABSTRACT

Currently the use of distributed energy resources, especially renewable generation, and demand response programs are widely discussed in scientific contexts, since they are a reality in nowadays electricity markets and distribution networks. In order to benefit from these concepts, an efficient energy management system is needed to prevent energy wasting and increase profits. In this paper, an optimization based aggregation model is presented for distributed energy resources and demand response program management. This aggregation model allows different types of customers to participate in electricity market through several tariffs based demand response programs. The optimization algorithm is a mixed-integer linear problem, which focuses on minimizing operational costs of the aggregator. Moreover, the aggregation process has been done via K-Means clustering algorithm, which obtains the aggregated costs and energy of resources for remuneration. By this way, the aggregator is aware of energy available and minimum selling price in order to participate in the market with profit. A realistic low voltage distribution network has been proposed as a case study in order to test and validate the proposed methodology. This distribution network consists of 25 distributed generation units, including photovoltaic, wind and biomass generation, and 20 consumers, including residential, commercial, and industrial buildings.

1. Introduction

The present paper is an extension of work originally proposed in *2017 IEEE Manchester PowerTech* [1]. The generation variation in Distributed Renewable Energy Resources (DRER) is a topic of introduction in a lot of research works, since they have a key role in nowadays power system [2]. By appropriate management on the consumption in demand side, energy efficiency and optimal energy usage should be addressed. Curtailment Service Provider (CSP), Virtual Power Player (VPP), and aggregator are entities that can provide reliable solutions for the management of consumption and generation resources, since these can be aggregated and represented as a unique resource in electricity markets [3-5].

In this context, an aggregator is responsible to optimally manage a certain number of resources in a region, and aggregate them as one resource. This simplifies the process of energy negotiation in electricity markets [6]. Moreover, if other players, such as Balance Responsible Parties (BRPs), exist in the network, the role of aggregator would be more efficient and important [7].

Nowadays, there are several European countries that employ the aggregator concept for electricity consumers [8]. As an example, France is one of these countries that accepted aggregated loads in every ancillary service program, and BRPs and aggregators have been reorganized based on [9], [10]:

- Performing electricity market negotiations, to calculate compensation costs by aggregator for BRP;
- Aggregator has no direct interaction with BRP, however, it establishes contract with an electricity supplier in order to have flexibility services.

In fact, an aggregator is accountable not only for DRERs, but also is responsible for Demand Response (DR) programs [10]. According to the Federal Energy Regulatory Commission (FERC) [11], DR program is referred as "Changes in electric use by demand-side resources from their normal consumption patterns in response to changes in the price of electricity, or to

*Corresponding Author: Pedro Faria, GECAD Research Group Polytechnic of Porto (ISEP/IPP), Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal, Email: pnsfaria@gmail.com

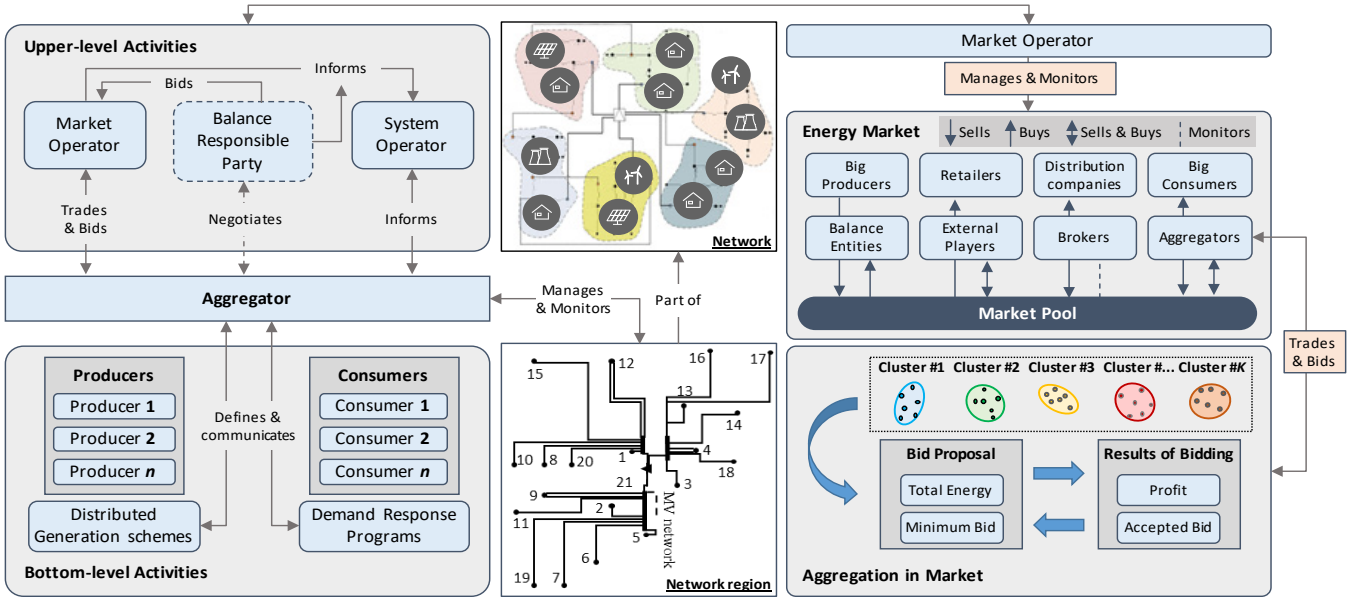


Figure 1. Overal architecture of the aggregation model.

incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized". The role of an aggregator in terms of DR programs is to gather all electricity consumer who can participate in DR programs, and present them as one. Therefore, it can be considered as a flexible player [12]. For this purpose, the aggregator can establish bidirectional contracts with end-users for DR programs to manage consumption resources, and consequently, to have flexibility in electricity market negotiations. In order to manage the generation of end-users, which are considered as prosumer (a consumer who is able to produce electricity), the aggregator can play the role of VPP, as [13]-[14] demonstrated before. It is clear that the generation capacity of these prosumers is not significant, thus, the network management would be difficult for system operators. Therefore, the need of a third party, namely an aggregator, is evident to gather all these small-scale consumption and generation resources, and participate in electricity market.

This paper represents an optimization based aggregation model for DRERs and DR programs managements, which enables small and medium resources to have active participation in the electricity markets. The aggregator controls demand-side customers by providing them several tariffs based DR programs, which brings flexibility in the electricity market negotiations. Moreover, this aggregator model gathers energy of resources and aggregated costs to be aware of available energy and minimum selling cost for defining remunerations, and also participate in the market with profit.

The rest of the paper is organized as follow. Section 2 details the aggregator model architecture considered for the aggregation. Section 3 describes the mathematical formulation considered for the optimization problem and aggregation process. Section 4 explains a case study that will test and validate the proposed method, and its results are expressed in Section 5. Finally, main conclusions of the work are proposed in Section 6.

2. Aggregator Model Architecture

This section focuses on how the presented aggregation model performs scheduling, aggregating and remuneration. The overall view of the presented model is illustrated by Figure 1. In this aggregation model, the consumption and generation resources are classified in several groups, where the output of the aggregation process will be the energy and cost of each group. As one can see in Figure 1 and also proposed in [15], the functionality of the aggregator is categorized in two sections of upper-level and bottom-level. In the upper level, the aggregator negotiates with players, such as market operator, BRP, and system operator; however, in the bottom level, it deals with demand-side users, namely small and medium scale consumers and producers.

The aggregator performs the scheduling process relying on external suppliers, Distributed Generation (DG) especially renewable resources, and DR programs. The customers who can execute DR programs would be able to establish contract with aggregator in three programs: load shifting, load reduction, and load curtailment. The load shifting model has been adapted from [16], and in this aggregation model it is considered as a free DR program. Load reduction and curtailment are the programs that aggregator takes them into account for scheduling and participating in the market. The aggregator considers a linear cost function for all external suppliers, DGs, load reduction and load curtailment. In this model, the aggregation process is done by K-Means Clustering algorithm by respect to the scheduled energy and its costs. In the aggregation process, only the resources that have been selected form the scheduling, are considered, and the rest that have no interaction in scheduling process, will not be considered. The aggregator categorizes the resources in several groups, and specifies a remuneration for each group, which called group tariff. This means the remuneration process should be calculated after the aggregation. The resources that are classified in a group, will be remunerated with same price. For this reason, the maximum price available in each group will be selected for group tariff. Therefore,

the cheapest resource in the group will be motivated to participate in aggregation, since the group tariff is greater than the price initially defined, and also the most expensive resource will be satisfied, since the group tariff is as same as the price that it proposed.

In this way, the aggregator is able to participate in the market with a bid for each group. In each bid, the aggregator deliberates the gathered energy from the resources and also the group tariff as the minimum rate. The energy in each group is related to the aggregation of scheduled resources of that related group, therefore, the aggregator can easily manage its activities. On the other hand, the aggregator will be able to have negotiation in the market by biding the available energy of each group with a certain price, where this price should be greater or equal to the group tariff for the aggregator to gain profits or at least obtain the amount expended for the resources.

3. Optimization problem

The mathematical formulation regarding the presented aggregation model, especially resource scheduling, will be presented in this section. The optimization problem developed for the aggregator scheduling contains several continuous and discrete variables, therefore, the problem is considered as a mixed-integer linear problem (MILP). The objective function considered for the aggregation model is to minimize its Operational Cost (OC) and is shown by (1). It should be noted that in this model it is supposed the technical verification of the network is the obligation of the network operator, and the aggregator is not responsible for this matter.

$$\begin{aligned} MinOC = & \sum_{s=1}^{S} P_{(s,t)}^{Sup} \cdot C_{(s,t)}^{Sup} + \sum_{p=1}^{P} P_{(p,t)}^{DG} \cdot C_{(p,t)}^{DG} \\ & + \sum_{c=1}^{Cs} \left[\begin{array}{l} P_{(c,t)}^{Red} \cdot C_{(c,t)}^{Red} + P_{(c,t)}^{Cut} \cdot C_{(c,t)}^{Cut} \\ + \sum_{d=1}^{T} P_{(c,t,d)}^{Shift} \cdot C_{(c,t,d)}^{Shift} \end{array} \right] \\ & \forall t \in \{1,...,T\} \end{aligned} \tag{1}$$

In this objective function, $P_{(s,t)}^{Sup}$ is purchased energy from external supplier, $P_{(p,t)}^{DG}$ denotes the attained energy from DG, $P_{(c,t)}^{Red}$ stands for DR load reduction, $P_{(c,t)}^{Cut}$ is for DR load curtailment, and $P_{(c,t,d)}^{Shift}$ represents DR load shifting.

There are several constraints that should be considered in the objective function. The first constraint stands for load balance, as (2) shows. In this equation, $P_{(c,t)}^{Load}$ presents the required demand of consumers.

Also, the technical limitations of all resources available in the proposed methodology should be considered. Therefore, (3) represents the generation limitations of external supplier in term of minimum and maximum ($P_{(s,t)}^{Sup\min}$, $P_{(s,t)}^{Sup\max}$), and (4) considers DG limitations ($P_{(p,t)}^{DG\min}$, $P_{(p,t)}^{DG\max}$).

$$\sum_{c=1}^{Cs} \left[\begin{array}{l} P_{(c,t)}^{Load} - P_{(c,t)}^{Red} - P_{(c,t)}^{Cut} \\ - \sum_{d=1}^{T} \left(P_{(c,t,d)}^{Shift} - P_{(c,d,t)}^{Shift} \right) \end{array} \right] - \sum_{s=1}^{S} P_{(s,t)}^{Sup} - \sum_{p=1}^{P} P_{(p,t)}^{DG} = 0 \quad \forall t \in \{1,...,T\} \tag{2}$$

$$P_{(s,t)}^{Sup\min} \le P_{(s,t)}^{Sup} \le P_{(s,t)}^{Sup\max} \quad \forall s \in \{1,...,S\}, \forall t \in \{1,...,T\} \tag{3}$$

$$P_{(p,t)}^{DG\min} \le P_{(p,t)}^{DG} \le P_{(p,t)}^{DG\max} \quad \forall p \in \{1,...,P\}, \forall t \in \{1,...,T\} \tag{4}$$

DR technical limitations, including load reduction, curtailment, shifting, are presented by (5)-(8).

$$P_{(c,t)}^{Red\min} \le P_{(c,t)}^{Red} \le P_{(c,t)}^{Red\max} \tag{5}$$

$$P_{(c,t)}^{Cut\min} \le P_{(c,t)}^{Cut} \le P_{(c,t)}^{Cut\max} \tag{6}$$

$$P_{(c,t)}^{Cut} = P_{(c,t)}^{Cut\max} \cdot X_{(c,t)}^{Cut}, \quad X_{(c,t)}^{Cut} \in \{0,1\} \quad \forall c \in \{1,...,C\}, \forall t \in \{1,...,T\} \tag{7}$$

$$P_{(c,t,d)}^{Shift\min} \le P_{(c,t,d)}^{Shift} \le P_{(c,t,d)}^{Shift\max} \tag{8}$$

Although load shifting may not be pleasant for end-users, it is an appropriate and practical tool for aggregator. Load shifting process may limit consumers use of devices in a certain period, however, it enables the aggregator to manage the consumption based on the offered generation capacity. For this purpose, the limitations of maximum energy that will be shifted out from a period ($P_{(c,t)}^{Shift_out}$), and enters to another period ($P_{(c,t)}^{Shift_in}$) are proposed in (9) and (10).

$$\sum_{d=1}^{T} P_{(c,t,d)}^{Shift} \le P_{(c,t)}^{Shift_out} \tag{9}$$

$$\sum_{d=1}^{T} P_{(c,d,t)}^{Shift} \le P_{(c,t)}^{Shift_in} \quad \forall c \in \{1,...,C\}, \forall t,d \in \{1,...,T\} \tag{10}$$

Moreover, (11) demonstrates the constraint regarding the groups tariff and their remuneration, which is the maximum price of group. The groups are separated based on the type of available resources (DG or DR).

$$\begin{aligned} & G_{(k,t)}^{DG} = max(C_{(p,t)}^{DG}) \\ & G_{(k,t)}^{DR} = max(C_{(c,t)}^{Red}, C_{(c,t)}^{Cut}), \forall c \in \{1,...,C\} \\ & \forall p \in \{1,...,P\}, \forall k \in \{1,...,K\}, \forall t \in \{1,...,T\} \end{aligned} \tag{11}$$

As a summary, the mathematical formulation for resources scheduling and their remuneration performed by the aggregator have been explained in this part. The methodology presented in this section will be employed in a case study in the next section.

4. Case Study

In order to examine the model represented in this paper, a case study is proposed. For this purpose, an low voltage distribution network of a university campus, in Porto, Portugal, is considered for the aggregator, which has been adapted from [17]. This distribution network is shown in the bottom of Figure 1 (Network region) and is considered as a part of main network. The network consists of underground electrical lines with 21 buses, where a MV/LV transformer in BUS #21, connects the campus network to the main network.

For this case study, we considered that there are 20 consumers, and 26 producers in the network. The consumers include 8 Residential (RE) buildings, 10 Commercial buildings in three scales of small (C-S), medium (C-M), and large (C-L), and 2 Industrial (IN) units, which are classified based on average daily consumption. Moreover, producers consist of renewable resources including 20 Photovoltaic (PV) units and 4 wind generators, 1 biomass, and external suppliers. The generation and consumption profiles of whole network considered for day-ahead scheduling in a winter day are shown on Figure 2.

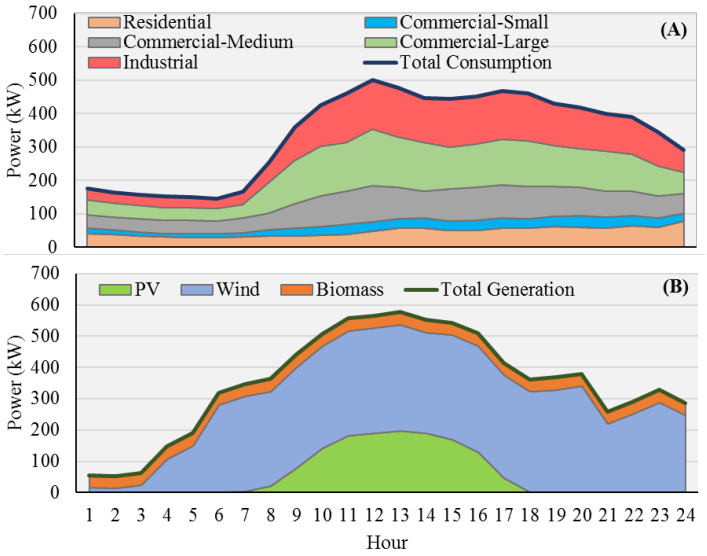


Figure 2. Day-ahead profiles of the network considered for case study: (A) Consumption, (B) Production.

As it can be seen in Figure 2 – (A), large commercial buildings and industrial units have occupied a huge part of consumption, and peak periods start from period #10 to #23. In this case study, it is presumed that the biomass production, and external suppliers have maximum capacity of 40 and 500 kW respectively. The external suppliers profile is not illustrated in Figure 2 – (B), since it is out of scope of figure and is a constant value during all periods. Moreover, it is considered that all producers would be able to contribute in the aggregation process, except external suppliers. Additionally, as you can see in Figure 2 – (B), since a winter day selected for the case study, PV producers have no significant generation, therefore, the aggregator should rely on wind, biomass, external suppliers and DR programs to prevent purchasing energy from the market. However, by comparing both parts of Figure 2, it is obvious that there are some periods that aggregator has more generation than consumption, therefore, it would be able to sell energy to the market and gain profits.

Regarding DR programs, Figure 3 demonstrates linear costs considered for each consumer based on its type. These costs are for load reduction and load curtailment, where 20% of the initial consumption is considered as maximum load reduction, and 15% for maximum load curtailment.

Furthermore, the linear costs considered for energy resources are shown on Figure 4. Each point in Figure 4 is the individual cost of each resource, where resource #1 to #20 are all PV, #21 to #24 are wind generators, #25 is biomass unit, and #26 illustrates external suppliers. It is should be mentioned that the costs demonstrated in Figure 4, are constant in all periods.

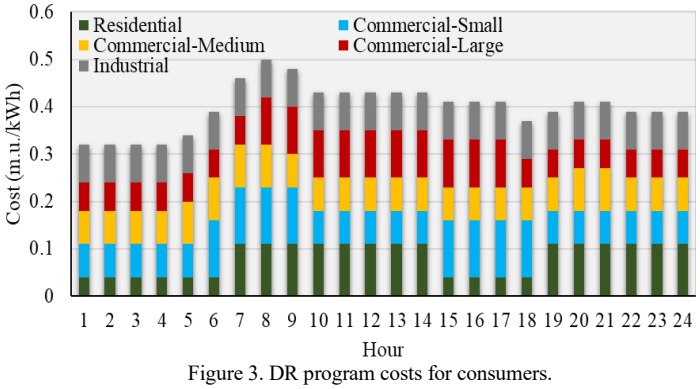


Figure 3. DR program costs for consumers.

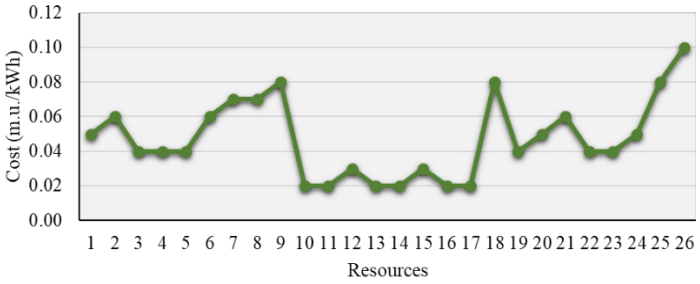


Figure 4. Individual cost for each energy resource.

Additionally, Figure 5 represents the day-ahead market prices considered for the aggregator in order to participate in market negotiations.

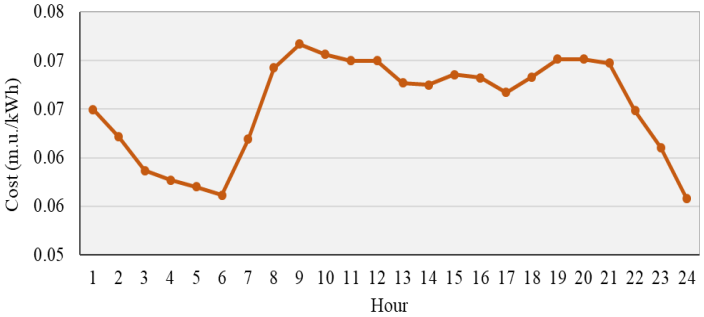


Figure 5. MIBEL market price for Portugues section in a winter day.

These prices are for a winter day in 2017 and have been adapted from Portuguese sector of Iberian Electricity Markets (MIBEL) [18]. In order to model the participation of the aggregator in the electricity markets, a market place should be taken into account, to guaranty its contribution in the competition. For this purpose, a market pool is an appropriate solution to ensure that third parties, such as aggregator, would be able to present energy bids.

5. Results

This section concerns the aggregation and scheduling results of the case study presented in the previous section. The optimization problem of aggregation and scheduling presented in this paper, has been solved through TOMLAB [19]. Additionally, the market negotiation results are represented, which shows how the aggregator utilizes these results for providing a bid. In the case study, we considered that the aggregator meets a drop from external suppliers in first four periods that can supply only 10% of their capacity. The reason of this lack of energy is considered as a fault or any other causes in the external suppliers. Figure 6 shows the network consumption before and after the scheduling of aggregator.

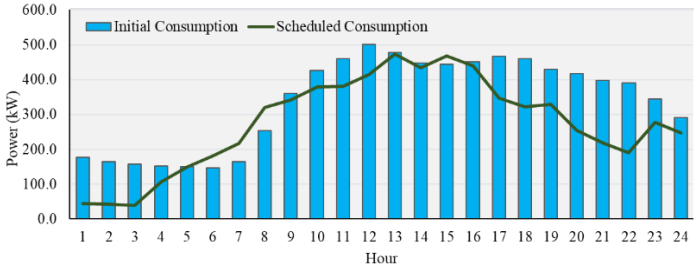


Figure 6. Total consumption of the network.

The scheduling results shown in Figure 6 are based on DG and available energy during each period. Additionally, there are several periods that scheduled consumption profile are greater or smaller than the initial profile. This is due to the utilization of DR programs by aggregator. With this in mind, Figure 7 illustrates more information regarding the generation and DR scheduling.

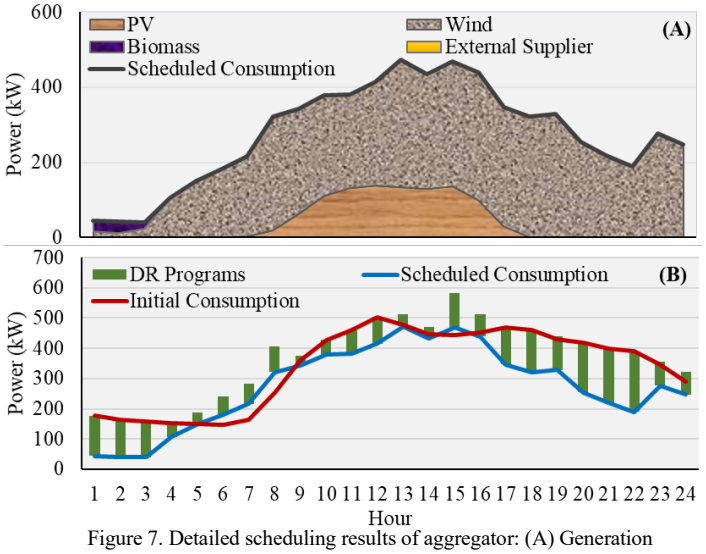


Figure 7. Detailed scheduling results of aggregator: (A) Generation scheduling, (B) DR scheduling.

As one can see in Figure 7 – (A), since the DG suppliers are considered as cheapest resources comparing with external suppliers, the aggregator utilizes the available DG energy, especially PV and wind, to supply the demand, and in the first four periods, it employs biomass generation to supply the loads. In other words, the aggregator reduced the consumption to the available DG energy in order to prevent purchasing energy from market for minimizing the costs. This means, in the periods that the DG generation is not adequate for the demand, aggregator applies DR programs to regulate the difference between the consumption and generation, as illustrated in Figure 7 – (B). The DR programs that aggregator employed to balance the network for each single period, are shown on Figure 8. The utilized DR programs include load reduction, load curtailment, and load shifting.

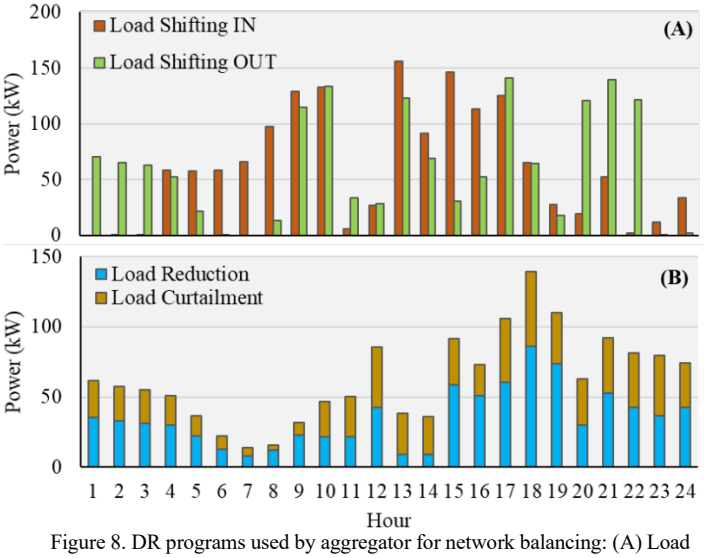


Figure 8. DR programs used by aggregator for network balancing: (A) Load shifting, (B) Load reduction and curtailment.

The incoming and outcoming consumption of each period during load shifting are shown on Figure 8 – (A), which occurred during low generation periods, and shifted to high generation periods. The load shifting enables the aggregator to manage the consumption and shift it to desired periods to prevent purchasing energy from the market, since it is more expensive comparing with DG resources.

Additionally, Table 1 shows the results of aggregation and remuneration processes for period number 12.

Table 1. Remuneration and aggregation results for a single period.

Group		1	2	3	4	5
DG	PV (kW)	0	36.23	19.64	37.54	43.86
	Wind (kW)	250.58	0	0	0	26.88
	Biomass (kW)	0	0	0	0	0
	tariff (m.u./kWh)	0.05	0.02	0.05	0.06	0.06
DR	Residential (kW)	0	0	0	0	0
	Commercial Small (kW)	0	0	0	0	10.09
	Commercial Medium (kW)	0	0	11.07	20.05	7.11
	Commercial Large (kW)	0	0	0	0	0
	Industrial (kW)	26.21	10.94	0	0	0
	tariff (m.u./kWh)	0.06	0.06	0.05	0.05	0.05
Total (kW)		500.18				

In Table 1, the total energy as well as the number of resources in each group have been calculated by aggregation computation, however, the group tariff has been indicated by remuneration calculation. Moreover, in order to calculate the profit of the aggregator after paying all resources, including DG and incentives for DR participation, (12) is proposed. This profit is the monetary benefit that aggregator gained after its operations.

$$Profit = C_{(t)}^{mcp} \cdot \left[\sum_{p \in k}^{P} P_{(p,t)}^{DG} + \sum_{c \in k}^{C} (P_{(c,t)}^{Red} + P_{(c,t)}^{Cut}) \right] - \left[G_{(k,t)}^{DG} \cdot \sum_{p \in k}^{P} P_{(p,t)}^{DG} + G_{(k,t)}^{DR} \cdot \sum_{c \in k}^{C} (P_{(c,t)}^{Red} + P_{(c,t)}^{Cut}) \right] \quad (12)$$

$$\forall k \in \{1,...,K\}, \forall t \in \{1,...,T\}$$

In (12), the $C_{(t)}^{mcp}$ denotes market clearing price, which is considered in this case study is equal to the market prices provided in Figure 5. The classification of the resources in the several groups enables the aggregator to provide lower group tariffs, comparing with the situation that all resources are in the same group. It is true that with classification of resources in several groups, high group tariff will be still remained, however, the chance of aggregator to reach some group tariff with lower rates will be increased. The financial profit gained by aggregator during period number 12, is shown on Table 2. In this single period, the aggregator has total energy of 500.18 kW, which has incoming of 35 monetary unit from the energy that sold to the market. However, it also paid 22.84 monetary unit for all resources, including DG units and DR incentives, and in total, 12.17 monetary unit will be the final profit of aggregator during period number 12.

Table 2. Gained profit by aggregator during market negotiations for one period.

Parameter	Value
Costs paid to all resources (m.u.)	22.84
Market clearing price (m.u./kWh)	0,0700
Income from market sell (m.u.)	35.00
Total aggregator profit (m.u.)	12.17

The profit of aggregator shown on Table 2, is for a single period (considered as one hour in a day in this case study), and even with a few number of consumers and generators, it could gain profit from market negotiations. This means that if the aggregator is responsible for a larger network He will be able to aggregate more energy capacity for clustering, and therefore, with great participation in market, which leads to obtain a satisfying amount of financial benefits. However, this profitability depends on the capabilities and offers of aggregator in market negotiations and existing competitions. Figure 9 demonstrates the financial results concerning the participation of aggregator in the electricity market for all periods of case study. These results are obtained after the scheduling and remuneration processes. It should be noted that only the resources that participated in these processes, are considered. The costs of each period in Figure 9 follows the same process represented in Table 2, which the gained profit is a subtract of costs paid to all resources and the income from market participation.

The last results of this section are related to a comparison that shows the impact of load shifting method for aggregator. For this purpose, it is considered that the aggregator is not capable to employ load shifting during scheduling process. The scheduling results, without load shifting, are illustrated in Figure 10. The results shown in Figure 10 (without load shifting) can be compared with the scheduling results demonstrated in Figure 7 (with load shifting).

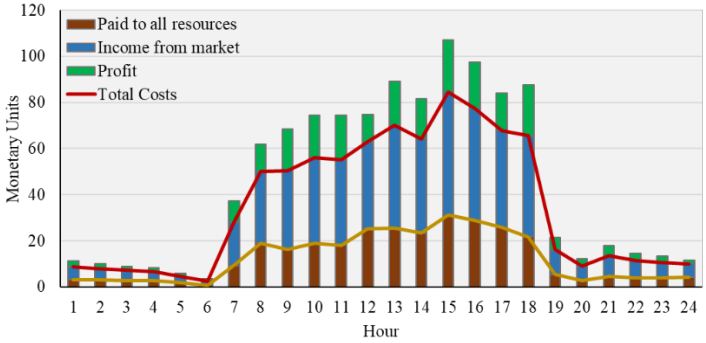


Figure 9. Detailed aggregator costs after scheduling and remuneration process for all periods.

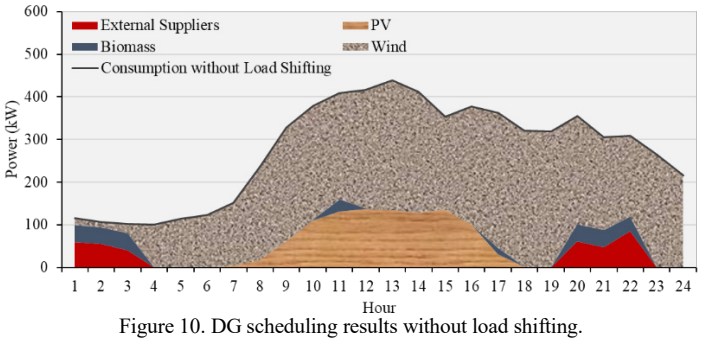


Figure 10. DG scheduling results without load shifting.

As one can see in Figure 10, in some periods the aggregator not only utilizes all available DG resources to supply the demand, but also, it is forced to use energy from external suppliers to feed all demand. By this way, since the electricity price of external suppliers are more expensive than the DG resources, the total costs of aggregator will be increased, and therefore, the gained profit will be decreased. However, as Figure 7 demonstrated, if the aggregator utilized load shifting scenario, and shift the load from the moments that there is no adequate DG energy, to the periods with high DG energy, its operational costs will be reduced, and obtained financial benefits will be increased.

6. Conclusions

This paper presented an aggregator model for distributed energy resource and demand response program management. The presented model considered the resources able to participate in electricity market negotiations through the aggregator. The aggregator has capability of demand-side flexibility by establishing several demand response contracts with consumers.

The main focus of the paper was given to a business model that aggregator utilized it to gather energy of resources and their costs, to define a fair remuneration tariff for all resources, as well as an affordable price for market participation. By this way, the aggregator guarantees that the small-scale resources, including distributed generation and demand response programs, will participate in the electricity market, and therefore, getting profits.

The results of case study demonstrate that the aggregator model is able to perform an optimal scheduling for distributed resources, in order to minimize the operational costs of the aggregator. This is done through implementing several DR programs. The final outcomes of aggregation and remuneration processes validated the proposed method, and proved that the aggregator can gain financial benefits from market negotiations, even by paying a fair tariff to all available resources.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 641794 (project DREAM-GO). This work also received funding from the following projects: NETEFFICITY Project (ANI | P2020 – 18015); and from FEDER Funds through COMPETE program and from National Funds through FCT under the project UID/EEA/00760/2013.

References

[1] J. Spinola, P. Faria, Z. Vale, "Model for the integration of distributed energy resources in energy markets by an aggregator", in 2017 IEEE Manchester PowerTech, Manchester UK, 2017. https://doi.org/10.1109/PTC.2017.7981234

[2] S. M. Nosratabadi, R.-A. Hooshmand, E. Gholipour, "A comprehensive review on microgrid and virtual power plant concepts employed for distributed energy resources scheduling in power systems," Renew. Sustain. Energy Rev., **67**, 341–363, 2017. https://doi.org/10.1016/j.rser.2016.09.025

[3] C. Battistelli, A. J. Conejo, "Optimal management of the automatic generation control service in smart user grids including electric vehicles and distributed resources," Electr. Power Syst. Res., **111**, 22–31, 2014. https://doi.org/10.1016/j.epsr.2014.01.008

[4] A. Roos, S. Ø. Ottesen, T. F. Bolkesjø, "Modeling Consumer Flexibility of an Aggregator Participating in the Wholesale Power Market and the Regulation Capacity Market," Energy Procedia, **58**, 79–86, 2014. https://doi.org/10.1016/j.egypro.2014.10.412

[5] O. Abrishambaf, P. Faria, L. Gomes, J. Spínola, Z. Vale J. Corchado, "Implementation of a Real-Time Microgrid Simulation Platform Based on Centralized and Distributed Management", Energies, **10**(6), 806-820, 2017. http://dx.doi.org/10.3390/en10060806

[6] D. J. Vergados, I. Mamounakis, P. Makris, E. Varvarigos, "Prosumer clustering into virtual microgrids for cost reduction in renewable energy trading markets," Sustain. Energy, Grids Networks, **7**, 90–103, 2016. https://doi.org/10.1016/j.segan.2016.06.002

[7] S. Rahnama, S. E. Shafiei, J. Stoustrup, H. Rasmussen, J. Bendtsen, "Evaluation of Aggregators for Integration of Large-scale Consumers in Smart Grid," IFAC Proc. Vol., **47**(3), 1879–1885, 2014. https://doi.org/10.3182/20140824-6-ZA-1003.00601

[8] EG3 Report - Smart Grid Task Force, "Regulatory Recommendations for the Deployment of Flexibility," 2015.

[9] Smart Energy Demand Coalition, "Mapping Demand Response in Europe Today 2015," 2015.

[10] Smart Energy Demand Coalition, "Enabling independent aggregation in the European electricity markets," 2015.

[11] Federal Energy Regulatory Commission, "Assessment of Demand Response & Advanced Metering," 2011.

[12] Smart Grid Task Force, "Regulatory Recommendations for the Deployment of Flexibility," 2015.

[13] S. Rahmani-Dabbagh, M. K. Sheikh-El-Eslami, "A profit sharing scheme for distributed energy resources integrated into a virtual power plant," Appl. Energy, **184**, 313–328, 2016. https://doi.org/10.1016/j.apenergy.2016.10.022

[14] P. Faria, J. Spínola, Z. Vale, "Aggregation and Remuneration of Electricity Consumers and Producers for the Definition of Demand-Response Programs," IEEE Trans. on Industrial Informatics, **12(**3), 952–961, 2016. https://doi.org/10.1109/TII.2016.2541542

[15] N. Mahmoudi, E. Heydarian-Forushani, M. Shafie-khah, T. K. Saha, M. E. H. Golshan, P. Siano, "A bottom-up approach for demand response aggregators' participation in electricity markets," Electr. Power Syst. Res., **143**, 121–129, 2017. https://doi.org/10.1016/j.epsr.2016.08.038

[16] P. Faria, Z. Vale, J. Baptista, "Constrained consumption shifting management in the distributed energy resources scheduling considering demand response," Energy Convers. Manag., **93**, 309–320, 2015. https://doi.org/10.1016/j.enconman.2015.01.028

[17] M. Silva, F. Fernandes, H. Morais, S. Ramos, Z. Vale, "Hour-ahead energy resource management in university campus microgrid," 2015 IEEE Eindhoven PowerTech, Eindhoven Netherlands, 2015. https://doi.org/10.1109/PTC.2015.7232449.

[18] MIBEL Electricity Market. Available online: http://www.omip.pt (accessed on 30 Nov. 2017)

[19] TOMLAB Optimization. Available online: https://tomopt.com/tomlab/ (accessed on 30 Nov. 2017)

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 234-238 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Medium Voltage Microgrid Test Setup and Procedures Implemented on a Real Pilot Project

Bruno Alberto Pacheco*,[1], Marcos Aurelio Izumida Martins[1], Cesare Quinteiro Pica[1], Nilo Rodrigues[2]

[1]*CERTI Foundation, Sustainable Energy Center, Florianópolis - SC, Brazil*

[2]*ENEL, Research & Development, Fortaleza - CE, Brazil*

ARTICLE INFO

Article history:
Received: 13 November, 2017
Accepted: 15 January, 2018
Online: 30 January, 2018

Keywords:
Microgrid
Energy Storage System
Distributed Generation

ABSTRACT

This paper presents various concepts related to the application of a microgrid pilot project in a residential condominium at Fortaleza / CE - Brazil, such as battery energy system, renewable and distributed generation, islanding recloser and all different units using interface based on power electronics. This papers main objective is to create information about microgrid operation and the interaction between its main equipment, such as power converters, utility energy distribution system and control units responsible for algorithms and changes in microgrids operation mode. This information is important for understanding the need for a test setup construction. To perform the test procedures, a temporary setup in a controlled environment within the microgrid is proposed. During the test periods, intentional power outages are required to evaluate the operating mode switching on each unit. The test setup described in this paper aims to mitigate the tests effects on other residential units inside the condominium.

1. Introduction

Typical electric power systems are known to operate by transmitting energy in bulk from large generating plants over long distances to large load centers. With the introduction of distributed generation and the growing evolution of power electronics for the renewable generation sources of electricity, such as solar and wind, this scenario of operation has been changing gradually, leaving behind the centralized form of generation and becoming now decentralized, producing energy closer to consumers [1, 2].

In this context, microgrids present a new concept of power generation and distribution. Microgrids are utility power grid subsystems composed of a set of distributed energy resources, controllable electric loads, energy storage elements and power control system. Microgrid main characteristic is the ability to operate in island conditions, i.e., operate isolated from the main electricity distribution grid preserving an energy quality for microgrid connected consumers [3-5].

In addition to both connected and islanded modes, this microgrid, under analysis, presents a third operation mode referred to as Maintenance Mode. This operation mode is activated when a fault, such as short circuit, happens inside the microgrid, and consequently making it impossible to operate islanded. Therefore, maintenance mode becomes essential while aiming towards a safe microgrid operation. Recent studies [6, 7] also described the importance of such mode and its benefits for microgrid operations.

While in maintenance mode, the microgrid cannot be powered by any of its distributed energy resources (DER), i.e., the central energy storage (grid-forming power converter) and its internal sources of distributed generation (grid-feeding power converter) are unable to generate energy. Maintenance mode is also triggered with a power distribution company request, followed by command from the operation center, due to maintenance in the local medium voltage distribution grid, as it is procedure to disable the microgrid and ensure a complete absence of power at the condominium internal distribution grid.

To perform real tests of the control system operation mode switching and interaction with the equipment, a test setup for the pilot microgrid is proposed, aiming to mitigate the impact on consumers loads. In this paper, the microgrid project in analysis will be presented first and then the proposed test setup and procedures.

2. Microgrid Architecture

This microgrid under study is located in a residential condominium at Fortaleza / CE – Brazil. The condominium is connected to the utility through medium voltage grid.

*Corresponding Author: Bruno Alberto Pacheco, CERTI Foundation, Email: bap@certi.org.br

The Microgrid is divided into several units according to their characteristics and devices. This unit delimitation allows a better understanding of microgrid operation. Those units are shown in Figure 1.

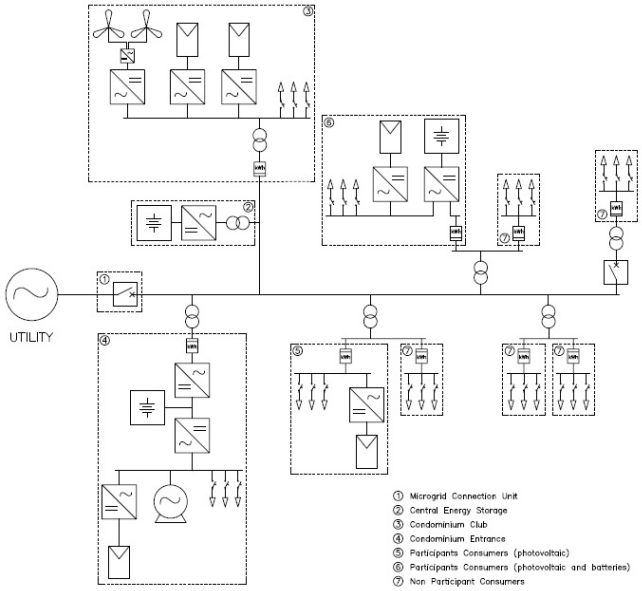


Fig. 1. Architecture of the real microgrid pilot project.

The main microgrid units are presented as it follows:

2.1. Central Energy Storage

This unit is composed by an energy storage system along with its power conversion and control devices. With a storage capacity of 105kWh and 250kW of power, this system is able to supply power for the microgrid while operating in islanded mode [8]. Figure 2 displays a simple electric diagram for this unit.

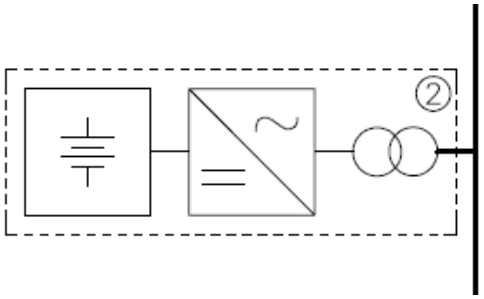


Fig. 2. Central Energy Storage.

The central storage system features a lithium-ion battery bank, grid-forming power converter, 380/13800 V power transformer, switchgear and their respective control and protection equipment.

During islanded mode, this unit function is to provide power for the loads and a voltage reference for all renewable distributed generation. While operation at connected mode, this unit is recharged and can also be discharged providing grid support services.

2.2. Consumers Units

Microgrids consumers are the systems main loads. They are strategically chosen residential consumers with distributed generation resources and some with local storage systems.

They are subdivided into Participant Consumers and Non Participant Consumers. Whilst operating in connected mode, both participant and non-participant act as loads for the system. During islanded operation, non-participant consumers are disconnected through smart meters. Therefore, only participant consumers remain connected at microgrid [8].

Figures 3 and 4 displays a simplified electrical diagram of those participant consumers. They are equipped with demand-response devices, solar generation, power inverters, smart meters, and some units are also equipped with a local storage system.

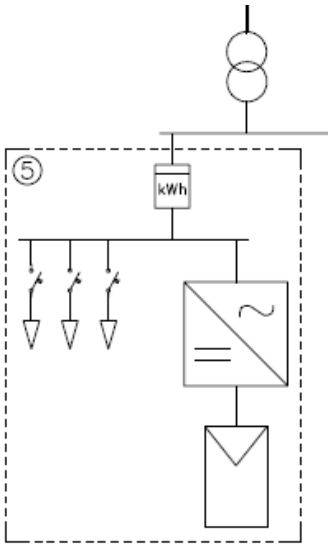


Fig. 3. Participant Consumers Units.

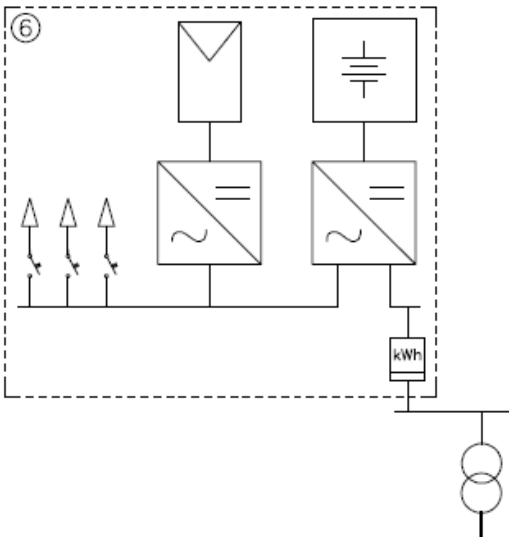


Fig. 4. Participant Consumers Units with a local storage system.

2.3. Club Unit

This residential condominium contains a club with common areas for its residents. The Club Unit (Figure 5) is also a microgrid consumer, however it has a larger distributed generations, composed with 24kWp of photovoltaic panels, as well as 7kW of wind generation.

Like the participating consumers, the Club is endowed with controllable loads through the demand-response devices. Another

important detail is that this unit is connected at microgrid in medium voltage and has its own substation.

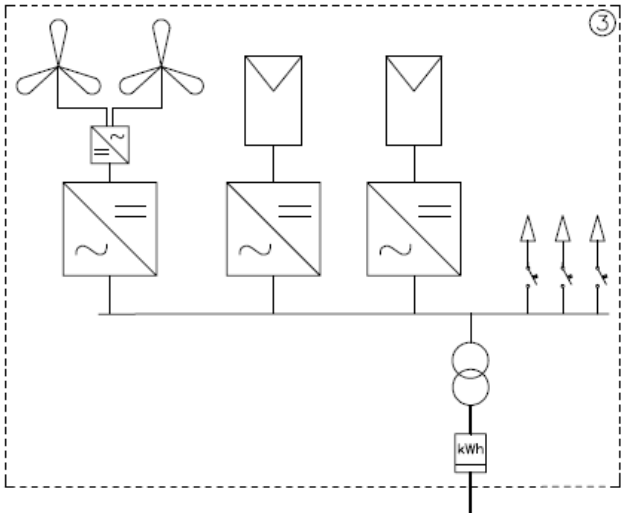


Fig. 5. Club Unit.

2.4. Condominium Entrance

This unit contains critical loads for condominium security and safety functions. Due to this fact, it is equipped with a diesel generator as a back-up unit.

A back-to-back converter is employed as a connection interface with the utility grid, allowing for smooth operation during power outage scenarios. Furthermore, this diesel generator can be used in a microgrid support scheme, relieving the system during a utility outage [8].

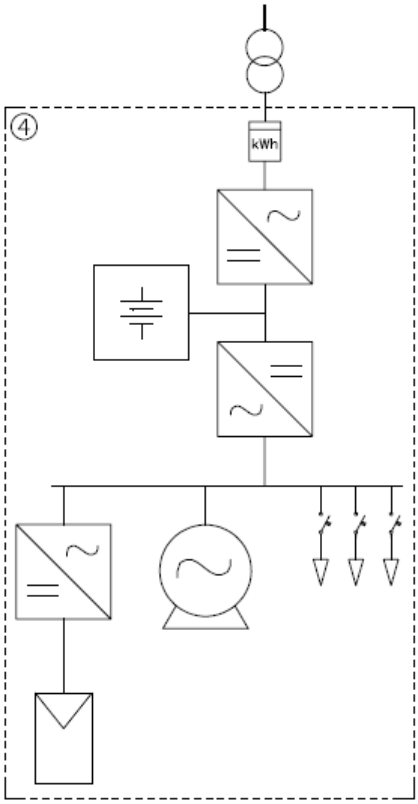


Fig. 6. Condominium Entrance.

This unit also has a solar generation system and controllable loads with demand-response devices as shown in Figure 6.

2.5. Microgrid Connection Unit

The Microgrid Connection Unit (Figure 7) consists of a medium voltage recloser and a set of control equipment for an energy outage detection in the distribution system, coordination of islanding and microgrid reconnection to the distribution system.

This unit is responsible for controlling the microgrid operation mode changes by informing other control units. It is also the communication interface with the utility grid SCADA (Supervisory Control and Data Acquisition) system.

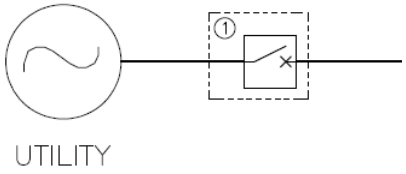


Fig. 7. Microgrid Connection Unit.

3. Microgrid Test Setup

In order to validate microgrid safe operation during its transitions through connected, islanded and maintenance modes, a power setup is proposed for conducting systemic tests (Figure 8).

The premise of this setup is to use the Club Unit area for real islanding, synchronization and reconnection microgrid tests. Since the Club Unit has a medium voltage connection, which is independent from the other condominium derivation line, any power supply interruption inside this area will not affect other microgrid units, reducing the inconvenience for all condominium residents.

As mentioned before, the Club unit has a larger amount of distributed generation and controllable loads. That can be used as a controlled test environment by simulating the residents homes energy consumption and generation. Thus, real tests of operation mode switching can be conducted in a safe way using the Club unit infrastructure.

In connected mode, the utility grid supplies the power for the loads and sets a voltage reference for the power electronics equipment of the generating units. The Central Energy Storage behaves like a grid-feeding power converter in this mode, only providing utility grid support when requested.

3.1. Islanding Tests

For islanding tests, the Microgrid Connection Unit (① in Figure 8) detects an outage in the distribution system due to under voltage protections. It then proceeds to disconnect the microgrid from the utility and sends a command to the Central Energy Storage, requesting a change in its mode of operation. The use of the temporary recloser (② in Figure 8) installed in the test setup input derivation is essential, in order to simulate an outage occurring when an intentional opening of this equipment is performed.

After disconnecting and receiving the command from the Microgrid Connection Unit, the Central Energy Storage operates as a grid-forming power converter, assuming the role of the main power supply of the microgrid and providing voltage reference for the generation interfaces.

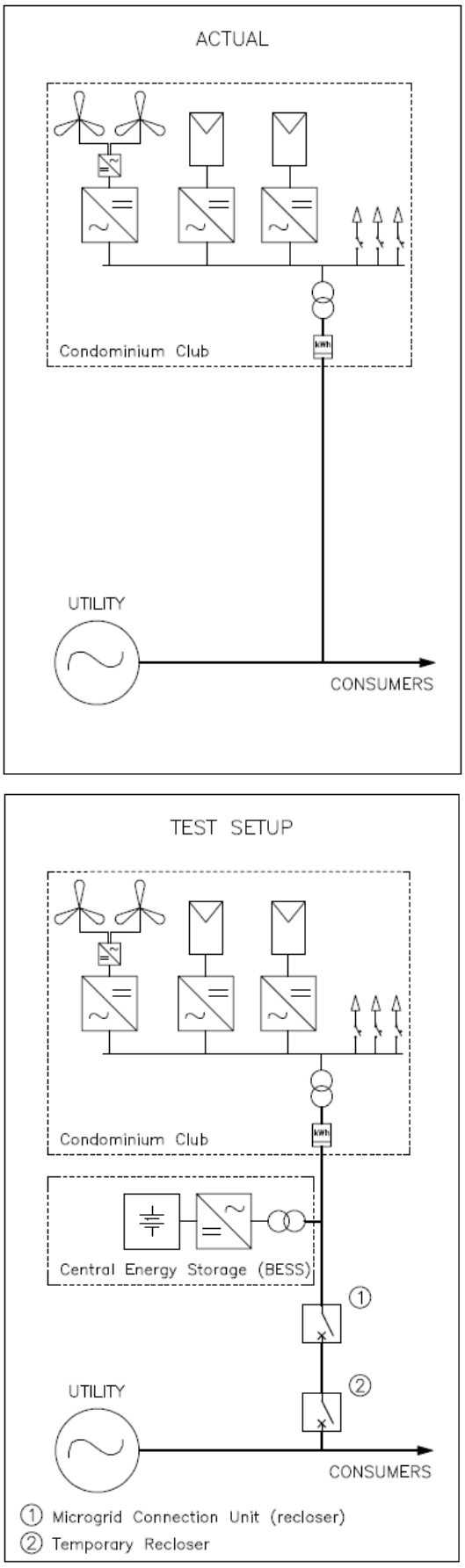


Fig. 8. Actual setup and Microgrid test setup.

3.2. Reconnecting Tests

When reconnecting the microgrid to the utility grid, the temporary recloser (② in Figure 8) is closed simulating the utility grid stable return after an outage.

The Microgrid Connection Unit detects the energy utility supply return while the Central Energy Storage synchronizes with the utility voltage reference. Microgrid is only reconnected to the main electricity distribution grid after synchronization is achieved.

Before the microgrid is connected to the utility grid, some of the indicators and parameters need to be checked, and the best time to close static switch should also be judged. This process is very important that can prevent the dangerous of transient phenomena during connecting the microgrid to utility grid. The good time to close the switch is when the voltage across the switch has to be very small (ideally zero), and the grid's frequency is higher than the microgrid's frequency. Because the current flowing through the switch is minimum, and the power flow direction is from utility grid to the microgrid in such a situation.

A safe and reliable grid-connection process will require to meet the following three conditions: a) the voltage through the static switch must be very small. b) microgrid operating frequency must be slightly smaller than the grid frequency. c) utility grid voltage must be ahead of the microgrid voltage [9].

3.3. Maintenance Tests

Maintenance mode is also tested in this setup with a request through utility operation center command. In such case, it must be verified that the Central Energy Storage is disconnected from the grid, not powering the microgrid when necessary.

Another advantage for this test setup is that all equipment employed will be used in the final installation. In this procedure it is possible to test the interaction between connected equipment in medium voltage, control system and power electronics interfaces. In addition, the Central Energy Storage is tested at its final installation place allows for other control and protection details to be verified.

4. Conclusions

In this work, the microgrid architecture on a pilot project at a residential condominium is presented, along with its main units, equipment and components according to its functionalities and technical characteristics.

Through the proposed test setup it is possible to obtain performance indicators on microgrid operating mode transitions and grid-forming power converter, as well as the integration between the control and protection equipment.

With the test setup execution, all operation modes shall be validated, along with transitions among them. The test results will improve the teams acknowledge around a correct microgrid functioning and a validation of its infrastructure and equipment.

Acknowledgment

The authors thank the Research Program and Technological Development of the electricity sector regulated by ANEEL and Enel Distribuição Ceará for the financial support to the project. This work is related to the project "Development Application of Pilot Microgrid Power Distribution with Distributed Generation

and Commercial Operations Model" under PD- 0039-0073 / 2014 ANEEL.

References

[1] H.M.A. Antunes, S.M. Silva and B.J.C. Filho, “Análise e Operação de uma Microrrede de Energia Elétrica,” VI Simpósio Brasileiro de Sistemas Elétricos, 2016.

[2] Gang HU et al. “Study on Modeling and Simulation of Photovoltaic Energy Storage Microgrid” 4th International Conference on Information Science and Control Engineering, p.692-695, 2017.

[3] R. H. Lasseter, “MicroGrids,” IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No.02CH37309), vol. 1, no. C, pp. 305–308, 2002.

[4] Daniel E. Olivares “A Centralized Energy Management System for Isolated Microgrids.” IEEE Transactions os Smart Grid, VOL. 5, July 2014.

[5] Benedetto Aluisio et al. “Embedding Energy Storage for Multi-Energy Microgrid Optimal Operation”, 2017.

[6] Sundari Ramabhotla et al. “Operation and Maintenance Cost Optimization in the Grid Connected Mode of Microgrid.” 2016 Ieee Green Technologies Conference, Kansas City, p.56-61, 2016.

[7] Richard B. Jones “How Reliable is your Microgrid: An insurer’s perspective on risk drivers for ditributed resources.”, Reston, July 2015.

[8] M.A.I. Martins et al., “Design and Implementation of a Microgrid Power Management Unit Using a Back-To-Back Converter in a Residential Condominium Connected at Medium Voltage,” COBEP SPEC IEEE Power Electronics Society, 2015.

[9] Meiqin Mao, Yinzheng Tao, Liuchen Chang, Yongchao Zhao and Peng Jin, "An intelligent static switch based on embedded system and its control method for a microgrid," IEEE PES Innovative Smart Grid Technologies, Tianjin, 2012, pp. 1-6.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 6, 236-241 (2017)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

A Study on Isomorphic Properties of Circulant Graphs

V. Vilfred Kamalappan *

Department of Mathematics, Central University of Kerala Periye, Kasaragod, Kerala, 671316 India

ARTICLEINFO

Article history:
Received: 13 October, 2017
Accepted: 06 November, 2017
Online: 20 December, 2017

Keywords:
Self-Complementary circulant graphs
Type-1 and Type-2 isomorphic circulant graphs
Cartesian product and factorization
prime and composite circulant graphs
Factorization theorem, Fundamental theorem

ABSTRACT

$C_n(R)$ denotes circulant graph $C_n(r_1, r_2, \ldots, r_k)$ of order n for a set $R = \{r_1, r_2, \ldots, r_k\}$ where $1 \leq r_1 < r_2 < \ldots < r_k \leq \left\lfloor \frac{n}{2} \right\rfloor$. Circulant graph $C_n(R)$ is said to have the Cayley Isomorphism (CI) property if whenever $C_n(S)$ is isomorphic to $C_n(R)$, there is some $a \in \mathbb{Z}_n^*$ for which $S = aR$. In this paper, isomorphic properties of circulant graphs that includes (i) Self-complementary circulant graphs; (ii) Type-2 isomorphism, a new type of isomorphism other than already known Adam's isomorphism of circulant graphs and (iii) Cartesian product and factorization of circulant graphs similar to the theory of product and factorization of natural numbers are studied. New abelian groups are obtained from these isomorphic circulant graphs. Type-2 isomorphic circulant graphs have the property that they are isomorphic graphs without Cayley Isomorphism (CI) property and thereby new families.

1 Introduction

This paper is an extension of work originally presented in ICMSAO2017 [1] and covers the author's study on a few isomorphic properties of circulant graphs that includes (i) Existence of self-complementary circulant graphs; (ii) Type-2 isomorphism, a new type of isomorphism other than already known Adam's isomorphism of circulant graphs that helps to obtain graphs without CI-property and abelian groups and (iii) Cartesian product and factorization of circulant graphs similar to the theory of product and factorization of natural numbers.

Beauty comes out of symmetry as well as asymmetry. Investigation of symmetries/asymmetries of structures yield powerful results in Mathematics. Circulant graphs form a class of highly symmetric mathematical (graphical) structures. In 1846 Catalan (cf. [2]) introduced circulant matrices and properties of circulant graphs have been investigated by many authors [1-20]. An excellent account of circulant matrices can be found in the book by Davis [2] and circulant graphs in the article [11].

If a graph G is circulant, then its adjacency matrix $A(G)$ is circulant. It follows that if the first row of the adjacency matrix of a circulant graph is $[a_1, a_2, \ldots, a_n]$, then $a_1 = 0$ and $a_i = a_{n-i+2}$, $2 \leq i \leq n$ [15, 17].

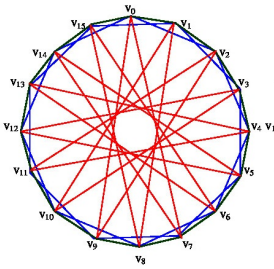
Through-out this paper, for a set $R = \{r_1, r_2, \ldots, r_k\}$, $C_n(R)$ denotes circulant graph $C_n(r_1, r_2, \ldots, r_k)$ where $1 \leq r_1 < r_2 < \ldots < r_k \leq \left\lfloor \frac{n}{2} \right\rfloor$. Only connected circulant graphs of finite order are considered, $V(C_n(R)) = \{v_0, v_1, v_2, \ldots, v_{n-1}\}$ with v_i adjacent to v_{i+r} for each $r \in R$, subscript addition taken modulo n and all cycles have length at least 3, unless otherwise specified, $0 \leq i \leq n-1$. However when $\frac{n}{2} \in R$, edge $v_i v_{i+\frac{n}{2}}$ is taken as a single edge for considering the degree of the vertex v_i or $v_{i+\frac{n}{2}}$ and as a double edge while counting the number of edges or cycles in $C_n(R)$, $0 \leq i \leq n-1$. Generally, write C_n for $C_n(1)$ and $C_n(1, 2, \ldots, \left\lfloor \frac{n}{2} \right\rfloor)$ for K_n. We will often assume, with-out further comment, that the vertices are the corners of a regular $n-gon$, labeled clockwise. Circulant graphs $C_{16}(1,2,7)$, $C_{16}(2,3,5)$, $C_{27}(1,3,8,10)$, $C_{27}(3,4,5,13)$ and $C_{27}(2,3,7,11)$ are shown in Figures 1 - 5. Now, let us consider the following definitions and results that are useful in the subsequent sections.

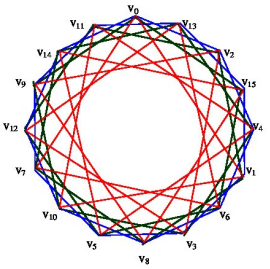
Definition 1.1. [17] *Let n and r be positive integers with $n \geq 4$ and $r < \frac{n}{2}$. Then, clearly, $C_n(r)$ consists of a collection of cycles $(v_0 v_r v_{2r} \ldots v_0)$, $(v_1 v_{1+r} v_{1+2r} \ldots v_1)$, $\ldots$, $(v_{r-1} v_{2r-1} v_{3r-1} \ldots v_{r-1})$. If $d = gcd(n, r)$, then there are d such disjoint cycles, each of length $\frac{n}{d}$. We say that each of these cycles is of period r, length $\frac{n}{d}$ and rotation $\frac{r}{d}$.*

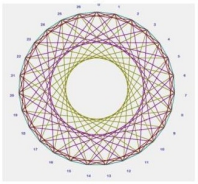
If $r = \frac{n}{2}$, then obviously $C_n(r)$ is just a $1-factor$. Since $C_n(R)$ is just the union of the cycles $C_n(r)$ for $r \in R$, we have a decomposition of $C_n(R)$.

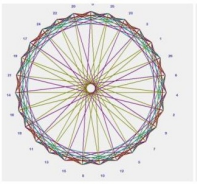
*Corresponding Author: V. Vilfred Kamalappan Email: vilfredkamalv@cukerala.ac.in, Contact No. 0091 9443577502

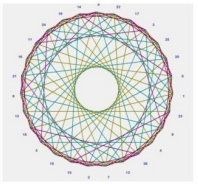
https://dx.doi.org/10.25046/aj020628


Fig. 1. $C_{16}(1,2,7)$


Fig. 2. $C_{16}(2,3,5)$


Fig. 3. $C_{27}(1,3,8,10)$


Fig. 4. $C_{27}(3,4,5,13)$


Fig. 5. $C_{27}(2,3,7,11)$

Theorem 1.2. [17] *Let $r \in R$. Then, in $C_n(R)$, the length of a cycle of period r is $\frac{n}{gcd(n,r)}$ and the number of disjoint periodic cycles of period r is $gcd(n,r)$.* □

Corollary 1.3. [17] *In $C_n(R)$, the length of a cycle of period r is n if and only if $gcd(n,r) = 1$, $r \in R$.* □

Remark 1.4. [17] *Let $|R| = k$. Then, the circulant graph $C_n(R)$ for a set $R = \{r_1, r_2, \ldots, r_k\}$ is $(2k-1)-regular$ if $\frac{n}{2} \in R$ and $2k-regular$ otherwise.* □

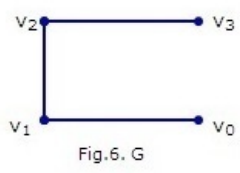
The following Lemmas are useful to obtain one-to-one mappings.

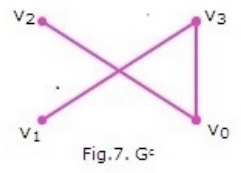
Lemma 1.5. [15] *Let A and B be two non-empty sets. Let $f : A \to B$ be a mapping. Then, f is one-to-one if and only if $f/_{A'}$ is one-to-one for every non-empty subset A' of A.* □

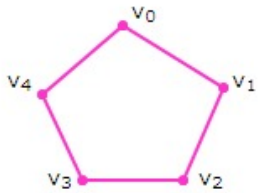
Lemma 1.6. [15] *Let A and B be non-empty sets and $A_1, A_2, \ldots, A_k$ be a partition of A (each A_i being non-empty, $i = 1, 2, \ldots, k$). Let $f : A \to B$ be a mapping. Then f is one-to-one if and only if $f/_{A_i}$ is one-to-one for every i, $i = 1, 2, \ldots, k$.* □

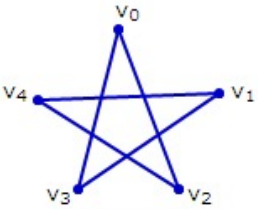
2 On self-complementary circulant graphs

In 1962, Horst Sachs [14] proved that the sufficient condition for the existence of a self-complementary circulant graph of order n is that every prime factor p of n should satisfy $p \equiv 1 \pmod 4$. He also conjectured that the sufficient condition is a necessary one. We have proved that the self-complementary circulant graph on n vertices does not exist if n has any prime factor which is not of the form $4m+1$, $m \in \mathbb{N}$. Thereby, we establish that the sufficient condition is also a necessary one. The proof is based on counting the number of disjoint cycles of a particular length in K_n and is given in this section [4, 17]. Graphs given in Figures 6 and 7 are self-complementary but not of circulant whereas graphs given in Figures 8 and 9 are self-complementary circulant graphs.


Fig.6. G


Fig.7. G


Fig.8. $C_5(1)$


Fig.9. $C_5(2)$

Theorem 2.1. [17] *If $C_n(R) \cong C_n(S)$, then there is a bijection f from R to S so that for all $r \in R$, $gcd(n,r) = gcd(n, f(r))$.*

Proof. The proof is by induction on the order of R. □

Theorem 2.2. [9] *If graph G of order n is self-complementary, then $n \equiv 0, 1 \pmod 4$.* □

Theorem 2.3. [17] *If $C_n(R)$ for a set $R = \{r_1, r_2, \ldots, r_k\}$ is self-complementary, then $n \equiv 1 \pmod 4$.*

Proof. Using Theorem 2.2, we get $n \equiv 0, 1 \pmod 4$. When n is even, a circulant graph is of even degree if and only if its complement is of odd degree since any circulant graph is a regular graph. Thus, self-complementary circulant graph of even order doesn't exist. Hence, we get the result. □

Theorem 2.4. [17] *If $C_n(R)$ for a set $R = \{r_1, r_2, \ldots, r_k\}$ is self-complementary, then $n = 4m+1$, $k = m$ and $|S_i|$ is even where $S_i = \{j : gcd(n,j) = gcd(n,i),\ 1 \le j \le \left\lfloor \frac{n}{2} \right\rfloor\}$, for all i, $i = 1, 2, \ldots, \left\lfloor \frac{n}{2} \right\rfloor$.*

Proof. Using Theorem 2.3, we get, $n = 4m+1$. This implies that the degree (of each vertex) of a self-complementary circulant graph of order $4m+1$ is $2m$ which implies $k = m$.

Let $C_{4m+1}(R)$ be a self-complementary circulant graph for $R = \{r_1, r_2, \ldots, r_m\}$. If it contains a cycle of period r (and of length $\frac{n}{gcd(n,r)}$, using Theorem 1.2), then its complement also contains a periodic cycle of period, say, s such that $\frac{n}{gcd(n,s)} = \frac{n}{gcd(n,r)}$, $1 \le r, s \le \left\lfloor \frac{n}{2} \right\rfloor$, using Theorem 1.2. This implies that $gcd(n,s) = gcd(n,r)$. Here we consider that the cycles of periods r_i and $n - r_i$ are the same in $C_n(r_1, r_2, \ldots, r_k)$, $1 \le i \le k$. Combining the above arguments, we get the result. □

Remark 2.5. *The above theorem states that if a self-complementary circulant graph of order n exists, then* $n = 4m+1$ *and so it is* $2m$*-regular and the number of periodic cycles of length* i *in* K_n *is always even for each* i, $1 \leq i \leq \left\lfloor \frac{n}{2} \right\rfloor$.

Theorem 2.6. [17] *Self-complementary circulant graph of order n doesn't exist when n has any prime factor of the form* $2(2m-1)+1$, $m \in \mathbb{N}$.

Proof. Using Theorem 2.3, $n = 4a+1$, $a+1 \in \mathbb{N}$.

Let $n = p_1^{n_1} p_2^{n_2} \dots p_j^{n_j}$ where $p_1, p_2, \dots, p_j$ are the (odd) prime factors of n. Let $p_i \neq 4m+1$ for at least one i, $1 \leq i \leq j$ and for any $m \in \mathbb{N}$. This implies, $p_i = 2(2q-1)+1$ for some $q \in \mathbb{N}$. Consider the circulant graph $C_{4a+1}(1, 2, \dots, 2a) \cong K_{4a+1} = K_n$. Let r be the natural number such that $r = \frac{n}{p_i}$.

Aim To find out all natural numbers lying between 1 and n such that g.c.d. of each one of them with n is exactly r.

Let $gcd(n, pr+s) = r$ where $p, s+1 \in \mathbb{N}$ such that $0 \leq s < r$ and $1 \leq pr+s \leq n = rp_i$. This implies that $s = 0$ and so $gcd(n, rp+s) = gcd(n, rp) = gcd(rp_i, rp) = r$ where $1 \leq pr \leq p_i r = n$. Thus we get $gcd(p, p_i) = 1$ where $p \leq p_i$ which is greater than 1. Therefore the possible values of p are $1, 2, \dots, p_i - 1$.

Thus $r, 2r, 3r, \dots, (p_i-1)r$ are the only numbers lying between 1 and n such that g.c.d. of n with each one of them is exactly r. This implies $r, 2r, 3r, \dots, (p_i-1)r$ are the possible periods of cycles of length p_i each, in $C_n(1, 2, \dots, \left\lfloor \frac{n}{2} \right\rfloor)$ since the length of a cycle of period rp in $C_n(1, 2, \dots, \left\lfloor \frac{n}{2} \right\rfloor)$ is $\frac{n}{gcd(n,rp)} = \frac{rp_i}{gcd(rp_i,rp)} = \frac{p_i}{gcd(p_i,p)} = p_i$ for $p = 1, 2, \dots, p_i - 1$, using Theorem 1.2.

In $C_n(1, 2, \dots, \left\lfloor \frac{n}{2} \right\rfloor)$,

the cycles of period r and $n - r(= (p_i-1)r)$ are the same,

the cycles of period $2r$ and $(p_i-2)r$ are the same,

...

the cycles of period $\frac{(p_i-1)r}{2}$ and $\frac{(p_i+1)r}{2}$ are the same.

Thus, there are $\frac{p_i-1}{2}$ number of possible distinct periodic cycles of (periods $r, 2r, \dots, \frac{p_i-1}{2}$ and) length p_i, each in $C_n(1, 2, \dots, \left\lfloor \frac{n}{2} \right\rfloor)$.

Now, $\frac{p_i-1}{2} = 2q-1$ is an odd number. This implies, any circulant graph of order n and its complementary circulant graph contain unequal number of periodic cycles of length p_i, each. This implies that self-complementary circulant graph of order n does n't exist when n contains any prime factor of the form $2(2q-1)+1$, $q \in \mathbb{N}$, by Remark 2.5. □

Thus, when $n = 9, 21, 33, 49, 57, 69, 77, 81, 93$, etc., self-complementary circulant graph doesn't exist on n vertices, by Theorem 2.6, even though in each case, $n \equiv 1 \pmod 4$.

Now, by combining Theorem 2.6 and the sufficient condition for the existence of a self-complementary circulant graph of order n, we get the following result.

Theorem 2.7. [17] *The necessary and sufficient condition for the existence of a self-complementary circulant graph of order n is that each prime factor p of n should satisfy* $p \equiv 1 (mod\ 4)$. □

3 On Isomorphism of Circulant Graphs

In this section, Type-2 isomorphism, a new type of isomorphism different from already known Adam's isomorphism of circulant graphs, main results related to it and families of new abelian groups obtained from isomorphic circulant graphs are presented. Type-2 isomorphic circulant graphs have the property that they are isomorphic graphs without Cayley Isomorphism (CI) property.

Definition 3.1. [12] *A circulant graph* $C_n(R)$ *is said to have the CI-property if whenever* $C_n(S)$ *is isomorphic to* $C_n(R)$, *there is some* $a \in \mathbb{Z}_n^*$ *for which* $S = aR$.

Lemma 3.2. [16] *Let S be a non-empty subset of* $\mathbb{Z}_n$ *and* $x \in \mathbb{Z}_n$. *Define a mapping* $\Phi_{n,x} : S \to \mathbb{Z}_n$ *such that* $\Phi_{n,x}(s) = xs$ *for every* $s \in S$ *under multiplication modulo n. Then,* $\Phi_{n,x}$ *is bijective if and only if* $S = \mathbb{Z}_n$ *and* $gcd(n, x) = 1$. □

Definition 3.3. [3] *Circulant graphs* $C_n(R)$ *and* $C_n(S)$ *for* $R = \{r_1, r_2, \dots, r_k\}$ *and* $S = \{s_1, s_2, \dots, s_k\}$ *are Adam's isomorphic if there exists a positive integer x relatively prime to n with* $S = \{xr_1, xr_2, \dots, xr_k\}_n^*$ *where* $< r_i >_n^*$, *the reflexive modular reduction of a sequence* $< r_i >$ *is the sequence obtained by reducing each* r_i *modulo n to yield* r_i' *and then replacing all resulting terms* r_i' *which are larger than* $\frac{n}{2}$ *by* $n - r_i'$.

Lemma 3.4. [16] *Let* $m, r, t \in \mathbb{Z}_n \in gcd(n, r) = m > 1$ *and* $0 \leq t \leq \frac{n}{m} - 1$. *Then the mapping* $\Theta_{n,r,t} : \mathbb{Z}_n \to \mathbb{Z}_n$ *defined by* $\Theta_{n,r,t}(x) = x + jtm$ *for every* $x \in \mathbb{Z}_n$ *under arithmetic modulo n is bijective where* $x = j + qm$, $0 \leq j \leq m-1$, $0 \leq q \leq \frac{n}{m} - 1$ *and* $j, q \in \mathbb{Z}_n$. □

Theorem 3.5. [16] *Let* $V(C_n(R)) = \{v_0, v_1, \dots, v_{n-1}\}$, $V(K_n) = \{u_0, u_1, \dots, u_{n-1}\}$, $r \in R$ *and* $gcd(n, r) = m > 1$. *Then the mapping* $\Theta_{n,r,t} : V(C_n(R)) \to V(K_n)$ *defined by* $\Theta_{n,r,t}(v_x) = u_{x+jtm}$ *and* $\Theta_{n,r,t}((v_x, v_{x+s})) = (\Theta_{n,r,t}(v_x), \Theta_{n,r,t}(v_{x+s}))$ *for every* $x \in \mathbb{Z}_n$, $x = j + qm$, $0 \leq j \leq m-1$, $0 \leq q, t \leq \frac{n}{m} - 1$ *and* $s \in R$, *under subscript arithmetic modulo n, for a set* $R = \{r_1, r_2, \dots, r_k, n - r_k, n - r_{k-1}, \dots, n - r_1\}$ *is one-to-one, preserves adjacency and* $\Theta_{n,r,t}(C_n(R)) \cong C_n(R)$ *for* $t = 0, 1, 2, \dots, \frac{n}{m} - 1$. □

Definition 3.6. [16] *For a given* $C_n(R)$ *and for a particular value of* t, $0 \leq t \leq \frac{n}{m} - 1$, *if* $\Theta_{n,r,t}(C_n(R)) = C_n(S)$ *for some* $S \subseteq [1, \frac{n}{2}]$ *and* $S \neq xR$ *for all* $x \in \Phi_n$ *under reflexive modulo n, then* $C_n(R)$ *and* $C_n(S)$ *are called Type-2 isomorphic circulant graphs w.r.t.* r, $r \in R$. *In this case, subsets R and S of* $\mathbb{Z}_n$ *are called Type-2 isomorphic subsets of* $\mathbb{Z}_n$ *w.r.t.* r.

Clearly, Type-2 isomorphic circulant graphs are circulant graphs without CI-property. We obtained the following results on Type-2 isomorphism.

Theorem 3.7. [16] *For $n \geq 2, k \geq 3, 1 \leq 2s-1 \leq 2n-1$, $n \neq 2s-1$, $R = \{2s-1, 4n-2s+1, 2p_1, 2p_2, \ldots, 2p_{k-2}\}$ and $S = \{2n-(2s-1),\ 2n+2s-1, 2p_1, 2p_2, \ldots, 2p_{k-2}\}$, circulant graphs $C_{8n}(R)$ and $C_{8n}(S)$ are Type-2 isomorphic (and without CI − property) where $gcd(p_1, p_2, \ldots, p_{k-2}) = 1$ and $n, s, p_1, p_2, \ldots, p_{k-2} \in \mathbb{N}$.* □

Theorem 3.8. [16] *For $R = \{2r-1, 2s-1, 2p_1, 2p_2, \ldots, 2p_{k-2}\}$, $n \geq 2, k \geq 3, 1 \leq t \leq [\frac{n}{2}], 1 \leq 2r-1 < 2s-1 \leq [\frac{n}{2}]$, $gcd(p_1, p_2, \ldots, p_{k-2}) = 1$ and $n, r, s, t, p_1, p_2, \ldots, p_{k-2} \in \mathbb{N}$, if $\Theta_{n,2,t}(C_n(R))$ and $C_n(R)$ are Type-2 isomorphic circulant graphs for some t, then $n \equiv 0\ (mod\ 8)$, $2r-1+2s-1 = \frac{n}{2}$, $t = \frac{n}{8}$ or $\frac{3n}{8}$, $2r-1 \neq \frac{n}{8}, 1 \leq 2r-1 \leq \frac{n}{4}$ and $n \geq 16$.* □

Definition 3.9. [1] *Let $Ad_n = \{\Phi_{n,x} : x \in \Phi_n\}$, $Ad_n(R) = \{\Phi_{n,x}(R) : x \in \Phi_n\} = \{xR : x \in \Phi_n\}$ and $Ad_n(C_n(R)) = T1_n(C_n(R)) = \{\Phi_{n,x}(C_n(R)) : x \in \Phi_n\} = \{C_n(xR) :\ x \in \Phi_n\}$ for a set $R = \{r_1, r_2, \ldots, r_k,\ n-r_k,\ n-r_{k-1}, \ldots, n-r_1\} \subseteq \mathbb{Z}_n$. Define '$\circ$' in $Ad_n(C_n(R))$ such that $\Phi_{n,x}(C_n(R)) \circ \Phi_{n,y}(C_n(R)) = \Phi_{n,xy}(C_n(R))$ and $C_n(xR) \circ C_n(yR) = C_n((xy)R)$ for every $x, y \in \Phi_n$, under arithmetic modulo n. Clearly, $Ad_n(C_n(R))$ is the set of all circulant graphs that are Adam's isomorphic to $C_n(R)$ and $(Ad_n(C_n(R)), \circ) = (T1_n(C_n(R)),\ \circ)$ is an abelian group and we call it as the Adam's group or Type-1 group on $C_n(R)$ under '$\circ$'.*

Definition 3.10. [1] *Let $V(C_n(R)) = \{v_0, v_1, \ldots, v_{n-1}\}$, $V(K_n) = \{u_0,\ u_1,\ \ldots,\ u_{n-1}\}$, $r \in R$, $m, q, t, t', x \in \mathbb{Z}_n$ such that $gcd(n, r) = m > 1$, $x = j + qm$, $0 \leq j \leq m-1$ and $0 \leq q, t, t' \leq \frac{n}{m} - 1$. Define $\Theta_{n,r,t} : \mathbb{Z}_n \to \mathbb{Z}_n$ and $\Theta_{n,r,t} : V(C_n(R)) \to V(K_n)$ such that $\Theta_{n,r,t}(x) = x + jtm$, $\Theta_{n,r,t}(v_x) = u_{x+jtm}$ and $\Theta_{n,r,t}((v_x, v_{x+s})) = (\Theta_{n,r,t}(v_x), \Theta_{n,r,t}(v_{x+s}))$ for every $x \in \mathbb{Z}_n$ and $s \in R$, under subscript arithmetic modulo n. Let $s \in \mathbb{Z}_n$, $V_{n,r} = \{\Theta_{n,r,t} : t = 0, 1, \ldots, \frac{n}{m} - 1\}$, $V_{n,r}(s) = \{\Theta_{n,r,t}(s) : t = 0, 1, \ldots, \frac{n}{m} - 1\}$ and $V_{n,r}(C_n(R)) = \{\Theta_{n,r,t}(C_n(R)) : t = 0, 1, \ldots, \frac{n}{m} - 1\}$. Define '$\circ$' in $V_{n,r}$ such that $\Theta_{n,r,t} \circ \Theta_{n,r,t'} = \Theta_{n,r,t+t'}$, $(\Theta_{n,r,t} \circ \Theta_{n,r,t'})(x)\ (= \Theta_{n,r,t}(\Theta_{n,r,t'}(x)) = \Theta_{n,r,t}(x + jt'm) = (x + jt'm) + jtm = x + j(t + t')m) = \Theta_{n,r,t+t'}(x)$ and $\Theta_{n,r,t}(C_n(R)) \circ \Theta_{n,r,t'}\ (C_n(R)) = \Theta_{n,r,t+t'}(C_n(R))$ for every $\Theta_{n,r,t}, \Theta_{n,r,t'} \in V_{n,r}$ where $t + t'$ is calculated under addition modulo $\frac{n}{m}$. Clearly, $(V_{n,r}(s), \circ)$ and $(V_{n,r}(C_n(R)), \circ)$ are abelian groups for every $s \in \mathbb{Z}_n$.*

$V_{n,r}(C_n(R))$ contains all isomorphic circulant graphs of Type-2 of $C_n(R)$ w.r.t. r, if exist. Let $T2_{n,r}(C_n(R)) = \{C_n(R)\} \cup \{C_n(S) : C_n(S)$ is Type-2 isomorphic to $C_n(R)$ w.r.t. $r\}$. Thus, $T2_{n,r}(C_n(R)) = \{C_n(R)\} \cup \{\Theta_{n,r,t}(\ C_n(R)) : \Theta_{n,r,t}(C_n(R)) = C_n(S)$ and $C_n(S)$ is Type-2 isomorphic to $C_n(R)$ w.r.t. $r,\ 0 \leq t \leq \frac{n}{m} - 1\} = \{\Theta_{n,r,0}(C_n(R))\} \cup \{\Theta_{n,r,t}(C_n(R)) : \Theta_{n,r,t}(C_n(R)) = C_n(S)$ and $C_n(S)$ is Type-2 isomorphic to $C_n(R)$, $0 \leq t \leq \frac{n}{m} - 1\} \subseteq V_{n,r}(C_n(R))$ and $(T2_{n,r}(C_n(R)),\ \circ)$ is a subgroup of $(V_{n,r}(C_n(R)),\ \circ)$. Clearly, $T1_n(C_n(R)) \cap T2_{n,r}(C_n(R)) = \{C_n(R)\}$. And $C_n(R)$ has Type-2 isomorphic circulant graph w.r.t. r if and only if $T2_{n,r}(C_n(R)) \neq \{C_n(R)\}$ if and only if $T2_{n,r}(C_n(R)) \cap \{C_n(R)\} \neq \Phi$ if and only if $|T2_{n,r}(C_n(R))| > 1$ [1].

Definition 3.11. [1] *For any circulant graph $C_n(R)$, if $T2_{n,r}(C_n(R)) \neq \{C_n(R)\}$, then $(T2_{n,r}(C_n(R)),\ \circ)$ is called the Type-2 group of $C_n(R)$ w.r.t. r under '$\circ$'.*

Theorem 3.12. [1] *Let p be an odd prime and $k \geq 3$. Then, for $i = 1$ to p, $d_i = (i-1)np + 1$ and $R_i = \{d_i,\ np^2 - d_i, np^2 + d_i, 2np^2 - d_i, 2np^2 + d_i, 3np^2 - d_i,\ 3np^2 + d_i, \ldots, (p-1)np^2 - d_i, (p-1)np^2 + d_i, np^3 - d_i, pp_1, pp_2, \ldots, pp_{k-2},\ p(np^3 - p_{k-2}),\ p(np^3 - p_{k-3}),\ \ldots,\ p(np^3 - p_1)\}$, circulant graphs $C_{np^3}(R_i)$ are Type-2 isomorphic (and without CI-property) where $gcd(p_1, p_2, \ldots, p_{k-2}) = 1$ and $n, p_1, p_2, \ldots, p_{k-2} \in \mathbb{N}$.* □

Theorem 3.13. [1] *Let p be an odd prime, $k \geq 3$, $1 \leq i \leq p$, $d_i = (i-1)np + 1$, $R_i = \{d_i,\ np^2 - d_i,\ np^2 + d_i,\ 2np^2 - d_i,\ 2np^2 + d_i,\ 3np^2 - d_i,\ 3np^2 + d_i,\ \ldots,\ (p-1)np^2 - d_i,\ (p-1)np^2 + d_i,\ np^3 - d_i,\ pp_1,\ pp_2,\ \ldots,\ pp_{k-2},\ p(np^3 - p_{k-2}),\ p(np^3 - p_{k-3}),\ \ldots,\ p(np^3 - p_1)\}$, $T2(R_i) = \{\Theta_{np^3,p,jn}(R_i) : j = 1, 2, \ldots, p\}$, $T2(C_{np^3}(R_i)) = \{\Theta_{np^3,p,jn}(C_{np^3}(R_i)) : j = 1, 2, \ldots, p\}$, $gcd(p_1, p_2, \ldots, p_{k-2}) = 1$ and $n, p_1, p_2, \ldots, p_{k-2} \in \mathbb{N}$. Then $T2_{np^3,p}(R_i) = T2(R_j)$, $T2_{np^3,p}(C_{np^3}(R_i)) = T2(C_{np^3}(R_j))$ and $(V_{np^3,p}(R_i), \circ)$, $(V_{np^3,p}(C_{np^3}\ (R_i)), \circ)$ and $(T2_{np^3,p}(R_i), \circ)$ are abelian groups, $1 \leq i, j \leq p$. Moreover, $(T2_{np^3,p}(C_{np^3}(\ R_i)), \circ)$ is the Type-2 group of order p on $C_{np^3}(R_i)$ w.r.t. $r = p$, $1 \leq i, j \leq p$.* □

Circulant graphs $C_{16}(1,2,7)$ and $C_{16}(2,3,5)$ are Type-2 isomorphic and $C_{27}(1,3,8,10)$, $C_{27}(3,4,5,13)$ and $C_{27}(2,3,7,11)$ are also. See Figures 1–5.

4 Cartesian Product and Factorization of Circulant Graphs

Just as integers can be factored into prime numbers, there are many results on decompositions of structures throughout mathematics [6]. The standard products - Cartesian, lexicographic, tensor, and strong - all belong to a class of products introduced by Imrich and Izbicki and called $B - products$ [8]. In this section, a few important results that are obtained in our study of Cartesian product and factorization of circulant graphs, similar to the theory of product and factorization of natural numbers, are presented (For details see [15]). Graphs $C_5 \square C_6$ and $C_{30}(5,6)$ are isomorphic and are given in Figures 12 and 13. One can see the difficulty of this study from this example.

Definition 4.1. [9] *The cross product or Cartesian product of two simple graphs $G(V, E)$ and $H(W, F)$ is the simple graph $G \square H$ with vertex set $V \times W$ in which two vertices $u = (u_1, u_2)$ and $v = (v_1, v_2)$ are adjacent if and only if either $u_1 = v_1$ and $u_2 v_2 \in F$ or $u_2 = v_2$ and $u_1 v_1 \in E$.*

Theorem 4.2. [15] *Let G be a connected graph of order n, $n > 2$. Then, $P_2 \square G$ is circulant if and only if $G \cong H$ or $P_2 \square H$ where H is a connected circulant graph of odd order.* □

Theorem 4.3. [15] *Let G be a connected graph of order $n \geq 2$. Then $C_4 \square G$ is circulant if and only if G is circulant of odd order.* □

Theorem 4.4. [15] *If G and H are connected graphs and $G \square H$ is circulant, then G and H are circulants.* □

Graphs $P_2 \Box C_3$ and $P_2 \Box C_4$ are given in Figures 10 and 11 and $C_5 \Box C_6$ and $C_{30}(5,6) \cong C_5 \Box C_6$ in 12 and 13, respectively.

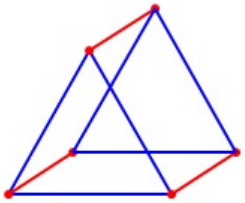


Fig.10. $P_2 \times C_3$

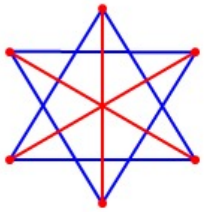


Fig.11. $C_6(2,3)$

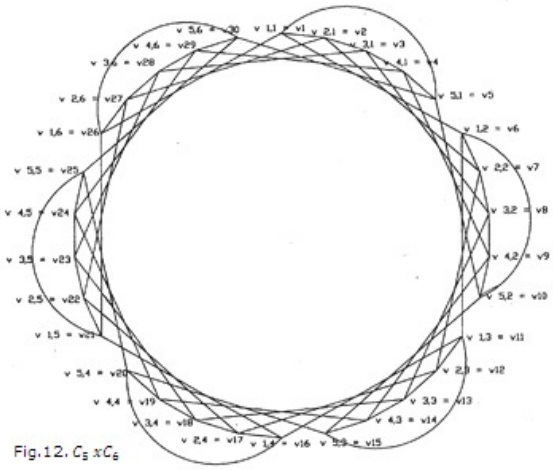


Fig.12. $C_5 \times C_6$

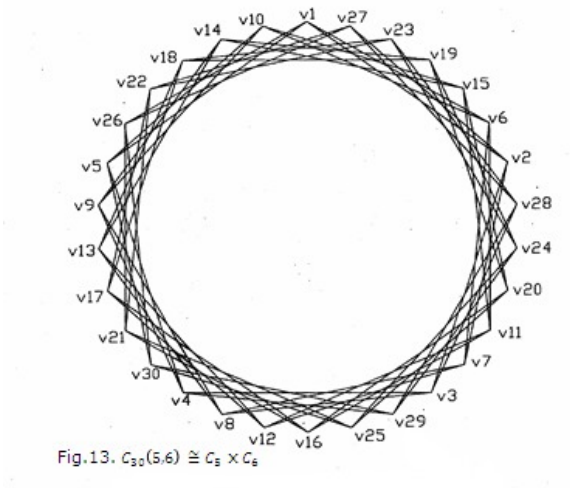


Fig.13. $C_{30}(5,6) \cong C_5 \times C_6$

Theorem 4.5. [15] *Let G and H be connected graphs, each of order > 2. Then $G \Box H$ is circulant if and only if G and H are circulants and satisfy one of the following conditions:*

(i) $G \cong C_m(R)$; $H \cong C_n(S)$ *and* $gcd(m,n) = 1$.

(ii) $G \cong C_{2m+1}(R)$; $H \cong C_{2n+1}(S)$, $P_2 \Box C_{2n+1}(S)$, $C_4 \Box C_{2n+1}(S)$ *or* $C_{2^k(2n+1)}(S)$ *and* $gcd(2m+1, 2^k(2n+1)) = 1$, $k \in \mathbb{N}$.

(iii) $G \cong P_2 \Box C_{2m+1}(R)$; $H \cong C_{2n+1}(S)$ *or* $P_2 \Box C_{2n+1}(S)$ *and* $gcd(2m+1, 2n+1) = 1$.

(iv) $G \cong C_{2^k(2m+1)}(R) \neq P_2 \Box C_{2^{k-1}(2m+1)}(T)$ *for any* $C_{2^{k-1}(2m+1)}(T)$; $H \cong C_{2n+1}(S)$ *and* $gcd(2^k(2m+1), 2n+1) = 1$, $k \in \mathbb{N}$.

(v) $G \cong C_4 \Box C_{2m+1}(R)$; $H \cong C_{2n+1}(S)$ *and* $gcd(2m+1, 2n+1) = 1$.

(vi) $G \cong C_{2^k(2m+1)}(R) \neq C_4 \Box C_{2^{k-2}(2m+1)}(T)$, $P_2 \Box C_{2^{k-1}(2m+1)}(U)$ *for any* $C_{2^{k-2}(2m+1)}(T)$ *and* $C_{2^{k-1}(2m+1)}(U)$; $H \cong C_{2n+1}(S)$ *and* $gcd(2^k(2m+1), 2n+1) = 1$, $k \in \mathbb{N}$, $k \geq 2$. □

Definition 4.6. [15] *A non-trivial graph G is said to be prime if $G = G_1 \Box G_2$ implies G_1 or G_2 is trivial; G is composite if it is not prime.*

Definition 4.7. [15] *If $C_m(R)$, $C_n(S)$ and $C_{mn}(T)$ are circulant graphs such that $C_m(R) \Box C_n(S) \cong C_{mn}(T)$, then we say that $C_m(R)$ and $C_n(S)$ are divisors or factors of $C_{mn}(T)$.*

Thus for any connected circulant graph, the graph and $C_1(\,) = K_1$ are always divisors and so we call them as *improper divisors* of the circulant graph. Divisors which are integer multiple of improper divisors also be called as improper divisors of the circulant graph. This doesn't arise since we consider divisors of connected graphs only. Divisor(s) other than improper divisors is called *proper divisor(s)* of the circulant graph.

Definition 4.8. [15] *A circulant graph whose only divisors are improper is called a prime circulant graph. Other circulant graphs are called composite circulant graphs.*

Theorem 4.9. [15] ***[Factorization Theorem On Circulant Graphs]***
Let m and n be relatively prime integers. If $R \subseteq [1, \frac{m}{2}]$, $S \subseteq [1, \frac{n}{2}]$ and $T \subseteq [1, \frac{mn}{2}]$ with $T = dnR \cup dmS$ for some d such that $gcd(mn, d) = 1$, then $C_{mn}(T) \cong C_m(R) \Box C_n(S)$. □

Theorem 4.10. [15] *If $n \neq 4$ and $1 \in R$, then $C_n(R)$ is a prime circulant.* □

Corollary 4.11. [15] *If $n \neq 4$ and R contains an integer relatively prime to n, then $C_n(R)$ is prime circulant.* □

Corollary 4.12. [15] *If n is a prime power other than 4 and $C_n(R)$ is connected, then $C_n(R)$ is prime circulant for all $R \neq \phi$.* □

Theorem 4.13. [15] ***[Fundamental Theorem of Circulant Graphs]***
Every connected circulant graph is the unique product of prime circulant graphs (uniqueness up to isomorphism). □

Remark 4.14. [13, 15] *If G is a connected graph such that $G \cong G_1 \Box G_2 \Box \ldots \Box G_k$, then the diameter of G, $dia(G) = \sum_{i=1}^{k} dia(G_i)$.*

Thus, we can find the diameter of any given circulant graph, provided diameters of its prime circulant graphs are known. Also the above relation helps to generate (circulant) graphs of bigger diameters.

Remark 4.15. [15]

1. *In prime factorization of connected circulants* $C_1(\ ) = K_1$ *and* $C_2 = P_2$ *act similar to* 1 *and* 2 *among the set of all natural numbers, respectively. Thus,* $C_1(\ )$ *is a unit, like* 1 *in number theory.*
2. *There exist two types of prime circulant graphs of order n, one with periodic cycle(s) of length n and the other without periodic cycle of length n.*
3. *The theory of factorization of circulants is similar to the theory of factorization of natural numbers and one of the very few well-known mathematical structures so vividly classified (expressed) in terms of prime factors. It can be applied in cryptography.*
4. *We developed VB programs POLY215.exe and POLY315.exe to show visually how the transformations* $\Theta_{n,r,t}$ *and* $\Phi_{m,n}$ *act on* C_n *and* $C_m \square C_n$*, respectively for different values of m and n,* $m, n \in \mathbb{N}$.
5. *An interesting problem is, for a given integer n, finding the number of prime (composite) circulant graphs of order either equal to n or less than or equal to n.*
6. *One can develop theories similar to the theory of Cartesian product and factorization of circulant graphs to the other standard products of circulant graphs.*

Conclusion This study covers a few isomorphic properties of circulant graphs. One can go for a similar study on Cayley graphs.

Acknowledgment The author expresses his sincere thanks to Prof. Lowell W Beineke, Indiana-Purdue University, U.S.A., Prof. Brian Alspach, University of Newcastle, Australia, Prof. M.I. Jinnah, University of Kerala, Trivandrum, India and Prof. V. Mohan, Thiyagarayar College of Engineering, Madurai, Tamil Nadu, India for their valuable suggestions and guidance and Dr. A. Christopher, Dr. P. Wilson and Mr. R. Satheesh of S.T. Hindu College, Nagercoil, India for their assistance to develop the VB programs to show how the transformations defined on circulant graphs to obtain their isomorphic graphs is taking place. He also expresses his gratitude to Central University of Kerala (CU Kerala), Periye - 671 316, Kasaragod, Kerala, India and St. Jude's College, Thoothoor - 629 176, Kanyakumari District, Tamil Nadu, India for providing facilities and CU Kerala and Lerroy Wilson Foundation, India (www.WillFoundation.co.in) for providing financial assistance to do this research work.

References

1. V. Vilfred, "A few properties of circulant graphs: Self-complementary, isomorphism, Cartesian product and factorization" in 7^{th} International Conference on Modeling, Simulation and Applied Optimization (ICMSAO2017), American University of Sharjah, Sharjah, April 2017.
2. P. J. Davis, Circulant Matrices, Wiley, New York, 1979.
3. A. Adam, "Research problem 2-10", J. Combinatorial Theory, **3**, 393, 1967.
4. B. Alspach, J. Morris and V. Vilfred, "Self-complementary circulant graphs", Ars Com., **53**, 187–191, 1999.
5. B. Alspach and T. B. Parsons, "Isomorphism of circulant graphs and digraphs", Discrete Math., **25**, 97–108, 1979.
6. R. Arratia, A. D. Barbour and Simon Tavare, "Random Combinatorial Structures and Prime Factorizations", A.M.S.Notices, **44**, 903–910, 1997.
7. F. T. Boesch and R. Tindell, "Circulant and their connectivities", J. Graph Theory, **8**, 487–499, 1984.
8. B. Elspas and J. Turner, "Graphs with circulant adjacency matrices", J. Combinatorial Theory, **9**, 297–307, 1970.
9. F. Harary, Graph Theory, Addison-Wesley, Reading, M.A., 1969.
10. W. Imrich and H. Izbicki, "Associative products of graphs", Monatsh. Math., **80**, 277–281, 1975.
11. I. Kra and S. Simanca, "On Circulant Matrices", AMS Notices, **59** (3), 368–377, 2012.
12. C. H. Li, "On isomorphisms of finite Cayley graphs-a survey", Discrete Math. **256**, 301–334, 2002.
13. G. Sabidussi, "Graph multiplication", Math. Z., **72**, 446–457, 1960.
14. H. Sachs, "Uber selbstkomplementare Graphen", Publ. Math. Debrecen, **9**, 270–288, 1962.
15. V. Vilfred, "A Theory of Cartesian Product and Factorization of Circulant Graphs", Hindawi Pub. Corp. - J. Discrete Math., Vol. 2013, Article ID 163740, 10 pages.
16. V. Vilfred, "New Abelian Groups from Isomorphism of Circulant Graphs", Proce. of Inter. Conf. on Applied Math. and Theoretical Computer Science, St.Xaviers Catholic Engineering College, Nagercoil, Tamil Nadu, India, xiii–xvi, 2013.
17. V. Vilfred, "$\sum$-labelled Graphs and Circulant Graphs", Ph.D. Thesis, University of Kerala, Trivandrum, India, 1994.
18. V. Vilfred and P. Wilson, "New Families of Circulant graphs without Cayley Isomorphism Property with $m_i = 5$", Int. Journal of Scientific and Innovative Mathematical Research, **3** (6), 39–47, 2015.
19. V. Vilfred and P. Wilson, "New Family of Circulant Graphs without Cayley Isomorphism Property with $m_i = 7$", IOSR Journal of Mathematics, **12**, 32–37, 2016.
20. H. Zhang, "Self-complementary symmetric graphs", J. Graph Theory, **16**, 1–5, 1992.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 239-243 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Approximate method of analysis of log-periodic antennas with in-phase currents

Boris Levin*

Holon Institute of Technology, Israel

ARTICLE INFO

Article history:
Received: 23 August, 2017
Accepted: 07 January, 2018
Online: 30 January, 2018

Keywords:
Antenna theory,
Directional antennas,
In-phase current distribution,
Method of electrostatic analogy,
Numerical calculation

ABSTRACT

Log-periodic antennas have high electrical characteristics over a wide frequency range. The large length of the antennas is their significant disadvantage caused by the fact that each active region consisting of several radiators operates in a narrow frequency range. To extend this range, it is proposed to use radiators with concentrated capacitive loads, providing in-phase currents with a linear amplitude distribution law. An approximate calculation of the fields and directional characteristics of a complex multielement radiating structure is performed using a new method of electrostatic analogy based on the assumption that the ratio of emf in the radiator centers is equal to the charge ratio on the conductors of the electrically neutral system. The problem of optimization of considered antennas by methods of mathematical programming is formulated.

1. Introduction

The use of concentrated loads included along the axis of a linear antenna allows solving the synthesis problem, i.e., to obtain an antenna with required characteristics, more precisely with characteristics closest to required [1]. The special case of this problem - the creation of an antenna that provides a high level of matching in a wide frequency range and a maximum of radiation in a plane perpendicular to the axis of the radiator – has a great practical importance.

A conventional linear radiator does not satisfy these requirements. A reactive component of its input impedance is great everywhere except of a vicinity of a series resonance. When the length of a radiator arm is greater than 0.7λ (λ is a wavelength), the radiation in the plane perpendicular to its axis decreases, since anti-phase segments form on the current curve.

The task of creating a wide-range radiator is to determine the types and quantities of loads that provide high characteristics of the antenna in the operating frequency range. This task was solved by applying concentrated capacitive loads (capacitors) installed along the axis of the radiator and forming an in-phase current with a given law of the amplitude distribution along the antenna. The solution was based on an understanding of the advantages of in-phase current distribution and Hallen's hypothesis on the use of capacitive loads, the values of which vary along the radiator axis in accordance with a linear or exponential law. The chosen approach confirmed the Hallen's hypothesis and demonstrated the effectiveness of the proposed approximate methods of calculating capacitive loads: the method of impedance long-line and the method of a long-line with loads. The loads calculated by these methods were used as the initial values for numerical solution of the problem by methods of mathematical programming.

Since the placement of concentrated capacitive loads along the radiator allows to obtain not only a high level of matching of the antenna with the cable, but also high directional characteristics in a wide frequency range, the application of these loads had been considered in radiators – first in a linear antenna and then in directional one.

It should be emphasized that the optimization problem, as a rule, is solved in two stages. At the first stage an approximate method of analysis, based on physical representations about the nature of the problem, is developing. At the second stage, approximate results are used as initial values for the numerical solution of the problem by methods of mathematical programming.

The proposed method of an approximate analysis of antennas characteristics allowed us to calculate the directional patterns of radiators with a given current distribution and to compare performances of antennas, in which various laws of current distribution are realized. This is, firstly, the linear metal antenna with a sinusoidal current distribution along its axis, and secondly, the linear antenna with an in-phase current and with a linear or exponential law of an amplitude distribution. Calculations show that the use of linear antennas with capacitive loads makes it possible to obtain in the wide frequency band the desired shape of the radiation pattern, which provides a substantial increasing the communication distance.

*Boris Levin, Email: levinpaker@gmail.com

To evaluate this shape, a special parameter has been introduced - the pattern factor, equal to an average radiation level in a given range of vertical angles, for example, at angles θ from 60° to 90°. This parameter shows in fact what part of the signal is propagating along the earth's surface, i.e., it is useful for increasing the communication distance. With the help of the proposed method, it is shown that the in-phase current distribution in linear and V-radiators provides not only a high level of matching, but also the high directivity and smooth variation of this directivity over a wide frequency range [2].

For an approximate calculation of the fields of complex multielement radiating structures, the method of electrostatic analogy was proposed. It is based on the assumption that the ratio of emf in the radiators centers is equal to the ratio of the charges located on the conductors of electrically neutral system. This allows proposing a simple and effective procedure of calculating directional characteristics of director and log-periodic antennas with concentrated capacitive loads. The calculation results confirm the possibility of increasing the frequency range of the director antennas and decreasing the length of the log-periodic antennas.

The results of applying this method to calculating the directional characteristics of director antennas are given in [3]. They are compared with the characteristics of director antenna, described in [4] and consisting from four metal elements: active radiator, reflector and two directors. When optimizing the metal antenna, the mathematical programming method was used. The application of the approximate method of electrostatic analogy to a metal antenna confirmed the results obtained in [4]. With the help of the presented procedure, it was shown also that in-phase current distribution in radiators of the director antennas provides the higher directivity in a wide frequency range.

The remainder of this paper is organized as follows. In Section II, the proposed method of analysis is applied to the calculation of fields and directional characteristics of an active region of the log-periodic antenna. The numerical results obtained in Section III for such antennas with in-phase currents show that the number of such regions can be substantially reduced. The section IV contains conclusions, which emphasize a theoretical interest and practical benefits of the work.

2. Method of Analysis

In accordance with a well-known method of analysis of a log-periodic antenna [5], we consider an active region of this antenna, consisting of three radiators (Fig. 1), and find fields of these radiators when emf *e* is located in the center of the middle radiator. In accordance with Kirchhoff' law other generators must be short-circuited. As shown in [5], this means that emf's in the centers of other radiators and at the ends of sections of a two-wires distributive long line in the points of these radiators location are shorted also. We regard that the arm length of the middle radiator is $L_0 = \lambda/4$, the arm length of the left (longer) radiator is equal to $L_1 = \lambda/(4\tau)$, and the arm length of the right (shorter) radiator is $L_2 = \lambda\tau/4$, where τ is a denominator of a geometric progression, according to which the radiators' dimensions are changed. A magnitude σ is the other antenna parameter, equal to $\sigma = 0.25(1-\tau)\cot\alpha$, where α is an angle between an antenna axis and a line passing through radiators ends. The value σ is the number of wavelengths between the half-wave radiator and the smaller neighboring radiator: $b_2 = \sigma\lambda$.

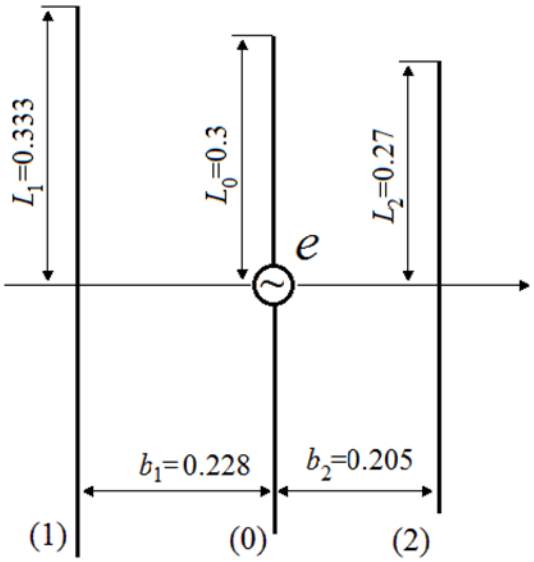


Figure 1: Antenna of three radiators

Respectively b_1 is equal to $\sigma\lambda/\tau$. As stated in [6], generalization of information in the literature leads to the conclusion that the minimum changes in the electrical characteristics of the log-periodic antenna with metal dipoles within the frequency range from f to $f\tau$ occur when $\sigma/\tau = 0.19$. We regard that this relation is valid for the considered antenna and that value τ for certainty is equal to 0.9. Then $L_1 = L_0/0.9 = 0.278\lambda$, $L_2 = L_0 \cdot 0.9 = 0.225\lambda$, $\alpha = 0.146$, $b_1 = 0.19\lambda$, $b_2 = 0.171\lambda$.

As already mentioned, we use the method of electrostatic analogy of two structures (of currents and charges), located on the same conductors, as an approximate method. We are assuming that the ratio of emf's in the radiator centers is equal to the ratio of charges located on these conductors. The positive charge, equal to Q_0, is placed on the conductor *0* (active radiator). The charges of other conductors *i* are negative. They are equal to $-Q_i$, and their sum is equal to $\sum_{i=1}^{2}(-Q_i) = -Q_0$, i.e., the sum of all charges is zero, and the conductors form an electrically neutral system. In this system

$$Q_i/Q_0 = C_{0i}\Big/\sum_{i=1}^{2} C_{0i}, \qquad (1)$$

where C_{0i} is the partial capacitance between conductors *0* and *i* (see, for example, [7]). From (1) it follows that the charges of the conductors *i* are directly proportional to the partial capacitances C_{0i} between these conductors and the conductor *0*. The mutual capacitance C_l per unit length of two parallel infinitely long conductors of the radius *a* is equal to

$$C_l = \pi\varepsilon/\ln(b/a), \qquad (2)$$

where ε b is the distance between the conductors. Multiplying this value into the conductor length,

we find in a first approximation the partial capacitance C_{0i} between these conductors.

We divide the active radiator (conductor *0*) into two parallel conductors (with numbers *01* and *02*), and the resulting structure of four conductors into two circuits: conductors with numbers *01* and *1* in the first scheme, and conductors with numbers *02* and *2* in the second scheme. In accordance with the method of electrostatic analogy of the two structures, i.e., in accordance with the physical content of the problem, we will assume that the ratio of the emf in the centers of the wires *01* and *02* is equal to the ratio of the partial capacitances C_{01} and C_{02}, i.e., $e_1 = 0.52e, e_2 = 0.48e$.

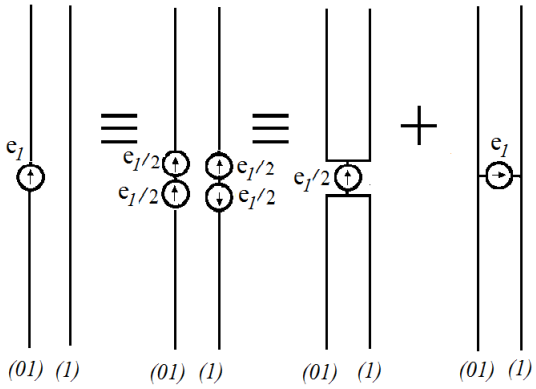


Figure 2: Successive transforming a structure of two parallel vertical conductors

Fig. 2 demonstrates a scheme for successive transforming a structure of two parallel vertical conductors (*01* and *1*), one of which is excited at the center, into a structure consisting of a two-conductor dipole and two open transmission lines. The scheme is based on the theory of a folded antenna (see for example, [1]). As can be seen from this figure, the current at the center of each dipole conductor and the current at the same point of each transmission line are respectively equal

$$J_{1d} = e_1/(4Z_{1d}), J_{1l} = e_1/(2Z_{1l}), \tag{3}$$

where Z_{1d} is the input impedance of this dipole, Z_{1l} is the input impedance of each long line. Accordingly, the current J_{01u} of the upper arm of the conductor *01* is equal to the difference of these currents, and the current of the lower arm of the conductor *01* - to the sum:

$$J_{01u} = J_{1d} - J_{1l}, J_{01l} = J_{1d} + J_{1l}. \tag{4}$$

The current J_{1u} of the upper arm of the conductor *1* on the contrary is equal to the sum of these currents, and the current J_{1l} of the lower arm of the conductor *1* – to their difference. The field amplitude and phase value of each radiator depend on its structure and location in an antenna. If the radiator arm is manufactured as the straight metallic wire, the current along it is distributed by the sinusoidal law

$$J(z) = J(0)\sin k(L - |z|). \tag{5}$$

Here z is a coordinate along the radiator axis, $J(0)$ is the current in its center, k is the propagation constant. In this case the far field of the radiator is equal to

$$E = \frac{AJ(0)}{Sin\theta}[\cos(kL\cos\theta) - \cos kL], \tag{6}$$

where $A = j\frac{60}{\sin kL}\cdot\frac{\exp(-jkR)}{R}$, R is the distance to the observation point. If concentrated capacitive loads located along the radiator axis allow realizing in it the in-phase current with an amplitude, distributed along this axis by the linear law,

$$J(z) = J(0)(L - |z|), \tag{7}$$

the far field of the radiator (see [2]) is equal to

$$E = \frac{AJ(0)\sin\theta}{\cos^2\theta}[1 - \cos(kL\cos\theta)]. \tag{8}$$

Consider an impact of the radiator location on the far field of the antenna by way of a specific example of the director antenna, shown in Fig. 1. The arm length of the middle (active) radiator is equal to 0.3 м, i.e., the wave length of the first (series) resonance is equal to 1.2 м. Radii of all conductors are the same and are equal to 0.001м. The magnitude τ is 0.9.

It is obvious that the maximal radiation of the antenna should be directed to the right, toward the radiator *2*. Since the radiator *1* is located from the left of the active radiator *0*, at a distance b_1 from it, its field lags behind the field of the active radiator, first on kb_1 in phase, that is by the time interval of the propagation of the signal from the active radiator to the passive radiator *1*, and, secondly, by $kb_1/\sin\theta$ in phase, i.e., by the time of the signal propagation in the opposite direction, from the radiator *1* to the active radiator (signal of the wire *1*, radiated at an angle θ, must come to the active radiator at the same angle θ, i.e., it must pass the distance $b_1/\sin\theta$, and not the distance b_1. The total change in phase is equal to $\psi_1 = -kb_1\cdot\frac{1+\sin\theta}{\sin\theta}$. Similarly, in the case of radiator *2*, this phase change is equal to $\psi_2 = -kb_2\cdot\frac{\sin\theta - 1}{\sin\theta}$.

The total field of the antenna structure shown in Fig. 1 with in-phase current distribution at angle θ on the base of aforesaid may be written in the form

$$E = \frac{AJ(0)\sin\theta}{2Z_{1d}\cos^2\theta}\sum_{i=1}^{2} e_i\{[1 - \cos(kL_0\cos\langle\theta - \theta_0\rangle)] + \exp(j\psi_i)[1 - \cos(kL_i\cos\theta)]\}. \tag{9}$$

The directivity magnitude is determined by the expression

$$D = |E(\pi/2)|^2 \Big/ \sum_{n=1}^{N}[|E(\theta_n)|^2\Delta\sin\theta_n], \tag{10}$$

where Δ is the interval between neighboring values θ_n, N is the number of these intervals between $\theta = 0$ and $\theta = \pi/2$.

3. Examples of Calcultations

The results of calculating the directivity magnitude for the structure, shown in Fig. 1, are given in Fig. 3 depending on an electrical length kL_0

$kL_0 = \pi$ to $kL_0 = 10\pi$.

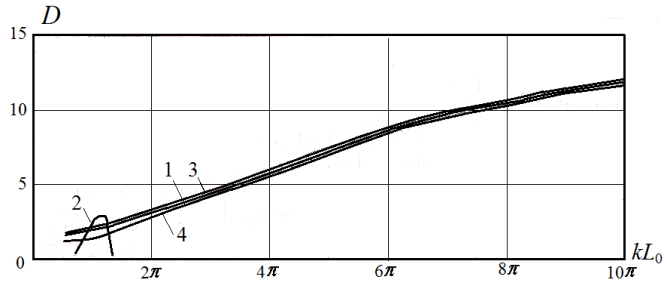


Figure 3: Directivity of structures of three radiators

Radiation of neighboring elements of log-periodic antenna with the arm length 0.27 m and 0.333 m respectively requires calculating fields in the structures presented in Fig. 4*a* and 4*b*. Results of these calculations are given in Fig. 3: curve 3 corresponds to Fig. 3*a*, curve 4 - to Fig. 3*b*. The directivity magnitudes in Fig. 3 for specific values kL correspond to the same frequencies, i.e., the same values of kL correspond to the elements of equal length in the all three schemes (for example, to the elements with the arm length 0.3 m). Fig. 3 shows that the directivity magnitudes at the same kL are close to each other. This is natural, since the directivity in each scheme increases slowly with increasing frequency. Small increasing the active radiator length causes at the same kL the small increase of the resonant wavelength, i.e., the small decrease of the resonance frequency and the directivity.

The obtained results show that each active radiator with in-phase current included in the structure of the log-periodic antenna provides high radiation directivity in a wide frequency range. Neighboring radiators have similar directivity magnitude. The direction of the radiation is the same. A signal propagates along the distributive line from short elements to the long dipoles (in Figures 1 and 4 to the left), the radiated signal propagates in the opposite direction. The total path difference is quite large. For example, for the signals radiated by a half-wave radiator and a smaller neighboring radiator it is equal to 0.38 λ, i.e., it is close to a half wavelength. Crossing wires of the distributive line in an interval between the elements permits to reduce dramatically this path difference.

Known log-periodic antenna with sinusoidal current distribution along the radiators have a property of automatic currents "cut-off", i.e., a separate antenna section (active region) radiates a signal in a narrow frequency band. Outside this band, outside the borders of the active region the signal decays rapidly. Wide frequency range is provided by a large antenna length, which is equal to the sum of the lengths of the active regions. Attempts to reduce the antenna length by disturbing a geometric progression and increasing a radiators number lead to a small dimension decrease and a sharp deterioration of electrical characteristics. The more effective methods are firstly a two-fold use of each active region by an application of linear-spiral radiators and secondly an employment of an asymmetrical log-periodic antenna with coaxial distribution line [6]. These methods allow decreasing the antenna length by 25-30%.

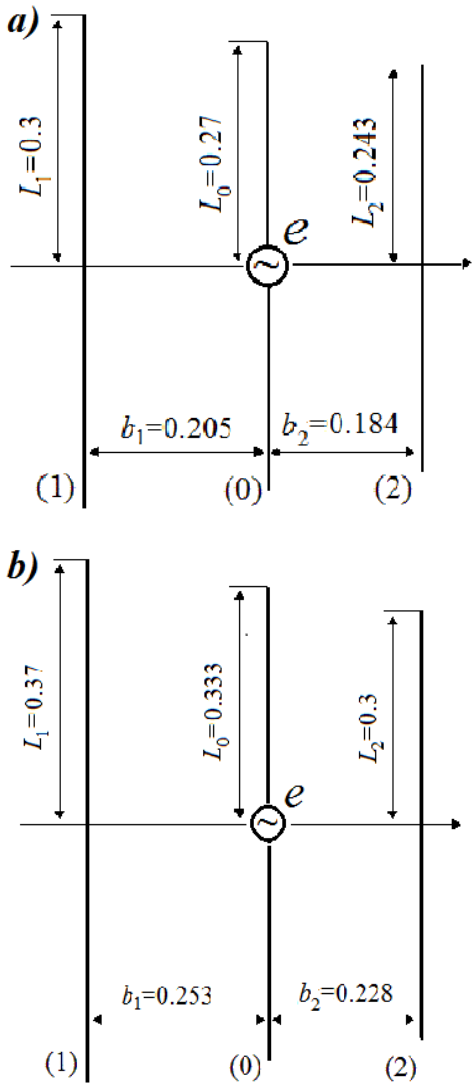


Figure 4: Structures of smaller (*a*) and greater (*b*) neighboring radiators

Replacement of the straight metal radiators by the radiators with concentrated capacitive loads allows obtaining a high directivity in a wide frequency range, using a simple structure of three radiators. Increasing the radiators number in the structure must allow to dramatically increasing its directivity.

4. Conclusion

In this work the following results are obtained:

1) The method of electrostatic analogy is proposed for calculating fields of complex multielement antennas. It is based on an assumption that a ratio of emf in radiators centers is equal to a ratio of charges located on conductors of an electrically neutral system. This method had been used for calculating directional characteristics of log-periodic antennas.

2) The given procedure permitted to compare the directional characteristics of log-periodic antennas with sinusoidal and in-phase current distribution in simple radiators. It was shown that replacement of the metal radiators by the radiators with concentrated capacitive loads allows providing, using a few

radiators, the high directivity in a wide frequency range, i.e., decreases sharply the antenna length. Increasing radiators number must allow increasing substantially the antenna directivity.

Results obtained by means of the method of electrostatic analogy may be used for solving optimization problem log-periodic antennas by methods of mathematical programming.

References

[1] 1. B. M. Levin. Antenna engineering. Theory and problems. A science publishers book: London, New York, 2017.

[2] 2. B. M. Levin. "Directional properties of linear and V-antennas" in 11 Intern. Confer. ICATT'17. Kiev (Ukraine), 2017, pp.104-109.

[3] 3. B. M. Levin. "Director antennas with in-phase currents" in 11 Intern. Confer. ICATT'17. Kiev (Ukraine), 2017, pp.110-113.

[4] 4. A. F. Chaplin, M. D. Buchazky, M. Yu. Mihailov. "Optimization of director-type antennas". Radiotechnics, 1983, no. 7, pp. 79-82 (in Russian).

[5] 5. R. L. Carrel. "The design of log-periodic dipole antennas. IRE Intern. Convention Record, part 1, 1961, pp. 61-75.

[6] 6. A. F. Yakovlev, A. E. Pyatnenkov. Wide-band directional antennas arrays from dipoles. S.-Petersburg: Research center of communication, 2007 (in Russian).

[7] 7. Yu. Ya. Iossel, E. S. Kochanov, M. G. Strunsky. Calculation of Electrical Capacitance. Leningrad: Energoisdat, 1981 (in Russian).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 242-251 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Efficiency Assessment of Plasma-Aided Porous Media Surface Finishing

Peter Dineff*, Dilyana Gospodinova, Ivaylo Ivanov

Technical university of Sofia, Faculty of electrical engineering, 1756, Bulgaria

ARTICLE INFO

Article history:
Received: 22 October, 2017
Accepted: 11 November, 2017
Online: 23 December, 2017

Keywords :
Processing efficiency assessment
Plasma-aided surface finishing
Porous media

ABSTRACT

The efficiency of a plasma-aided finishing process, namely capillary impregnation, can be predicted only by setting two basic parameters of the real wood porous surface and the actual impregnating solution – surface free energy and surface tension. In general, the following processing efficiency parameter was found and the rule is true: "The plasma aided or enhanced finishing of a porous media will be more successful and this media will be more susceptible to it as the difference between its surface free energy and the surface tension or the so-called penetration-spreading parameter is positive: PSP = $(\sigma_S - \gamma_L) > 0$. If not, wetting, wicking and finishing problems will occur.

1. Introduction

This paper is an extension of a work originally presented at the International conference on electrical machines, drives and power systems "*ELMA'17*" [1].

Dielectric-barrier discharges (DBD) have been known for more than a century – the first experimental investigations focused on ozone generation were reported by *Werner von Siemens* in 1857. In 1951, *Werner Eisby* founded his company *Vetaphone* on the patent of a new plasma surface technology, which he initially named *E-treatment* and later it was known worldwide as the *Corona technology*. In fact, *Corona* was a *Dielectric-barrier discharge* in air at atmospheric pressure. This high-frequency electric discharge was controlled by a dielectric barrier and organized in the *micro-discharge mode* – it created a pattern of streamers distributed over the whole surface. This *atmospheric pressure air-plasma* (APP) *pre-treatment* was necessary to obtain sufficient wetting and adhesion on plastic films or metallic foils before high-speed roll-to-roll converting processes to ensure perfect surface finishing - printing, laminating or coating. Adding a *Corona treater* on a blown film or cast film extrusion line was essential in order to obtain a perfect *surface free energy* (SFE) prior to the following converting process. This *plasma-aided finishing* has proved to be both highly effective, cost-effective and can take place in-line [2, 3].

In the 80s of the last century, the *APP pre-activation* came to facilitate the textile finishing. In textile manufacturing finishing refers to the processes, that convert the woven, non-woven or knitted cloth into a usable material or fabric to improve the aesthetic appearance and the functional characteristics of the finish textile. This *plasma-aided finishing* has proved to be both highly effective, cost-effective and can take place not only in-line, but also out-of-the-line. The finish textiles such as carpets, rugs, mats, curtains or clothing require the implementation of plasma-aided finishing *out of the production line* [3, 4].

At the beginning of this century, the *APP-activation* (APPA) was used to enhance the wood finishing. In wood manufacturing finishing refers to the process of refining or protecting wooden surfaces. Wood finishing is the final step of the manufacturing process that gives wood surfaces desirable characteristics, including enhanced appearance and increased resistance to moisture, to fire and to attack from destructive organisms such as fungi and insects. It is well known that the inactivated by hydrophobic contamination wood surface has low *SFE*. This *wood inactivation*, [5], leads to poor wetting, adhesion, and impregnation especially for water-based varnishes and printing inks, adhesives and coating substances, as well as flame retardant water solution. The plasma-aided finishing has proved to be both highly effective, cost-effective, [6, 7], and can take place not only in-line, but mostly out-of-the-line. The three-dimensional wooden buildings and structures require removal of the plasma-aided finishing "*in situ*" out of the production line [8].

The *APPA* is essential in surface finishing processes. Without plasma-chemical surface pre-treatment, the porous and non-porous substrates will not allow further finishing processes due to the low *SFE*. The *plasma surface activation* and *functionalization* are able

* Corresponding Author: Peter Dineff, Technical University - Sofia, Faculty of electrical engineering, 1756, Bulgaria, dineff_pd@abv.bg

to ensure that perfect finishing result we all want to see when looking at finish products. The objective of this kind of activation is to increase the *SFE* and subsequently the wettability, adhesion and wicking for laminating, printing, dyeing, gluing, impregnating and other finishing processes [9, 10].

Established *APPA* processes, such as plasma or flame activation, increase the *SFE* and decrease the contact angle, and therefore improve the processing characteristics. The *Sessile Drop Test* - a contact angle measurement technique, and the *Dine Test* - an ink test technique, were two widely used methods for *SFE* assessment and therefore for checking the *need for APPA* or the efficiency of such a pre-treatment processes [11, 12, 13].

The *SFE* of a solid cannot be measured directly because solids typically show no reaction to the exertion of *SFE*. Plasma, corona or flame treated and untreated surfaces are indistinguishable to the naked eye. Consequently, practical measurements of *SFE* involve the interaction of the solid with a test liquid to determine wetting surface tension as a measure of *SFE*. Standards such as DIN EN 828: 2013-04 (E) for adhesive finishing specify contact angle measurement as a method for determining the *SFE* of surfaces for checking the efficiency of cleaning or activating plasma pre-treatment processes.

The purpose of this paper is not only to collect new evidence and extend knowledge about the plasma-aided porous media finishing, especially *capillary impregnation,* with phosphorous flame retardant (FR) containing water solution for *flame retardancy* as finishing for wood, wooden and cellulosic materials, but also to provide a process efficiency assessment for the overall plasma enhanced process, not only for the *APPA pre-treatment*.

2. Assessment of plasma-activated finishing efficiency

2.1. Model of wetting-wicking theory and efficiency assessment

Plasma-aided technique was used to facilitate the surface finishing, more precisely the *surface impregnation* of porous media such as wood, wooden and cellulosic materials [1, 14-17].

Porous materials with low *SFE* (σ_S) were subjected to plasma-chemical surface pre-activation for increasing their *SFE* and enhance the finishing. The wetting phenomenon occurs in a different way on porous media. The surface curvature of a sessile liquid drop on a porous solid depends not only on wetting but also on wicking in depth. This is a phenomenon of wetting-wicking. Wicking and wetting are essential in porous media finishing, such as capillary impregnation for flame retardancy of wood, wooden materials,woven and non-woven textile Figure 1, [1].

Both wetting and wicking together determine the variation of the sessile drop curvature and the contact angle over time. This is the wetting-wicking and contact angle phenomenon. That is why it is imperative in the case of porous media to talk about *apparent contact angle* and *apparent surface free energy.* This process of contact angle change ends with the complete absorption of the flame retardant solution at almost constant value of the contact angle - *end contact angle*, much lower than the *initial contact angle,* Figure 2.

The "*wetting-wicking theory*", expressed in terms of the thermodynamic parameters, such as the *contact angle*, the *surface tension*, and the *surface free energy*, can be the most widely used approach to porous media in wetting science at future, also referred here as "*wetting-wicking phenomena*". Wetting-wicking is defined here as the ratio between the surface energies of the solid and the surface tension of the liquid [14].

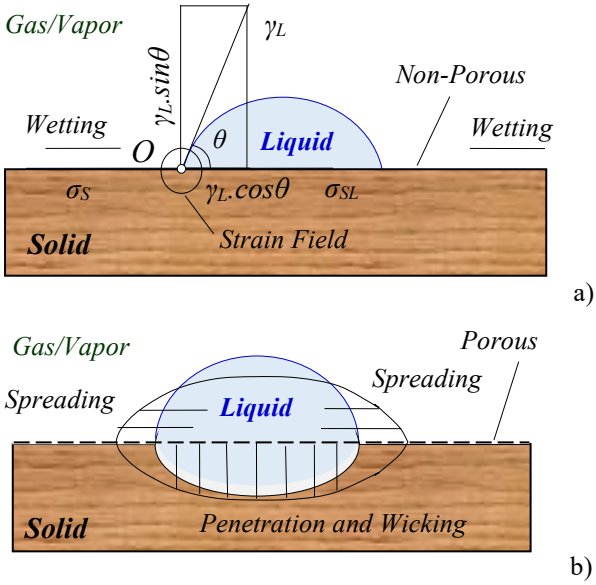


Fig. 1. Schematic illustration of two types of *wetting phenomena*: **a)** the *wetting phenomenon* on a *non-porous surface* when a liquid drop is placed on a smooth, non-porous and rigid solid, both exposed to gas/vapor, the system will not be in equilibrium and the liquid "wets out" or spreads over the solid, then the liquid exhibits a contact angle of zero against the solid (*Good,* 1993); **b)** *the wetting-wicking* or shortly *wicking phenomenon* on a *porous surface* can involve: *i* - spreading over the solid; *ii* - penetration or wicking into the porous solid (*Berg*, 1993).

According to *Young's equation*, which is proven since over 200 years (1805), there is a fundamental relationship between the *contact angle* θ, deg, the *surface free energy* (SFE) of the solid σ_S, mJ/m^2, the *surface tension* (SFT) of the liquid γ_L, mN/m, and the *interfacial tension* (IFT) σ_{SL} between liquid and non-porous or porous solid, mJ/m^2, Figure 1a, [1, 15]:

$$\sigma_S - \sigma_{SL} - \gamma_L \cos\theta = 0; \quad (1)$$

$$\Sigma = \gamma_L \cos\theta = \sigma_S - \sigma_{SL}, \quad (2)$$

where Σ is the *adhesion tension*, mJ/m^2; the adhesion tension reaches its maximum value at full wetting surface when the *IFT* reaches its minimum value: $\cos\theta = 1$ ($\theta = 0°$); $\sigma_{SL} = \sigma_S - \gamma_L$; and $\Sigma = \gamma_L$.

The liquid in porous materials is often pulled into surface pores and capillaries by *capillary action*. The capillary action or *capillarity* is defined as the movement of liquid within the spaces of a porous media due to the forces of adhesion, cohesion and surface tension. Capillarity manifests itself when the adhesion to the capillary walls is stronger than the cohesive forces between the liquid molecules. Its ability to fill-in defects, such as surface-breaking cracks, voids, pores and capillaries can be imputed to a great extent to penetration, sorption or wicking characteristics. In porous media, capillary pressure p_C is the force necessary to squeeze a droplet through a pore throat and to work against the *IFT* between solid and liquid phases. Capillary pressure p_C is proportional to the adhesion tension Σ and depends only on it at a certain characteristic pore size r [16]:

$$p_C = 2\,(\sigma_S - \sigma_{SL})/r = 2\,\Sigma/r, \quad (3)$$

where the capillary pressure p_C reaches its maximum value when the adhesion tension Σ reaches its maximum value at complete wetting: $cos\ \theta = 1$ ($\theta = 0°$); $\Sigma = \gamma_L$; and $p_C = 2\ \gamma_L/r$.

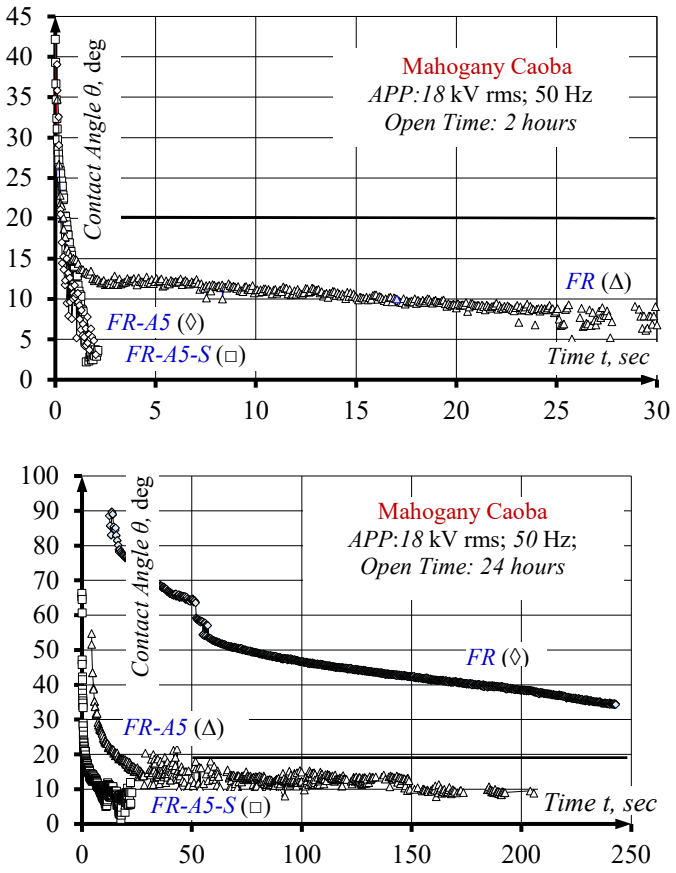


Fig. 2. Dynamics of change in *apparent contact angle* θ over time t depending on the used impregnating water solutions of phosphorous flame retardant with different surface tension: *FR* – 46.4 mN/m; *FR-A5* – 22.4 mN/m and *FR-A5-S* – 18.5 mN/m, and short and long aging time, 2 or 24 hours.

Sessile and pendant drop analyses allow *SFE* and *SFT* to be quantified and to propose two distinct relationships that illustrate in general this process in the wetting area, $\theta \in [0 \div 90]°$, – the *IFE* and adhesion tension Σ - proportional to capillary tension p_C, as a functions of the contact angle θ, Figure 3.

2.2. The best solution rule for finishing and plasma-aided finishing efficiency parameter

Both *SFE* and *SFT* are essential in porous media finishing phenomena, (1). *APPA* surface technique can be used to enhance surface finishing: materials with low *SFE* were subjected to oxidative *APPA* to increase it, Figure 3.

The difference between *SFE* of porous surface and *SFT* of impregnating solution was chosen as a finishing efficiency or *penetration-spreading parameter* (PSP) for a real porous solid such as wood: $PSP = (\sigma_S - \gamma_L)$. *Finishing* does not include evaporation of solution, dissolution or swelling of the solid by the water solution or any kind of chemical reaction between the solution and the solid substrate that changes the system composition. The higher the *SFE* of the solid, the better the finishing; the lower the *SFT*, the better the result of finishing. To sum up, large and positive difference between *SFE* and *SFT*: $\sigma_S - \gamma_L > 0$, (or $\sigma_S > \gamma_L$), will provide more effective porous surface finishing [1].

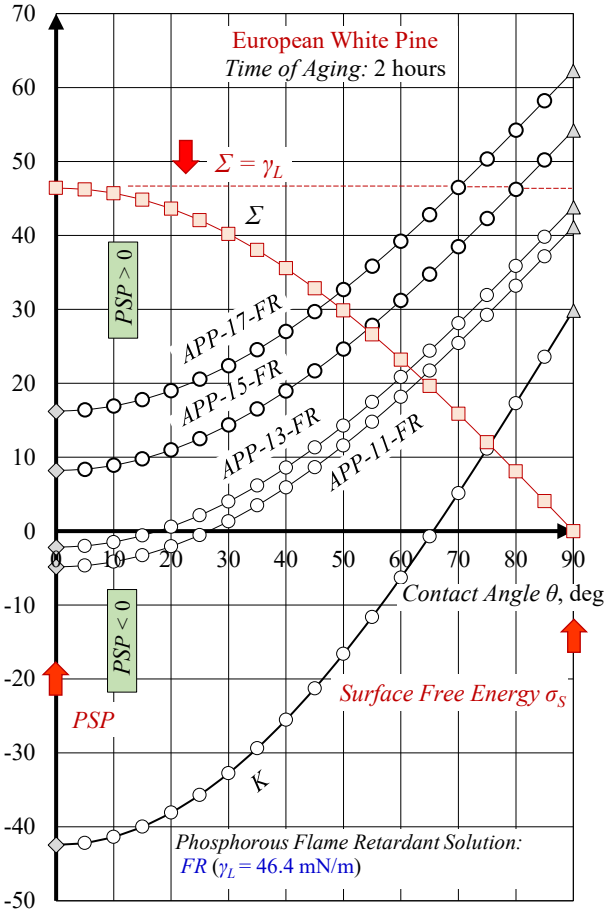


Fig. 3. Relationship between the "solid-liquid" interfacial energy σ_{SL} and the contact angle θ, measured by *Sessile drop technique*. Penetration-spreading parameter (*PSP*) is equal to the minimum *IFE* value at $cos\ \theta = 1$ ($\theta = 0°$): $PSP = \min \sigma_{SL}$. An effective finishing has a positive value of *PSP*: $PSP > 0$; a non-effective finishing has a value of less than or equal to zero: $PSP \leq 0$. Surface free energy σ_S is equal to maximum *IFE* value at $cos\ \theta = 0$ ($\theta = 90°$): $\sigma_S = \max \sigma_{SL}$.

Relationship between the adhesion tension Σ and the contact angle θ. The maximum adhesion tension Σ value is equal to the surface tension γ_L of the flame retardant water solution *FR*: $\gamma_L = 46.4$ mJ/m^2.

Atmospheric pressure plasma pre-activation at 50/60 Hz and different voltage: 11, 13, 15 and 17 kV rms: method – in situ; time of activation – 60 sec; glass barrier – 3 mm; air gap – 6 mm.

In general, the following processing efficiency parameter was found and the rule is true: "*The plasma aided or enhanced finishing of a porous media will be more successful and this media will be more susceptible to it as the difference between its SFE* σ_S *and the SFT* γ_L *or the so-called penetration-spreading parameter is positive*: $PSP = (\sigma_S - \gamma_L) > 0$. *If not, wetting, wicking and finishing problems will occur*" [1].

This new rule of processing efficiency is fully consistent with the well-known rule formulation: "*If the liquid has a dyne level (SFT) lower than the substrate dyne level (SFE),* $\gamma_L < \sigma_S$, *then the liquid will spread out over its entire surface in uniform wet layer. If the liquid dyne level is equal to or higher than a substrate dyne level,* $\gamma_L \geq \sigma_S$, *the liquid will become cohesive and tend to remain in droplets*" [18]:

$$\sigma_S > \gamma_L: \sigma_S - \gamma_L > 0 \text{ or } PSP > 0. \qquad (4)$$

According to *Young's equation*, there is a fundamental relationship between *IFE* and $PSP = (\sigma_S - \gamma_L)$:

$$\sigma_{SL} = (\sigma_S - \gamma_L) + \gamma_L (1 - \cos\theta) = PSP + \gamma_L - \gamma_L \cos\theta; \quad (5)$$

$$\min \sigma_{SL} = (\sigma_S - \gamma_L) = PSP;\ \theta = 0°: \cos\theta = 1, \quad (6)$$

or the minimum value of the "*solid-liquid*" *IFE* at $\theta = 0°$ is equal to the value of *PSP*: $\sigma_{SL} = PSP$, Figure 3, [1, 14].

Table 1. Contact angle, wetting and hydrophilicity change.

Contact angle θ, °	**0**	10	20	30	70	**90**
cos θ	1	$1 > \cos\theta > 0$				0
Interface free energy (*IFE*)	$\sigma_{SL} = PSP$					$\sigma_{SL} = \sigma_S$
Hydrophilicity / Hydrophobicity	Hydrophilicity				Hydro phobicity	
	Super-Hydrophilicity		Hydro philicity		Hydro phobicity	
Degree of wetting	*Spreading*	*Complete wetting*	*Wetting*	*Partial wetting*		*Incomplete wetting*
Wettability	*High wettability*					

The maximum value of the *IFE*, at $\theta = 90°$, is equal to the value of porous wood *SFE*: $\sigma_{SL} = \sigma_S$, Figures 3 and 4.

Therefore, the requirement for process efficiency finishing $PSP = (\sigma_S - \gamma_L) > 0$ can be transformed into a new requirement for effectiveness, Figure 3:

$$\sigma_S > \sigma_{SL} > PSP > 0, \quad (7)$$

which already refers directly to *IFE*: "*it is necessary for the entire range of variation of the static contact angle* ($0° \le \theta \le 90°$) in the area of the *high wetting*, *the value of the IFE to be greater than zero*: $\sigma_{SL} > 0$".

Plasma-aided capillary impregnation finishing will be more successful and the porous media will be more susceptible when the $PSP > 0$. This is the essence of the rule for obtaining effective plasma-enhanced capillary impregnation finishing.

Surface activation refers to the temporal increase of the *SFE* determined by the rise of the polar part of the *SFE*, Figure 4. This *APPA* process is based on the implantation of oxygen, leading to the formation of functional groups like *ether* (C-O), *hydroxyl* (C-OH), *carbonyl* (C=O), and *carboxyl* (O=C–OH). Such a plasma treatment enhances the affinity of the substrate for other substances and is especially needed for solids which typically have a low intrinsic *SFE*. This process is generally achieved using a standard *DBD* in open air and "in situ" or remote application Figure 5.

The *PS*-parameter depends only on two thermodynamic parameters of the finishing process – the *SFE* σ_S of the solid and the *SFT* γ_L of the liquid, which can be quantified individually using two well-known methods of contact angle measurement – the sessile drop and the pendant drop techniques [15, 16 and 17].

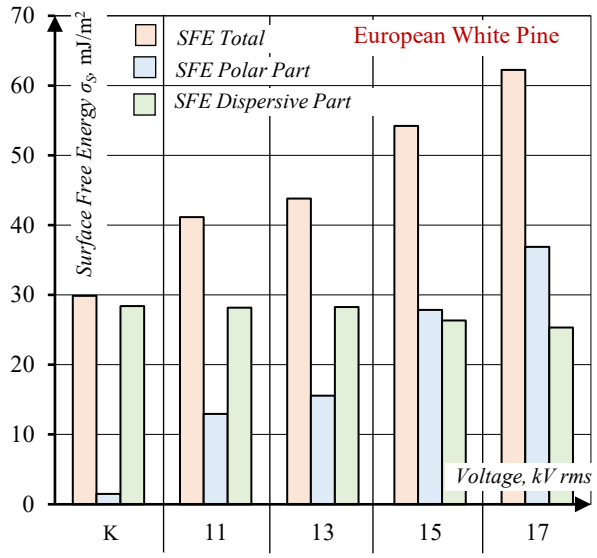


Fig. 4. Surface free energy σ_S and its dispersive and polar parts for *European white pine* after *APPA* and 2 hours of aging time and varied voltage: 11,13,15 and 17 kV rms, defined by Sessile drop technique and *theory of Wu*.

Working with the graphs and studying the relationship between the "*solid-liquid*" *IFE* σ_{SL} and *contact angle* θ in the area of high wettability, $0° \le \theta \le 90°$, it becomes clear that the graphs of the efficiency realizations of plasma-aided surface impregnation – *APP-15-FR* and *APP-17-FR* are located in the first quadrant above the abscissa, Figure 3.

One more graphic illustration has been brought about, Figure 5, depicting the efficiency parameter assessment of plasma-aided capillary impregnation with modified phosphorous *FR* solution by increasing the *SFE* and decreasing the *SFT* of the impregnating solution.

2.3. Aging, plasma activation durability and open time

It is very important to realize that the changes induced on the surface by an *APPA* are not permanent. There is aging, which is the effect of the *APPA* or *SFE* decreasing in time. Often, the activation effect will remain partially and a steady state will be reached after some hours, days or weeks. For *APPA*, the aging can be really an issue. Hence the typical *APPA* was performed "in-line", immediately before the finishing for which it is intended. Aging was usually regarded as an unwanted side-effect in plasma-aided finishing [17, 19].

The time of aging or *open time* is the period of time after an *APPA* has been applied, allowing for the finishing processes to take place – printing, bonding, painting, laminating and impregnating, during which an effective plasma-aided finishing can be achieved. Plasma pre-activation most often takes place "in-line" immediately prior to the onset of the finishing. The *time of aging* in this case is zero. The *time of aging* is different from zero when the plasma-aided capillary impregnation is applied "out-of-the-line", for example in the finishing of a wooden building construction, wood paneling, flooring or staircase. Most often, the entire surface is activated "*step by step*" for a certain time and then it is proceeded to capillary impregnation, Figure 6.

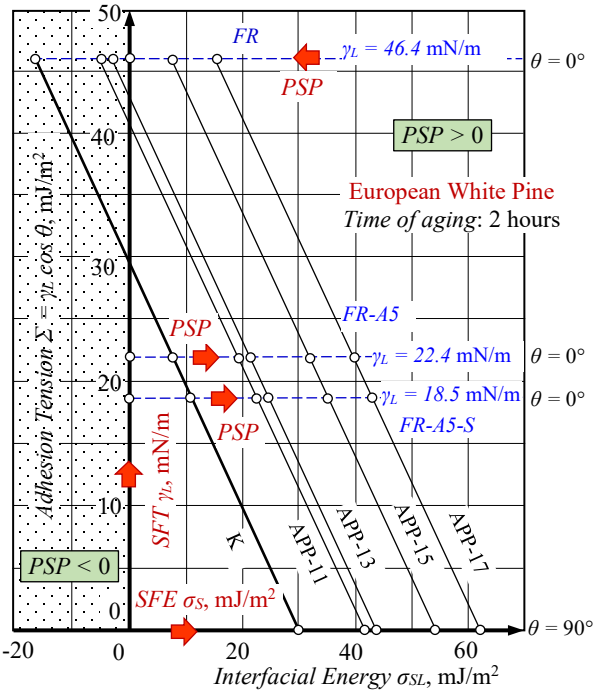


Fig. 5. Relationship between the adhesion tension Σ and the "solid-liquid" interfacial energy σ_{SL} that reveals at once the impact of: *i* – the variation of *SFE* as result of *APPA* and 2 hours of aging at varied voltage: 11, 13, 15 and 17 kV rms; *ii* – the variation of *SFT* as a result of *FR*-solution modification with surfactants: *FR-A5* – with micelle-forming anionic surfactant at 5 vol. % and *FR-A5-S* – with trisiloxane surfactant at 0.1 vol. %.

The industrialization of the plasma-aided finishing "out-of-the-line" during 2007÷09 allowed the empirical definition of two very important values for the time of aging – 2 hours or a short time of aging, allowing for successful finishing of large-scale wood wares such as paneling, flooring, wooden construction, and a day or 24 hours, a long time of aging, which is close to the *critical open time* when the effect of the plasma activation will be obliterated.

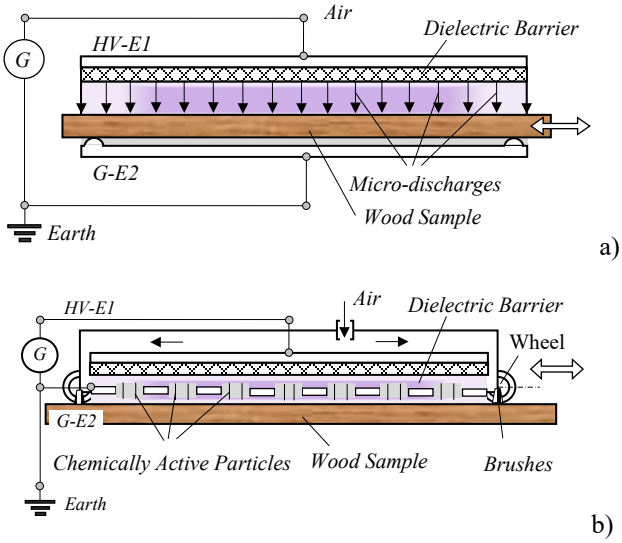


Fig. 6. Atmospheric pressure plasma pre-activation by dielectric barrier discharge in air: **a)** *APPA*, "in situ" activation where the sample is immersed in the cold plasma of the *DBD* (*V. Eisby*, 1951); **b)** *RAPPA*, remote activation where the sample interacts only with the chemically active particles produced by the *DBD* (*P. Dineff*, 2005).

HV-E1 – high-voltage electrode; *G-E2* – grounded electrode; *G* – AC electric generator (50/60 Hz).

Remember that the *SFE* level decay is extremely fast right after *APP*-activation "*in line*" – in fact an immediate loss of 10 mJ/m^2 is possible. This is due to the contact with process rolls, surface migration of additives and interfacial transfers between treated and untreated surfaces within the finished wound roll. Generally, the *SFE* of a film should be between 3 and 10 mJ/m^2 greater than the *SFT* of the ink or impregnating liquid in order to ensure acceptable performance, Figure 6, [11, 18].

The highest plasma-activation effect to the finishing, expressed by maximum *SFE* value increase, was observed with "*in situ*" *APPA*, "in line" application, Figure 6, [1, 18].

THE OBJECTIVE of this paper is to study the efficiency assessment of plasma-aided capillary impregnation as finishing for flame retardancy of wood at industrial frequency (50 Hz) and voltage of 15/18 kV rms, after short and long aging time, 2 and 24 hours, and the change of the wood impregnability monitored by the penetration-spreading parameter (PSP) and the processing efficiency rule: *PSP* > 0.

3. Experimental investigation

The development of a plasma-enhanced capillary impregnation finishing for wood flame retardancy goes through three stages:

- *The first stage* includes the selection of an effective *APPA* mode that provides the necessary high level of *SFE* after a short aging of 2 hours, Figures 3 and 9;
- *The second stage* includes the selection of an anionic micelle-forming water surfactant in a concentration to ensure the required low level of the *SFT*: 30 mN/m > γ_L > 20 mN/m, Figure 10;
- *The third stage* includes the selection of an effective *APPA* mode that will provide the necessary high level of *SFE* after a long time of aging, a day or 24 hours, and low level of *SFT,* by some adequate surfactants and concentrations, Figure 11.

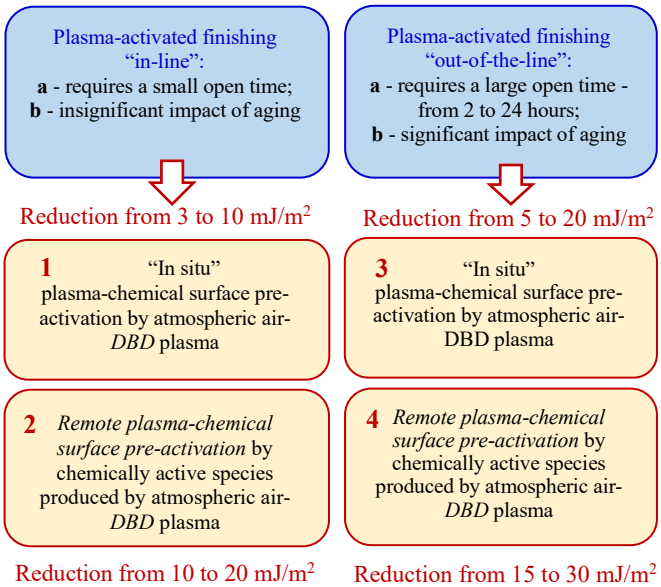


Fig. 7. Surface free energy reduction before finishing due to: *i* - contact with other non-activated surfaces, *ii* – aging after plasma activation, and *iii* – different modes of plasma surface activation provide different activation options: *APPA*, "in situ" activation immerced in cold plasma or *RAPPA*, remote activation with chemically active particles.

This experimental study was carried out on four kinds of wood samples: *European white pine* (*Pinus Sylvestris L.*) from Bulgaria and three woods from Mexican rain forest: *Tzalam* (*Lysiloma Bahamensis L.*), *Mexican white cedar* (*Cupressus lusitaca L.*), *Mahogany caoba* (*Swietenia mahogany L.*), according to the well-known method [1, 15-17]. The selected wood species form the following range of decreasing density, in kg/m^3: *Tzalam* – 780; *Mahogany caoba* – 650; *Mexican white cedar* – 470; *European white pine* – 371. There are two hard wood species with high density, $d > 540$ kg/m^3: *Tzalam* and *Mahogany caoba*, and two soft wood species with low density, $d < 540$ kg/m^3: *Mexican white cedar* and *European white pine*.

On the basis of our own former experience in plasma-aided capillary impregnation of wood and wooden materials, [9-12], an oxidative *APPA* has been applied on the test samples for 60 sec. This study employs a well-known experimental methodology, already used by us [1, 14-17].

All experimental studies pertain to *APPA* of the wood specimens in a specific burning mode of the air *DBD* at atmospheric pressure (~ 101 kPa) at the first peak of surface density of real power p_a = 26 W/m^2 and the second peak p_a = 8 W/m^2 in asymmetric rectangular coplanar electrode system with 6 mm air gap – mode *A* and *B* respectively, Figure 7.

The *APPA* was carried out with "*in situ*" application of *DBD* for 60 sec, Figure 5a. The effective operating modes of *DBD* at voltages above 13 kV rms (18.3 kV peak value) determine a surface density of real power above 4 W/m^2, Figure 7.

The aim of this study was to verify the possibility to evaluate the *PS*- efficiency parameter values for plasma and surfactant enhanced capillary impregnation trough calculation of *SFE* and *SFT* by drop shape analysis. *Drop shape analysis* is an image analysis method for determining the contact angle from the shadow image of a sessile drop and the *SFT* from the shadow image of a pendant drop. There are number of theories and models for *SFE* calculation with the help of contact angle data. The method according to *Wu* for calculating *SFE* and its two parts – polar and disperse, is suitable for high polar or plasma activated surfaces, Figure 4.

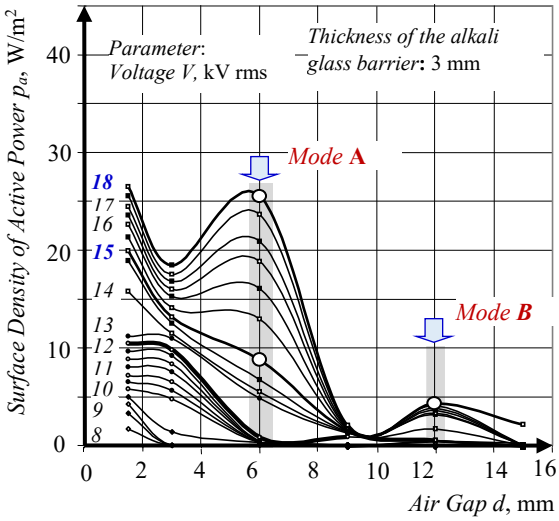


Fig. 8. Technological discharge characteristic "surface density of active power p_a – air gap size *d*" of non-equilibrium *DBD* burning at industrial frequency (50 Hz) in asymmetric rectangular coplanar electrode system with one glass barrier.

All methods described are included in the KRÜSS Drop shape analysis programs. *Drop Shape Analyzer* (DSA100 and DSA30, KRÜSS GmbH, Germany) was used for the precise measurement of contact angle *SFE* and also for measurement of the *SFT* of liquids using the pendant drop method. Measuring range (referred to image analysis) for *Sessile drop technique*: contact angle – 1°÷180°; surface free energy – 0.01÷1 000 mJ/m^2; and for *Pendant drop technique*: surface tension – 0.01÷2000 mN/m. Measurement accuracy: contact angle – 0.3°; surface tension – 0.3 mN/m [7].

Modes of *APPA* for wood samples: *mode A* – cathode-directed streamer mode at air gap of 6 mm; *mode B* – anode-directed streamer mode at air gap of 12 mm, Figure 8.

A halogen-free, phosphorous and nitrogen containing flame retardant (FR) based on ortho-phosphorous acid, urea and ammonia has been produced and studied as water solution. The impregnating *FR* water solution was based on it: dry substance of 30 wt. %; phosphorus content of about 13 wt. %, pH = 7÷8 and density of 1.15÷1.14 g/cm^3.

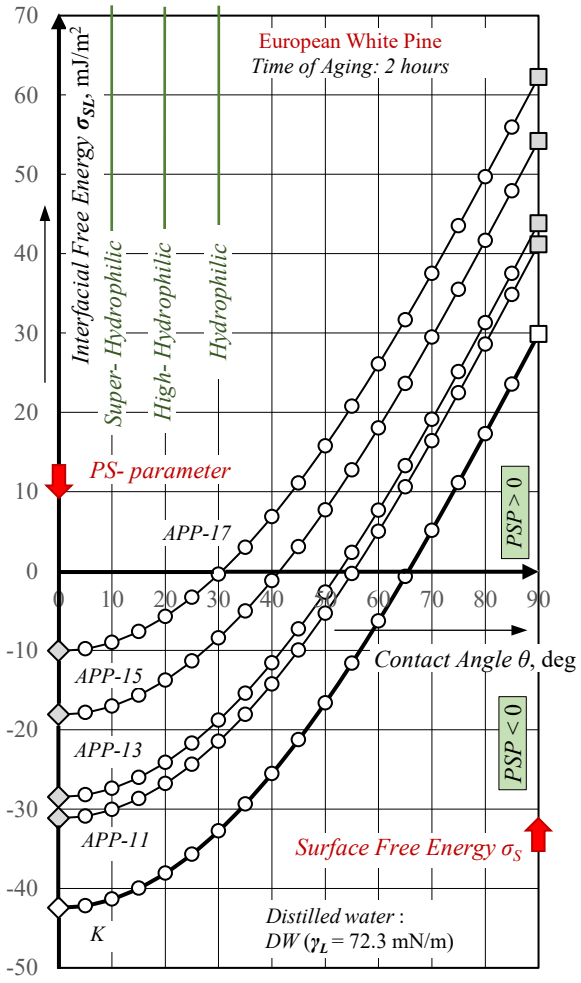


Fig. 9. The use of distilled water as a test-liquid with $SFT - \gamma_L = 72.3$ mN/m, does not carry adequate information on the efficiency of the finishing of *FR* capillary impregnation – no effective mode of *APPA* has been found at voltages up to 17 kV rms for *European white pine*: $PSP = \sigma_S - \gamma_L < 0$. All *SFE* values were determined according to the *method of Wu* for short aging time of two hours.

The non-equilibrium air (oxidative) *APPA* was combined with anionic micelle-forming (AS, "Anticrystalin A", *Chimatech, Ltd.*,

Bulgaria) and trisiloxane-ethoxylate surfactant (Y-17113, *Momentive Performance Materials* GmbH & Co. KG, Germany) enhanced capillary impregnation with well-known experimental methodology [10, 11, 12].

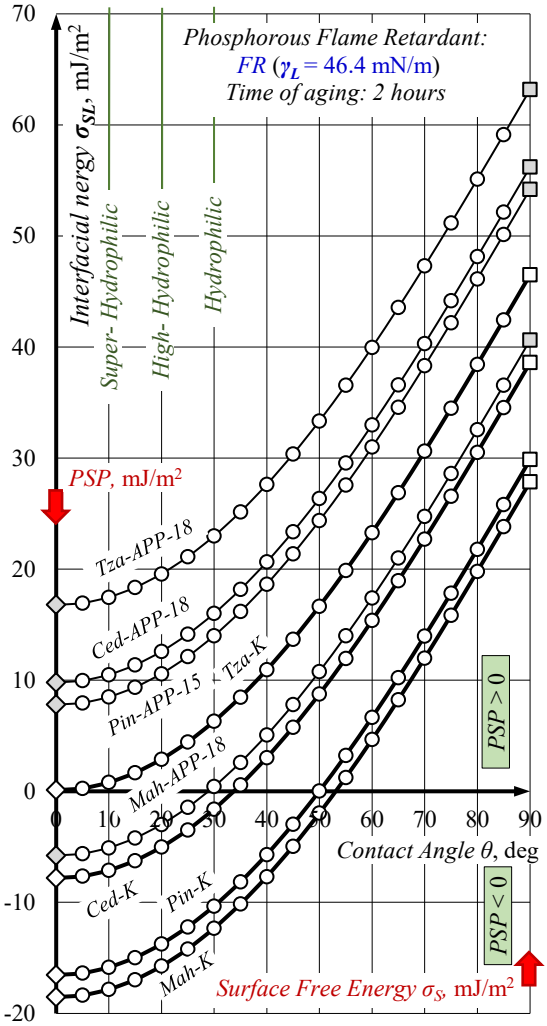


Fig. 10. Effective mode selection of *APPA* at an air gap of 6 mm and short aging time of two hours for plasma-aided capillary impregnation finishing was found only for *European white pine* (*Pin-LF-15*), *Mexican white cedar* (*Ced-LF-18*), and *Tzalam* (*Tza-LF-18*): *PSP* > 0. No steady *APPA* mode for *Mahogany caoba* (*Mah-LF-18*) was found: *PSP* = 0.1 mJ/m² ≈ 0. Plasma pre-activation modes provided trought the declared voltage – 15/18 kV rms, get the maximum *SFE* value for each species of studied woods.

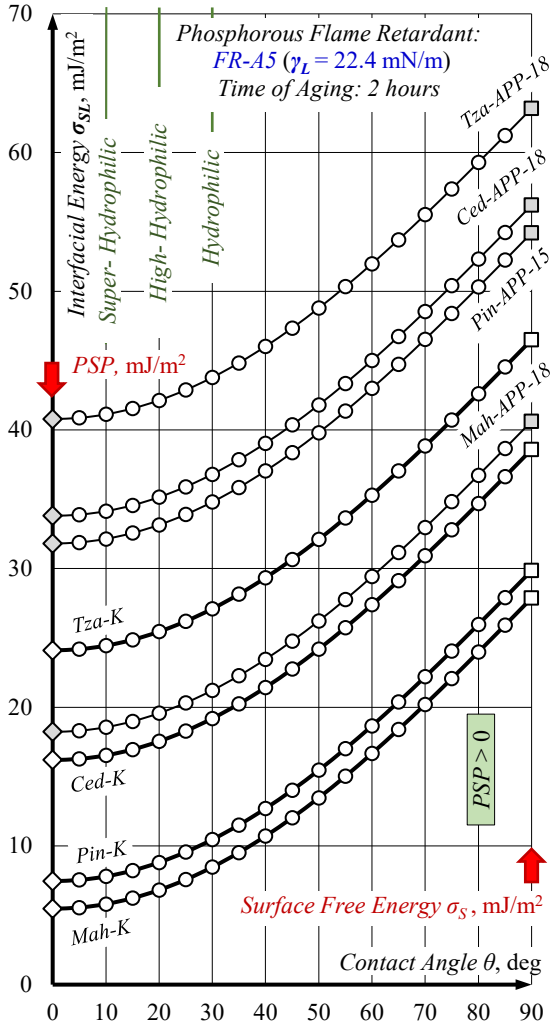


Fig. 11. Effective mode selection of *APPA* at an air gap of 6 mm and short aging time of two hours for plasma-aided capillary impregnation finishing was found for all wood species by decreasing the *SFT* of the basic phosphoric *FR* solution with *anionic micelle-forming water surfactant* at 5 vol. % (*FR-A5*): γ_L = 22.4 mN/m. The rule of processing efficiency was met for all applications of capillary impregnation – non-activated (*-K*) and *APP*-activated (*-APP-XX*) at different voltages: 11, 13, 15 and 17 kV rms: *PSP* > 0.

4. Experimental rezults and discusion

This study meets another issue which can be formulated as follows: "*is it possible a simple drop test with distilled water* (DW) *to adequately replace the entire PSP study conducted with a real impregnating FR-water solution*".

The answer can be found in the information provided in Figure 3 and 8. The use of distilled water as a test-liquid does not carry adequate efficiency assessment of the finishing – the plasma-aided capillary impregnation. The effective mode selection (*APP-XX*: *APPA* at 11, 13, 15, and 17 kV rms; 50/60 Hz) can be determined only by the use of *FR* solution – a basic *FR* water solution containing phosphor- and nitrogen flame retardant with *SFT*, γ_L = 46.4 < 72.3 mN/m, and the relationship between the "*solid-liquid*" *IFE* (σ_{SL}) and static contact angle θ measured by Sessile drop technique – *IFE* acquires a value equal to the *PS-Parameter* at *complete wetting*, i.e. contact angle equal to zero: $\sigma_{SL} = PSP$ ($\theta = 0$).

Distilled water has a too high *SFT*: γ_L = 72.3 > 46.4 >> 22.4 mJ/m². *APPA* cannot compensate it by increasing *SFE*, so a rule process efficiency should be enforced: $\sigma_S < \gamma_L$; *PSP* < 0, Figure 8. Normally, it fails to predict the efficiency of the plasma-aided finishing with modified γ_L = 22.4 mJ/m², or non-modified γ_L = 46.4 mJ/m², *FR*-solution, Figures 3 and 8.

All modes of *APPA* with a voltage higher than 13 kV rms (18.3 kV peak value) provide processing efficiency finishing of capillary impregnation: *PSP* > 0, Figure 3.

At the first stage, the *PSP*- testing indicates that the voltage of the *APPA* for European white pine samples should be 15 kV rms (σ_S = 54.2 kJ/m²) for a maximum processing efficiency of the *FR* (γ_L = 46.4 mJ/m²) impregnation finishing: max *PSP* = 7.8 mJ/m²

> 0, Figure 9.

The most effective voltage for the other studied species of wood was determined using the same procedure and it is 18 kV rms. They are arranged in order of decreasing *PS-parameter* according to their density, mJ/m^2: *Tzalam* – 16.78, *Mexican white cedar* – 9.8 and *European white pine* – 7.8. However, no steady *APPA* mode for *Mahogany caoba* was found: *PSP* = 0.1 mJ/m^2 ≈ 0. This result indicates that there are wood species for which a processing efficiency finishing – a capillary impregnation, cannot be applied, Figure 9.

At the second stage, the *PSP*- testing of the all studied wood species indicates that the modified *FR*- solution with less *SFT* γ_L = 22.4 mJ/m^2, performs the rule of processing efficiency after a short time of aging, *PSP* > 0, Figure 10.

At the third stage, the *PSP*- testing of all studied wood samples indicates that the *APPA* and the finishing of capillary impregnation with the modified impregnating *FR* solution (FR-A5), γ_L = 22.4 mJ/m^2, satisfy the rule of finishing processing efficiency: *PSP* > 0, Figure 11.

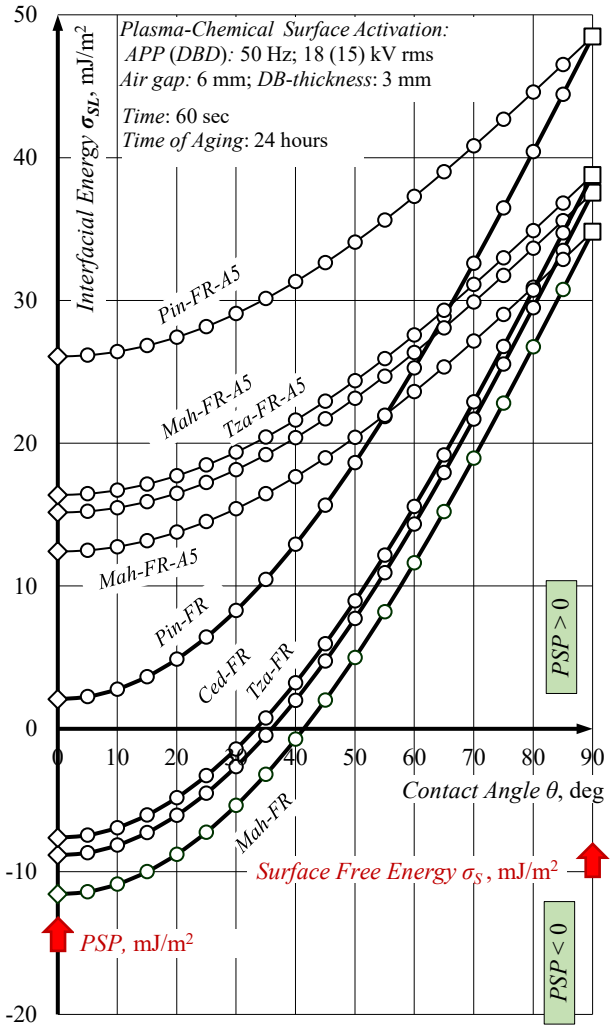


Fig. 12. Effective mode of *APPA* at air gap of 6 mm and short and long aging time - 2 and 24 hours, for plasma-aided capillary impregnation finishing was found for all studed wood samples by decreasing the *SFT* of the *FR* impregnating solution with anionic micelle-forming surfactant at 5 vol. % (*-FR-A5*): γ_L = 22.4 mN/m. The rule of processing efficiency was met for all applications of plasma-aided finishing with this modified solution: *PSP* > 0.

This study proves our claim that the plasma-aided capillary impregnation can be applied "out-of-the-process-line" after a long time of aging 24 hours, between *APPA* and the capillary impregnation finishing.

The plasma-aided capillary impregnation of porous woods has resulted in the formation of a *FR*-functional coating with a certain thickness, continuity and defect-free surface. In porous media, the *capillary pressure* p_C is proportional to the *adhesion tension* Σ and depends only on it at certain characteristic pore size, (3).

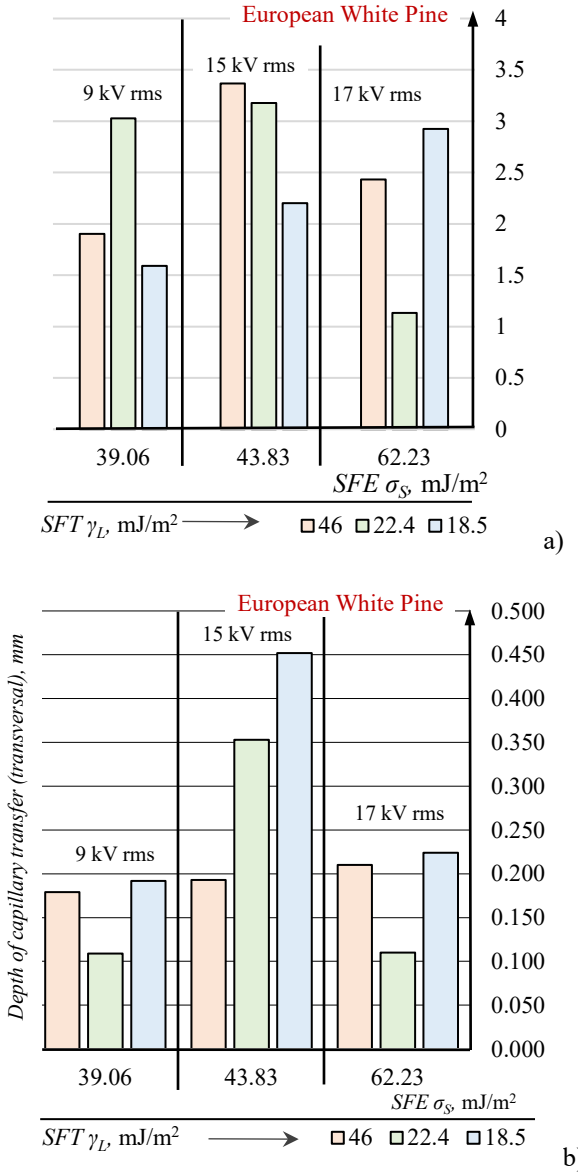


Fig. 13. Effect of *APPA mode* (voltage) – 9, 15 and 17 kV rms (50/60 Hz; air gap: 6 mm), the short time of aging – 2 hours, the used impregnating phosphorous *FR* solutions and consumption rate – constant rate: 0.139 l/m^2 (dm^3/m^2) on the depth of "capillary" transfer to the wood capillary direction: *longitudinal* – capillary displacement (**a**) and *transversal* – wicking through the capillary walls (**b**). The best quality of the plasma-aided capillary impregnation for flame retardancy of *European white pine* is at 15 kV rms.

It was expected that the largest *DBD* voltage of 17 kV rms should ensure greater depth of wicking transfer as the *SFE* and its greatest polar part, which determines greater value of *PS*-efficiency parameter. Although the *SFE* and its polar part are greater, Figure 4, just like the *PS*-parameter, Figure 3, the depth of

capillary transfer – longitudinal to the capillary structure and transversal to it, or through the capillary walls, is less than 15 kV rms, Figure 12. What is typical for this mode (15 kV rms) of plasma and surfactant enhanced capillary impregnation, is that the polar part of the *SFE* becomes greater than the dispersive part of *SFE*, Figure 4. The transition to more intensive mode of *APPA* (17 kV rms) does not give a positive result in spite of increasing *SFE* and its polar part, and *PS*-efficiency parameter.

Both the capillary displacement and the wicking through the capillary walls are two visible forms of maximum solution transfer, Figure 12. The maximum *capillary displacement* was determined by a maximum value of capillary pressure p_C or adhesion tension Σ respectively. This is determined by the maximum basic *FR* solution, Figure 5 and 13. On the contrary, the maximum *wicking through the capillary walls* was obtained at the minimum value of adhesion tension Σ; this is provided by the modified *FR-A5-S* solution. Wicking, however, is essential for the impregnation process. More extensive studies on different species of wood are needed to confirm this experimental fact, which has been observed only for European white pines so far.

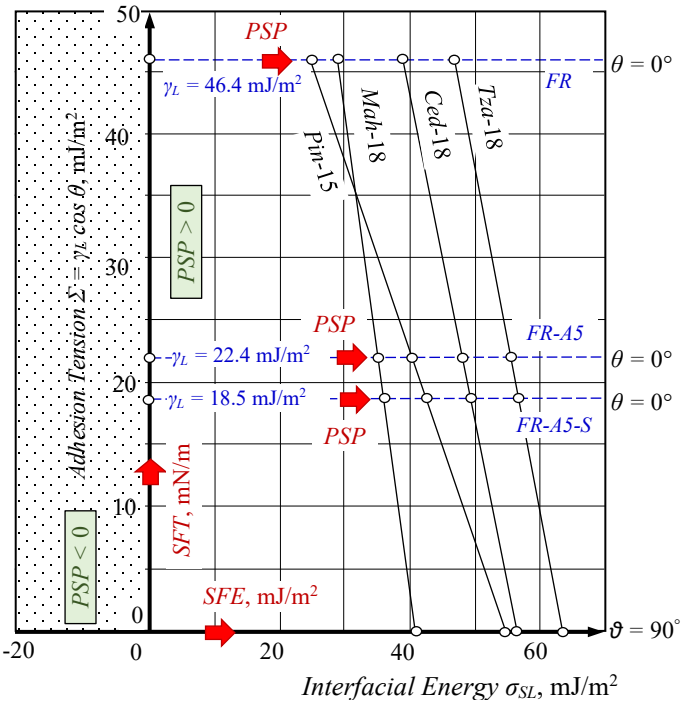


Fig. 14. Relationship between the *adhesion tension* Σ and the "*solid-liquid*" *interfacial energy* σ_{SL} that reveals at once the impact of: *i* – the variation of *SFE* as result of *APPA* at short time of aging (2 hours), air gap of 6 mm and voltage 15 kV rms; *ii* – the variation of *SFT* as a result of *FR*-solution modification with surfactants: *FR-A5* – anionic micelle-forming surfactant at 5 vol. %; *FR-A5-S* – trisiloxane neutral surfactant at 0.1 vol. %.

In conclusion, an attention has to be paid to the fact that the transfer of the *FR*-solution in depth is mainly carried out in the more complicated way through the capillary walls. The modification of the basic *FR*-solution through the combination of anionic micelle-forming and trisiloxane surfactant thus acquires a new meaning.

The graphical representation of the relationship between *adhesion tension* Σ and "*solid-liquid*" *IFE* σ_{SL} fully illustrates the two pathways for achieving the finishing efficiency *PSP* of plasma-aided capillary impregnation by increasing the *SFE* of porous media and reducing the *SFT* of the impregnating *FR* solution, Figures 5 and 13.

By selecting the voltage, the quality of the *APPA* expressed by the process efficiency parameter *PSP* was controlled: the *SFE* of the European white pine samples was increased; the *SFT* of the impregnating *FR* solution was decreased. The process efficiency parameter *PSP* meets the rule of plasma-aided finishing efficiency, *PSP* > 0, at and above 15 kV rms, Figure 5. For all wood species, the maximum process efficiency parameter *PSP* is at a voltage of 18 kV rms, Figure 13.

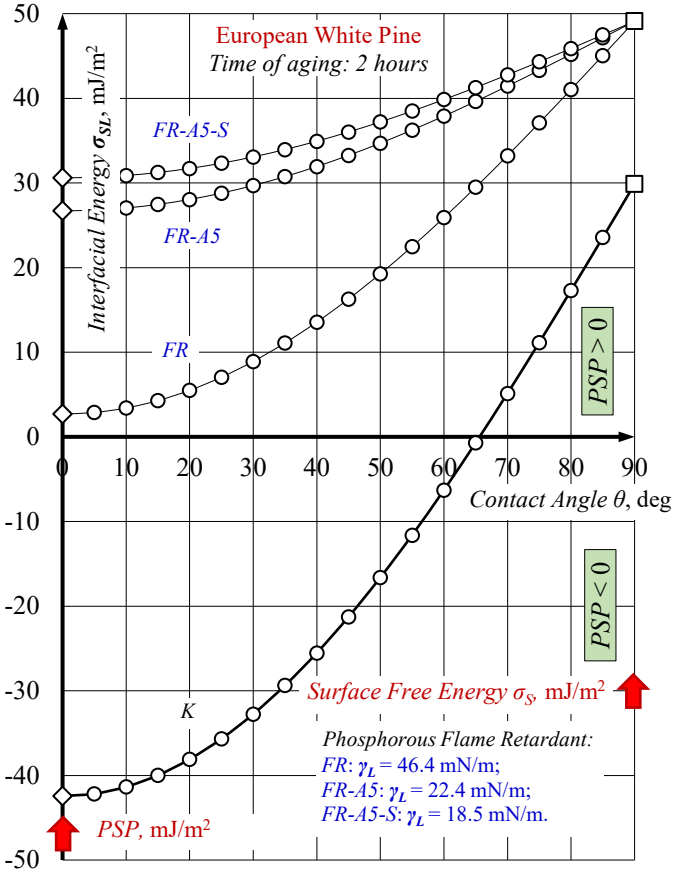


Fig. 15. Effective mode selection of plasma-aided capillary finishing of *European white pine* after two hours aging time from air *RAPPA* at air gap of 6 mm by a modification of the phosphoric *FR*-solution with an *anionic micelle-forming water surfactant* at 5 vol. % (*FR-A5*) and neutral trisiloxane surfactant at 0.1 vol. %. The rule of processing efficiency was met for all plasma-aided applications of capillary impregnation.

The new created plasma-aided finishing can find its wide application only if it can be envisaged and managed so as to observe the process efficiency rule: *PSP* > 0, Figure 15.

The prerequisite for this is the possibility for *SFE* measurement "*in situ*" out of the lab. We believe that this approach can be successful, as the market now has a technical device for measuring the *SFE* "in situ" – pocket goniometers, mobile or hand-held surface and contact angle analyzers.

There is also an innovative "mobile" analyzer with two test liquids using a dispenser with two parallel drops with "one click" for direct analysis of the static contact angles and the results obtained from the *SFE*.

Conclusion

Plasma, surfactant or plasma and surfactant aided finishing were used to enhance the finishing of capillary impregnation and thus the flame retardancy of wood: wood materials and constructions – timber, lumber, plywood, cardboard, paper, flooring, paneling, staircase and wooden structure with low *SFE* were subjected to *APPA* and *RAPPA* for increasing their *SFE*.

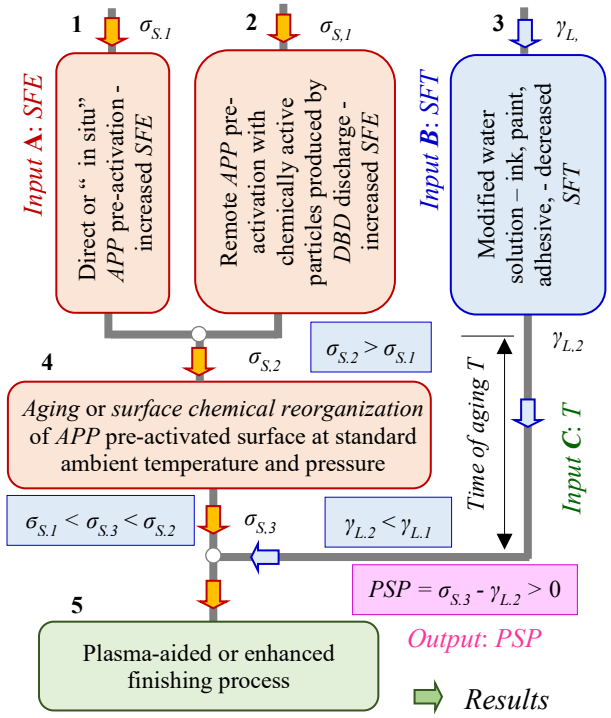


Fig. 16. Schematic representation of the plasma-aided impregnation finishing model – flow of information and processes involved in the processing efficiency and results to be obtained. Management of the plasma-aided finishing with two input factors – *SFE* of the porous solid and *SFT* of the liquid. The time of aging is set depending on the *APPA* and impregnation used technology.

In general, a processing efficiency parameter named *PSP* was found: "*Plasma, surfactant or plasma and surfactant enhanced finishing of capillary impregnation will be more successful and the material will be more susceptible when the PSP is positive: PS* > 0".

This process efficiency parameter was defined by the difference between the *SFE* of the porous wood and the *SFT* of the impregnation *FR*- solution: $PS = \sigma_S - \gamma_L$. This efficiency parameter is equal to the value that gets the *IFE* when the water solution wets the surface completely: $PS = \sigma_{SL}$, $\theta = 0°$ ($\cos\theta = 1$).

The efficiency of a plasma-aided impregnation finishing can be predicted only by setting both basic parameters of the real porous wood surface and the actual *FR*-impregnating solution – the *SFE* and the *SFT*. Successful plasma aided or enhanced finishing, including the capillary impregnation, can be expected if the process efficiency parameter takes a positive value, *PSP* > 0, after a prescribed aging time – 2 or 24 hours.

The evolution of plasma-enhanced wicking phenomena in time after *APPA* – the short and long-time aging of plasma activated surface, after 2 and 24 hours, was investigated for four species of wood by the *PS*-efficiency parameter change. All species of wood were ranked according to the plasma aided capillary impregnation process efficiency as per its *PSP*.

The existing methods and instruments for measuring the *SFE* of solids and *SFT* of liquids, including portable measurement systems, [20], allow this approach to determine the process efficiency of the plasma-enhanced finishing such as impregnation, printing, painting, laminating to be recommended for application "in-line" and "out-of-the-line".

References

[1] Dineff, P.; Gospodinova, D.; Ivanov, I. Processing efficiency of plasma-aided porous media finishing. Prossidings of the 15-th international conference on electrical machines, drives and power systems (ELMA), IEEE 2017, IEEE Part Number CFP17L07-ART, pp 334-338.

[2] Fridman, A. *Plasma chemistry*. Cambridge University Press, Cambridge, England, New York; 2008.

[3] *Plasma technologies for textiles*. Edited by Shishoo, R. Published by Woodhead Publishing Limited, Abington, Cambridge, England, 2007.

[4] *Plasma Technology for Hyperfunctional Surfaces. Food, Biomedical and Textile Application*, Edited by Rauscher, H.; Perucca, M.; Buyle, G. WILEY-VCH Verlag GmbH & Co., Weinheim Germany, 2010.

[5] Sernek, M. Comparative analysis of inactivated wood surfaces. Doctoral thesis, 2002. Glasser, W.; Kamke, F. (Co-Chairs). Virginia Polytechnic Institute and State University. Blacksburg, Virginia, USA.

[6] Viöl, W. Method for modifying wooden surfaces by electrical discharges at atmospheric pressure. WO039944A1 (EP1233854B1; US6818102). Priority Date: 01.12.1999; Publication Date: 16.11. 2004.

[7] Cernak, M.; Rahel, J. Apparatus and method for treatment of wood, wood fibres and wood-based materials. WO Patent WO2008085139A1 (EP2095694A1), Priority date: 05.12.2006; Publication date: 17.07.2008.

[8] Dineff, P.; Kostova, L. Method for plasma chemical surface modification. WO Patent WO2006133524A2 (BG 66022B1); Priority Date: 14.06.2005; Publication date: 12.04.2007.

[9] Podgorski, L.; Chevet, B.; Onic, L.; Merlin, A. Modification of wood wettability by plasma and corona treatments. *Internat. J. Adhes. Adhes* 2000, 20(2), pp 103÷111. Doi: 10.1016/S0143-7496(99)00043-3.

[10] Avramidis, G.; Wascher, R.; Militz, H.; Viöl, W. Impact of air-plasma treatment at atmospheric pressure on wood and wood extractives. Int. Wood P. 2016. 7(2) pp 76-79. 8-th European Conference on Wood Modification. Doi: 10.1080/20426445.2016.1162429.

[11] Nolan, M. Dyne testing –why test in the first place? Converting Supply, Inc., Ontario, Canada, 2011.

[12] Why test inks cannot tell the full truth about surface free energy. Application report No 272, Krüss GmbH, Hamburg Germany, 2014. www.kruss.de.

[13] Wolkenhauer, A.; Avramidis, G.; Militz, H.; Viöl, W. Plasma treatment of heat treated beech wood - investigation on surface free energy. Holzforschung 2008. 62(4), pp 472-474. Doi: 10.1515/HF.2008.074.

[14] Dineff, P.; Gospodinova, D.; Ivanov, I. Study on processing efficiency of plasma-aided capillary impregnation of porous media through a penetration-spreading parameter. Porceedings of the X-th international conference advanced manufacturing technologies & XI-th international conference advanced materials and operations (AMT&AMO 2016). 2016, pp 54-65.

[15] Dineff, P.; Ivanov, I.; Gospodinova, D. Efficiency processing parameter for plasma-aided surface impregnation: I. Penetration-spreading parameter. 19-th international symposium on electrical apparatus (SIELA) 2016. IEEE DOI: 10.1109/SIELA.2016. 7542991.

[16] Dineff, P.; Ivanov, I.; D. Gospodinova, "Efficiency Processing Parameter for Plasma-Aided Surface Impregnation: II. Interfacial Energy, Adhesion and Capillary Pressure. 19-th international symposium on electrical apparatus (SIELA) 2016; IEEE DOI: 10.1109/SIELA.2016.7542992.

[17] Dineff, P.; Ivanov, I.; Gospodinova, D.; Stanimirova, R. Efficiency Processing Parameter for Plasma-Aided Surface Impregnation: III. Plasma Surface Activation. 19-th international symposium on electrical apparatus (SIELA) 2016; IEEE DOI: 10.1109/SIELA.2016.7542993.

[18] Corona and plasma for extrusion. Vetaphone A/S, Denmark, 2017. www.vetaphone.com.

[19] Wolkenhauer, A., Avramidis, G.; Hauswald, E.; Militz, H., Viöl, W. Sanding vs. plasma treatment of aged wood: a comparison with respect to surface energy. Intern. J. Adhes. Adhes. 2009. 29(1), pp 18-22. Doi: 10.1016/j.ijadhadh.2007.11.001.

[20] Mobile surface anlyzer – MSA. Krüss GmbH, 2017. www.kruss.de.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 244-251 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Measuring modifiability in model driven development using object oriented metrics

Nwe Nwe[*,1], Ei Thu[2]

[1]*Computer University, Monywa, Myanmar*

[2]*University of Computer Studies Mandalay, Myanmar*

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 02 January, 2018
Online: 30 January, 2018

Keywords :
Model transformation
Drools rule-based
Modifiability

ABSTRACT

Model driven development is an important role in software engineering. It consists of multiple transformation functions. This development is a paradigm for writing and implementing computer program quickly, effectively, at minimum cost and reducing development efforts because it transforms design model to object-oriented code. Our approach is rule-based model driven development in which textual Umple model is used as primary artifact and transformed to mobile applications. In this model driven development, evaluation of quality of transformation is critical. This paper has presented a set of metrics to assess the quality attribute of modifiability and evaluated using these object-oriented metrics. Results represent our approach achieves high efficiency in quality of modifiability.

1. Introduction

Changing with technology, mobile devices and mobile applications are necessary thing for every person and every sector. The burst on the availability of mobile devices is powering a growing mobile application. According to this fact, mobile applications, which are used in these devices, are critical in an industrial development. To fulfill the demand of these things, mobile application development preferably uses model-driven engineering than traditional software development process widely and effectively. It focuses on model for the development of software. There is lower the overall cost of building large internal applications, there is lower the risk of large application, speed time to build large applications and expand the pool of resources that can work on large application are main strategic objectives to use model driven development. Reduction of both direct and indirect development efforts, which enables scripters to contribute to enterprise development and enables task-oriented management of development are benefit of model driven development. It is a superset of model driven development because it goes beyond the traditional development.

In this case, the rule based model driven development of mobile application using Drool Knowledge-based Rule was presented in [1]. They also measured assessment of transformability using object oriented metrics. Drools Knowledge-based is a business rule management system with a forward and backward chaining inference based rules engines [2].

Moreover, this model-driven development is a development paradigm that uses model as the primary artifacts of the development process. It transforms source model to target model/code according to changing requirement and software reused more rapidly than traditional software development. A model transformation consists of multiple transformation functions. These transformation functions transform target language elements from the source language. Most of researchers concentrate on model to model transformation using intermediate meta-model or model to code transformation [3-6]. There is no quality, there is no efficiency in everything. Quality issue also change scale and become more important. The process transforms to new model or code related to quality of final software product and the quality of the model used to generate it. The consistency of source and target model and the assessment of quality of transformation is the critical issue in model transformation domains.

There are many attributes to evaluate quality in software engineering. Among them, most of these quality attributes can be applied to software artifacts in general. However, in [7] authors describe two quality criteria important in model transformation. They are transformability and modifiability. According to their

[*]Corresponding Author: Nwe Nwe, Email: nwenwemdy08@gmail.com

work, it is necessary to extend the assessment to other quality attribute of modifiability, maintainability and reusability using object-oriented metrics. Modifiability is the extent to which a model transformation can be adapted to provide different or additional functionality. The main reason for modifying a model transformation is changing requirements. Another reason is that the (domain specific) language in which the source and/ or target model are described which may be subject to changes. Modifiability captures the amount of effort needed to modify a model transformation. It is the combination of the modularity and reusability, an essential aspect of software engineering that promotes software maintainability. Moreover, it enables transform without affecting other parts of a program that are not directly connected to the changes. This paper is organized as follows: Section II explains the basic concepts of related work. Section III presents the contribution of rule based model transformation and quality attributes of modifiability. Section IV describes our experimental results and comparison results. Finally, section V concludes our approach and evaluation of modifiability using object oriented metrics.

2. Related work

There is increasing attention towards the generation of source code from modeling languages. Several researchers propose model driven approach for the different aspects of mobile applications. The model driven development of mobile applications using Drools knowledge-based Rule was describes in [1]. They developed mobile application by applying Drool rule based. Their work is closely related with JUSE4 android application [8,9]. Moreover, they attempted to address the consistency of source and target model and the assessment of transformability by measuring the accuracy of consistency between source and target model and assessing the transformability using object oriented metrics. According to their work, it is necessary to extend the assessment to other quality attributes of model driven development using object-oriented metrics. Authors performed a comparative study on C++, C# and Java programs using object-oriented metrics in [10]. It consists of class size, complexity, coupling cohersion, inheritance, encapsulation, polymorphism and reusability.

An evaluation of the quality of model transformation was defined in [11]. They made the quality of model transformation measurable. They presented the quality attributes and a set of metrics to assess these quality attributes. A calculation of metrics values using the same set of standard metrics for three software system of different sizes was described in [12].

In [5], authors presented quality goal in MDE and states that the quality of models is affected by the quality of modeling languages, tools, modeling processes, the knowledge and experience of modelers and the quality assurance techniques applied. In [3, 4], authors defined the meta-model and model transformation rule for model driven android application development. In [13], authors also defined ATLAS transformation rules for UML sequence diagram to generate enterprise java bean code (EJB). In [6], authors presented enhance code generation tool for android source code based on UML class and sequence diagram. In [14], authors specified meta-model with Ecore and transformation rules with Xpand templates for entity relationship diagram to generate android SQLite database model.

In model driven transformation, the approaches are quite different in their respective use of input model. Most of these approaches are based on graphical modeling or textual modeling languages. In contrast to our approach, the previous approaches applied pre-defined meta-modeling while our approach automatically parses and extracts syntax form input model of Umple [15]. Moreover, our approach has specified transformation rules in object pattern matching approach. JUSE4Android is also based on textual modeling languages. Unlike our approach, it is adding annotation into JUSE model and transform into android source according to the predefined meaning of annotation. Therefore, they generated source code contains some more files in their project. The authors [16, 17] have proposed the approach for empirical evaluation of model driven engineering in multiple dimensions. Their case studies include qualitative (expert judgements) and quantitative data (metrics) evaluations. They suppose that the productivity and defect detection rate are the popular metrics for measuring automation degree of MDD processes. Some quality goals such as well-establishment and precision are especially important in MDE [18 -20]. In [21], authors also developed open source tool aims to address quality measurement and prediction process to achieve automatically. In [22], authors presented the most recent challenges faced in the process to make model transformation more sophisticated. According to the literature, it is necessary to extend the assessment to other quality attribute of modifiability using object-oriented metrics. We conducted the comparative study for measuring the modifiability of MDD generated source code using object-oriented metrics. Therefore, we obtain more reliable findings.

3. Rule based model transformation framework

Model transformations become essential with the evolution of model driven development. It is a mechanism of automating the manipulation of models. A transformation is the automatic generation of a target model from source model using transformation definition. These transformations definition is a set of transformation rules that define how a model in the source language can transform into target language. These rules are descriptions of how one or more constructs in the source language can be transformed into one or more constructs in the target language.

In this section, we provide an overview of the framework and their underlying architecture. The proposed architecture is divided into three major parts corresponding to the main capabilities of the proposed framework. These components are parser, transformer and code generator. The parser receives an input model, written in Umple language, tokenizes it and passes it to the next component transformer. The transformer is a knowledge

based rule engines. It has received the tokens previously obtained and transforms them into internal representation consistent with target source code model using predefined set of Drools mapping rules. It is more correctly classified as a production rule system. It is a kind of Rule engine and also an Expert system, the validation and expression evaluation Rule Engines. It is declarative programming and allows to present what to do. The key advantage of rule engine is that using rules can make it easy to express solution to difficult problems and consequently have those solutions verified. The code generator translates the internal representation into target artifacts: source code as Java, XML and android activity class. Each component is tested independently to ensure that the input is processed correctly and the resulted output is validated [1]. Figure 1 describes the overall architecture of the proposed system. We have used Umple as input and transformed to mobile application. In this case of model transformation, we have applied Drool Rule based transformation.

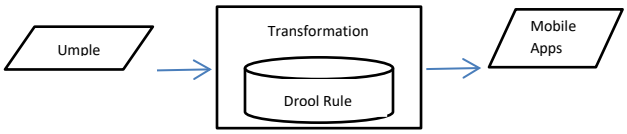


Figure 1 Overall Architecture of model transformation

3.1. Rule-based Inference System

In this system, we present a rule-based model driven approach to generate android application from text-based modeling language. Rule languages and inference engines incorporate reasoning capabilities are used in mobile application development system. A rule is made up of a collection of conditions associated with a sequence of actions to be applied to each collection of facts matching the rule condition. The proposed model transformation rules are based on Drools rule inference engine [2]. It is improved to reach the generation of mobile applications source code and introducing new concern in model driven mobile engineering. The core of the Drool suites and advanced Inference Engine using are improved Rete algorithm for object pattern matching. Rules are stored in the production memory, while facts are maintained in the working memory. The production memory remains the same during an analysis session, i. e, rules cannot be added or removed or changed. The contents of the working memory on the other hand can change. Facts may be modified, removed or added by executing rules or from external sources. After changing in the working memory, the inference engine is triggered and it works out which rules become "true" for the given facts. If there are multiple selected rules, their execution order will be managed via the Agenda, using a conflict resolution strategy.

3.2. Drools Transformation Rule

Drools rules are defined using Java-like language. It is a Business Logic integration Platform (BLiP). With the runtime, we create a working memory. The syntax of rule is shown as follows:

Rule
Rule <Rule Name>
When <Condition>
then <Action>

Rule: A rule is nothing but the logic that will be applied to incoming data. It has two main parts; when and then.

When: works out the condition on which Rule will be fired.

Then: the action; if the rules met the condition, they define what work this rule performs.

Step 1: Create a.drl (droolRule.drl)file where we will define the rules.

Step 2: Create Person POJO class.

The proposed rule engine consists three parts: umple2model, umple2view and umple2controller according to android model, view and controller perspective. Table 1 shows the sample form of Drools transformation rule for simple variable declaration for Account Title. Umple2Model.drl transforms incoming abstract syntax model (ASM) into plain java object (POJO). Umple2View.drl transforms ASM into android user interface XML file and Umple2Controller.drl transforms ASM into android activity class. The code generator receives the POJO model for model layer, XML model for view layer and android model for controller layer. The generator use the java development tool (JDT-core) to generate POJO class and android class source code. It is also used the JDOM to generate XML user interface file.

Table 1: Transformation Rule Sample

```
Umple          String AccountTitle;

Rule "VariableDeclaration"
Dialect "java"
when
$st : SyntaxTree(status==SyntaxTree.VAR_DECLARE)
then
TypeDeclaration type=AST2Android.Variable_Decl($st.getType());
CompilationUnit cu=$st.getCu();
$st.setStatus(SyntaxTree.ACTIVITY_CREATE);
$st.setType(type);
$st.setCu(cu);
update($st);
end
```

```
POJO (Model Layer) String AccountTitle;
PublicvoidsetAccountTitle(String at){
accountTitle=at;}
public String getAccountTitle(){
return accountTitle;}
}

XML (View Layer)
<EditText android:id=
"@+id/ txtaccountTitle"
android:layout_height=
"wrap_content"
android:layout_width=
"wrap_content"/>

Android (Controller Layer)
private EditText txtaccountTitle;
private String accountTitle;
txtaccount-
Title=(EditText)findViewById(R.id.accountTitle);
```

4. Experimental results and comparison

In model driven development, there are two important criteria to evaluate the quality of model transformation. They are transformability and modifiability. The consistency of source and target model and the assessment of transformability are evaluated in [1]. According to these results, we extend to evaluate the assessment of quality of modifiability in this model driven transformation.

4.1. Modifiability

Changes made for the requirements are rendered quality of the code in the models the code. This fact becomes challenges in quality of model driven development [quality]. To address these issues, this paper has proposed the evaluation of modifiability of model driven transformation using object oriented metrics. This modifiability is decomposed into traceability of model elements and well-designated or not being too complex. Moreover, the extent to which a model transformation can be adapted to provide different or additional functionality. The main reason for modifying a model transformation is changing requirements. Another reason is that the (domain specific) language in which the source and/ or target model is described which may be subject to changes. Modifiability captures the amount of effort need to modify a model transformation. It controls the visibility of system development. Such controls contribute to modularity, an essential aspect of software engineering that promotes software maintainability. In object-oriented programming, we note that these classes form the modules of programs. From the modularity perspective, modules should be as independent as possible with minimal coupling.

We have also performed a comparative study using both our proposed system and JUSE4 Android [8, 9] based on object oriented metrics. These metrics indicate quality of source code directly. We evaluate the quality of model driven for model transformation of generated source code quality and prior approach's generated source code quality. The results are used to evaluate a model is complete or suitable for automation or a modeling technique is appropriate for a target transformation. Therefore we have identified metrics to examine in this process.

4.2. Metrics

We describe the metrics for assessing the quality attributes for model transformation. Those metrics are applicable to language definition and characteristics of languages. For modifiability, we determine encapsulation, polymorphism and reusability as the quality criteria. These are described in table 2.

Table 2: Object oriented quality criteria

Quality of Criteria for Comparative Study				
No	Quality Criteria	Metrics	Acronyms	Desired Results
1	Encapsulation	Methods of hiding factor	MHF	High
		Attribute hiding factor	AHF	High
2	Polymorphism	Number of method overridden by a subclass	NMO	High
		Polymorphism factor	PF	High
3	Reusability	Reuse ratio	RP	High
		Specialization ratio	SR	High

A. Data Collection

To evaluate the modifiability of transformation, we have collected the metric values by using Eclipse metrics Plug-in [23]. It is an open source metrics calculation tools which measures various metrics and detects cycle in package and type dependencies. At first, we have generated the android application from different approaches of proposed and prior approach respectively. In the next step, we enable Eclipse Metrics Plug-in on each generate source code that give common solution. Finally, we extracted the mean, standard deviation and maximum metric values for each generate source code. In our comparative study, we have collected the average metric values from the proposed

and prior generated source code with respect to the quality criteria. Table 3 shows extracted metric values.

4.3. Encapsulation

It is the bringing together of a set of fields and methods into an object definition and hiding their internal working from the users of the object. By encapsulation, the way an object or its fields and methods are structured which is not visible to the users of the object. It is also facilitated by bundling and information hiding. It enhances the software maintainability. Encapsulation increases the cohesiveness of data and methods through bundling and reduces the strength of coupling between software components through information hiding. For encapsulation, we have measured method hiding factor (MHF) and attribute hiding factors (AHF) using the following equation 1 and 2. MHF and AHF are indicators to show how well methods and attributes are hidden inside classes. The results are presented in table 3. The comparisons of MHF values and AHF values of proposed system and prior systems are shown in the following figure 2 and 3 respectively. These metrics are measured at system level and high metric values are expected. The results are compared with prior approach and our proposed approach. This result means that we achieve the higher method hiding factor and attribute hiding factors.

Let V (M)= number of classes where the method M is visible, then

$$MHF= 1-\frac{\sum V(M)/(Total\ numbers\ of\ classes-1)}{Number\ of\ methods\ in\ all\ classes} \quad (1)$$

Let V (A) = number of classes where the attribute A is visible, then

$$AHF= 1-\frac{\sum V(A)/(Total\ numbers\ of\ classes-1)}{Number\ of\ attributrd\ in\ all\ classes} \quad (2)$$

By using the equation 1, we have calculated the MHF value. The result described that the ratio of number of classes where the visible method M is higher, the MHF value is lower.

By using the equation 2, we have calculated the AHF value. The result described that the ratio of number of classes where the visible attribute A is higher, the AHF value is lower.

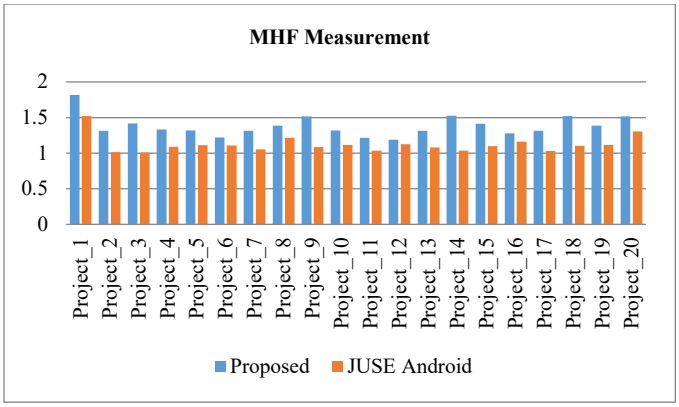


Figure 2 Comparison results of method of hiding factors

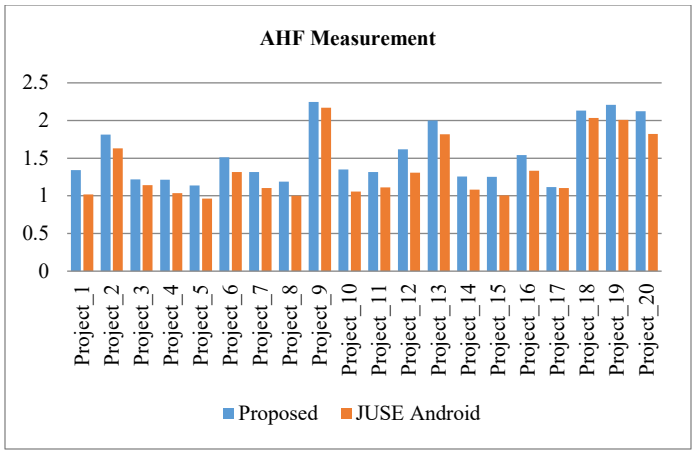


Figure 3 Comparison results of attribute hiding factors

4.4. Polymorphism

It is the ability of objects to respond to the same message but with the appropriate method based on their class definitions. For polymorphism, we have measured the number of method overridden by a sub class (NMO) and polymorphism factors (PF). Results are described in figure 4 and 5 respectively.

For NMO, we have determined the number of methods in a subclass overridden from its base class by using equation 3. Moreover, we determine the PF by using following equation 4. By using this equation 4, we present the PF value in table 3. To be specific, NMO is a class-level metric, which refers to the number of methods overridden by a single subclass, while PF is a system level metric, which measures the degree of method overriding in the whole type tree. These values are desired to be high. The results are compared with prior approach and our proposed approach. This result means that we achieve the higher number of methods overridden by a subclass and polymorphism factor.

NMO= the number of methods in a subclass overridden from its base class (3)

$$PF= \frac{\sum_{i=0}^{TC} M0(Ci)}{\sum_{i=0}^{TC}[(Mn\ Ci)x\ DC\ (Ci)]} \quad (4)$$

Where TC = the total number of classes

M_n (C_i)= Number of new methods of the class C_i

M_o (C_i)= Number of overriding methods of the class C_i

DC(C_i)= Number of Descendant of the class C_i

We have calculated the NMO value by using the number of methods in a subclass overridden from its base class and applying equation 3. They are presented in table 3.

We have calculated the PF value by using equation 4 and the results are described in table 3. The results present that number of overriding methods of the class is higher, the PF value is higher.

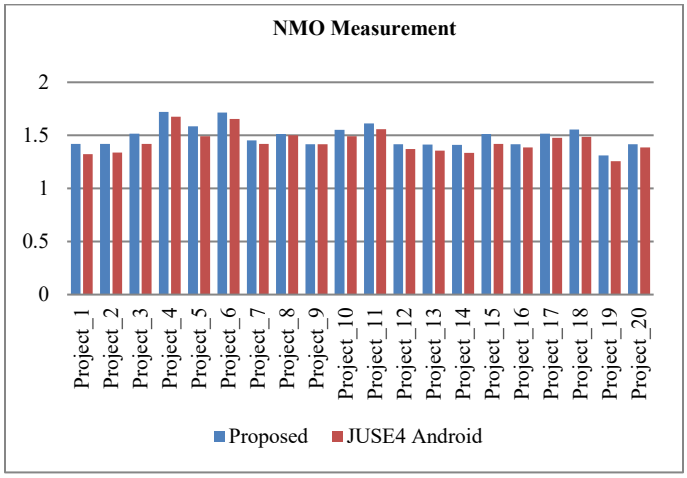


Figure 4 Comparison results of number of methods overridden by a sub class

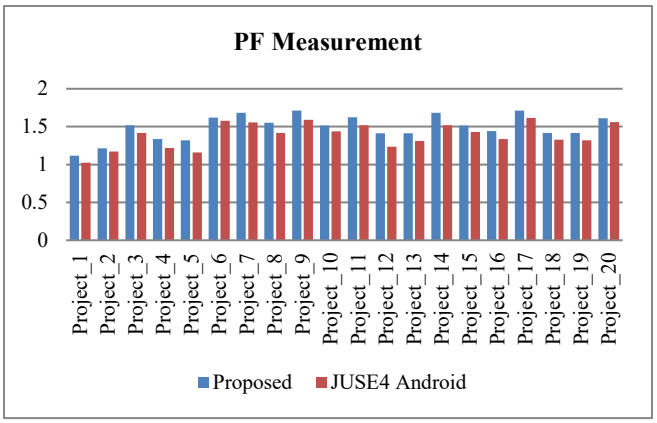


Figure 5 Comparison results of polymorphism factors

4.5. Reusability

Reusability is the extent to which a model transformation can be reused by other model transformations. It refers to as-is reuse. It is especially relevant for model transformations when a source model has to be transformed into different target models or vice versa. For reusability, we have measured reuse ratio (RR) and specialization ratio (SR) and results are presented in figure 6 and 7.

We have determined the RR and SR by using the following equation 5 and 6. RR and SR are both system level reusability metrics. They are calculated as the ratios of subclass to all classes and to super classes, respectively. The results are presented in table 3. We have expected to be highly reused, large reusability metrics values are desirable.

$$RR = \frac{(Total\ number\ of\ Super\ classes)}{(Total\ number\ of\ classes)} \quad (5)$$

$$SR = \frac{(Total\ number\ of\ Sub\ classes)}{(Total\ number\ of\ super\ classes)} \quad (6)$$

We have calculated the RR and SR values by applying equations 5 and 6 respectively. There are more subclasses, the higher the RR and SR values.

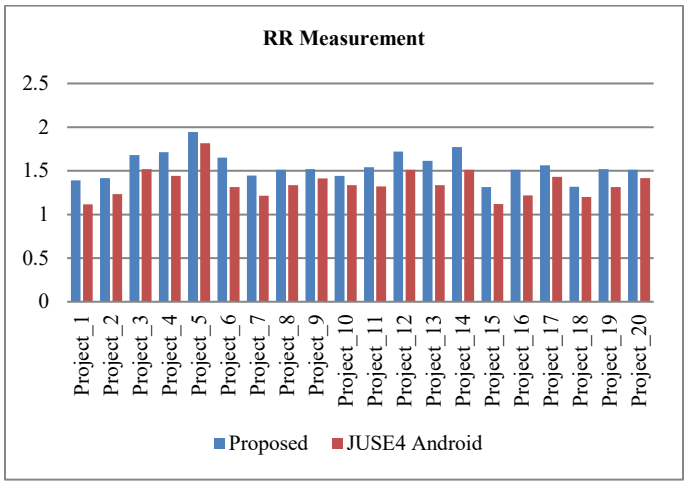


Figure 6 Comparison results of reuse ratios

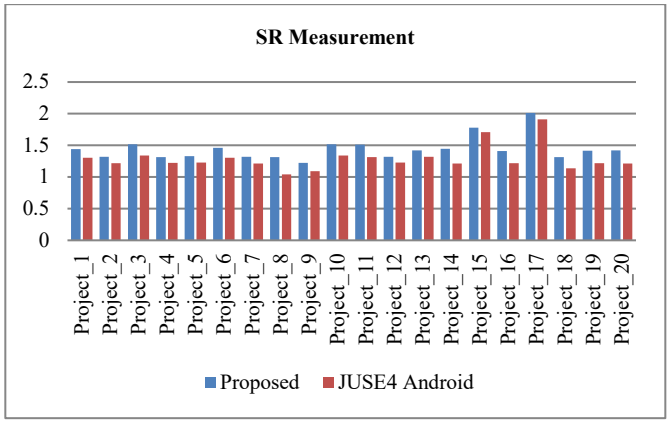


Figure 7 Comparison results of specialization ratios

5. Results and discussion

We have presented comparison results of our measurement. Expected results for our evaluation are shown in table 2. The larger a model transformation, the harder it is to understand and modify. Moreover, the number of signature and equations per function has a negative effect on consistency. If more similar signatures or equations have to be written, it is more likely that a different style is used.

The comparison results of our measurements are described in table 3. We have employed the result from MHF and AHF metrics for encapsulation, NMO and PF metrics for polymorphism and RR and SR for reusability to compare our approach and prior approach.

We have used private, protected and public keywords to control the accessibility to the method and attributes inside a class. According to these facts, we have planned quality attributes of our system to achieve higher modifiability.

Table 3: Metric values collected by Eclipse Metrics Plug in

Project	MHF		AHF		NMO		PF		RR		SR	
	Proposed	Prior	Proposed	Prior	Proposed	Prior	Proposed	Prior	Proposed	Prior	Proposed	Prior
Project_1	1.817	1.52	1.341	1.017	1.418	1.324	1.118	1.024	1.391	1.117	1.441	1.305
Project_2	1.312	1.012	1.814	1.631	1.418	1.339	1.214	1.172	1.415	1.234	1.318	1.216
Project_3	1.418	1.01	1.218	1.141	1.516	1.418	1.516	1.418	1.681	1.519	1.515	1.341
Project_4	1.332	1.091	1.216	1.037	1.721	1.675	1.338	1.219	1.712	1.441	1.312	1.225
Project_5	1.318	1.113	1.137	0.964	1.584	1.492	1.321	1.158	1.944	1.814	1.331	1.226
Project_6	1.218	1.108	1.512	1.314	1.714	1.654	1.62	1.578	1.651	1.315	1.461	1.306
Project_7	1.314	1.052	1.315	1.103	1.454	1.418	1.681	1.554	1.445	1.215	1.317	1.211
Project_8	1.386	1.216	1.189	1.003	1.512	1.5	1.551	1.418	1.512	1.335	1.312	1.041
Project_9	1.514	1.084	2.247	2.171	1.416	1.415	1.712	1.589	1.518	1.412	1.224	1.091
Project_10	1.317	1.117	1.351	1.056	1.551	1.491	1.512	1.438	1.441	1.337	1.516	1.338
Project_11	1.215	1.034	1.314	1.114	1.612	1.558	1.623	1.519	1.541	1.319	1.511	1.314
Project_12	1.188	1.127	1.618	1.306	1.416	1.371	1.412	1.237	1.721	1.512	1.317	1.227
Project_13	1.315	1.081	2.001	1.818	1.412	1.356	1.412	1.311	1.615	1.336	1.418	1.318
Project_14	1.523	1.034	1.258	1.084	1.411	1.336	1.681	1.518	1.771	1.512	1.446	1.215
Project_15	1.412	1.098	1.252	1.004	1.513	1.418	1.512	1.429	1.314	1.118	1.781	1.711
Project_16	1.278	1.161	1.541	1.331	1.417	1.387	1.441	1.337	1.512	1.217	1.412	1.219
Project_17	1.314	1.033	1.118	1.105	1.516	1.477	1.711	1.616	1.561	1.431	2.012	1.911
Project_18	1.521	1.102	2.132	2.034	1.554	1.486	1.417	1.327	1.318	1.201	1.316	1.138
Project_19	1.386	1.117	2.21	2.007	1.312	1.258	1.415	1.318	1.518	1.312	1.416	1.216
Project_20	1.517	1.305	2.124	1.823	1.416	1.387	1.612	1.559	1.512	1.416	1.418	1.213

By applying the experimental results, our approach has higher value than prior approach in encapsulation, polymorphism and reusability. Moreover, our expected result is that the high metric values are preferable.

These experimental results show how well methods and attributes are hidden inside classes. Therefore, our approach can help system develop methods and attributes which are hidden inside classes more efficiently. This means that our approach is more efficient than others. However, these results describes that they are a little bit higher than prior approach. By using these results, we will enhance our approach to achieve more efficiently and effectively model driven development process.

6. Conclusion

Model transformations become essential with the evolution of model driven development. It is an automatic generation of a target model from source model by using transformation definition. Modifiability is key issues in quality of this transformation. To address this issue, we have evaluated the modifiability of quality of model transformation using object oriented metric. This modifiability is decomposed into traceability of model elements and well-designated or not being too complex. Moreover, the extent to which a model transformation can be adapted to provide different or additional functionality. The main reason for modifiability of a model transformation is changing requirements. In this paper, we have performed the comparative study on our approach and prior approach to determine modifiability to develop more efficient mobile application system. We have determined the encapsulation, polymorphism and reusability as quality metrics. These metrics are measured at system level. We have used private, protected and public keywords to control the accessibility to the method and attributes inside a class. According to these facts, we have planned quality attributes of our system to achieve higher modifiability. The determination of experimental results represent that we achieve high score from comparison of our approach and prior approach. This means that our system is more traceability and

well-designated. Using these findings, we will enhance our approach to achieve higher efficiency and quality. In the future work, we will investigate more quality attribute for high accuracy of system development. Moreover, we will also evaluate the impact of transformation rules.

References

[1] Thu, E. E, Nwe N, "Model Driven Development of Mobile Applications Using Drools Knowledge-based rule" proceeding of SERA 2017, June 7-9, 2017, London, UK.

[2] De Lay, E, Jacobs D, "Rules-based Analysis with JBoss Drools: Adding Intelligence to Automation", ICALEPCS 2011, Proceeding of ICALEPCS Genoble, France

[3] Son, H.S, Kim, W.Y and Chul, R.Y, "MOF based Code Generation Method for Android Platform", International Journal of Software Engineering and Its application, Vol 7, No 3, Hongik University, Sehong Campus, Korea, 2013.

[4] Son, H. S, Kim, J.S, Chul, R. Y, SMTL Oriented Model Transformation Mechanism for Heterogeneous Smart Mobile Models, International Journal of Software Engineering and its Applications, Volume 7, No3, Sejong Campus, Korea, 2013.

[5] Parada, A.G, Lisane, B, A Model Driven Approach for Android Application Development, Brazilian, Symposium on Computing System Engineering (SBESC), Pelotas, Brazil, 2012.

[6] Parada, A.G.; Milena, R.S.; Automating mobile application development: UML-based code generation for Android and Window phone, Journal of Theoretical and Applied Information (RITA), volume 22, Pelotas, Brazil, 2015.

[7] Solheim, L and Neple, T, Model Quality in the Context of Model-Driven Development, Proceeding of 2nd International Workshop on Model- Driven Enterprise Information Systems (MDEIS'06), pp. 27-35, 2006.

[8] Silva, L.P, Abreu,F.B.: A Model-Driven Approach for Mobile Business Information Systems Applications, ACM/IEEE 17th International Conference on Model Driven Engineering Languages and Systems, MODELS 2014, QUASAR/ISTAR/ISCTE-IUL, Lisboa, Portugal.

[9] Silva, L.P, Abreu,F.B.: Model-Driven GUI Generation and Navigation for Android BIS Apps, MODELSWARD 2014, Portugal.

[10] Baowen, X, Di.W.: A metrics-based Comparative Study on Object-Oriented Programming Languages, 27th International Conference on Software Engineering and Knowledge Engineering, SEKE' 2015, USA.

[11] van Amstel, M.F; Lange, C.F.J.; van den Bran, M.G. J.: Metrics for analyzing the quality of model transformations, Proceeding s 12th ECOOP Workshop on Quantitative Approaches on Object Oriented Software Engineering (QAOOSE 08, Paphos, Cyprus, July 8, 2008.

[12] Rudiger, L, Jonas, L.: Comparing Software Metrics Tools, International Symposium on Software Testing and Analysis, ISSTA'2008, USA.

[13] Omar, E.B, Bragun, B, Automatic Code Generation by Model Transformation from Sequence Diagram of System's Internal Behavior International Journal of Information Technology, Hassan 1st University, Morocco, 2012.

[14] Steeg, C.C, Gotz, F, Model Driven, Data Management in Android with the Android Content Provider, https:// code.goodle.com/archieve/p/mdsd-android-content-provider/Bingen, Germany, 2011.

[15] https://github.com/umple/umple

[16] Mohagheghi, P, Aagedal, J, Evaluating Quality in Model-Driven Engineering, Proceeding of 29th International Conference on Software Engineering Workshops (ICSEW'07), 2007

[17] Mohagheghi,p.: An Approach for Empirical Evaluation of Model Driven Engineering in Multiple Dimensions, MODELPLEX, Oslo, Norway, 2010.

[18] Chidamber, S.R, Kamerer, C.F.: A metrics suite for object-oriented design, IEEE Transaction of Software Engineering, pp-(20)6, 1994: 476-493.

[19] Henderson-Sellers, B.: Object-oriented metrics: measures of complexity, Prentice Hall, 1995.

[20] Badreddin, O, Lethbridge, T.C, Forward, A, A Novel Approach to Versioning and Merging Model and Code Uniformaly, in MODELSWARD 2014, Proceeding of the 2nd International Conference on Model-Driven Engineering and Software Development, 2014.

[21] Kocaguneli, Ekrem, et al.: Prest: An Intelligent Software Metrics Extraction, Analysis and Defect Prediction Tool, 21th International Conference on Software Engineering and Knowledge Engineering, SEKE'2009, USA.

[22] Pallavi, K, Suita, P.U: Model Transformation, Concept, Current Trends and Challenges, International Journal of Computer Applications, Volume 119, No 14, Nasik, Manarashtra, India, 2015.

[23] http://metrics.sourceforge.net

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 252-260 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Constant Envelope DCT-based OFDM System with M-ary PAM Mapper over Fading Channels

Rayan Hamza Alsisi [*] , Raveendra Kolarramakrishna Rao

Faculty of Engineering, Electrical and Computer Engineering, The University of Western Ontario, London, ON, N6A 5B9, Canada

ARTICLEINFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 31 January, 2018

Keywords:
Bit Error Rate analysis
Discrete Cosine Transform
Constant Envelope signals

ABSTRACT

Constant Envelope Discrete Cosine Transform based Orthogonal Frequency Division Multiplexing (CE-DCT-OFDM) system with M-ary Pulse Amplitude Modulation (PAM) mapper is considered. In the system phase modulation is used to achieve constant envelope signals that have 0 dB Peak-to-Average-Power Ratio (PAPR). Transmission of such signals permit high power amplifiers in the system to operate with maximum power efficiency. The performance of CE-DCT-OFDM system is examined over Additive White Gaussian Noise (AWGN) and over fading channels. Closed-form expressions for Bit Error Rate (BER) over Ricean and Rayleigh channels are derived. The performances of CE-DCT-OFDM and conventional DCT-OFDM systems are compared as a function of Input power Back-Off (IBO) and Signal-to-Noise Ratio (SNR) for the Traveling-Wave Tube Amplifier (TWTA) model. Results show that CE-DCT-OFDM system offers superior BER performance compared to DCT-OFDM system and has other advantages as well.

1 Introduction

Fast Fourier Transform based Orthogonal frequency division multiplexing (FFT-OFDM) is widely adapted in a variety of communication standards due to its attractive properties such as high spectral efficiency and low complexity of the receiver, particularly, over multipath fading channels [2]. In an FFT-OFDM system, complex orthogonal exponential functions are used as basis functions. Instead, orthogonal cosinusoidal functions can be utilized as basis to create multicarrier system. Such a system utilizes Discrete Cosine Transform (DCT) [3] and is referred to as DCT-OFDM system. Several researchers have been investigating the use of DCT in OFDM system [4,5,6], as it has several advantages over conventional FFT-OFDM system. They are:

1. DCT is well known to have excellent spectral compaction and energy concentration properties. As a result, the channel estimation and also the system performance can be improved in noisy environments [7].
2. DCT is widely adopted in image/video coding standards (e.g. JPEG). Using IDCT for modulation and DCT for demodulation in an OFDM system, results in better integrated system design and reduced overall implementation cost are possible [8].
3. DCT uses real arithmetic compared to complex arithmetic in the case of FFT. This reduces signal processing complexity and power consumption, especially, when M-ary Pulse Amplitude Modulation (MPAM) mapper is used in DCT-OFDM system [7].
4. In the presence of frequency offset, due to the energy-compaction property of DCT, the inter-carrier interference (ICI) coefficients in DCT-OFDM system are concentrated around the main coefficient. As a result, DCT-OFDM system is robust to Carrier frequency offset (CFO) [7].
5. When MPAM mapper is used in DCT-OFDM system, it requires half of bandwidth required by an FFT-OFDM system, with the same number of subcarriers [9].

One of the major drawbacks of an OFDM system is

[*]Corresponding Author Name: Rayan Hamza Alsisi, Email: ralsisi@uwo.ca,
This paper is an extension of work originally presented in (SysCon) [1]

https://dx.doi.org/10.25046/aj030131

the high PAPR of transmitted signals in an FFT-OFDM system. When high PAPR signals are amplified using non-linear power amplifier, severe signal distortion will occur. Therefore, power amplifier with suitable power backoff is required in the system. Without appropriate power backoff, the system suffers from spectral broadening, intermodulation distortion, and, consequently, performance degradation. The problem can be mitigated by increasing the power backoff, but this results in poor power efficiency. In mobile devices with limited battery supply power efficiency is required to be as high as possible [10]. Several techniques have been suggested to mitigate the problem of high PAPR in an OFDM system such as coding, partial transmission sequences, clipping, tone reservation, and filtering [11,12,13]. These techniques offer a variety of trade-offs in terms of complexity, performance and spectral efficiency.

An alternative approach to completely eliminate the PAPR problem in an OFDM system is based on signal transformation. In this technique, signal transformation occurs at the transmitter prior to modulation and an inverse transformation at the receiver prior to demodulation. In [14,15,16,17] phase modulation and demodulation are considered in OFDM systems. Such systems are characterized by constant envelope signal with 0 dB PAPR, and hence suitable for power amplification close to the saturation level of nonlinear power amplifier. While FFT-OFDM systems with phase modulation have been extensively studied in the literature, DCT-OFDM system with phase modulation has not received much attention. In this paper, therefore, DCT-OFDM system with phase modulation referred to as CE-DCT-OFDM is presented and examined. The intent of this paper is to present a generalized model of CE-DCT-OFDM system that can be used to examine its performance. The BER analysis of CE-DCT-OFDM system in AWGN channel is presented and then the analysis is extended to the case of fading channel, as over practical communication channels signal fading is always present.

This paper is organized as follows. Section II describes the generation of DCT-OFDM signal. Section III introduces phase modulation in DCT-OFDM system. CE-DCT-OFDM system with MPAM mapper is described in Section IV, and its performance is analysed in AWGN channel. Section V deals with performance analysis of CE-DCT-OFDM system over fading channels. Finally, the paper is concluded in Section VI.

2 Baseband DCT-OFDM Signal

The process of generating DCT-OFDM signal is shown in Figure 1. The signal can be represented by

$$f(t) = \sum_{n=0}^{N-1} C_n \, \varphi_n(t), 0 \le t < T, \tag{1}$$

$$\varphi_n(t) = \begin{cases} \sqrt{\frac{2}{T}} cos 2\pi f_n t, & 0 \le t < T, \\ 0, & otherwise \end{cases} \tag{2}$$

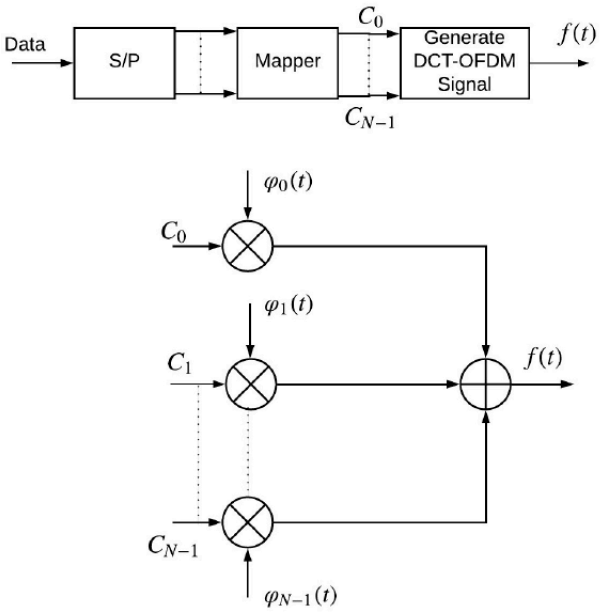


Figure 1: Block diagram of DCT-OFDM signal generator

The cosinusoidal function $\varphi_n(t)$ is the n^{th} orthogonal signal with frequency $f_n = n/2T$ and repersents the n^{th} subcarrier. The subcarrier spacing is $1/2T$. The subcarriers are orthonormal over $0 \le t \le T = NT_s$. That is,

$$\int_0^T \varphi_n(t)\varphi_k(t)dt \tag{3}$$

$$= \int_0^T \sqrt{\frac{2}{T}} cos 2\pi f_n t * \sqrt{\frac{2}{T}} cos 2\pi f_k t \, dt \tag{4}$$

$$= \begin{cases} 1 & n = k \\ 0 & otherwise \end{cases} \tag{5}$$

The time duration of the OFDM symbol is $T = NT_s$; T_s is the time duration of DCT-OFDM symbol. It is noted that $T_s = kT_b$ and $M = 2^k$. T_b denotes the bit duration. $C_n (n = 0, 1, .., N-1)$ are N independent data symbols obtained from MPAM signal constellation. The DCT-OFDM signal can thus be represented by

$$f(t) = \sqrt{\frac{2}{T}} \sum_{n=0}^{N-1} C_n \cos \pi nt/T, 0 \le t < T, \tag{6}$$

3 DCT-OFDM Signal with Phase Modulation

The phase modulated bandpass signal can be represented by

$$s(t) = A_c cos(2\pi f_c t + \phi(t)) \tag{7}$$

Where A_c and f_c are the carrier amplitude and frequency. The phase in (7) is proportional to $f(t)$ and is given by

$$\phi(t) = h_p f(t) \tag{8}$$

where h_p is the modulation index.

In DCT-OFDM system, $f(t)$ is real for MPAM mapper [9]. The advantage of DCT-OFDM system with phase modulation is that the transmitted signals have peak and average powers the same and hence, their PAPR is 0 dB. Figure 2 shows a comparison of instantaneous powers of DCT-OFDM and CE-DCT-OFDM signals.

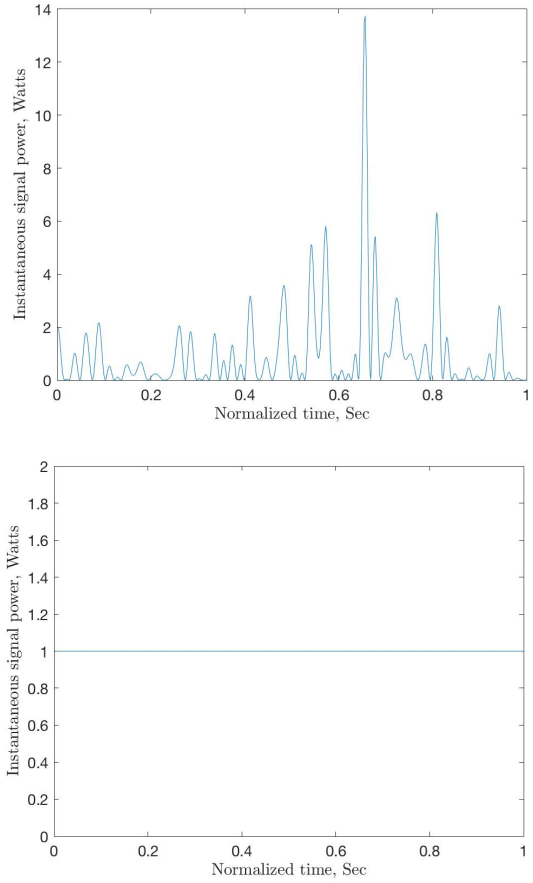


Figure 2: Instantaneous signal power: (a) DCT-OFDM signal and (b) CE-DCT-OFDM signal.

4 CE-DCT-OFDM Transmitter

The block diagram of CE-DCT-OFDM transmitter is shown in Figure 3. The output of the system can be written as:

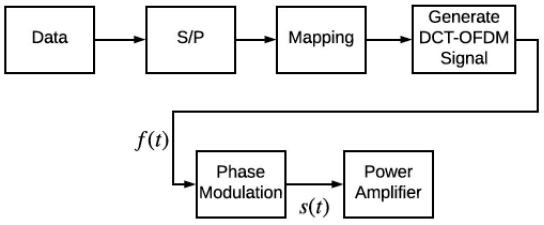


Figure 3: Block diagram of CE-DCT-OFDM transmitter

$$s(t) = A_c cos\left(2\pi f_c t + \sqrt{\frac{2}{T\sigma_s^2}} h_p \sum_{n=0}^{N-1} C_n cos(\pi nt/T)\right) \quad (9)$$

where $0 \le t \le T$, A_c is the signal amplitude and f_c is the carrier frequency. $\sigma_s^2 = (M^2-1)/3$ is the variance of the data symbols [16]. $\{C_n\}$ are MPAM data symbols, $C_n \in \{\pm 1, \pm 3, \cdots, \pm(M-1)\}$ for all n. The message signal is given by: $f(t) = \sqrt{2/T\sigma_s^2} \sum_{n=0}^{N-1} C_n cos(\pi nt/T)$. The average power of $s(t)$ is $P_s \left(= \int_0^T s(t)^2 dt/T = A_c^2/2\right)$ and the signal energy is $E_s = P_s T = A_c^2 T/2$. For k bits of information per symbol per transmission, the average bit energy is $E_b = A_c^2 T/2N \log_2 M = A_c^2 T/2Nk$.

4.1 Bandwidth Considerations

Phase modulated signals are complex to analyze for their bandwidth. However, simple observations can be used to get rough idea about the bandwidth of CE-DCT-OFDM signals. Using Maclaurin series, the PM signal described in equation (7) can be written as

$$s(t) = A_c(cos2\pi f_c t - h_p f(t) sin2\pi f_c t - \frac{h_p^2}{2!} f^2(t) cos2\pi f_c t + \cdots) \quad (10)$$

When h_p is small, the first two terms in the series are sufficient to represent CE-DCT-OFDM signal. That is,

$$s(t) \approx A_c cos2\pi f_c t - A_c h_p f(t) sin2\pi f_c t \quad (11)$$

This represents the narrowband case and the bandwidth of the signal is at least $2W$, where W is the bandwidth of $f(t)$. As h_p becomes larger, the bandwidth of the signal broadens. A useful expression for bandwidth of the signal is given by the root-mean-square (RMS) bandwidth [18] which is equal to $max(2h_p, 2)WHz$. The bandwidth of the message signal $f(t)$ is $W = (N/2T)Hz$

4.2 BER Analysis over AWGN Channel

The CE-DCT-OFDM receiver consists of a phase demodulator followed by the standard DCT-OFDM demodulator to recover the transmitted data symbols as shown in Figure 4. Each block will be analyzed below.

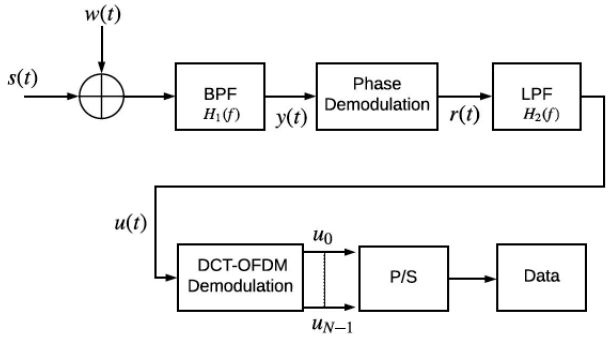


Figure 4: Block diagram of CE-DCT-OFDM receiver

4.2.1 Phase Demodulation

The noise $w(t)$ is modelled as additive white Gaussian with zero mean and power spectral density $N_0/2$. The received signal $s(t)+w(t)$ is fed to a BPF with transfer function $H_1(f)$ shown in Figure 5. The filter has a center frequency f_c and bandwidth B. It is noted that only a negligible amount of input signal power lies outside the frequency band $f_c - B/2 \leq |f| \leq f_c + B/2$. The bandwidth B is in excess of twice the message bandwidth W by an amount that depends on the deviation ratio of the signal $s(t)$. Thus, it is noted that the BPF allows the CE-DCT-OFDM signal without any distortion. The filtered narrow band noise $n(t)$ can be represented as

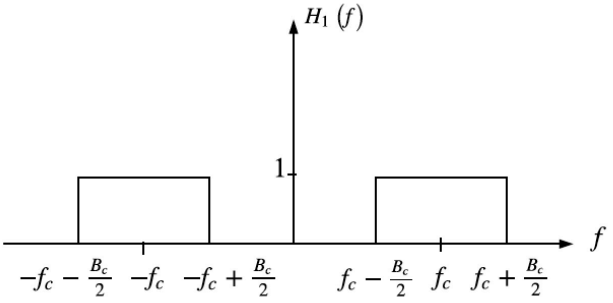


Figure 5: Ideal bandpass filter characteristic

$$n(t) = n_I(t)cos2\pi f_c t - n_Q(t)sin2\pi f_c t \quad (12)$$

where $n_I(t)$ and $n_Q(t)$ are the in-phase and quadrature components of the zero-mean filtered Gaussian noise. $n(t)$ can also be represented as

$$n(t) = x(t)cos[2\pi f_c t + \Psi] \quad (13)$$

where

$$x(t) = \sqrt{[n_I(t)^2 + n_Q(t)^2]} \quad (14)$$

and

$$\Psi = tan^{-1}(n_Q(t)/n_i(t)) \quad (15)$$

The bandpass filter output $y(t)$ can be written as:

$$y(t) = A_c cos[2\pi f_c t + \phi(t)] + x(t)cos[2\pi f_c t + \Psi(t)] \quad (16)$$

Equation (16) in polar form is given by:

$$y(t) = v(t)cos[2\pi f_c t + \theta(t)] \quad (17)$$

where $v(t)$ represents the envelope and $\theta(t)$ is the phase angle which can be written as [18,19]

$$\theta(t) = \phi(t) + \varepsilon(t) \quad (18)$$

where

$$\varepsilon(t) = tan^{-1}\left\{\frac{x(t)sin[\Psi(t)+\phi(t)]}{A_c + x(t)cos[\Psi(t)+\phi(t)]}\right\} \quad (19)$$

is the noise signal. With the assumption of a high Carrier-to-Noise Ratio (CNR), $A_c >> x(t)$, equation (18) becomes

$$\theta(t) \approx \phi(t) + \frac{x(t)}{A_c} sin[\Psi(t) - \phi(t)] \quad (20)$$

The output of the phase demodulator is given by

$$r(t) = k_p \theta(t) \quad (21)$$

where k_p is the gain constant. With the large CNR assumption and choosing $k_p = 1/h_p$, equation (21) becomes.

$$r(t) = f(t) + n_d(t) \quad (22)$$

where

$$n_d(t) = \frac{k_p x(t)}{A_c} sin[\Psi(t) + \phi(t)] \quad (23)$$

The power spectral density $S_{Nd}(f)$ of $n_d(t)$ is related to the power spectral density $S_{NQ}(f)$ of $n_Q(t)$. That is [19,20],

$$S_{Nd}(f) = \left\{\frac{k_p}{A_c}\right\}^2 S_{NQ}(f) \quad (24)$$

where

$$S_{NQ}(f) = \begin{cases} N_0, & |f| \leq \frac{B}{2} \\ 0, & otherwise \end{cases} \quad (25)$$

The phase demodulator output is applied to a low pass filter of bandwidth equal to message bandwidth W. It is used to pass the message signal and reject out-of-band noise from $n_d(t)$. The ideal transfer function of the filter is

$$H_2(f) = \begin{cases} 1, & |f| \leq W \\ 0, & otherwise \end{cases} \quad (26)$$

The output of the low pass filter can be written as

$$u(t) = f(t) + n_u(t) \quad (27)$$

The power spectral density $S_{Nu}(f)$ of output noise $n_u(t)$ at the output of low pass filter is given by

$$S_{Nu}(f) = \begin{cases} N_0 k_p^2 / A_c^2, & |f| \leq W \\ 0, & otherwise \end{cases} \quad (28)$$

and the average output noise power is

$$\int_{-W}^{W} \frac{N_0 k_p^2}{A_c} df = \frac{2WN_0 k_p^2}{A_c^2} \quad (29)$$

4.2.2 DCT-OFDM Receiver

The DCT-OFDM receiver, as depicted in Figure 4, is composed of two stages: a demodulator and a detector. The demodulator projects the incoming signal using orthonormal bases and generates a vector whilst the detector applies detection algorithm to estimate the transmitted information symbols.

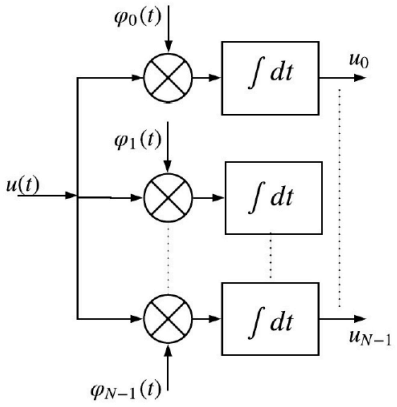


Figure 6: DCT-OFDM signal demodulation using correlators

The input to the bank of correlators (DCT-OFDM demodulator) in Figure 6 is the signal $u(t)$. The output of the demodulator is the vector $\bar{u}$. The j-th element of $\bar{u}$, can be expressed as:

$$u_j = \int_0^T u(t)\varphi_j(t)dt \tag{30}$$

$$= \int_0^T [f(t) + n_u(t)]\varphi_j(t)dt \tag{31}$$

$$= C_j/\sqrt{\sigma_s^2} + N_{uj}; j = 1, 2, .., N-1 \tag{32}$$

The mean of u_j is,

$$E[u_j] = E[C_j/\sqrt{\sigma_s^2} + N_{uj}] = E[C_j/\sqrt{\sigma_s^2}] = C_j/\sqrt{\sigma_s^2} \tag{33}$$

where $C_j \in \{\pm 1, \pm 3, ..., \pm(M-1)\}$. The mean is independent of the noise. However, the variance of u_j is dependent on noise and is given by:

$$Var[u_j] = \sigma_{uj}^2 \tag{34}$$

$$= E[(u_j - C_j)^2] \tag{35}$$

$$= E[(N_{uj})^2] \tag{36}$$

$$= E\left[\int_0^T n_u(t)\varphi_j(t)dt \int_0^T n_u(z)\varphi_j(z)dz\right] \tag{37}$$

$$= E\left[\int_0^T \int_0^T \varphi_j(t)\varphi_j(z) \cdot n_u(t)n_u(z)dtdz\right] \tag{38}$$

Taking the expectation operation inside the integrals, we can write

$$\sigma_{uj}^2 = \int_0^T \int_0^T \varphi_j(t)\varphi_j(z)E\left[n_u(t) \cdot n_u(z)\right]dtdz \tag{39}$$

$$= \int_0^T \int_0^T \varphi_j(t)\varphi_j(z)R_n(t,z)dtdz \tag{40}$$

where $R_{nu}(t,z)$ is the autocorrelation function of the noise process. The variance of u_j can be shown to be given by:

$$\sigma_j^2 = \frac{2WN_0k_p^2}{A_c^2}\int_0^T \varphi_j^2(t)dt \tag{41}$$

$$= \frac{2WN_0k_p^2}{A_c^2} \tag{42}$$

4.2.3 Probability of Bit Error

The symbol error rate (SER) can be shown to be given by [21]:

$$SER = \frac{M-1}{M}2P\left(N_{uj} > \frac{1}{\sqrt{\sigma_s^2}}\right) \tag{43}$$

$$= 2\frac{M-1}{M}\int_{1/\sqrt{\sigma_s^2}}^{\infty} \frac{1}{\sqrt{2\pi\left(2WN_0/A_c^2h_p^2\right)}} \tag{44}$$

$$e^{\left(-x^2/\left[2\left(2WN_0/A_c^2h_p^2\right)\right]\right)}dx$$

$$= 2\frac{M-1}{M}\int_{1/[2WN_0\sigma_s^2/A_c^2h_p^2]^{0.5}}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \tag{45}$$

$$= 2\left(\frac{M-1}{M}\right)Q\left(\sqrt{\frac{6h_p^2\log_2(M)E_b}{(M^2-1)N_0}}\right) \tag{46}$$

It is noted that for $h_p = 1$, (46) represents the SER for MPAM system [21]. For high CNR, the only significant symbol errors are those that occur at adjacent signal levels. The BER of CE-DCT-OFDM system, thus, can be approximated as [21]

$$BER \approx \frac{SER}{\log_2(M)}$$

$$\approx 2\left(\frac{M-1}{M\log_2(M)}\right)Q\left(\sqrt{\frac{6h_p^2\log_2(M)E_b}{(M^2-1)N_0}}\right) \tag{47}$$

The BER performance given by (47) is a function of E_b/N_0, signal-to-noise ratio, h_p, modulation index, and M, the number of amplitude levels in the MPAM mapper in the CE-DCT-OFDM system. The performance of CE-DCT-OFDM system with $h_p = 0.7$ for various values of M are illustrated in Figure 7, which shows that BER increases as M increases for fixed value of modulation index. For example at BER = 10^{-5} and $h_p = 0.7$, the SNR required for $M = 16$ is 14 dB more than that required for $M = 4$.

Figure 8 depicts BER performance of CE-DCT-OFDM system for 16-PAM mapper for $h_p = 0.3, 0.7$ and 1.2. It is observed that BER decreases as h_p increases for a fixed value of SNR. For example at BER = 10^{-5} the SNR required for $h_p = 0.3$ is 12 dB more than that required for $h_p = 1.2$. The BER performance of the system can be controlled by varying h and M as shown in Figure 9. For example, the system with $M = 16$ and $h_p = 1.7$ outperforms the system with $M = 4$ and $h_p = 0.2$ by nearly 9 dB at BER = 10^{-5}.

Figure 10 compares simulation results to theoretical BER given by (47) for CE-DCT-OFDM system using $N = 64$ subcarriers, and $M = 4$. For E_b/N_0 greater

than 15 dB, and for a small modulation index $h_p = 0.1$, simulation result is nearly the same as theoretical result. For large modulation index, for example $h_p = 0.8$, the theoretical BER is not as accurate as simulation result but still is within 1 dB of the former.

For the traveling-wave tube amplifier (TWTA) model, BER performance of CE-DCT-OFDM system is compared with that of DCT-OFDM system. The undesirable effects of TWTA nonlinearities can be reduced by increasing the input power backoff (IBO). For a given DCT-OFDM signal, we need to adjust the average input power so that the peaks of the signal are rarely clipped. That is, we will have to apply an IBO to the signal prior to amplification. Computer simulations are used to study the performance of the systems using nonlinear TWTA with various IBO levels. Figure 11 compares BER performance of 64 subcarrier 8-PAM CE-DCT-OFDM system and 8-PSK DCT-OFDM system with TWTA using IBO of 0 dB, 8 dB and 12 dB. At high SNR, CE-DCT-OFDM system provides significant performance improvement due primarily to the 0 dB backoff. The DCT-OFDM system with 0 dB IBO has an error floor at BER of 0.09. At the BER 10^{-3}, the IBO that results in the best DCT-OFDM system performance is 12 dB, with $E_b/N_0 = 16$ dB. However, the CE-DCT-OFDM system achieves this BER= 10^{-3} with $E_b/N_0 = 12$ dB which implies an advantage of 4 dB.

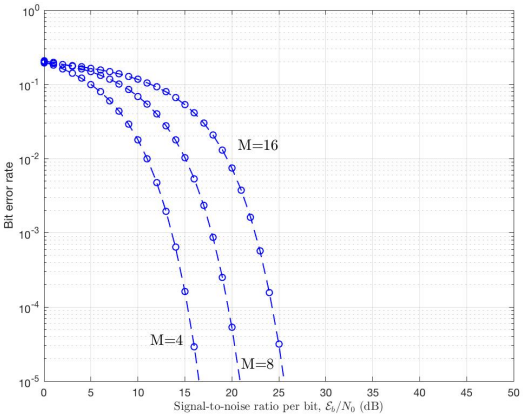


Figure 7: BER performance of CE-DCT-OFDM system over AWGN channel, as a function of M for $h_p = 0.7$

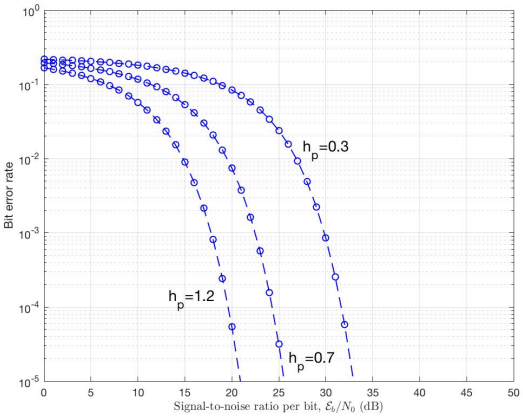


Figure 8: BER performance of CE-DCT-OFDM system over AWGN channel, as a function of h_p for $M = 16$.

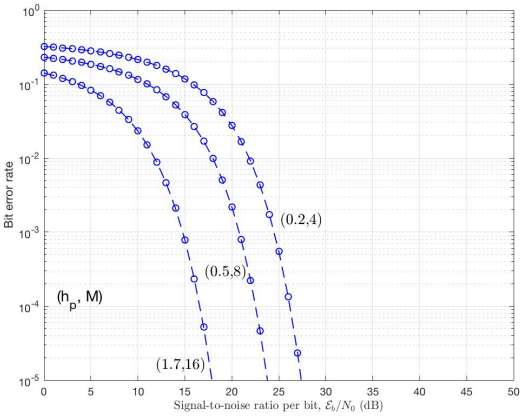


Figure 9: BER performance of CE-DCT-OFDM system over AWGN channel, as a function of (h_p, M).

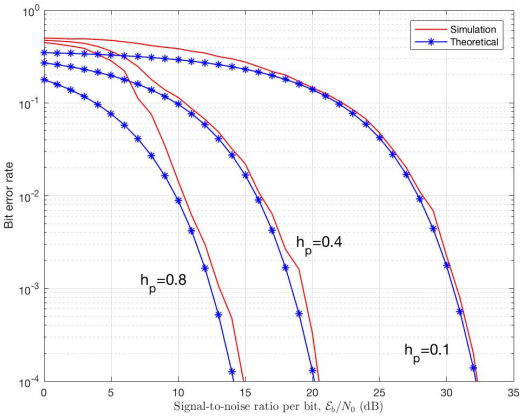


Figure 10: BER performance of 64 subcarrier CE-DCT-OFDM system with $M = 4$ over AWGN channel, as a function of h_p.

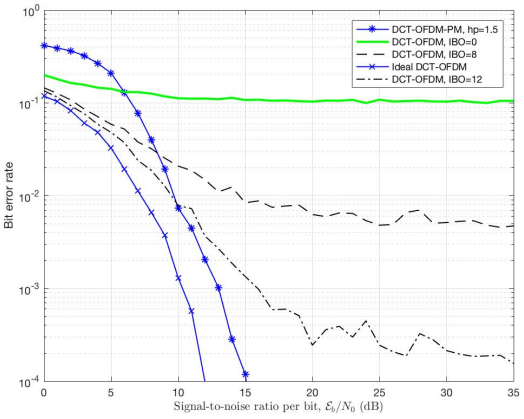


Figure 11: Comparison of 64 subcarrier 8-PAM CE-DCT-OFDM and DCT-OFDM systems for model TWTA for various of IBO.

5 Performance over Fading Channels

The received CE-DCT-OFDM signal over a fading channel can be expressed as:

$$r(t) = h(t) * s(t) + n(t) \tag{48}$$

where $h(t)$ represents the impulse response of the fading channel given by $h(t) = \alpha\delta(t)$. The instantaneous

SNR per bit and the average SNR per bit can be represented as $\gamma = \alpha^2 E_b/N_0$ and $\bar{\gamma} = E\left\{\alpha^2\right\} E_b/N_0$, respectively. To obtain the bit error rate (P_b) of CE–DCT-OFDM system over such a fading channel, the conditional BER is averaged over the Probability Density Function (PDF) of γ and can be written as [22]:

$$P_b = \int_0^\infty P(\gamma) p_\gamma(\gamma) d\gamma \tag{49}$$

where $P_b(\gamma)$ is given by:

$$P_b(\gamma) = 2\left(\frac{M-1}{M\log_2(M)}\right) Q\left(\sqrt{D\gamma}\right) \tag{50}$$

where $D = \frac{6h_p^2 \log_2(M)}{M^2-1}$. It is noted that $Q(z)$ in (50) is the well-known Q-function and it can also be written as:

$$Q(z) = \frac{1}{\pi}\int_0^{\pi/2} exp\left(-\frac{z^2}{2sin^2(\theta)}\right) d\theta \tag{51}$$

5.1 Rayleigh Fading Channel

For Rayleigh fading channel, the PDF of γ is given by [22]:

$$p_\gamma(\gamma) = \frac{1}{\bar{\gamma}} exp\left(\frac{\gamma}{\bar{\gamma}}\right), \gamma \geq 0 \tag{52}$$

Using (50)-(52) in (49), the average BER over Rayleigh fading channel can be shown to be given by

$$P_b = 2\left(\frac{M-1}{M\log_2(M)}\right)\frac{1}{\pi\bar{\gamma}}\int_0^{\pi/2}\int_0^\infty exp\left(-\frac{D\gamma}{2sin^2(\theta)} - \frac{\gamma}{\bar{\gamma}}\right) d\gamma d\theta \tag{53}$$

Upon simplification (53) [23], we get:

$$P_b = \left(\frac{M-1}{M\log_2(M)}\right)\left[1 - \sqrt{\frac{D\bar{\gamma}/2}{1+D\bar{\gamma}/2}}\right] \tag{54}$$

The BER given by (54) for CE-DCT-OFDM system over Rayleigh fading channel is a function of h_p, modulation index, M, number of levels in MPAM, and E_b/N_0, signal-to-noise ratio. The BER of CE-DCT-OFDM system for $M = 4$ is plotted as a function of h_p, and E_b/N_0 as shown in Figure 12. It is observed that BER increases as h_p decreases for a fixed value of SNR. For example at BER = 10^{-5} the SNR required for $h_p = 0.7$ is 7 dB more than that required for $h_p = 1.5$.

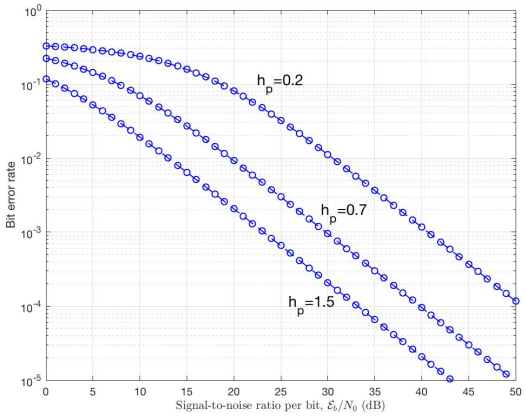


Figure 12: BER performance of CE-DCT-OFDM system over Rayleigh fading channel, as a function of h_p for $M = 4$.

5.2 Ricean Fading Channel

For the Ricean fading channel, the PDF of γ is given by [22]:

$$p_\gamma(\gamma) = \frac{(1+K)e^{-K}}{\bar{\gamma}} exp\left[-\frac{(1+K)\gamma}{\bar{\gamma}}\right] I_0\left[2\sqrt{\frac{(K+K^2)\gamma}{\bar{\gamma}}}\right], \gamma \geq 0 \tag{55}$$

Using (50), (51) and (55) in (49), P_b can be written as:

$$P_b = \frac{2}{\pi}\left(\frac{M-1}{M\log_2(M)}\right)\frac{(1+K)e^{-K}}{\bar{\gamma}}\int_0^{\pi/2}\int_0^\infty exp\left[-\frac{D\gamma}{2sin^2(\theta)} - \frac{(1+K)\gamma}{\bar{\gamma}}\right] I_0\left[2\sqrt{\frac{(K+K^2)\gamma}{\bar{\gamma}}}\right] d\gamma d\theta \tag{56}$$

Integrating (56) [23], BER of CE-DCT-OFDM system over Ricean fading channel can be written as:

$$P_b = \frac{2}{\pi}\left(\frac{M-1}{M\log_2(M)}\right)\int_0^{\pi/2}\frac{(1+K)sin^2(\theta)}{(1+K)sin^2(\theta)+D\bar{\gamma}/2} exp\left[-\frac{KD\bar{\gamma}/2}{(1+K)sin^2(\theta)+D\bar{\gamma}/2}\right] d\theta \tag{57}$$

The BER given by (57) is a function of h_p, M, K, Rice distribution parameter and E_b/N_0. The BER performance $K = 7$ dB and $M = 4$ is illustrated in Figure 13, are a function of h_p. It observed that there is improvement in BER as h_p increases, for example at BER = 10^{-5} the SNR required for $h_p = 1.5$ is 17 dB less than that required for $h_p = 0.2$.

The effect of parameter K on BER performance is illustrated in Figure 14, for $M = 4$ and $h_p = 0.5$. It is noted that the BER decreases as K increases, for example at BER = 10^{-5} and $h_p = 0.5$ the SNR required for $K = 2$ dB is 30 dB more than that required for $K = 18$ dB.

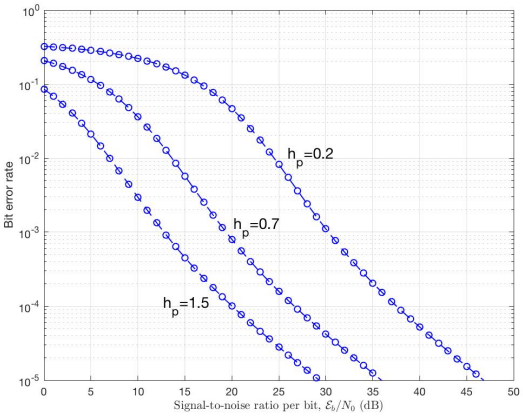


Figure 13: BER performance of CE-DCT-OFDM system over Ricean fading channel, as a function of h_p for $M = 4$.

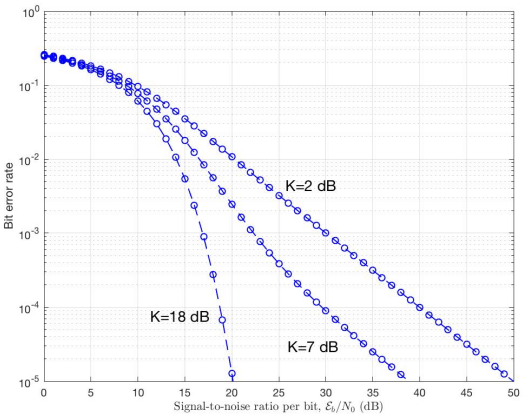


Figure 14: BER performance of CE-DCT-OFDM system over Ricean fading channel, as a function of K for $M = 4$ and $h_p = 0.5$.

6 Conclusions

A generalized description of CE-DCT-OFDM system is presented. In this system, phase modulation is used to eliminate the problem of PAPR. BER analyses of this system over AWGN and flat fading channels are presented and closed-form expressions for BER have been obtained. Improved BER performance is seen with increased value of modulation index at fixed transmission power. It is also observed that BER performance can be controlled by varying h_p and M as well. Simulation performance of CE-DCT-OFDM system over AWGN channel is also presented and compared with theoretical results. The results show that for a small modulation index, simulation result is nearly the same as theoretical result. With nonlinear the results show that CE-DCT-OFDM system has better BER performance than the conventional DCT-OFDM system when TWTA amplifier is used.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The first author would like to gratefully thank the Islamic University of Madinah and the ministry of higher education in Saudi Arabia for their support and scholarship.

References

1. Rayan H. Alsisi, Raveendra K. Rao, "Performance of constant envelope DCT based OFDM system with M-ary PAM mapper in AWGN channel," IEEE International Systems Conference (SysCon), pp.1-7, Montreal, April 2017. http://ieeexplore.ieee.org/document/7934716/
2. S. H. Han and J. H. Lee, "An overview of peak-to-average power ratio reduction techniques for multicarricer transmission," IEEE Wireless Commun., vol.12, no.2, pp.56-65, April 2005. http://ieeexplore.ieee.org/document/1421929/
3. Gi Hyun Kim, Honey Durga Tiwari, Chan Mo Kim, Yong Beom Cho, and Younggoo Kwon, "Implementation of DCT based OFDM system," International SoC Design Conference, IEEE, vol.2, pp.41-44, Busan, November 2008. http://ieeexplore.ieee.org/document/4815679/
4. C. Tang and D. Mandyam,"Performance of OFDM modem with alternative basis functions, " IEEE Radio and Wireless Symposium, pp.551-554, San Diego, October 2006. http://ieeexplore.ieee.org/document/1615216/
5. G. Mandyam, "Sinusoidal transforms in OFDM systems, " IEEE Trans. On Broadcasting, vol.50, no.2, pp.172-184, June 2004. http://ieeexplore.ieee.org/document/1304952/
6. R. Merched, "On OFDM and single carrier frequency domain systems based on trigonometric transforms," IEEE Signal Processing Letters, vol.13, no.8, pp.473-476, August 2006. http://ieeexplore.ieee.org/document/1658060/
7. N. Al-Dhahir, H. Minn, and S. Satish, "Optimum DCT-based multicarrier transceivers for frequency-selective channels," IEEE Trans. Commun., vol.54, no.5, pp.911-921, May 2006. http://ieeexplore.ieee.org/document/1632105/
8. S. Satish, N. Al-Dhahir, and H. Minn, "A DCT-Based Broadband Multicarrier Transceiver,", SoutheastCon, Proceedings of the IEEE, pp.175-180, Memphis, 2006. http://ieeexplore.ieee.org/document/1629345/
9. J. Tan and G. L. Stuber, "Constant Envelope Multi-Carrier Modulation," in Proceedings of IEEE Military Communications Conference, vol.1, pp.607-611, Anaheim, October 2002. http://ieeexplore.ieee.org/document/1180513/
10. Y. Tsai and G. Zhang, "Orthogonal Frequency Division Multiplexing with Phase Modulation and Constant Envelope Design," in Proc. of IEEE Milcom, pp.2658-2664, Atlantic City, October 2005. http://ieeexplore.ieee.org/document/1606068/
11. T.A. Wilkinson, A.E. Jones, "Minimization of the Peak to Mean Envelope Power Ratio of Multicarrier Transmission Schemes by Block Coding," IEEE VTC, pp.825-829, Chicago, July 1995. http://ieeexplore.ieee.org/document/504983/
12. B. S. Krongold and D. L. Jones, "An active-set approach for OFDM PAR reduction via tone reservation," IEEE Trans. Signal Processing, vol.52, no.2, pp.495-509, February 2004. http://ieeexplore.ieee.org/document/1261335/
13. J. Armstrong, "Peak-to-average power reduction for OFDM by repeated clipping and frequency domain filtering," IEE Electron. Lett., vol.38, no.5, pp.246-247, February 2002. http://ieeexplore.ieee.org/document/990223/
14. C.-D. Chung and S.-M. Cho, "Constant-envelope orthogonal frequency division multiplexing modulation," APCC/OECC Conference, vol.1, pp.629-632, Beijing, October 1999. http://ieeexplore.ieee.org/document/824966/
15. R. Pacheco and D. Hatzinakos, "Error rate analysis of phase-modulated OFDM (OFDM-PM) in AWGN channels," IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, pp.IV337-IV-340, Toulouse, May 2006. http://ieeexplore.ieee.org/document/1660974/
16. Steve C Thompson, Ahsen U. Ahmed, John G. Proakis and James R. Zeidler, "Constant Envelope OFDM Phase Modulation: Spectral Containment Signal Space Properties and Performance," IEEE MILCOM, pp.1129-1135, Monterey, 2004. http://ieeexplore.ieee.org/document/1495013/

17. S. C. Thompson, J. G. Proakis, and J. R. Zeidler, "Constant Envelop Binary OFDM Phase Modulation," IEEE MILCOM, pp.621-626, Boston, October 2003. http://ieeexplore.ieee.org/document/1290175/

18. J. G. Proakis and M. Salehi, Communication Systems Engineering. New Jersey: Prentice-Hall, 1994.

19. R. Zeimer and W. Tranter, Principles of Communications: Systems, Modulation, and Noise, 4th ed. New York: John Wiley, 1995.

20. J. J. Downing, Modulation Systems and Noise. Prentice-Hall, 1964.

21. J. G. Proakis, Digital Communications, 4th ed. New York: McGraw- Hill, 2001.

22. M. K. Simon and M.-S. Alouini, Digital Communication over Fading Channels. New York: John Wiley and Sons, Inc., 2000.

23. I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series, and Products, vol. 6, California, Academic Press, 2000.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 252-260 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Heuristic Based Approach for Voltage Stability Improvement using FACTS Devices

Hajer Jmii*, Asma Meddeb, Souad Chebbi

LATICE Laboratory, University of Tunis, Tunis 1008, Tunisia

ARTICLE INFO

Article history:
Received: 30 October, 2017
Accepted: 17 November, 2017
Online: 23 December, 2017

Keywords:
Heuristic method
SVC
TCSC
UPFC
Voltage Stability

ABSTRACT

The aim of this paper is to develop a hybrid device for voltage stability enhancement using two kinds of FACTS (Flexible AC Transmission System) namely SVC (Static Var Compensator) and TCSC (Thyristor Controlled Series Capacitor). The idea behind the proposed method is to maintain safe and satisfactory power system operation in a lesser costing manner by taking advantage of the performances of SVC and TCSC at the same time. We propose to evaluate the efficacy of the combined device to UPFC, as it is a hybrid FACTS and it is the most versatile compensator. For purpose of identifying the placement of the devices, we opt for a heuristic based approach. The methodology is tested with the IEEE 14-Bus system using the software EUROSTAG, and the simulation results reveal the efficiency of the proposed method for enhancing voltage stability.

1. Introduction

Today's power system is more and more operating close to its stability limits due to the unceasing growing of the power demand [1, 2]. Therefore, transmission lines are prone to the overloading, especially when the system is unable to cope up with power transition [3]. In addition, in some cases, the transmission of the energy is a hard task because of the long distance between the generation station and the loads. This may increase power losses and threats to the voltage stability. Voltage stability is defined as the system's ability to retain voltages at all the buses in the whole network within the specified boundaries after a disturbed condition [4]. Several system collapses have been reported in recent years [5] making the need for rapid and accurate control systems more and more insistent [6]. As power electronics components continue to develop, a variety of control devices as FACTS (Flexible AC Transmission System) have been prospered [7, 8]. The introduction of such a technology in the power system provides the control of the transmission line impedance, the voltage magnitude, and the phase angle. UPFC, SVC, and TCSC are three of the main FACTS devices which have acquired a well-recognized term for higher and smoother controllability in power systems.

The use of FACTS technology for system stability support has been studied in a large number of works [9-11]. A. Motiebirjandi *et al.* [12] studied the impact of UPFC on damping oscillations of the generator rotor. For this purpose, authors proposed system critical modes and residue factor methods for the optimal placement. In addition, they applied particle swarm optimization (PSO) to optimize the parameters of the UPFC. Reference [13] proposed the use of SVC to improve power system transient stability and damp oscillations in case of three-phase short-circuit. In reference [14], the authors demonstrated the performance of UPFC compared to SVC, in terms of improving power system stability. In order to enhance transient stability, the authors in [15] compared the performances of UPFC to different FACTS devices, namely TCSC, STATCOM (Static Synchronous Compensator) and SVC. B. Bhattacharyya et al. [16] were interested in the cost implication and power loss of installing UPFC alongside with SVC and TCSC. After determining the optimal emplacement and parameters of the FACTS using specific algorithms, authors have reported the gain obtained when adding the hybrid compensator to the other FACTS. Likewise, authors in [17] discussed the use of SVC, TCSC, and UPFC in the improvement of dynamic and transient system stability. They compared the three FACTS based on their mathematical models and operation modes. It was found that UPFC provided the most rapid control and the highest performances in stabilizing the system. P. Pandey *et al.* [18] studied the contribution of the SVC and UPFC in enhancing the voltage profile of a grid connected distributed generation system. They demonstrated through simulations, the satisfactory operation of the two FACTS especially the hybrid device. Reference [19] presented the application of a heuristic based procedure to

*Hajer Jmii, University of Tunis, Tunis 1008, Tunisia, +21694014865, hajerjmii@yahoo.com

parameters, type and location problems of UPFC, SVC and TCSC. Simulations showed satisfactory results of the proposed method.

This paper is an extension of work initially published in the Proceedings of the International Conference on Sciences of Electronics, Technologies of Information and Telecommunications [20]. The aim of this work is to improve power system voltage stability by using a hybrid device composed of SVC and TCSC. The benefits of this idea are to achieve a safe and secure power system operation in a less costly way. We choose to compare the performance of the proposed device to UPFC, as it belongs to the family of hybrid FACTS and it is the most powerful tool in the current control systems. The placement of the devices is identified by applying a heuristic based approach and the proposed method is tested on the IEEE 14-Bus system via the software EUROSTAG.

Firstly, in section 2 we present statistics about FACTS devices. Second, the voltage instability phenomenon is explained mathematically. Then, the mathematical models and structures of the proposed FACTS are detailed. Section 5 outlines the proposed approach. In section 6, we focus on the case study and the discussion of simulation results. Finally, we summarize the main points of this paper in the conclusion.

2. Statistics about FACTS Devices

Since the development of power electronics, a great interest has been granted to FACTS devices. The research considers this technology as a substantial and a timely topic. For this reason, an important number of studies discussing the application of FACTS to power systems have been published. In power system stability field, the publications related to FACTS are also numerous as described in Figure 1. It was found that the interest to the flexible controllers is more and more increased which reflects the efficiency of such a solution for instability problems. According to reference [21], when looking at the statistics for SVC, TCSC and UPFC publications, we noted that they are dominant compared to the other FACTS. Nonetheless, SVC is reaping the major concern in researcher's studies with 114 publications until 2004.

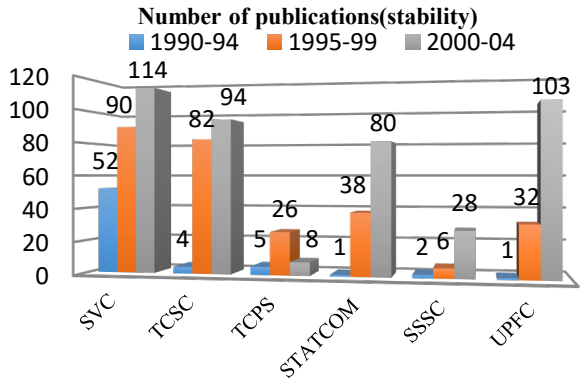


Figure 1. Statistics of the publications related to the FACTS

Actually, these statistics are reflecting the installed FACTS systems in the world. Reference [22] presented approximate results of a survey on worldwide integrated FACTS. It should be mentioned that SVC is the most exploited FACTS with a total power of 90.000MVA. While 2.000MVA are distributed on 10 incorporated TCSC into power systems, there are only 3 real networks equipped with UPFC with 250MVA of generated power.

To understand the reasons for the widespread use of SVC in comparison to the others FACTS particularly UPFC, we must look at system planner's choices. Recognizing that the investment cost is always a considerable constraint, economical solutions for power system instability are generally preferred. The cost of a FACTS device is highly dependent on the complexity of its model which is determined by the number of semiconductors used. Thereby, consisting of two voltage source converters, UPFC has the highest cost among the various FACTS [22]. According to reference [23], it is estimated at 0.33 million$ for 1MVAR generated power while the cost of SVC is approximately 0.19 million$ and the instrument and investment costs of TCSC are estimated at 0.22 million$ for 1MVAR generated power. Thus, when the network is already equipped with SVC, we can get well-improved stability by adding TCSC and at the same time economize 0.11million$ for each MVAR generated.

Otherwise, the cost-efficiency is not always to install the least costly FACTS device, but rather to choose the appropriate one based on the type of instability problem that we want to solve. As these problems are, on the one hand, unpredictable and on the other hand, closely linked, it is required to invest in equipment which can assure the control of more than one network parameter. Currently, UPFC has the potential to act on three parameters, namely: phase angle, line impedance, and bus voltage either simultaneously or separately. So, confronted with these constraints, we thought to find a solution for network instability, combining the economical side with the wide application area and efficiency side. Hence, our idea is to create an equivalent of UPFC by joining the action of TCSC, as a series FACTS, to that of SVC which assure the shunt compensation. With this approach, power system stability is enhanced at lower cost.

3. Voltage Instability

In power system stability field, a particular interest is accorded to the control of voltage and reactive power. The main issue is to avoid voltage instability which can result in a widespread problem known as voltage collapse [7]. For this purpose and in order to guarantee the reliability and the efficiency of power system operation, the following points must be respected in voltage and reactive power control:

- Voltages of all network buses must be within the permissible interval. According to reference [24], it is specified as [0.94pu; 1.06pu] for MV network.
- The reactive power flow in transmission lines must be minimized as much as possible. This leads to reducing line losses: RI^2 and XI^2.

We can illustrate the voltage instability by considering the system of Figure 2. This 2-bus network consists of a voltage source supplying a load through a transmission line.

The apparent impedance at bus 2 is expressed by:

$$S_2 = P_2 + jQ_2 = V_2 I^* \quad (1)$$

Where:

$$\bar{I} = \frac{\overline{V_1} - \overline{V_2}}{jX}$$

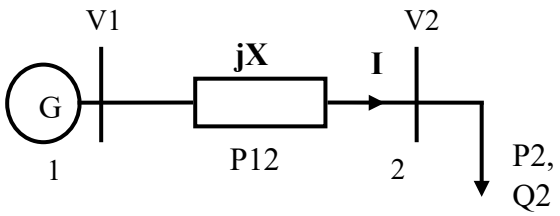


Figure 2. Equivalent model of a 2-bus system

Therefore equation (1) may be expressed as :

$$S_2 = V_2 \times \frac{V_1 \cos(\delta) + jV_1 \sin(\delta) - V_2}{jX} \quad (2)$$

$$S_2 = \frac{V_1 V_2 \sin(\delta)}{X} + j\frac{V_1 V_2 \cos(\delta) - V_2^2}{X} \quad (3)$$

The active and reactive powers supplied to the load are :

$$P_2 = \frac{V_1 V_2 \sin(\delta)}{X} = P_{\max} \sin(\delta) \quad (4)$$

$$Q_2 = \frac{V_1 V_2 \cos(\delta) - V_2^2}{X} \quad (5)$$

From equations (4) and (5) and using: $\cos^2(\theta) + \sin^2(\theta)$, we obtain:

$$P_2^2 + (Q_2 + \frac{V_2^2}{X})^2 = \frac{V_1^2 V_2^2}{X^2} \quad (6)$$

Therefore the voltage at bus 2 can be extracted as:

$$V_2^2 = \frac{1}{2}\left[V_1^2 - 2Q_2X \pm V_1(V_1^2 - \frac{4P_2^2X^2}{V_1^2} - 4Q_2X)^{1/2}\right] \quad (7)$$

Where :

V_1 : Busbar 1 voltage

V_2 : Busbar 2 voltage

I : Line current

X: Line reactance

δ : Generator internal angle

P_2 and Q_2 : Active and reactive powers supplied to the load.

Equation (7) describes the evolution of the voltage at bus 2 in function of the load increase at the same bus. We can deduce that in addition to the active and reactive powers, bus 2 voltage depends on the line reactance.

4. FACTS Models

In order to compare the performances of the FACTS devices, it is required to describe their functionalities supported with mathematical models and equivalent schemes. That is what we are going to present in this subsection.

4.1. TCSC

TCSC is one of the most important FACTS equipment, which has been used to modify the series impedance of the transmission line in order to improve system stability. Typically, it consists of a thyristor-controlled reactor in parallel with a fixed capacitor, which is equivalent to an adjustable reactance [25, 26]. The reactance of a transmission line equipped with a TCSC is expressed as follows:

$$X_{jk} = X_{jk}^{line} + X_{jk}^{TCSC} \quad (8)$$

Figure 3 shows a TCSC integrated between bus j and bus k with its equivalent model while Figure 4 describes the configuration of the series FACTS.

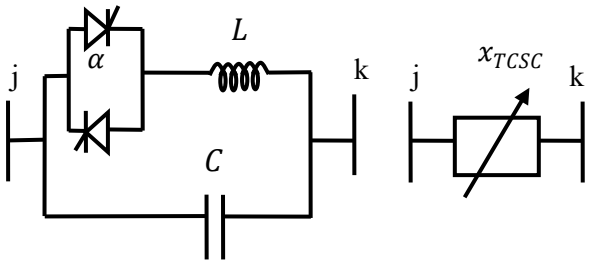


Figure 3. TCSC configuration

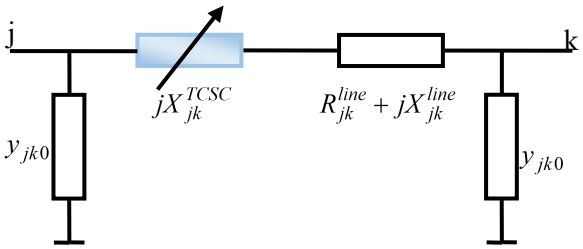


Figure 4. Equivalent model of TCSC

Before the integration of the TCSC in the transmission line jk, the admittance matrix Y is given by equation (9) [27]:

$$Y = \begin{pmatrix} \underline{y}_{jk} + \frac{\underline{y}_{jk0}}{2} & -\underline{y}_{jk} \\ -\underline{y}_{jk} & \underline{y}_{jk} + \frac{\underline{y}_{jk0}}{2} \end{pmatrix} \quad (9)$$

Where:

$$y_{jk} = G_{jk} + jB_{jk} \quad (10)$$

$$G_{jk} = \frac{R_{jk}^{line}}{(R_{jk}^{line})^2 + (X_{jk}^{line})^2} \quad (11)$$

$$B_{jk} = -\frac{X_{jk}^{line}}{(R_{jk}^{line})^2 + (X_{jk}^{line})^2} \quad (12)$$

The addition of the TCSC modifies the admittance matrix as follows:

$$Y' = \begin{pmatrix} \underline{y}'_{jk} + \frac{\underline{y}_{jk0}}{2} & -\underline{y}'_{jk} \\ -\underline{y}'_{jk} & \underline{y}'_{jk} + \frac{\underline{y}_{jk0}}{2} \end{pmatrix} \quad (13)$$

Where:

$$y'_{jk} = G'_{jk} + jB'_{jk} \quad (14)$$

$$G'_{jk} = \frac{R_{jk}^{line}}{(R_{jk}^{line})^2 + (X_{jk})^2} \quad (15)$$

$$B'_{jk} = -\frac{X_{jk}}{(R_{jk}^{line})^2 + (X_{jk})^2} \quad (16)$$

On EUROSTAG, to model a series and variable admittance, we must create two fictive buses in a transmission line and insert a current injector to each bus, as described in Figure 5. The two fictive buses must be connected to the network by lines with high reactance value to avoid injector shutdown in case of an opening line.

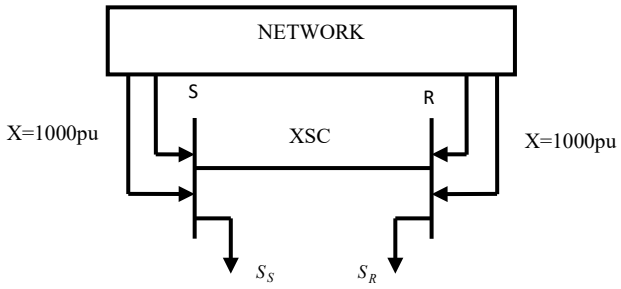


Figure 5. TCSC model on EUROSTAG

4.2. SVC

SVC is a static reactive power compensator whose output is adjusted for exchanging a capacitive or inductive current with the network to typically control bus voltage [28, 29]. In the steady state as well as in transient regime, this device is able to maintain voltage within the desired limits. Figure 6 shows the dynamic model of SVC. It can be modeled as variable shunt admittance with a thyristor controller. However, by neglecting the losses of SVC, we can consider it as ideal, so the admittance is purely imaginary and is described by the equations (17) and (18):

$$G_{SVC} = 0 \quad (17)$$

$$y_{SVC} = jB_{SVC} \quad (18)$$

The susceptance can be capacitive or inductive. Indeed, in the case of reactive power excess, SVC absorbs the increased amount through the inductor and in the opposite case; the capacitor covers the reactive demand.

The capacitive power injection model of SVC at the rated voltage is given by equation (19) [30]:

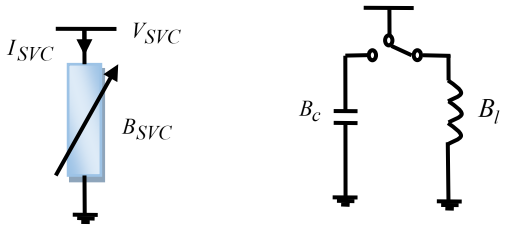


Figure 6. SVC model

$$Q_{SVC} = -V_N{}^2 B_{SVC} \quad (19)$$

Where B_{SVC} must be controlled according to equation (20):

$$B_{SVC}{}^{\min} \le B_{SVC} \le B_{SVC}{}^{\max} \quad (20)$$

$B_{SVC}^{\max}$ designates the capacitive limit state while $B_{SVC}^{\min}$ designates the inductive one. If SVC susceptance reaches its limits without maintaining the voltage of the bus where it is connected, it loses the capability of voltage control and it becomes similar to a fixed susceptance.

The connection of SVC to a bus j, as illustrated in Figure 7, changes the admittance matrix only at the element Y'_{ii}, in fact, the admittance of the compensator is added according to the following equation:

$$Y'_{ii} = Y_{ii} + y_{SVC} \quad (21)$$

The admittance matrix is so expressed by equation (22).

$$Y = \begin{pmatrix} y_{ij} + \frac{y_{ij0}}{2} + y_{SVC} & -y_{ij} \\ -y_{ij} & \frac{y_{ij0}}{2} + y_{ij} \end{pmatrix} \quad (22)$$

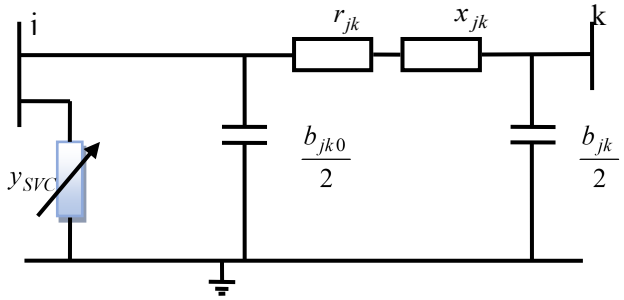


Figure 7. SVC connected to a bus

EUROSTAG adopts the model of Figure 8 and represents SVC as an impedance injector connected to a bus of the electrical network.

4.3. UPFC

UPFC is the most recognized hybrid compensator, it is the third generation of FACTS devices, it is the first and only device that has the ability to control simultaneously or separately voltage, line impedance and the phase shift of the bus voltage. UPFC consists

of two transformers, one connected in series and the other in shunt with the transmission line as shown in Figure 9 [31].

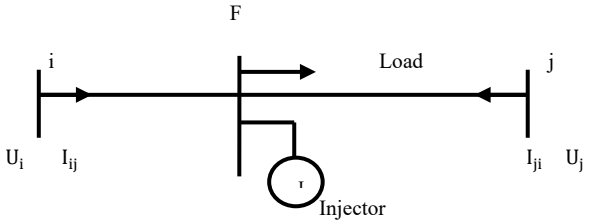


Figure 8. SVC model on EUROSTAG

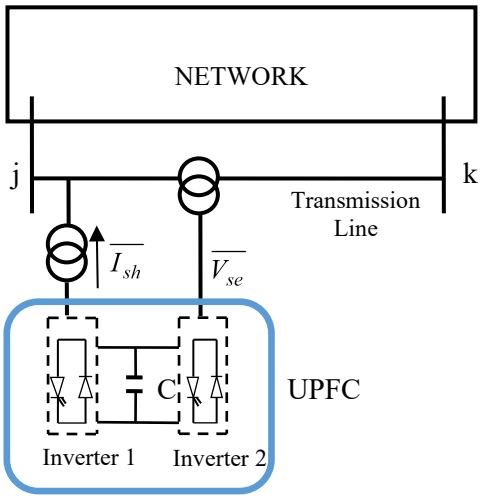


Figure 9. UPFC inserted in a power system

The inverter 2 injects the voltage $\overline{V}_{se}$ which is controllable in amplitude and in phase so that it can perform the serial compensation function of the active power. On the other hand, the inverter 1 is used, through the continuous connection, to provide the active power required for the inverter 2, it serves to compensate the reactive power since it can absorb or inject the active power into the network. Indeed, UPFC allows both the active power control and the line voltage control. It can switch from one to the other function instantaneously by changing the control of the inverters separated by a capacitor.

The equivalent circuit of UPFC is represented in Figure 10 [32]. From this scheme, we can extract the active and reactive power flows of the shunt and series converters which are expressed by the equations (23) - (25).

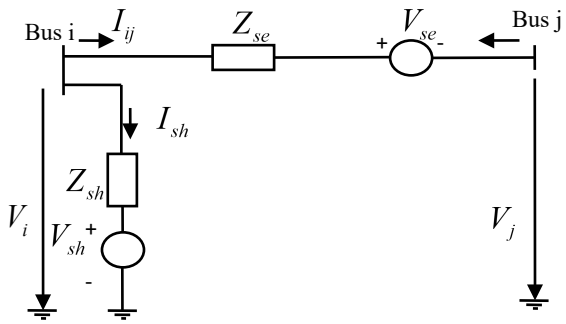


Figure 10. Equivalent circuit of UPFC

$$P_{sh} = V_i^2 g_{sh} - V_i V_{sh}(g_{sh}\cos(\theta_i - \theta_{sh}) + b_{sh}\sin(\theta_i - \theta_{sh})) \quad (22)$$

$$Q_{sh} = -V_i^2 b_{sh} - V_i V_{sh}(g_{sh}\sin(\theta_i - \theta_{sh}) - b_{sh}\cos(\theta_i - \theta_{sh})) \quad (23)$$

$$P_{ij} = V_i^2 g_{ij} - V_i V_j(g_{ij}\cos\theta_{ij} + b_{ij}\sin\theta_{ij}) - V_i V_{se}(g_{ij}\cos(\theta_i - \theta_{se}) + b_{ij}\sin(\theta_i - \theta_{se})) \quad (24)$$

$$Q_{ij} = -V_i^2 b_{ij} - V_i V_j(g_{ij}\sin\theta_{ij} - b_{ij}\cos\theta_{ij}) - V_i V_{se}(g_{ij}\sin(\theta_i - \theta_{se}) - b_{ij}\cos(\theta_i - \theta_{se})) \quad (25)$$

Where:

$$g_{sh} + jb_{sh} = \frac{1}{Z_{sh}}$$

$$g_{ij} + jb_{ij} = \frac{1}{Z_{se}}$$

$$\theta_{ij} = \theta_i - \theta_j$$

V_i and V_j: voltages at buses i and j

P_{sh} and Q_{sh}: active and reactive power flows of the shunt inverter

P_{ij} and Q_{ij}: active and reactive power flows of the series inverter

V_{sh} and V_{se}: shunt and series voltage sources

Z_{sh} and Z_{se}: shunt and series coupling transformer impedances.

The modeling of the UPFC shunt part on EUROSTAG is simple; it is represented by a current injector. As for the modeling of the series part, we must open the line where we want to insert UPFC and place at its extremities two current injectors (Figure 11). The opening of the line is assured by a high reactance value.

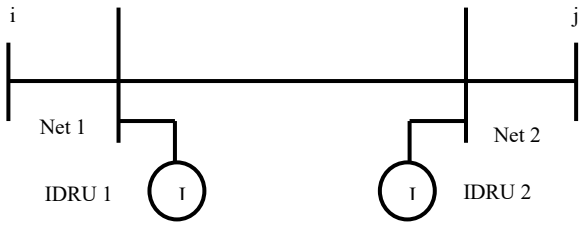


Figure 11. UPFC model on EUROSTAG

5. Proposed Approach

The main purpose of our paper is to find an economical and efficient way for sustaining voltage stability by comparing the performances of SVC-TCSC to those of UPFC. In order to achieve

this, it is important to take into consideration the identification of the proper placement of the FACTS in the system.

There is a variety of methods for finding optimal solutions [19, 33]. However, in practice, the time factor is more important than optimally solving the problems so that engineers and system planners often resort to heuristic methods. Heuristic methods known also as approximate methods are used for the resolution of optimization problems. These methods guarantee to find in a polynomial time at least a good and a feasible solution, but not necessarily the optimal one [34].

There are different types of heuristic algorithms, as shown in Figure 12, among which we are interested in reduction heuristics. The principle of such a method is to simplify the problem and to reduce the domain of the solutions. This implies to define the boundaries of the problem by identifying the properties that must satisfy a good solution.

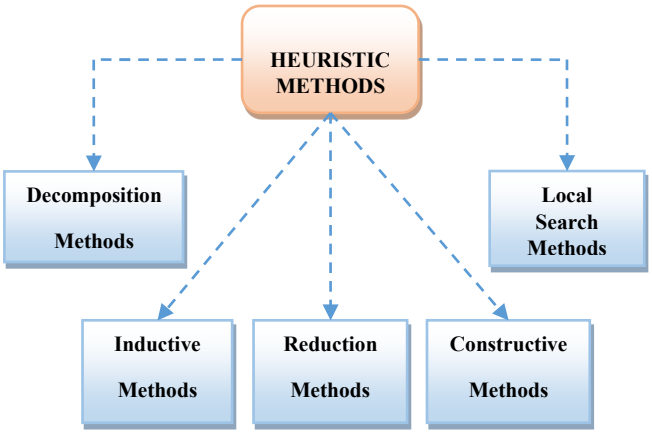


Figure 12. Types of heuristic algorithms

Figure 13 shows the flowchart of the proposed method. The normal state of the system must be obtained first of all so that we can analyze the load flow results and determine the weakest zone of the system that needs to be compensated. In this way, the location problem is simplified and the domain of solutions is reduced to the weakest zone. Based on the load flow analyses, a preliminary selection of the FACTS placement is made. Then, the system with the FACTS devices is executed. The evaluation of the selected location is performed by comparing the voltage magnitude to the specified lower and upper boundaries. A proper FACTS placement must maintain the voltage of all the buses within the permissible limits. If this criterion is not respected, another location has to be considered, and the evaluation based on the defined boundaries is repeated. Once the best location of the devices is identified, the next step is to perform the comparison between the proposed FACTS. The comparison takes into consideration the voltage magnitude, the oscillations, and the power flows.

With this method, we can get satisfactory results and avoid complicated and slow routines of the optimization methods.

6. Simulation

6.1. Test System

The proposed case study is the classical system IEEE 14-bus network, it consists of:

- Two generator buses and eleven load buses.
- Fifteen transmission lines.
- A three-winding transformer and two step-up transformers with two windings.
- Three synchronous compensators connected to buses 3, 6 and 8.

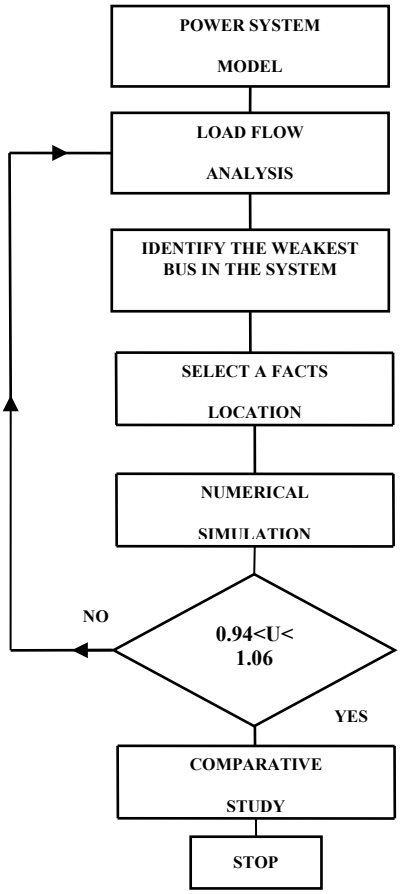


Figure 13. Flowchart of the proposed method.

All data relating to this test network are extracted from reference [32] and we used EUROSTAG software package [35] for the simulations. This software is a powerful tool dedicated to dynamic simulations. Its major advantage is the high rapidity of its algorithm. Irrespective of the nature of the disturbance and the system size, it can rapidly visualize the behavior of the network until it returns to an equilibrium state. Furthermore, this simulation is not accompanied by any deterioration in the accuracy of the calculations. In addition, it gives the possibility to define the modeling level adapted to the type of the performed study.

6.2. Simulation Results

6.2.1. Identification of FACTS Location

To compare the performances of UPFC and the proposed SVC-TCSC, first of all, it is mandatory to identify the appropriate placement of the FACTS. This step has to provide a firm basis for comparison. Therefore, we plan a step change in load at time t= 200s from 5% of the initial load by a step of 5%. The procedure

continues until reaching the maximum loading capability of the system.

The results of the base state signaled that bus 14 has got the lowest voltage amplitude as shown in Figure 14; hence we summarized the voltage levels of bus 14 during various step changes in Table 1.

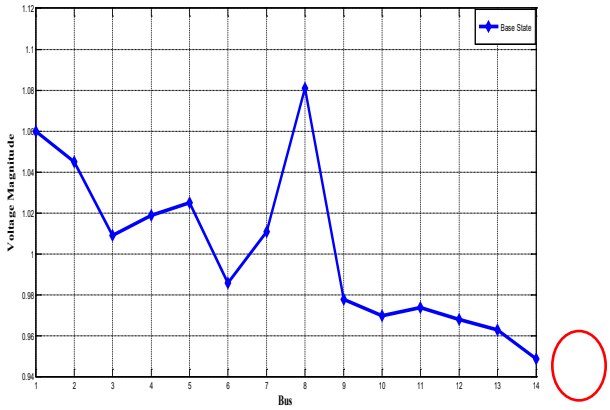


Figure 14. Voltage profile of IEEE 14-bus system in normal state

Table 1. Network behavior under incremental load increase.

Network Settings	Load Increase in %					
	5%	*10%*	*20%*	*25%*	*30%*	*33%*
U14 (pu)	0.966	0.962	0.952	0.948	0.946	Voltage collapse
State System	Stable					

We note the occurrence of a voltage dip in bus 14, which is increased as more the loads are augmented. The voltage level is still acceptable and the network is maintaining its stability until 32% of load increase. When reaching the case of 33%, EUROSTAG fails to execute the simulation which means that the system is collapsing.

After identifying the weakest bus of the system, we make a preliminary selection of the FACTS location. Hybrid or series controller both must be integrated through a transmission line. According to the network architecture, as schematized in Figure 15, bus 14 is related to bus 13 and bus 9, so we have two possibilities to place the device.

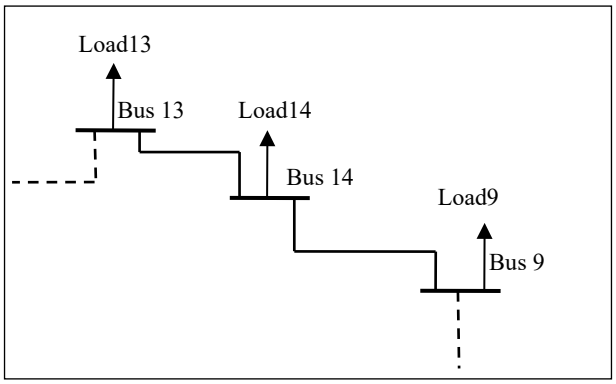


Figure 15. Scheme of the architecture of the IEEE 14-bus system

First of all, we integrate UPFC in the middle of line 13-14. Then, we apply an increase of 20% of the total load of the system. As shown in Figure 16, there is no improvement in the magnitude of bus 10 voltage. In addition, lower amplitudes equal to 0.925pu and 0.977pu are obtained respectively in bus 14 and bus 9, compared to the uncompensated system. Thus, it is concluded that line 13-14 is not a good location for the FACTS to perform the comparative study.

Now, we integrate UPFC through line 9-14 and we rerun the system. A meaningful improvement of the voltage magnitude is observed. According to Figure 16, UPFC has increased the voltage amplitude of bus 14 from 0.952pu to 0. 1.007pu with less severe oscillations, bus 10 from 0.976pu to 0.985pu and bus 9 from 0.982pu to 0.991pu. As we have got satisfactory results, the line 9-14 is retained as a good placement to insert the proposed FACTS and then to compare their potentialities.

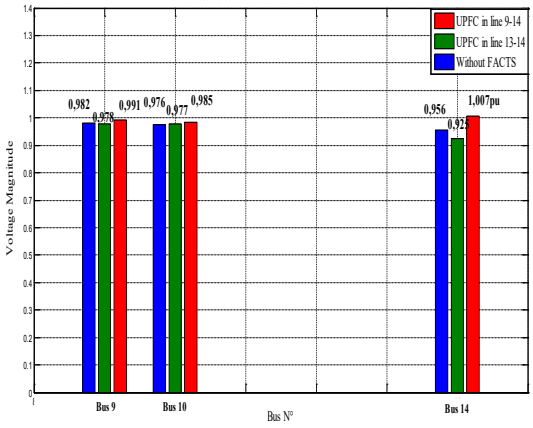


Figure 16. Voltage profile with different UPFC locations.

6.2.2. *Performance comparison between the FACTS*

We connect UPFC and SVC-TCSC to the network, separately. Firstly, UPFC was integrated into the middle of line 9-14 with a capacity of 60MVAR for each one of its inverters, and then, with keeping the same dimensions, we inserted TCSC through the line 9-14 and SVC to bus9.

Based on the results of the prior section, we apply a total load increase of 30% of the initial load. In this case, the test network is under heavy loaded conditions.

We are interested in the comparison, to the evolution of bus voltage as well as the active and reactive power flows in transmission lines.

Taking bus 14 as an example as shown in Figure 17, it can be seen that both UPFC and SVC-TCSC have a bearing on the voltage magnitude by increasing it significantly, while it was under the minimum acceptable value before compensation. However, UPFC showed a higher capability of enhancing voltage level, from 0.90pu to 0.96pu. In addition, we obtained a rapid damping of oscillations after only 7s either by UPFC or SVC-TCSC.

The impact of the FACTS on the oscillatory regime is clearer in the voltage curve of bus1, plotted in Figure 18. Nonetheless, we note that the damping action of SVC-TCSC is more

considerable compared to that of UPFC, in fact, it reduced the highest oscillation from 1.08pu to 1.071pu and rapidly established the steady state with well-amortized oscillations.

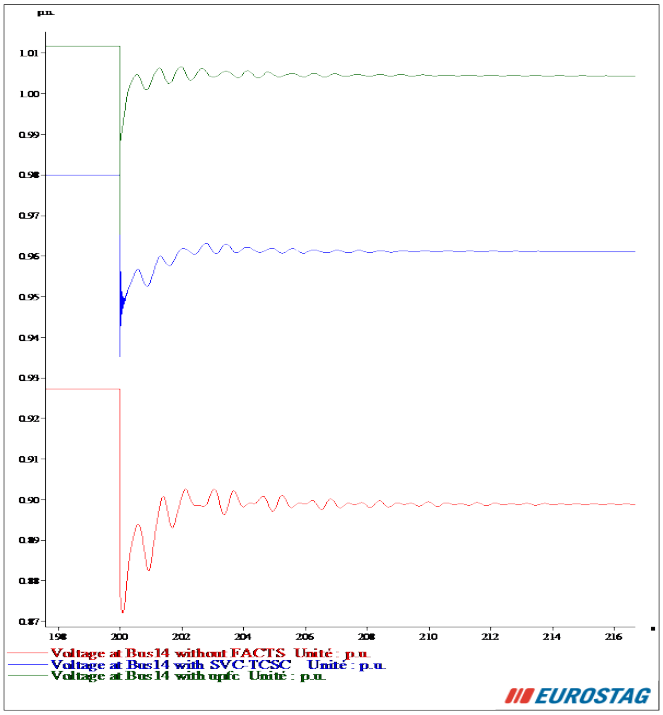


Figure 17. Voltage of bus14 with and without FACTS.

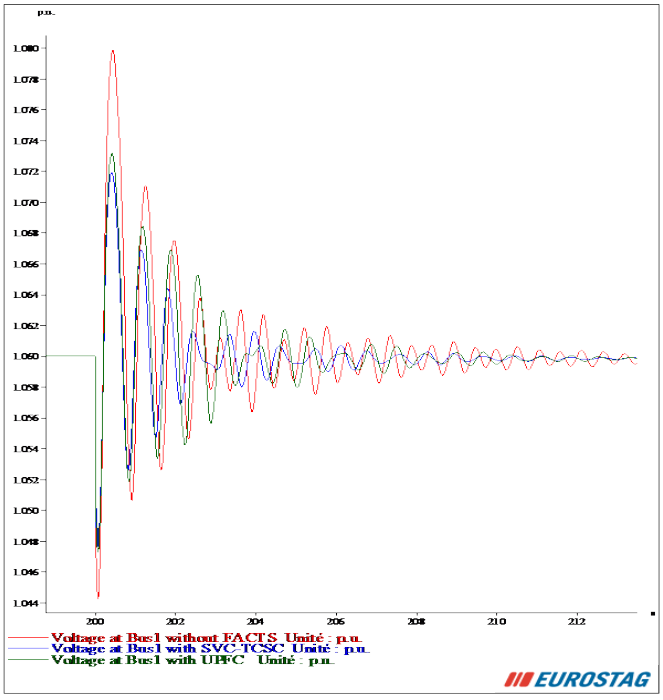


Figure 18. Voltage of bus1 with and without FACTS.

According to the results of Table 2, an excellent enhancement in the active power has been observed with the action of UPFC. However, we didn't note any important change in its level when used SVC-TCSC. As for the reactive power flow, it encountered a considerable decrease by using UPFC; it was decremented from 5.7 MVAR to 2.5MVAR through the line 9-10 as an example. This was observable also when introducing the combination SVC-TCSC. The amount of reactive power in line 6-12 is reduced by about 12% and in line 1-2, from 20MVAR to 19.2MVAR. Nevertheless, we recorded an increase of 10.5% in the transmitted MVARs across line 9-10.

Figure 19 shows the active power transmitted in line 1-2 with and without FACTS. When focusing on the transient regime, it must be mentioned that even in the temporal evolution of powers, the damping action of SVC-TCSC is more efficient; it gives more attenuated oscillations and joins the stable state in a shorter time.

Table 2. Active and reactive power flows with and without FACTS.

Simulation case	Line 9-10		Line 1-2		Line 6-12	
	Active Power MW	*React. Power MW*	*Active Power MW*	*React. Power MW*	*Active Power MW*	*React. Power MW*
Without FACTS	8.3	5.7	209	20	9.8	3.4
With SVC-TCSC	8.3	6.3	209	19.2	10	3
With UPFC	9.2	2.5	216	18.1	10.8	1.5

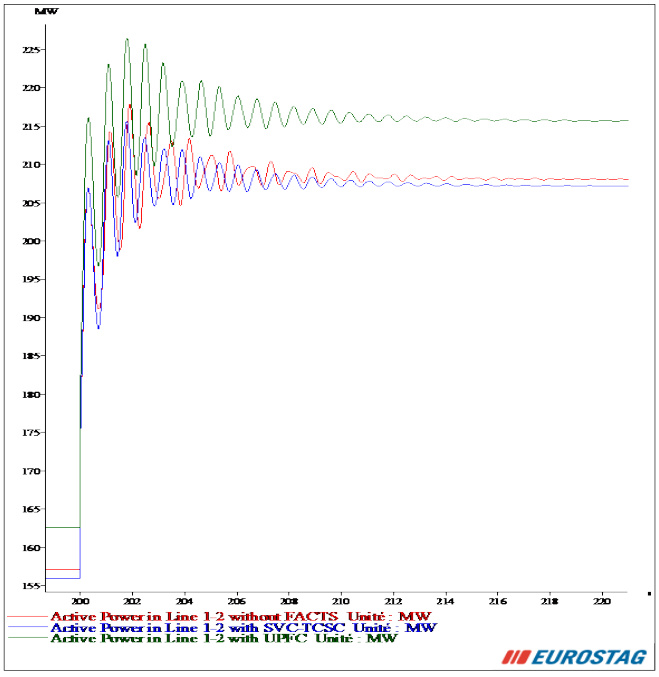


Figure 19. Active power flow in line 1-2 with and without FACTS.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

7. Conclusion

This paper presented a hybrid method for voltage stability enhancement using SVC and TCSC. The advantage of the proposed approach is the lesser cost of achieving a safe power

system operation. The IEEE 14-bus system was utilized to test the performance of the combination SVC-TCSC and compare it to the UPFC, the most pricey and multi-functional FACTS. The identification of the best placement of the FACTS devices was performed using a simple heuristic method. The simulation results demonstrated that the hybrid FACTS based on SVC and TCSC is able to significantly improve the voltage profile of the whole network in addition to the satisfactory action in amortizing the oscillations. The findings are encouraging to rather improve the performance of the SVC-TCSC in controlling power flow in lines and introduce other techniques for optimal location.

References

[1] B. Singh, "Applications of FACTS Controllers in Power Systems for Enhance the Power System Stability: A State-of-The-Art," International Journal of Reviews in Computing, Vol. 6, 15th July 2011.

[2] P. Kundur, "Power system stability and control," McGraw-Hill, Inc., 1993.

[3] K. Damor, . D. Patel, V. Agrawal, H. Patel, "Improving Power System Transient Stability by using FACTS Devices," International Journal of Engineering Research & Technology (IJERT), Vol. 3, Issue 7, July 2014.

[4] S. Rath, B. Sahu, P. Dash, "Power System Operation and Control using FACTS Devices," International Journal of Engineering Research & Technology (IJERT), Vol. 1, Issue 5, July 2012.

[5] S. Dheebika, R. Kalaivani, "Optimal Location of SVC, TCSC and UPFC Devices for Voltage Stability Improvement and Reduction of Power Loss using Genetic Algorithm," International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE), 2014.

[6] Sh. Ahmad, F.M. Albatsh, S. Mekhilef, H. Mokhlis, "An Approach to Improve Active Power Flow Capability by Using Dynamic Unified Power Flow Controller," *IEEE Trans*., 2014.

[7] A. Gupta, P.R. Sharma, "Static and Transient Stability Enhancement of Power System by Optimally Placing UPFC (Unified Power Flow Controller)," *IEEE trans.*, 2013.

[8] H. Jmii, A. Meddeb, S. Chebbi, "Efficiency Limits of SVC in Improving Voltage Stability," 3rd Inter. conf. (GEEE), April, 2016, Hammamet, Tunisia.

[9] P. Sivachandran, T. Hariharan, R. Pushpavathy, "Improvement of Power System Stability using D-FACTS Controllers: A Review," ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 2, February 2015.

[10] B. Rao, P. Chanti, N. Lavanya, S. Sekhar, Y. kumar, "Power System Stability Enhancement Using Fact Devices," International Journal of Engineering Research and Applications, pp.339-344, Vol. 4, Issue 4(Version 1), April 2014.

[11] M. Kamarposhti, M. Alinezhad, H. Lesani, Nemat Talebi, "Comparison of SVC, STATCOM, TCSC, and UPFC Controllers for Static Voltage Stability Evaluated by Continuation Power Flow Method," IEEE Electrical Power & Energy Conference, 2008.

[12] A. Motiebirjandi, D. Fateh, "Optimal placement method of multi UPFCs to damp power system oscillations," International Transactions on Electrical Energy System, 2017.

[13] M. Benaissa, S. Hadjeri, S.A. Zidi, "Impact of PSS and SVC on the Power System Transient Stability," Advances in Science, Technology and Engineering Systems Journal, pp. 562-568, Vol. 2, No. 3, 2017.

[14] A. Meddeb, H. Jmii, S. Chebbi, "UPFC and SVC Devices for Transient Stability Enhancement," 4th International Conference on Automation, Control Engineering and Computer Science (ACECS) Proceedings of Engineering and Technology – PET, pp.82-87, Vol.19, 2017.

[15] S. Titus, B.J. Vinothbabu, I. Nishanth, "Power System Stability Enhancement under Three Phase Fault with FACTS Devices TCSC, STATCOM and UPFC," International Journal of Scientific and Research Publications, Vol. 3, Issue 3, March 2013.

[16] B. Bhattacharyya, V.K. Gupta, S. Kumar, "UPFC with series and shunt FACTS controllers for the economic operation of a power system," Ain Shams Engineering Journal, 2014.

[17] S. Varma, "FACTS Devices for Stability Enhancements," International Conference on Green Computing and Internet of Things, 2015.

[18] P. Pandey, B. Bag, "A Comparative study on UPFC and SVC towards Voltage Profile improvement of a Grid Connected Distributed Generation System," *IEEE tran.s*, 2015.

[19] P. Stefanov, A. Savić, G. Dobrić, "Power System Optimization Using Parallel Scenario Algorithm," IEEE International Energy Conference, 2014.

[20] H. Jmii, A. Meddeb, S. Chebbi, "An approach for improving voltage stability by combination of SVC and TCSC," 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp.134–141, 2016.

[21] M.A. Abido, "Power System Stability Enhancement using FACTS Controllers: A Review," The Arabian Journal for Science and Engineering, vol.34, no1B, Apr 2009.

[22] X.P. Zhang, C. Rehtanz, B. Pal, "Flexible AC Transmission Systems: Modeling and Control," 2nd Edition, Springer, Feb 2012.

[23] M. Sadi, S. Ali, "A Comprehensive Analysis Of Transient Stability Enhancement Methods Of Electric Power System," *IEEE trans*, 2015

[24] Characteristics of voltage in public distribution networks, Standard EN 50160, 2007.

[25] M. Zellagui, H.A. Hassan, A. Chaghi, A. Ghorbani, "A Comparative Study of Ground Fault Analysis for a Practical Case of a Transmission Line Equipped with Different Series FACTS Devices," Automatika Journal, pp 262–274, 2015.

[26] D. Murali, M. Rajaram, N. Reka, "Comparison of FACTS Devices for Power System Stability Enhancement," International Journal of Computer Applications, Vol. 8, No.4, October 2010.

[27] Y. Amrane, M. Boudour, M. Belazzoug, "A new hybrid technique for power systems multi-facts optimization design," International Transactions on Electrical Energy Systems, 2014.

[28] C.V. The., K.L. Minh., T.T. Quoc, N.B. Khue., "FACTS devices applications on power system to improve the angle stability," IEEE, 2006.

[29] S. Varma, "FACTS Devices for Stability Enhancements," *IEEE trans.*, 2015.

[30] P. Kumkratug, "Application of Lyapunov Theory and Fuzzy Logic to Control Shunt FACTS Devices for Enhancing Transient Stability in Multimachine System," Journal of Electrical Engineering & Technology, vol. 7, no. 5, pp. 672-680, 2012.

[31] M. Haque, "Evaluation of First Swing Stability of a Large Power System with Various FACTS Devices," IEEE Transactions on Power Systems, Vol. 23, No. 3, August 2008.

[32] F. Milano, "Power System Modelling and Scripting," Springer, ETSII, University of Castilla - La Mancha 13071, Ciudad Real Spain 2010.

[33] R. Dubey, S. Dixit, G. Agnihotri, "Optimal Placement of Shunt FACTS Devices Using Heuristic Optimization Techniques: An Overview," International Conference on Communication Systems and Network Technologies, 2014.

[34] R. Marti, G. Reinelt, "Exact and Heuristic Methods in Combinqtorial Optimization,"Available:https://books.google.tn/books?id=4Lk8pPYYI3cC&printsec=frontcover&hl=fr#v=onepage&q&f=false.

[35] Eurostag, Eurostag Software Release Notes, Tractebel-EDF, Release 5.1, Dec 2010.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 261-266 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Enhancement of Voltage Stability through Optimal Location of UPFC

Hajer Jmii*, Asma Meddeb, Souad Chebbi

LATICE Laboratory, University of Tunis, Tunis 1008, Tunisia

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 02 December, 2017
Online: 23 December, 2017

Keywords:
Numerical method
Optimal location
UPFC
Voltage Stability

ABSTRACT

This paper proposed the identification of the proper placement of UPFC (Unified Power Flow Controller) using a series of methodological numeric simulations to improve voltage stability. For this purpose, the critical zone of the system is determined, then, comparative analyses depending on different emplacements of UPFC are performed. The dynamic model of UPFC and the proposed method are tested on the IEEE 14-bus system. The finding shows that the proper location of UPFC helps in improving voltage profiles and increasing the maximum loading capacity.

1. Introduction

Power systems have been forced to operate close to their stability limits and around their full capacities. This is due to the economic and environmental pressures imposed on the power generation and the construction of new transmission lines [1-4]. On the other hand, the amount of power transmitted across transmission lines is limited for safety and stability considerations. The system is prone to be collapsed when the power flowing exceeds the limits in case of an arbitrary incident [5]. System planners are so more and more focusing on how to manage the obstacles facing the power transmission.

Over the last decades, FACTS (Flexible AC Transmission Systems) devices have opened new perspectives for managing the transmission of the power [6]. Different types of FACTS controllers are available for this aim, but the most powerful and prominent one is UPFC. UPFC is able to control the three parameters of the transmission system: phase angle, impedance, and voltage. In addition, it provides voltage support, improves transient stability and amortizes oscillations [7, 8].

However, according to references [9, 10], the major disadvantage of UPFC is the expensive cost of installation and reactive power generation which is due to the great efforts of its voltage source converters. Therefore, an optimal location of the device helps in achieving good performances and enhancing power system stability with a reduced investment cost [11].

Previous studies have been focused on the optimal placement of UPFC and the evaluation of its impact on system stability using various methods and techniques [12-14]. The authors in reference [15] presented an evolutionary algorithm-based approach to identify the optimal placement and settings of the hybrid device. Reference [16] has investigated the use of UPFC to increase the load ability margin and improve the system stability. The emplacement of the FACTS is decided by the stability indices which determine the critical line, the appropriate line for introducing UPFC. In reference [17], H. Wang investigated the impact of UPFC on the stability of a multi-machine power system. A. Ray *et al.* [18] studied the application of UPFC for controlling power flow in the transmission line and improving the voltage profile. They computed the VSI (Voltage Stability Index) for each line to integrate UPFC into the line which has got the highest VSI value. In reference [19], authors opted for the CPF (Continuation Power Flow) and LSI (Line Stability Index) for the identification of the suitable location of UPFC. Through simulation results, it was demonstrated that optimal placement of the hybrid FACTS improves the static and transient stability of power system when subjected to large disturbances. S. Lee *et al.* [20] addressed the determination of UPFC location and rating for voltage stability improvement using the analysis of the Jacobian matrix. Intelligent techniques were applied in reference [21] for the optimal position of UPFC along with series FACTS to minimize the operation cost of the power system. In [22], authors proposed two evolutionary optimization techniques to optimally place and size the hybrid controller in view of reducing the active power losses.

This paper is an extension of work initially published in the Proceedings of the International Conference on Control Engineering &Information Technology [23]. In this work, we proceed with an efficient method for the identification of the proper placement of UPFC, for the purpose of improving voltage

*Hajer Jmii, University of Tunis, Tunis 1008, Tunisia, +21694014865, hajerjmii@yahoo.com

stability. We plan a series of methodological dynamic simulations to find the critical zone of IEEE 14-bus case study. The zone having the lowest voltage magnitudes is taken as the best location of UPFC. Then, we perform comparative studies to justify the placement selected. The dynamic model of UPFC and the numerical simulation method are implemented within the software EUROSTAG.

2. Necessity of FACTS devices

Consider a single machine infinite bus system interconnected through a transmission line having reactance X, as shown in Figure 1. The power transferred across the transmission line is given in equation (1) [24, 25].

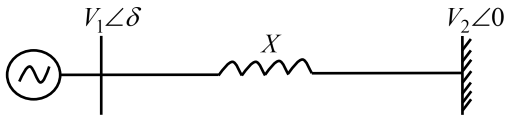


Figure 1. Single machine infinite bus system

$$P = \frac{V_1 V_2}{X}\sin\delta \quad (1)$$

V_1 and V_2 are the voltages of the synchronous generator and the infinite bus respectively where: $V_1\angle\delta$ and $V_2\angle 0°$.

The power variation is a sinusoidal function of the difference between the two voltage angles as shown in Figure 2. The maximum power that can be transmitted across the line corresponds to $\delta = 90°$ and by considering $V_1 = V_2 = V$, it is given by:

$$P_{\max} = \frac{V_2}{X} \quad (2)$$

$P_{\max}$ depends on the value of the reactance X which imposes the theoretical limit of the steady state transmitted power. However, the line resistance R creates power losses quantified by RI^2 and thus sets the practical limit of the line. The physical characteristics of the conductor could also limit the power transmission, which is known as the thermal limit.

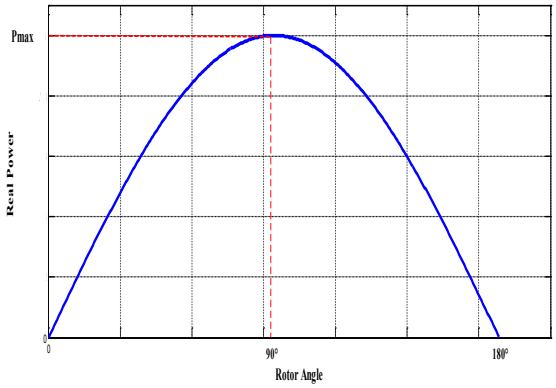


Figure 2. Power-angle curve

From equation (1), we note that the power can be controlled by two parameters; it can be enhanced by increasing the voltage magnitude of the synchronous machine and the infinite bus or by increasing the angular difference of the voltages. However, there are limits imposed on either the voltage magnitude or the angular difference, which have to be considered. The voltage variation must not exceed $\pm 5\,\%$ of the nominal voltage while the difference between the angles is limited to less than $30° - 35°$ to maintain the transient stability [26, 27].

Therefore, it is inferred that the power transmission is limited by:

- Thermal limitations.
- Physical limitations.
- Stability.

It is possible to deal with the thermal limitations by renovating the transmission lines and opting for a higher current rating conductor. Nonetheless, this solution does not necessarily guarantee voltage within the acceptable boundaries or the flow and the controllability of the power, but it is feasible by line compensation. Electro-mechanical devices used for line compensation cannot achieve rapid compensation and are prone to the wear quicker than the static equipments. The solid-state based technology, FACTS, provides the control of one or more parameters to magnify the loading capability and to develop controllability. As the current in a transmission line has the property to be controlled, it is possible to use a FACTS device for flowing power across the line during normal or disturbed conditions. This returns to the ability of FACTS devices to control all power flow parameters, namely phase angle, bus voltage and line impedance. In other words, the FACTS technology gives the possibility to maintain acceptable voltage magnitude and power flow.

3. UPFC

The UPFC is a hybrid device consisting of two voltage source converters interconnected through a common DC link capacitor, as shown in Figure 3 [28-30]. The inverter 1 is connected in shunt with the transmission line while the inverter 2 is series connected. It injects a series connected. It injects a series voltage controllable in phase and magnitude. The decomposition of this voltage leads to a quadrature component responsible for controlling the real power and an in-phase component which controls the reactive power in the line. Similarly, the injected shunt voltage is decomposed into two components. The quadrature component regulates the voltage of the DC link as for the in-phase component; it controls the exchange of the reactive power with the system.

The model of UPFC is represented in Figure 4 [30, 31]. From this scheme, we can extract the active and reactive power flows of the series and shunt converters which are expressed by the equations (3) - (6).

$$P = \frac{VV_k}{(1+K)X}\sin(\theta_{jk} + \theta) \quad (3)$$

$$Q = \frac{1}{(1+K)^2 X}[V^2 - VV_k\cos(\theta_{jk} + \theta)] \quad (4)$$

$$-\frac{1}{(1+K)^2 X}[{V_k}^2 - VV_k\cos(\theta_{jk} + \theta)]$$

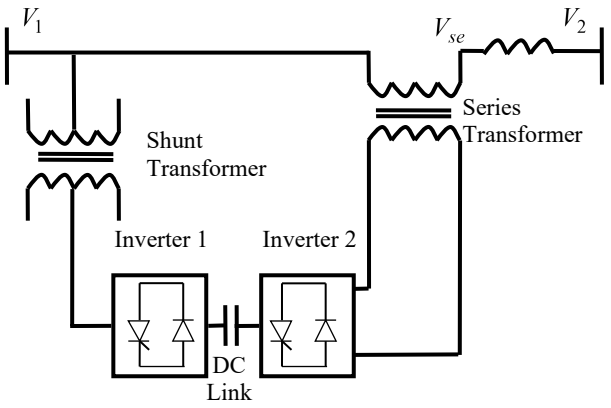


Figure 3. UPFC incorporated in a power system

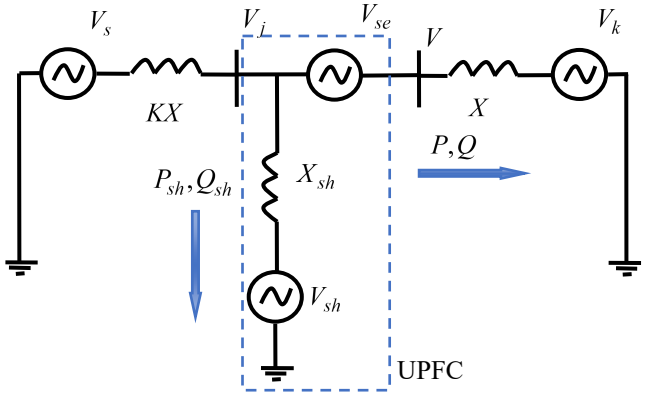


Figure 4. Model of UPFC

Where:

$$V^2 = V_j^2 + V_{se}^2 + 2V_jV_{se}\cos\phi_{se}\ \ \phi_{se} = \tan^{-1}\frac{\gamma}{\beta},\ \ \theta = \tan^{-1}\frac{\gamma}{1+\beta}$$

Similarly, the active and reactive powers of the shunt inverter are expressed by:

$$P_{sh} = \frac{V_jV_{sh}}{X_{sh}}\sin\varepsilon \quad (5)$$

$$Q_{sh} = \frac{V_j^2}{X_{sh}} - \frac{V_jV_{sh}}{X_{sh}}\cos\varepsilon \quad (6)$$

Where: $\varepsilon = \tan^{-1}\frac{\xi}{1+\eta}$ and $V_{sh}^2 = (1+\eta)^2V_j^2 + \xi^2V_j^2$

KX: Short circuit level;

$\phi(x)$: Gaussian function

γ : Relative quadrature series voltage;

β : Relative in-phase series voltage;

ξ : Relative quadrature shunt voltage;

η : Relative in-phase shunt voltage;

V : Right YPFC voltage;

V_{se} : Series voltage;

V_{sh} : Shunt voltage;

P : Real power;

Q : Reactive power;

P_{sh} : Shunt real power;

Q_{sh} : Shunt reactive power.

The modeling of the UPFC shunt part on EUROSTAG is simple; it is represented by a current injector. As for the modeling of the series part, we must open the line where we want to insert UPFC and place at its extremities two current injectors (Figure 5). The opening of the line is assured by a high reactance value.

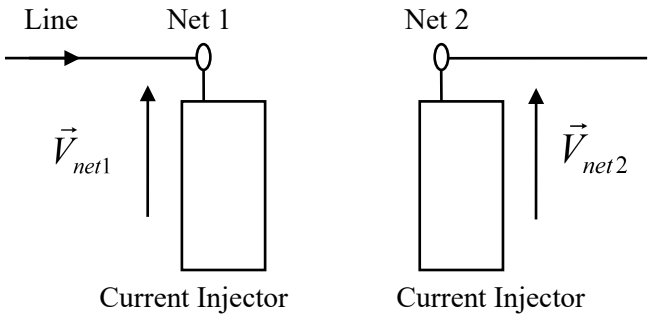


Figure 5. Model of UPFC in EUROSTAG

4. Simulation

4.1. Case Study

To test the proposed method, we used the IEEE 14-bus test system. This system contains two generators each one is equipped with voltage and speed regulators and three synchronous compensators to produce reactive power. It also has two transformers with two windings, a three-winding transformer, fifteen transmission lines and eleven loads. The data of the above mentioned system are taken from reference [31].

EUROSTAG [32] adopts the model considering the voltage behind transient reactance for representing all the generators and constant impedance to model loads and transformers.

4.2. Simulations Results

4.2.1. Identification of the Critical Zone:

The IEEE 14-bus network can be decomposed into a transmission system (69 KV) referred by « zone A » and a distribution one (13.8 KV) referred by « zone B » interconnected through three step-down transformers. To identify the critical zone, we proceeded by an incremental load increase of each zone separately. The part of the network experiencing voltage collapse rapidly is considered as the critical zone.

According to the steady state, we noted that bus 14 is the most insecure bus; it has got the lowest magnitude compared to the other buses. For this reason, we choose to adopt this bus as an index for voltage stability assessment.

Figure 6 shows the main results of increasing load. In zone A, the voltage of bus 14 underwent a progressive decrease of 0.002pu from one case to another until collapsing in the case of 49%. As for zone B, it could not withstand over than 39% of load increase where bus 14 has got the limit voltage amplitude. While the increase of the total system load leads to a severe voltage drop which ends with a collapse when reaching 25%.

Noteworthy that distribution system is the critical zone. All thanks to voltage and speed regulators of synchronous machines which reinforce transmission system stability.

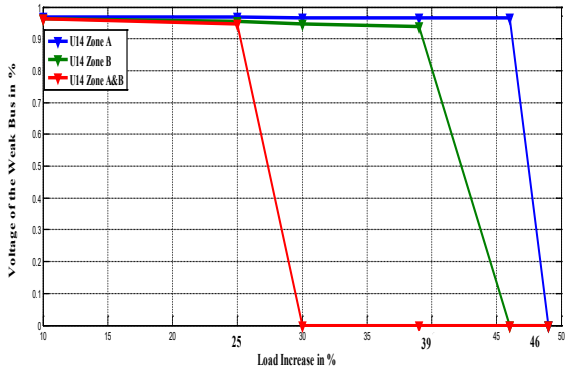


Figure 6. Identification of the critical zone of IEEE 14-Bus system.

4.2.2. UPFC Application:

Having identified the critical zone, we placed UPFC between buses 9 and 14, sized at 60MVAR for each one of its converters. Line 9-14 is close to the weakest bus so that we expect satisfying results when introducing UPFC into it. To evaluate the contribution of the controller, we rerun the system under 20 % of load increase at time t=250s.

As we can see in Figure 7, UPFC was able to considerably increase bus14 voltage amplitude just after perturbation and with well-damped oscillations compared to the uncompensated system. This result indicates the voltage support supplied by the UPFC. Note that we obtained similar results for the other buses of the network.

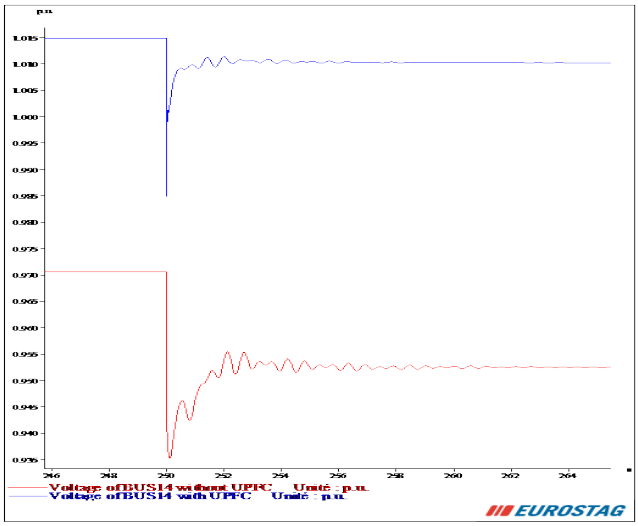


Figure 7. Temporal evolution of voltage at bus14 for 20 % of load increase with and without UPFC

The test network along with UPFC at different locations is tested now for comparison purpose. In addition to line 9-14, we choose two placements in the transmission system: lines 1-2 and 2-3 which are intended for transmitting the highest amount of reactive and active powers. Thereby, enhancing power flowing in these lines can be effective in improving voltage stability of the whole network. Figure 8 and Figure 9 illustrate the temporal evolution of voltage at bus1 and bu2 respectively. It should be noted that the implementation of UPFC in line 1-2 has considerably damped voltage oscillations compared to the other cases. While, this location had no effect on the voltage magnitude as shown in Figure 10. UPFC in line 9-14 gave best voltage levels at all network buses than when it is incorporated in line 2-3, where we obtained the lowest voltage magnitudes.

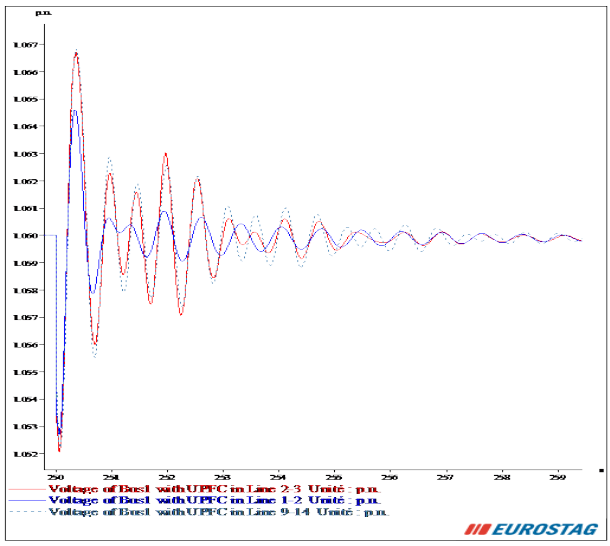


Figure 8. Temporal evolution of voltage at bus 1 for 20% of load increase with different locations of UPFC.

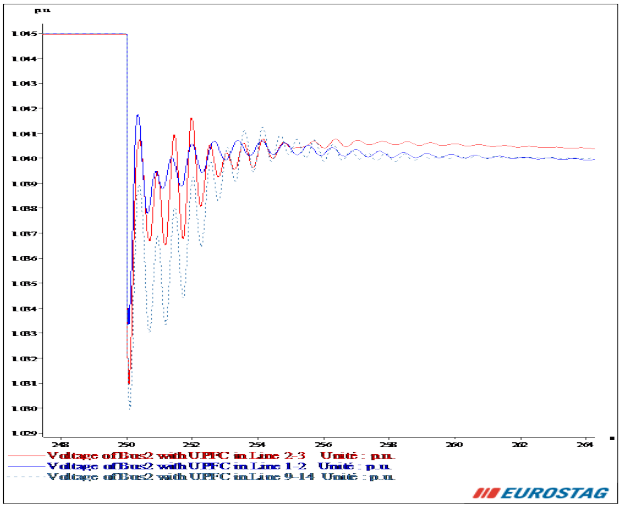


Figure 9.Temporal evolution of voltage at bus 2 for 20% of load increase with different locations of UPFC.

Figure 11 illustrates the active power losses of the system. It is clearly observable that the active power losses had the lowest value in the case of UPFC inserted through line 9-14, it is about 13.5MW. However, a large amount of active losses is detected with line 2-3 placement and about 14 MW in line 1-2 case.

Now, we evaluate the impact of UPFC location on the loadability margin of the power system. From Table 1, we can note that the contribution of the controller in improving the maximum loading capability was very meaningful when installing in the distribution system. In fact, it was able to increase the loadability margin in zone A from 48% to 59% and from 39% to 70% in zone B. As for the whole network, it became able to

maintain a stable state up to 33.5% of load increase while its maximum loadability was about only 25% before compensation. Nevertheless, when introduced in the transmission system, UPFC was less effective in increasing loadability margin. That was observable notably in line 2-3, where only the loading capacity of zone A; reached a higher percentage compared to the case without UPFC.

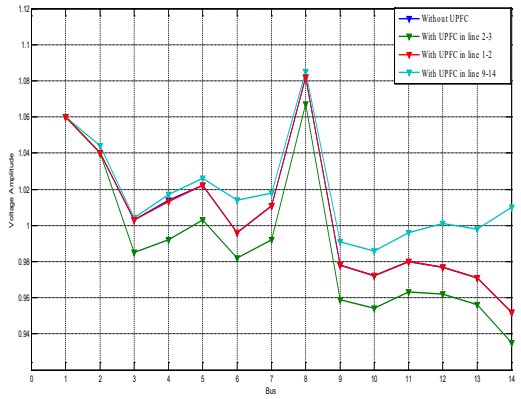


Figure 10.Voltage profiles of IEEE 14-Bus system for 20% load increase with different locations of UPFC

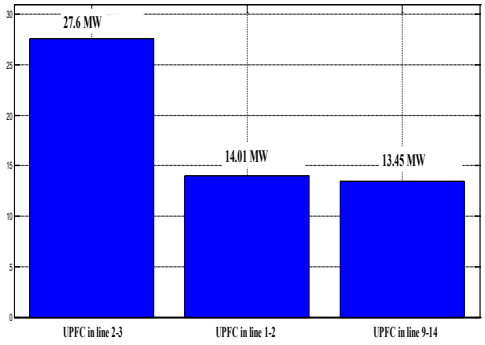


Figure 11. Active power losses for 20% load increase with Different UPFC Locations.

Table 1. Load ability Margins with Different Locations of UPFC.

Network Zone	Loadability Margin in %			
	Without UPFC	With UPFC in line 2-3	With UPFC in line 1-2	With UPFC in line 9-14
Zone A	48	52	55	59
Zone B	39	39	39	70
Zones A&B	25	25	32	33.5

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

5. Conclusion

In this paper, the impact of the proper installation of UPFC on improving voltage stability has been investigated. The placement of UPFC is done by using a numerical simulation based method. The identification of the critical zone of the system can determine the best location of the hybrid device. Then, the performance of the proper placement is compared to different selected locations in the system for justification purpose. It was found that UPFC incorporated through line 9-14 in the critical zone, helps in improving voltage magnitude, reducing active power losses as well as increasing maximum loading capacity of the system. However, the integration of the controller in the transmission system gave rapid and well- damped oscillations. Thereby, the optimal placement of the UPFC is highly dependent on the objective of the integration.

References

[1] H. Jmii, A. Meddeb, S. Chebbi, "Efficiency Limits of SVC in Improving Voltage Stability,"3rd Inter, conf, GEEE, April, 2016, Hammamet, Tunisia.

[2] B. Singh, "Applications of FACTS Controllers in Power Systems for Enhance the Power System Stability: A State-of-The-Art," International Journal of Reviews in Computing, Vol. 6, 15th July 2011.

[3] A. Gupta, P.R. Sharma," Optimal location and setting of SVC and TCSC devices using non-dominated sorting particle swarm optimization," Electric Power Systems Research, 2009.

[4] P. Kundur., "Power system stability and control,"McGraw-Hill, Inc., 1993.

[5] S. Amara, H. Hsan, "Power system stability improvement by FACTS devices: a comparison between STATCOM, SSSC and UPFC," 1st International Conference on Renawble Energies and Vehicular Technology, 2012.

[6] M.A. Abido, "Power System Stability Enhancement using FACTS Controllers: A Review," The Arabian Journal for Science and Engineering, vol.34, no1B, Apr 2009.

[7] Gyugyi L., " Unified power flow control concept for flexible AC transmission systems," Proceedings of IEE Part-C IEE Proceedings of Generation, Transmission and Distribution, 139:323–331, 1992.

[8] J. Singh, H. Qazi, M. Ghandhari, " Load curtailment minimization by optimal placement of Unified Power Flow Controller," International Transactions on Electrical Energy systems, 2016.

[9] V.G. Mathad, B.F. Ronad, S.H. Jangamshetti, "Review on Comparison of FACTS Controllers for Power System Stability Enhancement," International Journal of Scientific and Research Publications, vol. 3, Issue 3, March 2013

[10] X.P. Zhang, Ch. Rehtanz, B. Pal, " Flexible AC Transmission Systems: Modeling and Control," 2nd Edition, Springer, Feb 2012.

[11] Q.H. Wu, Z. Lu, M.S. Li, T.Y. Ji, " Optimal Placement of FACTS Devices by A Group Search Optimizer with Multiple Producer," *IEEE trans*, 2008.

[12] S. Gupta, R. Tripathi, R. Shukla, " Voltage stability improvement in power systems using FACTS controllers: State-of-the-art review," https://www.researchgate.net/publication/251987213, 2010.

[13] M. Benaissa, S. Hadjeri, S.A. Zidi, "Impact of PSS and SVC on the Power System Transient Stability," Advances in Science, Technology and Engineering Systems Journal, pp. 562-568 , Vol. 2, No. 3, 2017.

[14] M. Sanjari, O. Mousavi, G. Gharehpetian, "Assessing the risk of blackout in the power system including HVDC and FACTS devices," European Transactions on Electrical Power, 2012.

[15] H. Shaheen, G. Rashed, S. Cheng, "Application of differential evolution algorithm for optimal location and parameters setting of UPFC, 19:911–932, 2009.

[16] A. Jupta, P.R. Sharma,"Optimal Placement of FACTS Devices for Voltage Stability Using Line Indicators," *IEEE trans*, 2012.

[17] H. Wang,"Applications of modelling UPFC into multi-machine power systems," IEE Proceedings online no. 19990170, DOI: 10.1049/ip-gtd:19990170, 1999.

[18] A. Ray, J.O. Chandle , "Voltage Stability Enhancement during Excess Load Increments through Optimal Location of UPFC Devices," IEEE

International Conference on Technological Advancements in Power & Energy, 2015.

[19] A. Jupta, P.R. Sharma, " Static and transient stability enhancement of power system by optimally placing UPFC (Unified Power Flow Controller)," *IEEE trans*, 2013.

[20] S. Lee, J. Liu, C. Chu, " Modelling and locating unified power-flow controllers for static voltage stability enhancements," International Transactions on Electrical Energy systems, 2012.

[21] B. Bhattacharyya, V.K. Gupta, S. Kumar, "UPFC with series and shunt FACTS controllers for the economic operation of a power system," Ain Shams Engineering Journal, 2014.

[22] H.I. Shaheen, G.I. Rashed, S.J.Cheng, "Optimal Location and Parameters Setting of Unified Power Flow Controller Based on Evolutionary Optimization Techniques," *IEEE trans*, 2007.

[23] H. Jmii, A. Meddeb, S. Chebbi, "Proper Placement of UPFC for the Improvement of Voltage Stability of IEEE-14 Bus System," 4th International Conference on Control Engineering & Information Technology (CEIT) Tunisia, Hammamet, December 2016.

[24] R. Mathur, R. Varma, "Thyristor-Based FACTS Controllers for Electrical Transmission Systems," J. Wiley & sons.

[25] S. Rath, B. Sahu, P. Dash, "Power System Operation and Control using FACTS Devices," International Journal of Engineering Research & Technology (IJERT), Vol. 1, Issue 5, July 2012.

[26] I. Khan, M. Mallick, M. Rafi, M. Mirza, "Optimal placement of FACTS controller scheme for enhancement of power system security in Indian scenario," Journal of Electrical Systems and Information Technology 2, pp.161–171, 2015.

[27] Aboreshaid, S.,Billinton, R., "Probabilistic evaluation of voltage stability," IEEE Trans.PowerSyst.14(February(1)), pp. 342–348, 1999.

[28] S. Morsli, "Robust Control of UPFC in MultiMachines Power Network," M.S. thesis, Dept. Electron. Eng., Oran Univ., Algeria, 2012.

[29] A. Meddeb, H. Jmii, S. Chebbi, " Comparison of UPFC, TCSC and SVC for Improving Voltage Stability," 3rd Inter, conf, ACECS , March, 2016, Hammamet, Tunisie.

[30] M. Farrag, G. Putrus, "An on-line training radial basis function neural network for optimum operation of the UPFC," EUROPEAN TRANSACTIONS ON ELECTRICAL POWER, 21, pp.27–39, 2010.

[31] F. Milano, "Power System Modelling and Scripting," Springer, ETSII, University of Castilla - La Mancha 13071, Ciudad Real Spain 2010.

[32] Eurostag, Eurostag Software Release Notes, Tractebel-EDF, Release 5.1, Dec 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 261-273 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Short CCA-Secure Attribute-Based Encryption

Hiroaki Anada[*,1], Seiko Arita[2]

[1]*Department of Information Security, University of Nagasaki, 851–2195, Japan*

[2]*Graduate School of Information Security, Institute of Information Security, 221–0835, Japan*

ARTICLEINFO

Article history:
Received: 16 November, 2017
Accepted: 16 January, 2018
Online: 31 January, 2018

Keywords:
Role-based access control
Attribute-based encryption
Direct chosen-ciphertext security
Twin Diffie-Hellman

ABSTRACT

Chosen-ciphertext attacks (CCA) are typical threat on public-key encryption schemes. We show direct chosen-ciphertext security modification in the case of attribute-based encryption (ABE), where an ABE scheme secure against chosen-plaintext attacks (CPA) is converted into an ABE scheme secure against CCA by individual techniques. Our modification works in the setting that the Diffie-Hellman tuple to be verified in decryption is in the target group of a bilinear map. The employed techniques result in expansion of the secret-key length and the decryption cost by a factor of four, while the public-key and the ciphertext lengths and the encryption cost remain almost the same.

1 Introduction

Access control is one of the fundamental processes and requirements in cybersecurity. Attribute-based encryption (ABE) invented by Sahai and Waters [1], where attributes mean authorized credentials, enables to realize access control which is functionally close to role-based access control (RBAC), but *by encryption*. In key-policy ABE (KP-ABE) introduced by the subsequent work of Goyal, Pandey, Sahai and Waters [2], a secret key is associated with an access policy over attributes, while a ciphertext is associated with a set of attributes. In a dual manner, in ciphertext-policy ABE (CP-ABE) [2, 3, 4], a ciphertext is associated with an access policy over attributes, while a secret key is associated with a set of attributes. In a KP-ABE or CP-ABE scheme, a secret key works to decrypt a ciphertext if and only if the associated set of attributes satisfies the associated access policy. The remarkable feature of ABE is attribute privacy; that is, in decryption, no information about the access policy and the identity of the secret key owner in the case of KP-ABE (or, the attributes and the identity of the secret key owner in the case of CP-ABE) leaks except the fact that the set of attributes satisfies the access policy. Since the proposals, it has been studied to attain certain properties such as indistinguishability against chosen-plaintext attacks (IND-CPA) in the standard model [4] and adaptive security against adversary's choice of a target access policy [5].

In this paper[1], we work through resolving a problem of constructing a shorter ABE scheme that attains indistinguishability against chosen-ciphertext attacks (IND-CCA) in the standard model. Here CCA means that an adversary can collects decryption results of ciphertexts of its choice through adversaries' attacking. Note that "provable security" of a cryptographic primitive is now a must requirement when we employ the primitive in a system, where it means that an appropriately defined security is polynomially reduced to the hardness of a computational problem. Moreover, the CCA security of an encryption scheme is preferable to attain because the CCA security is one of the theoretically highest securities and hence the scheme can be used widely.

To capture the idea of our approach, let us recall the case of identity-based encryption (IBE). The CHK transformation of Canetti, Halevi and Katz [7] is a generic tool for obtaining IND-CCA secure IBE scheme. It transforms any hierarchical IBE (HIBE) scheme that is selective-ID IND-CPA secure [8] into an IBE scheme that is adaptive-ID IND-CCA secure [8]. A point of the CHK transformation is that it introduces a dummy identity vk that is a verification key of a one-time signature. Then a ciphertext is attached with vk and a signature σ, which is generated each time one executes encryption. In contrast, the direct chosen-ciphertext security technique for IBE of

[*]Hiroaki Anada, 1–1–1, Manabino, Nagayo-cho, Nishisonogi-gun, Nagasaki, +81-95-813-5113 & anada@sun.ac.jp

[1]This paper is an extension of the work originally presented in SMARTCOMP 2017 [6]. The schemes of KP-ABKEM and KP-ABE have been newly proposed.

https://dx.doi.org/10.25046/aj030132

Boyen, Mei and Waters [9] is individual modification for obtaining an IND-CCA secure IBE scheme. It converts a HIBE scheme that is adaptive-ID IND-CPA secure into an IBE scheme that is adaptive-ID IND-CCA secure. Though the technique needs to treat each scheme individually, the obtained scheme attains better performance than that obtained by the generic tool (the CHK transformation). Let us transfer into the case of ABE. The transformation in [10] is a generic tool for obtaining IND-CCA secure ABE scheme. It transforms any ABE scheme (with the delegatability or the verifiability [10]) that is IND-CPA secure into an ABE scheme that is IND-CCA secure. A point of their transformation is, similar to the case of IBE, that it introduces a dummy attribute vk that is a verification key of a one-time signature. Then a ciphertext is attached with vk and a signature σ. Notice here that discussing direct chosen-ciphertext security modification for ABE (in the standard model) is a missing piece. One of the reasons seems that there is an obstacle that a Diffie Hellman tuple to be verified is in the target group of a bilinear map. In that situation, the bilinear map looks of no use.

1.1 Our Contribution

A contribution is that we fill in the missing piece; we demonstrate direct chosen-ciphertext security modification in the case of the Waters CP-ABE scheme [4] and the KP-ABE scheme of Ostrovsky, Sahai and Waters [11] To overcome the above obstacle, we employ the technique of the Twin Diffie-Hellman Trapdoor Test of Cash, Kiltz and Shoup [12]. In addition, we also utilize the algebraic trick of Boneh and Boyen [13] and Kiltz [14] to reply for adversary's decryption queries.

1.2 Related Works

Waters [4] pointed out that IND-CCA security would be attained by the CHK transformation. Gorantla, Boyd and Nieto [15] constructed a IND-CCA secure CP-ABKEM in the random oracle model. In [10] the authors proposed a generic transformation of a IND-CPA secure ABE scheme into a IND-CCA secure ABE scheme. Their transformation is considered to be an ABE-version of the CHK transformation, and it is versatile. Especially, it can be applied to non-pairing-based scheme.

The Waters CP-ABE [4] can be captured as a CP-ABKEM: the blinding factor can be considered as a random one-time key. This Waters CP-ABKEM is IND-CPA secure because the Waters CP-ABE is proved to be IND-CPA secure. For theoretical simplicity, we demonstrate an individual conversion of the Waters CP-ABKEM into a CP-ABKEM which is IND-CCA secure. Then we provide a CP-ABE scheme which is IND-CCA secure. As for KP-ABE, we demonstrate an individual conversion of KP-ABKEM of Ostrovsky, Sahai and Waters [11], which is IND-CPA secure, into a KP-ABKEM which is IND-CCA secure. Then we provide a KP-ABE scheme which is IND-CCA secure.

Finally, we note that there is a remarkable work of CP-ABE schemes and KP-ABE schemes with *constant-size* ciphertexts [16, 17]. Our direct chosen-ciphertext security modification is not constant-size ciphertexts but a different approach for easier implementation in engineering.

1.3 Organization of the Paper

In Section 2, we survey concepts, definitions and techniques needed. In Section 3, we revisit the concept, the algorithm and the security of the twin Diffie-Hellman technique. In Section 4, we construct a CCA-secure CP-ABKEM from the Waters CPA-secure CP-ABKEM [4], and provide a security proof. Also, we describe the encryption version, a CCA-secure CP-ABE. In Section 5, we construct a CCA-secure KP-ABKEM from the Ostrovsky-Sahai-Waters CPA-secure KP-ABKEM [11], and provide a security proof. Also, we describe the encryption version, a CCA-secure KP-ABE. In Section 6, we compare efficiency of our CP-ABE and KP-ABE schemes with the original schemes, and also, with the schemes obtained by applying the generic transformation [10] to the original schemes. In Section 7, we conclude our work.

2 Preliminaries

The security parameter is denoted λ. A prime of bit length λ is denoted p. A multiplicative cyclic group of order p is denoted $\mathbb{G}$. The ring of exponent domain of $\mathbb{G}$, which consists of integers from 0 to $p-1$ with modulo p operation, is denoted $\mathbb{Z}_p$.

2.1 Bilinear Map

We remark first that our description in the subsequent sections is in the setting of a symmetric bilinear map for simplicity, but we can employ an asymmetric bilinear map instead for better efficiency as is noted in Section 6. Let $\mathbb{G}$ and $\mathbb{G}_T$ be two multiplicative cyclic groups of prime order p. Let g be a generator of $\mathbb{G}$ and e be a bilinear map, $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$. The bilinear map e has the following properties:
1. Bilinearity: for all $u, v \in \mathbb{G}$ and $a, b \in \mathbb{Z}_p$, we have $e(u^a, v^b) = e(u, v)^{ab}$.
2. Non-degeneracy: $e(g, g) \neq \mathrm{id}_{\mathbb{G}_T}$ (: the identity element of the group $\mathbb{G}_T$).

Parameters of a bilinear map are generated by a probabilistic polynomial time (PPT) algorithm $\mathtt{Grp}$ on input λ: $(p, \mathbb{G}, \mathbb{G}_T, g, e) \leftarrow \mathtt{Grp}(\lambda)$.

Hereafter we assume that the group operation in $\mathbb{G}$ and $\mathbb{G}_T$ and the bilinear map $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$ are computable in PT in λ.

2.2 Access Structure

Let $\mathcal{U} = \{\chi_1, \ldots, \chi_u\}$ be a set of attributes, or simply set $\mathcal{U} = \{1, \ldots, u\}$ by numbering. An *access structure*, which corresponds to an access policy, is defined as a collection $\mathbb{A}$ of non-empty subsets of $\mathcal{U}$; that is, $\mathbb{A} \subset 2^{\mathcal{U}} \setminus \{\phi\}$. An access structure $\mathbb{A}$ is called *monotone* if for any $B \in \mathbb{A}$ and $B \subset C$, $C \in \mathbb{A}$ holds. The sets in $\mathbb{A}$ are called authorized sets, and the sets not in $\mathbb{A}$ are called unauthorized sets. We will consider in this paper only monotone access structures.

2.3 Linear Secret-Sharing Scheme

We only describe a linear secret-sharing scheme (LSSS) in our context of attribute-based schemes. A secret-sharing scheme Π over the attribute universe $\mathcal{U}$ is called linear over $\mathbb{Z}_p$ if:
1. The shares for each attribute form a vector over $\mathbb{Z}_p$,
2. There exists a matrix M of size $l \times n$ called the share-generating matrix for Π and a function ρ which maps each row index i of M to an attribute in $\mathcal{U} = \{1, \ldots, u\}$: $\rho : \{1, ..., l\} \to \mathcal{U}$.

To make shares, we first choose a random vector $\vec{v} = (s, y_2, \ldots, y_n) \in \mathbb{Z}_p^n$: s is a secret to be shared. For $i = 1$ to l, we calculate each share $\lambda_i = \vec{v} \cdot M_i$, where M_i denotes the i-th row vector of M and $\cdot$ denotes the formal inner product. LSSS $\Pi = (M, \rho)$ defines an access structure $\mathbb{A}$ through ρ.

Suppose that an attribute set S satisfies $\mathbb{A}$ ($S \in \mathbb{A}$) and let $I_S = \rho^{-1}(S) \subset \{1, \ldots, l\}$. Then, let $\{\omega_i \in \mathbb{Z}_p; i \in I_S\}$ be a set of constants (*linear reconstruction constants*) such that if $\{\lambda_i \in \mathbb{Z}_p; i \in I_S\}$ are valid shares of a secret s according to M, then $\sum_{i \in I_S} \omega_i \lambda_i = s$. It is known that these constants $\{\omega_i\}_{i \in I_S}$ can be found in time polynomial in l: the row size of the share-generating matrix M. If S does not satisfy $\mathbb{A}$ ($S \notin \mathbb{A}$), then no such constants $\{\omega_i\}_{i \in I_S}$ exist.

2.4 Attribute-Based Key Encapsulation Mechanism

Ciphertext-policy attribute-based key encapsulation mechanism (CP-ABKEM). A CP-ABKEM consists of four PPT algorithms (Setup, Encap, KeyGen, Decap)[2].

Setup$(\lambda, \mathcal{U})$**.** A setup algorithm Setup takes as input the security parameter λ and the attribute universe $\mathcal{U} = \{1, \ldots, u\}$. It returns a public key PK and a master secret key MSK.

Encap$(\mathbf{PK}, \mathbb{A})$**.** An encapsulation algorithm Encap takes as input the public key PK and an access structure $\mathbb{A}$. It returns a random string κ and its encapsulation ψ. Note that $\mathbb{A}$ is contained in ψ.

KeyGen$(\mathbf{PK}, \mathbf{MSK}, S)$**.** A key generation algorithm KeyGen takes as input the public key PK, the master secret key MSK and an attribute set S. It returns a secret key SK_S corresponding to S. Note that S is contained in SK_S.

Decap$(\mathbf{PK}, \mathbf{SK}_S, \psi)$**.** A decapsulation algorithm Decap takes as input the public key PK, an encapsulation (we also call it a ciphertext according to context) ψ and a secret key SK_S. It first checks whether $S \in \mathbb{A}$, where S and $\mathbb{A}$ are contained in SK_S and ψ, respectively. If the check result is False, it puts $\hat{\kappa} = \perp$. It returns a decapsulation result $\hat{\kappa}$.

Chosen-Ciphertext Attack on CP-ABKEM. According to previous works (for example, see [15]), the chosen-ciphertext attack on a CP-ABKEM is formally defined as the indistinguishability game (IND-CCA game). In this paper, we consider the *selective game on a target access structure* (IND-sel-CCA game); that is, the adversary $\mathcal{A}$ declares a target access structure $\mathbb{A}^*$ *before* $\mathcal{A}$ receives a public key PK, which is defined as the following experiment.

$\mathbf{Experiment}^{\text{ind-sel-cca}}_{\mathcal{A},\mathtt{CP\text{-}ABKEM}}(\lambda, \mathcal{U})$
$\mathbb{A}^* \leftarrow \mathcal{A}(\lambda, \mathcal{U}), (\mathrm{PK}, \mathrm{MSK}) \leftarrow \mathbf{Setup}(\lambda, \mathcal{U})$
$\epsilon \leftarrow \mathcal{A}^{\mathbf{KeyGen}(\mathrm{PK},\mathrm{MSK},\cdot),\mathbf{Decap}(\mathrm{PK},\mathrm{SK}_\cdot,\cdot)}(\mathrm{PK})$
$(\kappa^*, \psi^*) \leftarrow \mathbf{Encap}(\mathrm{PK}, \mathbb{A}^*), \kappa \leftarrow KeySp(\lambda), b \leftarrow \{0, 1\}$
If $b = 1$ then $\tilde{\kappa} = \kappa^*$ else $\tilde{\kappa} = \kappa$
$b' \leftarrow \mathcal{A}^{\mathbf{KeyGen}(\mathrm{PK},\mathrm{MSK},\cdot),\mathbf{Decap}(\mathrm{PK},\mathrm{SK}_\cdot,\cdot)}(\tilde{\kappa}, \psi^*)$
If $b' = b$ then return Win else return Lose.

In the above experiment, two kinds of queries are issued by $\mathcal{A}$. One is key-extraction queries. Indicating an attribute set S_i, $\mathcal{A}$ queries its key-extraction oracle $\mathrm{KeyGen}(\mathrm{PK}, \mathrm{MSK}, \cdot)$ for the secret key SK_{S_i}. Here we do not require any input attribute sets S_{i_1} and S_{i_2} to be distinct. Another is decapsulation queries. Indicating a pair (S_j, ψ_j) of an attribute set and an encapsulation, $\mathcal{A}$ queries its decapsulation oracle $\mathrm{Decap}(\mathrm{PK}, \mathrm{SK}_\cdot, \cdot)$ for the decapsulation result $\hat{\kappa}_j$. Here an access structure $\mathbb{A}_j$, which is used to generate an encapsulation ψ_j, is implicitly included in ψ_j. In the case that $S \notin \mathbb{A}$, $\hat{\kappa}_j = \perp$ is replied to $\mathcal{A}$. Both kinds of queries are at most q_k and q_d times in total, respectively, which are polynomial in λ.

The access structure $\mathbb{A}^*$ declared by $\mathcal{A}$ is called a *target access structure*. Two restrictions are imposed on $\mathcal{A}$ concerning $\mathbb{A}^*$. In key-extraction queries, each attribute set S_i must satisfy $S_i \notin \mathbb{A}^*$. In decapsulation queries, each pair (S_j, ψ_j) must satisfy $S_j \notin \mathbb{A}^* \vee \psi_j \neq \psi^*$.

The *advantage* of the adversary $\mathcal{A}$ over CP-ABKEM in the IND-CCA game is defined as the following probability:

$$\mathbf{Adv}^{\text{ind-sel-cca}}_{\mathcal{A},\mathtt{CP\text{-}ABKEM}}(\lambda, \mathcal{U}) \stackrel{\text{def}}{=} \Pr[\mathbf{Experiment}^{\text{ind-sel-cca}}_{\mathcal{A},\mathtt{CP\text{-}ABKEM}}(\lambda, \mathcal{U}) \text{ returns Win}].$$

CP-ABKEM is called *selectively secure against chosen-ciphertext attacks* if, for any PPT adversary $\mathcal{A}$ and for any attribute universe $\mathcal{U}$, $\mathbf{Adv}^{\text{ind-sel-cca}}_{\mathcal{A},\mathtt{CP\text{-}ABKEM}}(\lambda, \mathcal{U})$ is negligible in λ. Here we must distinguish the two cases; the case that $\mathcal{U}$ is small (i.e. $|\mathcal{U}| = u$ is bounded by a polynomial of λ) and the case that $\mathcal{U}$ is large (i.e. u

[2] In Gorantla, Boyd and Nieto [15], they say *encapsulation-policy* attribute-based-KEM (EP-AB-KEM) instead of saying ciphertext-policy attribute-based KEM here.

is not necessarily bounded by a polynomial of λ). We assume the *small case* in this paper.

In the indistinguishability game against *chosen-plaintext attack* (IND-CPA game), the adversary $\mathcal{A}$ issues no decapsulation query (that is, $q_{\mathrm{d}} = 0$).

Ciphertext-Policy Attribute-Based Encryption Scheme (CP-ABE). In the case of the encryption version (i.e. CP-ABE), Encap(PK, $\mathbb{A}$) and Decap(PK, SK_S, ψ) are replaced by PPT algorithms Encrypt(PK, $\mathbb{A}$, m) and Decrypt(PK, SK_S, CT), respectively, where m and CT mean a message and a ciphertext, respectively.

The IND-CCA game for CP-ABE is defined in the same way as for CP-ABKEM above, except the following difference. In Challenge phase, the adversary $\mathcal{A}$ submits two equal length messages (plaintexts) m_0 and m_1. Then the challenger flips a coin $b \in \{0,1\}$ and gives an encryption result CT of m_b to $\mathcal{A}$. In Guess phase, the adversary $\mathcal{A}$ returns $b' \in \{0,1\}$. If $b' = b$, then $\mathcal{A}$ wins in the IND-CCA game. Otherwise, $\mathcal{A}$ loses.

Key-Policy Attribute-Based Key Encapsulation Mechanism (KP-ABKEM) and Encryption Scheme (KP-ABE). The *key-policy* case is analogously defined as the case of the ciphertext-policy case. We state only the syntax and the security experiment of the key-policy ABKEM.

Setup($\lambda, \mathcal{U}$)**.** A setup algorithm Setup takes as input the security parameter λ and the attribute universe $\mathcal{U} = \{1, \ldots, u\}$. It returns a public key PK and a master secret key MSK.

Encap(**PK**, S)**.** An encapsulation algorithm Encap takes as input the public key PK and an attribute set S. It returns a random string κ and its encapsulation ψ. Note that S is contained in ψ.

KeyGen(**PK**, **MSK**, $\mathbb{A}$)**.** A key generation algorithm KeyGen takes as input the public key PK, the master secret key MSK and an access structure $\mathbb{A}$. It returns a secret key $\mathrm{SK}_{\mathbb{A}}$ corresponding to S. Note that $\mathbb{A}$ is contained in $\mathrm{SK}_{\mathbb{A}}$.

Decap(**PK**, **SK**$_{\mathbb{A}}$, ψ)**.** A decapsulation algorithm Decap takes as input the public key PK, an encapsulation (we also call it a ciphertext according to context) ψ and a secret key $\mathrm{SK}_{\mathbb{A}}$. It first checks whether $S \in \mathbb{A}$. If the check result is False, it puts $\hat{\kappa} = \perp$. It returns a decapsulation result $\hat{\kappa}$.

Chosen-Ciphertext Attack on KP-ABKEM. The *selective game on a target attribute set* (IND-sel-CCA game) is defined by the following experiment.

$\mathbf{Experiment}^{\text{ind-sel-cca}}_{\mathcal{A},\mathsf{KP-ABKEM}}(\lambda, \mathcal{U})$
$S^* \leftarrow \mathcal{A}(\lambda, \mathcal{U}), (\mathrm{PK}, \mathrm{MSK}) \leftarrow \mathbf{Setup}(\lambda, \mathcal{U})$
$\epsilon \leftarrow \mathcal{A}^{\mathbf{KeyGen}(\mathrm{PK},\mathrm{MSK},\cdot),\mathbf{Decap}(\mathrm{PK},\mathrm{SK}_{\cdot},\cdot)}(\mathrm{PK})$
$(\kappa^*, \psi^*) \leftarrow \mathbf{Encap}(\mathrm{PK}, S^*), \kappa \leftarrow \mathit{KeySp}(\lambda), b \leftarrow \{0,1\}$
If $b = 1$ then $\tilde{\kappa} = \kappa^*$ else $\tilde{\kappa} = \kappa$
$b' \leftarrow \mathcal{A}^{\mathbf{KeyGen}(\mathrm{PK},\mathrm{MSK},\cdot),\mathbf{Decap}(\mathrm{PK},\mathrm{SK}_{\cdot},\cdot)}(\tilde{\kappa}, \psi^*)$
If $b' = b$ then return Win else return Lose.

2.5 Target Collision Resistant Hash Functions

Target collision resistant (TCR) hash functions [18] are treated as a family. Let us denote a function family as $\mathit{Hfam}(\lambda) = \{H_\mu\}_{\mu \in \mathit{HKey}(\lambda)}$. Here $\mathit{HKey}(\lambda)$ is a hash key space, $\mu \in \mathit{HKey}(\lambda)$ is a hash key and H_μ is a function from $\{0,1\}^*$ to $\{0,1\}^\lambda$. We may assume that H_μ is from $\{0,1\}^*$ to $\mathbb{Z}_p$, where p is a prime of length λ.

Given a PPT algorithm $\mathcal{CF}$, a collision finder, we consider the following experiment (the target collision resistance game).

$\mathbf{Experiment}^{\mathrm{tcr}}_{\mathcal{CF},\mathit{Hfam}}(\lambda)$
$m^* \leftarrow \mathcal{CF}(\lambda), \mu \leftarrow \mathit{HKey}(\lambda), m \leftarrow \mathcal{CF}(\mu)$
If $m^* \neq m \wedge H_\mu(m^*) = H_\mu(m)$
then return Win else return Lose.

Then we define *$\mathcal{CF}$'s advantage over Hfam in the game of target collision resistance* as follows.

$$\mathbf{Adv}^{\mathrm{tcr}}_{\mathcal{CF},\mathit{Hfam}}(\lambda) \stackrel{\mathrm{def}}{=} \Pr[\mathbf{Experiment}^{\mathrm{tcr}}_{\mathcal{CF},\mathit{Hfam}}(\lambda) \text{ returns Win}].$$

We say that *Hfam* is *a TCR function family* if, for any PPT algorithm $\mathcal{CF}$, $\mathbf{Adv}^{\mathrm{tcr}}_{\mathcal{CF},\mathit{Hfam}}(\lambda)$ is negligible in λ.

TCR hash function families can be constructed based on the existence of a one-way function [18].

3 The Twin Diffie-Hellman Technique Revisited

A 6-tuple $(g, X_1, X_2, Y, Z_1, Z_2) \in \mathbb{G}^6$ is called a *twin Diffie-Hellman tuple* if the tuple is written as $(g, g^{x_1}, g^{x_2}, g^y, g^{x_1 y}, g^{x_2 y})$ for some elements x_1, x_2, y in $\mathbb{Z}_p$. In other words, a 6-tuple $(g, X_1, X_2, Y, Z_1, Z_2)$ is a twin Diffie-Hellman tuple (twin DH tuple, for short) if $Y = g^y$ and $Z_1 = X_1^y$ and $Z_2 = X_2^y$.

The following lemma of Cash, Kiltz and Shoup will be used in the security proof to decide whether a tuple is a twin DH tuple or not.

Lemma 1 ("Trapdoor Test"[12]) *Let X_1, r, s be mutually independent random variables, where X_1 takes values in $\mathbb{G}$, and each of r, s is uniformly distributed over $\mathbb{Z}_p$. Define the random variable $X_2 = X_1^{-r} g^s$. Suppose that $\hat{Y}, \hat{Z}_1, \hat{Z}_2$ are random variables taking values in $\mathbb{G}$, each of which is defined independently of r. Then the probability that the truth value of $\hat{Z}_1{}^r \hat{Z}_2 = \hat{Y}^s$ does not agree with the truth value of $(g, X_1, X_2, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ being a twin DH tuple is at most $1/p$. Moreover, if $(g, X_1, X_2, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ is a twin DH tuple, then $\hat{Z}_1{}^r \hat{Z}_2 = \hat{Y}^s$ certainly holds.*

Note that Lemma 1 is a statistical property. Especially, Lemma 1 holds without any number theoretic assumption. To be precise, we consider the following experiment of an algorithm *Cheat* with *unbounded computational power (not limited to PPT)*, where *Cheat*, given a triple (g, X_1, X_2), tries to complete a 6-tuple $(g, X_1, X_2, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ which passes the "Trapdoor Test" but which is *not* a twin DH tuple.

$$\begin{aligned}&\textbf{Experiment}^{\text{twinDH-test}}_{Cheat,\mathbb{G}}(\lambda)\\&(g,X_1)\leftarrow\mathbb{G}^2,(r,s)\leftarrow\mathbb{Z}_p^2,X_2=X_1^{-r}g^s\\&\mathbb{G}^3\ni(\hat{Y},\hat{Z}_1,\hat{Z}_2)\leftarrow Cheat(g,X_1,X_2)\\&\text{If }\hat{Z}_1{}^r\hat{Z}_2=\hat{Y}^s\wedge(g,X_1,X_2,\hat{Y},\hat{Z}_1,\hat{Z}_2)\\&\qquad\text{is NOT a twin DH tuple,}\\&\qquad\text{then return Win else return Lose}\end{aligned}$$

Let us define the advantage of *Cheat* over $\mathbb{G}$ as follows.

$$\begin{aligned}&\textbf{Adv}^{\text{twinDH-test}}_{Cheat,\mathbb{G}}(\lambda)\\&\stackrel{\text{def}}{=}\Pr[\textbf{Experiment}^{\text{twinDH-test}}_{Cheat,\mathbb{G}}(\lambda)\text{ returns Win}].\end{aligned}$$

Now we are ready to complement Lemma 1.

Lemma 2 (Complement for "Trapdoor Test" [12]) *For any algorithm* Cheat *with unbounded computational power,* $\textbf{Adv}^{twinDH\text{-}test}_{\text{Cheat},\mathbb{G}}(\lambda)$ *is at most* $1/p$.

For a proof of Lemma 2, see Appendix A.

4 Securing the Waters CP-ABKEM against Chosen-Ciphertext Attacks

In this section, we describe our direct chosen-ciphertext security technique by applying it to the Waters CP-ABE [4].

Overview of Our Modification The Waters CP-ABE is proved to be secure in the IND-sel-CPA game [4]. We convert it into a scheme that is secure in the IND-sel-CCA game by employing the Twin Diffie-Hellman technique of Cash, Kiltz and Shoup [12] and the algebraic trick of Boneh and Boyen [13] and Kiltz [14].

In encryption, a ciphertext becomes to contain additional two elements (d_1,d_2), which function in decryption as a "check sum" to verify that a tuple is certainly a twin DH tuple.

In security proof, the Twin Diffie-Hellman Trapdoor Test does the function instead. It is noteworthy that we are unable to use the bilinear map instead because the tuple to be verified is in the target group. In addition, the algebraic trick enables to answer for adversary's decryption queries. Note also that the both technique become compatible by introducing random variables.

Key Encapsulation and Encryption. The Waters CP-ABE can be captured as a CP-ABKEM: the blinding factor of the form $e(g,g)^{\alpha s}$ in the Waters CP-ABE can be considered as a random one-time key. So we call it the Waters CP-ABKEM hereafter and denote it as $\texttt{CP-ABKEM}_{\text{cpa}}$. Likewise, we distinguish parameters and algorithms of $\texttt{CP-ABKEM}_{\text{cpa}}$ by the index $_{\text{cpa}}$. For theoretical simplicity, we first develop a KEM CP-ABKEM.

4.1 Our Construction

Our CP-ABKEM consists of the following four PPT algorithms (Setup, Encap, KeyGen, Decap). Roughly speaking, the Waters original scheme $\texttt{CP-ABKEM}_{\text{cpa}}$ (the first scheme in [4]) corresponds to the case $k=1$ below excluding the "check sum" (d_1,d_2).

Setup$(\lambda,\mathcal{U})$**.** Setup takes as input the security parameter λ and the attribute universe $\mathcal{U}=\{1,\ldots,u\}$. It runs $\texttt{Grp}(\lambda)$ to get $(p,\mathbb{G},\mathbb{G}_T,g,e)$, where $\mathbb{G}$ and $\mathbb{G}_T$ are cyclic groups of order p, $e:\mathbb{G}\times\mathbb{G}\to\mathbb{G}_T$ is a bilinear map and g is a generator of $\mathbb{G}$. These become public parameters. Then Setup chooses u random group elements $h_1,\ldots,h_u\in\mathbb{G}$ that are associated with the u attributes. In addition, it chooses random exponents $\alpha_k\in\mathbb{Z}_p,k=1,\ldots,4,a\in\mathbb{Z}_p$ and a hash key $\eta\in HKey(\lambda)$. The public key is published as $\text{PK}=(g,g^a,h_1,\ldots,h_u,e(g,g)^{\alpha_1},\ldots,e(g,g)^{\alpha_4},\eta)$. The authority sets $\text{MSK}=(g^{\alpha_1},\ldots,g^{\alpha_4})$ as the master secret key.

Encap$(\textbf{PK},\mathbb{A})$**.** The encapsulation algorithm Encap takes as input the public key PK and an LSSS access structure $\mathbb{A}=(M,\rho)$, where M is an $l\times n$ matrix and ρ is the function which maps each row index i of M to an attribute in $\mathcal{U}=\{1,\ldots,u\}$. Encap first chooses a random value $s\in\mathbb{Z}_p$ that is the encryption randomness, and chooses random values $y_2,\ldots,y_n\in\mathbb{Z}_p$. Then Encap forms a vector $\vec{v}=(s,y_2,\ldots,y_n)$. For $i=1$ to l, it calculates $\lambda_i=\vec{v}\cdot M_i$, where M_i denotes the i-th row vector of M. In addition, Encap chooses random values $r_1,\ldots,r_l\in\mathbb{Z}_p$. Then, a pair of a random one-time key and its encapsulation (κ,ψ) is computed as follows.

$$\begin{aligned}&\text{Put }C'=g^s;\text{For }i=1\text{ to }l:C_i=g^{a\lambda_i}h_{\rho(i)}^{-r_i},D_i=g^{r_i};\\&\psi_{\text{cpa}}=(\mathbb{A},C',((C_i,D_i);i=1,\ldots,l)),\tau\leftarrow H_\eta(\psi_{\text{cpa}});\\&\text{For }k=1\text{ to }4:\kappa_k=e(g,g)^{\alpha_k s};d_1=\kappa_1^\tau\kappa_3,d_2=\kappa_2^\tau\kappa_4;\\&(\kappa,\psi)=(\kappa_1,(\psi_{\text{cpa}},d_1,d_2)).\end{aligned}$$

KeyGen$(\textbf{MSK},\textbf{PK},S)$**.** The key generation algorithm KeyGen takes as input the master secret key MSK, the public key PK and a set S of attributes. KeyGen first chooses a random $t_k\in\mathbb{Z}_p,k=1,\ldots,4$. It generates the secret key SK_S as follows.

$$\begin{aligned}&\text{For }k=1\text{ to }4:K_k=g^{\alpha_k}g^{at_k},L_k=g^{t_k}\\&\quad\text{For }x\in S:K_{k,x}=h_x^{t_k};\\&\text{SK}_S=((K_k,L_k,(K_{k,x};x\in S));k=1,\ldots,4).\end{aligned}$$

Decap$(\textbf{PK},\psi,\textbf{SK}_S)$**.** The decapsulation algorithm Decap takes as input the public key PK, an encapsulation ψ for an access structure $\mathbb{A}=(M,\rho)$ and a private key SK_S for an attribute set S. It first checks whether $S\in\mathbb{A}$. If the result is False, put $\hat{\kappa}=\perp$. Otherwise, let $I_S=\rho^{-1}(S)\subset\{1,\ldots,l\}$ and let $\{\omega_i\in\mathbb{Z}_p;i\in I_S\}$ be a set of linear reconstruction constants. Then, the decapsu-

lation $\hat{\kappa}$ is computed as follows.

Parse ψ into $(\psi_{\mathrm{cpa}} = (\mathbb{A}, C', ((C_i, D_i); i = 1, \ldots, l)), d_1, d_2)$;
$\tau \leftarrow H_\eta(\psi_{\mathrm{cpa}})$;
For $k = 1$ to 4 :

$$\hat{\kappa_k} = e(C', K_k) / \prod_{i \in I_S} (e(L_k, C_i) e(D_i, K_{k,\rho(i)}))^{\omega_i} = e(g,g)^{\alpha_k s}$$

If ${\hat{\kappa_1}}^{\tau} \hat{\kappa_3} \neq d_1 \vee {\hat{\kappa_2}}^{\tau} \hat{\kappa_4} \neq d_2$,
then put $\hat{\kappa} = \perp$, else put $\hat{\kappa} = \hat{\kappa_1}$.

4.2 Security and Proof

Theorem 1 *If the Waters CP-ABKEM$_{cpa}$ [4] is selectively secure against chosen-plaintext attacks and an employed hash function family Hfam has target collision resistance, then our CP-ABKEM is selectively secure against chosen-ciphertext attacks. More precisely, for any given PPT adversary $\mathcal{A}$ that attacks CP-ABKEM in the IND-sel-CCA game where decapsulation queries are at most q_d times, and for any small attribute universe $\mathcal{U}$, there exist a PPT adversary $\mathcal{B}$ that attacks CP-ABKEM$_{cpa}$ in the IND-sel-CPA game and a PPT target collision finder $\mathcal{CF}$ on Hfam that satisfy the following tight reduction.*

$$\begin{aligned} &\mathbf{Adv}^{ind\text{-}sel\text{-}cca}_{\mathcal{A},CP\text{-}ABKEM}(\lambda, \mathcal{U}) \\ \leq &\mathbf{Adv}^{ind\text{-}sel\text{-}cpa}_{\mathcal{B},CP\text{-}ABKEM_{cpa}}(\lambda, \mathcal{U}) + \mathbf{Adv}^{tcr}_{\mathcal{CF},Hfam}(\lambda) + \frac{q_d}{p}. \end{aligned}$$

Proof. Given any adversary $\mathcal{A}$ that attacks our scheme CP-ABKEM in the IND-sel-CCA game, we construct an adversary $\mathcal{B}$ that attacks the Waters scheme CP-ABKEM$_{\mathrm{cpa}}$ in the IND-sel-CPA game as follows.

Commit to a Target Access Structure. $\mathcal{B}$ is given $(\lambda, \mathcal{U})$ as inputs, where λ is the security parameter and $\mathcal{U} = \{1, \ldots, u\}$ is the attribute universe. $\mathcal{B}$ invokes $\mathcal{A}$ on input $(\lambda, \mathcal{U})$ and gets a target access structure $\mathbb{A}^* = (M^*, \rho^*)$ from $\mathcal{A}$, where M^* is of size $l^* \times n^*$. $\mathcal{B}$ uses $\mathbb{A}^*$ as the target access structure of itself and outputs $\mathbb{A}^*$.

Set up. In return to outputting $\mathbb{A}^*$, $\mathcal{B}$ receives the public key $\mathrm{PK}_{\mathrm{cpa}}$ for CP-ABKEM$_{\mathrm{cpa}}$, which consists of the following components.

$$\mathrm{PK}_{\mathrm{cpa}} = (g, g^a, h_1, \ldots, h_u, e(g,g)^\alpha).$$

To set up a public key PK for CP-ABKEM, $\mathcal{B}$ *herein* needs a challenge instance: $\mathcal{B}$ queries its challenger and gets a challenge instance $(\tilde{\kappa}, \psi^*_{\mathrm{cpa}})$. It consists of the following components.

$\tilde{\kappa} = e(g,g)^{\alpha s^*}$ OR a random one-time key $\kappa \in \mathit{KeySp}(\lambda)$,
$\psi^*_{\mathrm{cpa}} = (\mathbb{A}^*, C'^* = g^{s^*}, ((C_i^*, D_i^*); i = 1, \ldots, l^*))$.

Then $\mathcal{B}$ makes the rest of parameters of PK as follows.

Choose $\eta \leftarrow \mathit{HKey}(\lambda)$ and take $\tau^* \leftarrow H_\eta(\psi^*_{\mathrm{cpa}})$;
Put $e(g,g)^{\alpha_1} = e(g,g)^\alpha$;
Choose $\gamma_1, \gamma_2 \leftarrow \mathbb{Z}_p$, put $e(g,g)^{\alpha_2} = e(g,g)^{\gamma_2} / e(g,g)^{\alpha_1 \gamma_1}$;
Choose $\mu_1, \mu_2 \leftarrow \mathbb{Z}_p$, put $e(g,g)^{\alpha_3} = e(g,g)^{\mu_1} / e(g,g)^{\alpha_1 \tau^*}$,

$$e(g,g)^{\alpha_4} = e(g,g)^{\mu_2} / e(g,g)^{\alpha_2 \tau^*}.$$

Note we have implicitly set relations in the exponent domain:

$$\begin{aligned} &\alpha_2 = \gamma_2 - \alpha_1 \gamma_1, \quad \alpha_3 = \mu_1 - \alpha_1 \tau^*, \\ &\alpha_4 = \mu_2 - \alpha_2 \tau^* = \mu_2 - (\gamma_2 - \alpha_1 \gamma_1) \tau^*. \end{aligned} \tag{1}$$

A public key PK for CP-ABKEM become:

$$\mathrm{PK} = (\mathrm{PK}_{\mathrm{cpa}}, e(g,g)^{\alpha_2}, e(g,g)^{\alpha_3}, e(g,g)^{\alpha_4}, \eta).$$

Then $\mathcal{B}$ inputs PK into $\mathcal{A}$. Note that PK determines the corresponding MSK uniquely.

Phase 1. $\mathcal{B}$ answers for two types of $\mathcal{A}$'s queries as follows.

(1) Key-Extraction Queries. In the case that $\mathcal{A}$ issues a key-extraction query for an attribute set $S \subset \mathcal{U}$, $\mathcal{B}$ has to simulate $\mathcal{A}$'s challenger. To do so, $\mathcal{B}$ issues key-extraction queries to $\mathcal{B}$'s challenger for S *repeatedly up to four times*. As replies, $\mathcal{B}$ gets four secret keys of the Waters CP-ABKEM$_{\mathrm{cpa}}$ for a single attribute set S:

$$\mathrm{SK}_{\mathrm{cpa},S,k} = (K_{\mathrm{cpa},k}, L_{\mathrm{cpa},k}, (K_{\mathrm{cpa},k,x}; x \in \mathcal{S})), k = 1, \ldots, 4.$$

We remark that, according to the randomness in the key-generation algorithm of the Waters CP-ABKEM$_{\mathrm{cpa}}$, all four secret keys $\mathrm{SK}_{\mathrm{cpa},S,1}, \ldots, \mathrm{SK}_{\mathrm{cpa},S,4}$ are random and mutually independent. To reply a secret key SK_S of our CP-ABKEM to $\mathcal{A}$, $\mathcal{B}$ converts the four secret keys as follows.

$$\begin{aligned} &K_1 = K_{\mathrm{cpa},1}, && L_1 = L_{\mathrm{cpa},1},\ K_{1,x} = K_{\mathrm{cpa},1,x}, x \in S; \\ &K_2 = g^{\gamma_2} K_{\mathrm{cpa},2}^{-\gamma_1}, && L_2 = L_{\mathrm{cpa},2}^{-\gamma_1},\ K_{2,x} = K_{\mathrm{cpa},2,x}^{-\gamma_1},\ x \in S; \\ &K_3 = g^{\mu_1} K_{\mathrm{cpa},3}^{-\tau^*}, && L_3 = L_{\mathrm{cpa},3}^{-\tau^*},\ K_{3,x} = K_{\mathrm{cpa},3,x}^{\tau^*},\ x \in S; \\ &K_4 = g^{\mu_2 - \gamma_2 \tau^*} K_{\mathrm{cpa},4}^{\gamma_1 \tau^*}, && L_4 = L_{\mathrm{cpa},4}^{\gamma_1 \tau^*},\ K_{4,x} = K_{\mathrm{cpa},4,x}^{\gamma_1 \tau^*},\ x \in S. \end{aligned}$$

Then $\mathcal{B}$ replies $\mathrm{SK}_S = ((K_k, L_k, (K_{k,x}; x \in S)); k = 1, \ldots, 4)$ to $\mathcal{A}$.

(2) Decapsulation Queries. In the case that $\mathcal{A}$ issues a decapsulation query for (S, ψ), where $S \subset \mathcal{U}$ is an attribute set and $\psi = (\psi_{\mathrm{cpa}}, d_1, d_2)$ is an encapsulation concerning $\mathbb{A}$, $\mathcal{B}$ has to simulate $\mathcal{A}$'s challenger. To do so, $\mathcal{B}$ computes the decapsulation result $\hat{\kappa}$ as follows.

If $S \notin \mathbb{A}$ then put $\hat{\kappa} = \perp$,
else
  $\tau \leftarrow H_\eta(\psi_{\mathrm{cpa}})$;
  $\hat{Y} = e(C', g)^{\tau - \tau^*}, \hat{Z}_1 = d_1 / e(C', g)^{\mu_1}, \hat{Z}_2 = d_2 / e(C', g)^{\mu_2}$;
  If ${\hat{Z}_1}^{\gamma_1} \hat{Z}_2 \neq \hat{Y}^{\gamma_2}$ (: call this checking TwinDH-Test)
    then put $\hat{\kappa} = \hat{\kappa_1} = \perp$
  else
    If $\tau = \tau^*$ then abort (: call this case Abort)
    else $\hat{\kappa} = \hat{\kappa_1} = {\hat{Z}_1}^{1/(\tau - \tau^*)}$.

Challenge. In the case that $\mathcal{A}$ queries its challenger for a challenge instance, $\mathcal{B}$ makes a challenge instance as follows.

$$\begin{aligned} &\text{Put } d_1^* = e(C'^*, g)^{\mu_1}, d_2^* = e(C'^*, g)^{\mu_2}; \\ &\text{Put } \psi^* = (\psi^*_{\mathrm{cpa}}, d_1^*, d_2^*). \end{aligned}$$

Then $\mathcal{B}$ feeds $(\tilde{\kappa}, \psi^*)$ to $\mathcal{A}$ as a challenge instance.
Phase 2. The same as in Phase 1.
Guess. In the case that $\mathcal{A}$ returns $\mathcal{A}$'s guess $\tilde{b}$, $\mathcal{B}$ returns $\tilde{b}$ itself as $\mathcal{B}$'s guess.

In the above construction of $\mathcal{B}$, $\mathcal{B}$ can perfectly simulate the real view of $\mathcal{A}$ until the case Abort happens, except for a negligible case, and hence the algorithm $\mathcal{A}$ works as designed. To see the perfect simulation with a negligible exceptional case, we are enough to prove the following seven claims.

Claim 1 *The reply $SK_S = ((K_k, L_k, (K_{k,x}; x \in S)); k = 1, \ldots, 4)$ for a key-extraction query of $\mathcal{A}$ is a perfect simulation.*

Proof. We must consider the implicit relations (1). For the index 2, we have implicitly set the randomness $t_2 = t_{\mathrm{cpa},2}(-\gamma_1)$ and we get:

$$\begin{aligned}
K_2 &= g^{\gamma_2} K_{\mathrm{cpa},2}^{-\gamma_1} = g^{\gamma_2} (g^{\alpha_1} g^{a t_{\mathrm{cpa},2}})^{-\gamma_1} \\
&= g^{\gamma_2} (g^{\alpha_1} g^{a t_2/(-\gamma_1)})^{-\gamma_1} = g^{\gamma_2 - \alpha_1 \gamma_1} g^{a t_2} = g^{\alpha_2} g^{a t_2}, \\
L_2 &= L_{\mathrm{cpa},2}^{-\gamma_1} = (g^{t_{\mathrm{cpa},2}})^{-\gamma_1} = g^{t_2}, \\
K_{2,x} &= K_{\mathrm{cpa},2,x}^{-\gamma_1} = (h_x^{t_{\mathrm{cpa},2}})^{-\gamma_1} = h_x^{t_2}, x \in S.
\end{aligned}$$

For the index 3 and 4, see Appendix B.

Claim 2 *$(e(g,g),\ e(g,g)^{\alpha_1},\ e(g,g)^{\alpha_2},\ \hat{Y},\ \hat{Z}_1,\ \hat{Z}_2)$ is a twin Diffie-Hellman tuple if and only if $(e(g,g),\ e(g,g)^{\alpha_1 \tau} e(g,g)^{\alpha_3},\ e(g,g)^{\alpha_2 \tau} e(g,g)^{\alpha_4},\ e(C', g),\ d_1,\ d_2)$ is a twin Diffie-Hellman tuple.*

Proof. This claim can be proved by a short calculation. See Appendix C. □

Claim 3 *If $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ is a twin Diffie-Hellman tuple, then $(\hat{Y}, \hat{Z}_1, \hat{Z}_2)$ certainly passes the* TwinDH-Test*: $\hat{Z}_1^{\gamma_1} \hat{Z}_2 = \hat{Y}^{\gamma_2}$.*

Proof. This claim is a direct consequence of Lemma 1. □

Claim 4 *Consider the following event which we name as* Overlook$_i$*:*

In the i-th TwinDH-Test*, the following condition holds:*

$$\begin{cases} \hat{Z}_1^{\gamma_1} \hat{Z}_2 = \hat{Y}^{\gamma_2} \text{ holds and} \\ (e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}, \hat{Y}, \hat{Z}_1, \hat{Z}_2) \\ \quad \text{is NOT a twin DH tuple.} \end{cases}$$

Then, for at most q_d times decapsulation queries of $\mathcal{A}$, the probability that at least one Overlook$_i$ *occurs is negligible in λ. More precisely, the following inequality holds:*

$$\Pr[\bigvee_{i=1}^{q_d} \textsc{Overlook}_i] \leq q_d/p. \tag{2}$$

Proof. To apply Lemma 2, we construct an algorithm $Cheat_{\lambda,\mathcal{U}}$ with unbounded computational power, which takes as input $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2})$ and returns $(\hat{Y}, \hat{Z}_1, \hat{Z}_2)$ employing the adversary $\mathcal{A}$ as a subroutine. Fig. 1 shows the construction.

First, note that the view of $\mathcal{A}$ in $Cheat_{\lambda,\mathcal{U}}$ is the same as the real view of $\mathcal{A}$ and hence the algorithm $\mathcal{A}$ works as designed.

Second, note that the return $(\hat{Y}, \hat{Z}_1, \hat{Z}_2)$ of $Cheat_{\lambda,\mathcal{U}}$ is randomized in TABLE. Hence:

$$\begin{aligned}
&\sum_{i=1}^{q_\mathrm{d}} \frac{1}{q_\mathrm{d}} \Pr[\textsc{Overlook}_i] = \frac{1}{q_\mathrm{d}} \sum_{i=1}^{q_\mathrm{d}} \Pr[\textsc{Overlook}_i] \\
&= \mathbf{Adv}^{\text{twinDH-test}}_{Cheat_{\lambda,\mathcal{U}},\mathbb{G}}(\lambda).
\end{aligned} \tag{3}$$

Third, applying Lemma 2 to $Cheat_{\lambda,\mathcal{U}}$, we get:

$$\mathbf{Adv}^{\text{twinDH-test}}_{Cheat_{\lambda,\mathcal{U}},\mathbb{G}}(\lambda) \leq 1/p. \tag{4}$$

Combining (3) and (4), we have:

$$\begin{aligned}
&\Pr[\bigvee_{i=1}^{q_\mathrm{d}} \textsc{Overlook}_i] \leq \sum_{i=1}^{q_\mathrm{d}} \Pr[\textsc{Overlook}_i] \\
&\leq q_\mathrm{d} \mathbf{Adv}^{\text{twinDH-test}}_{Cheat_{\lambda,\mathcal{U}},\mathbb{G}}(\lambda) \leq \frac{q_\mathrm{d}}{p}.
\end{aligned}$$

Claim 5 *The probability that* Overlook$_i$ *never occurs in* TwinDH-Test *for every i and* Abort *occurs is negligible in λ. More precisely, the following inequality holds:* □

$$\Pr[\Big(\bigwedge_{i=1}^{q_d} \neg \textsc{Overlook}_i\Big) \wedge \textsc{Abort}] \leq \mathbf{Adv}^{tcr}_{\mathcal{CF},Hfam}(\lambda). \tag{5}$$

Proof. This claim is proved by constructing a collision finder $\mathcal{CF}$ on *Hfam*. See Appendix D. □

Claim 6 *The reply $\hat{\kappa}$ to $\mathcal{A}$ as an answer for a decapsulation query is correct.*

Claim 7 *The challenge instance $\psi^* = (\psi^*_{cpa}, d_1^*, d_2^*)$ is correctly distributed.*

Proof. These two claims are proved by a direct calculation. See Appendices E and F, respectively. □

Evaluation of the Advantage of $\mathcal{B}$. Now we are ready to evaluate the advantage of $\mathcal{B}$ in the IND-sel-CPA game. That $\mathcal{A}$ wins in the IND-sel-CCA game means that $(\tilde{\kappa}, \psi^* = (\psi^*_{\mathrm{cpa}}, d_1^*, d_2^*))$ is correctly guessed. This is equivalent to that $(\tilde{\kappa}, \psi^*_{\mathrm{cpa}})$ is correctly guessed because ψ^*_{cpa} determines the consistent blinding factor $\kappa^* = e(g,g)^{\alpha s^*}$ uniquely. This means that $\mathcal{B}$ wins in the IND-sel-CPA game.

Therefore, the probability that $\mathcal{B}$ wins is equal to the probability that $\mathcal{A}$ wins, Overlook$_i$ never holds in TwinDH-Testfor each i and Abort never occurs. So

Given $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2})$ as input :
Set up
Initialize the inner state, put TABLE = ϕ;
Get a target access structure $\mathbb{A}^* \leftarrow \mathcal{A}(\lambda, \mathcal{U})$;
Compute the base $g \in \mathbb{G}$ from $(e(g,g), e)$;
Choose $a \in \mathbb{Z}_p$ and $h_1, \ldots, h_u \in \mathbb{G}$;
Put $\mathrm{PK}_{\mathrm{cpa}} = (g, g^a, h_1, \ldots, h_u, e(g,g)^{\alpha_1})$;
Get $(\kappa^*, \psi^*_{\mathrm{cpa}}) \leftarrow \mathbf{Encap}_{\mathrm{cpa}}(\mathrm{PK}_{\mathrm{cpa}}, \mathbb{A}^*)$;
Choose $\eta \leftarrow \mathit{HKey}(\lambda)$ and compute $\tau^* \leftarrow H_\eta(\psi^*_{\mathrm{cpa}})$;
Compute discrete logarithms $\alpha_1, \alpha_2 \in \mathbb{Z}_p$ of $e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}$ to the base $e(g,g)$;
Choose $\mu_1, \mu_2 \leftarrow \mathbb{Z}_p$, put $\alpha_3 = \mu_1 - \alpha_1\tau^*, \alpha_4 = \mu_2 - \alpha_2\tau^*$;
Put $\mathrm{PK} = (\mathrm{PK}_{\mathrm{cpa}}, e(g,g)^{\alpha_2}, e(g,g)^{\alpha_3}, e(g,g)^{\alpha_4}, \eta), \mathrm{MSK} = (g^{\alpha_1}, g^{\alpha_2}, g^{\alpha_3}, g^{\alpha_4})$;
Give PK to $\mathcal{A}$;
Phase 1
In the case that $\mathcal{A}$ makes a key-extraction query for $S \subset \mathcal{U}$;
Reply SK_S to $\mathcal{A}$ in the same way as **KeyGen** does using MSK;
In the case that $\mathcal{A}$ makes a decapsulation query for $(\mathbb{A}, \psi = (\psi_{\mathrm{cpa}}, d_1, d_2), S)$;
Reply $\hat{\kappa}$ to $\mathcal{A}$ in the same way as **Decap** does using MSK;
Compute $\hat{Y} = e(C', g)^{\tau - \tau^*}, \hat{Z}_1 = d_1 / e(C', g)^{\mu_1}, \hat{Z}_2 = d_2 / e(C', g)^{\mu_2}$;
Update TABLE = TABLE $\cup (\hat{Y}, \hat{Z}_1, \hat{Z}_2)$;
Challenge
In the case that $\mathcal{A}$ makes a challenge instance query;
Put $d^*_1 = e(C'^*, g)^{\mu_1}, d^*_2 = e(C'^*, g)^{\mu_2}, \psi^* = (\psi^*_{\mathrm{cpa}}, d^*_1, d^*_2)$;
Choose $\kappa \leftarrow \mathit{KeySp}(\lambda), b \leftarrow \{0,1\}$;
If $b = 1$ then put $\tilde{\kappa} = \kappa^*$ else put $\tilde{\kappa} = \kappa$;
Reply $(\tilde{\kappa}, \psi^*)$ to $\mathcal{A}$;
Phase 2
The same as in Phase 1;
Return
In the case that $\mathcal{A}$ returns its guess b^*;
Choose one triple $(\hat{Y}, \hat{Z}_1, \hat{Z}_2)$ from TABLE at random;
Return $(\hat{Y}, \hat{Z}_1, \hat{Z}_2)$.

Figure 1: An Algorithm $\mathit{Cheat}_{\lambda,\mathcal{U}}$ with Unbounded Computational Power for a Proof of Claim 4.

we have:

$$
\begin{aligned}
&\Pr[\mathcal{B} \text{ wins}] \\
&= \Pr[(\mathcal{A} \text{ wins}) \wedge \Big(\bigwedge_{i=1}^{q_{\mathrm{d}}} \neg\textsc{Overlook}_i\Big) \wedge (\neg\textsc{Abort})] \\
&= \Pr[\mathcal{A} \text{ wins}] \\
&\quad - \Pr[(\mathcal{A} \text{ wins}) \wedge \neg\Big(\Big(\bigwedge_{i=1}^{q_{\mathrm{d}}} \neg\textsc{Overlook}_i\Big) \wedge (\neg\textsc{Abort})\Big)] \\
&\geq \Pr[\mathcal{A} \text{ wins}] - \Pr[\neg\Big(\Big(\bigwedge_{i=1}^{q_{\mathrm{d}}} \neg\textsc{Overlook}_i\Big) \wedge (\neg\textsc{Abort})\Big)] \\
&= \Pr[\mathcal{A} \text{ wins}] \\
&\quad - (\Pr[\bigvee_{i=1}^{q_{\mathrm{d}}} \textsc{Overlook}_i] + \Pr[\Big(\bigwedge_{i=1}^{q_{\mathrm{d}}} \neg\textsc{Overlook}_i\Big) \wedge \textsc{Abort}]).
\end{aligned}
$$

Substituting (2), (5) and advantages into the above, we have:

$$
\begin{aligned}
&\mathbf{Adv}^{\text{ind-sel-cpa}}_{\mathcal{B},\mathtt{CP\text{-}ABKEM}_{\mathrm{cpa}}}(\lambda, \mathcal{U}) \\
&\geq \mathbf{Adv}^{\text{ind-sel-cca}}_{\mathcal{A},\mathtt{CP\text{-}ABKEM}}(\lambda, \mathcal{U}) - \frac{q_{\mathrm{d}}}{p} - \mathbf{Adv}^{\mathrm{tcr}}_{\mathcal{CF},\mathit{Hfam}}(\lambda).
\end{aligned}
$$

This is what we should prove in Theorem 1. □

4.3 Encryption Scheme from KEM

It is straightforward to construct our encryption scheme CP-ABE from CP-ABKEM. The IND-sel-CCA security of CP-ABE is proved based on IND-sel-CPA security of the Waters *KEM* $\mathtt{CP\text{-}ABKEM}_{\mathrm{cpa}}$.

Setup$(\lambda, \mathcal{U})$. The same as Setup of CP-ABKEM.

Encrypt$(\mathrm{PK}, \mathbb{A}, m)$. The same as Encap of CP-ABKEM except that Encrypt multiplies m by the blinding factor κ in the group $\mathbb{G}_T$. Encrypt returns $\mathrm{CT} = (C = m\kappa, \psi = (\psi_{\mathrm{cpa}}, d_1, d_2))$.

KeyGen$(\mathrm{MSK}, \mathrm{PK}, S)$. The same as KeyGen of CP-ABKEM.

Decrypt$(\mathrm{PK}, \mathrm{CT}, \mathrm{SK}_S)$. The same as Decap of CP-ABKEM except that Decrypt divides out C by the decapsulated blinding factor $\hat{\kappa}$. Decrypt returns the result $\hat{m}$.

4.4 Security and Proof

Theorem 2 *If the Waters* $\mathtt{CP\text{-}ABKEM}_{cpa}$ *[4] is selectively secure against chosen-plaintext attacks and an employed hash function family* Hfam *has target collision resistance, then our* CP-ABE *is selectively secure against chosen-ciphertext attacks. More precisely, for any given PPT ad-*

versary $\mathcal{A}$ that attacks CP-ABE in the IND-sel-CCA game where decryption queries are at most q_d times, and for any small attribute universe $\mathcal{U}$, there exist a PPT adversary $\mathcal{B}$ that attacks CP-ABKEM$_{cpa}$ in the IND-sel-CPA game and a PPT target collision finder $\mathcal{CF}$ on Hfam that satisfy the following inequality.

$$\begin{aligned}&\mathbf{Adv}^{ind\text{-}sel\text{-}cca}_{\mathcal{A},CP\text{-}ABE}(\lambda,\mathcal{U})\\ \leq& 2\Big(\mathbf{Adv}^{ind\text{-}sel\text{-}cpa}_{\mathcal{B},CP\text{-}ABKEM_{cpa}}(\lambda,\mathcal{U})+\mathbf{Adv}^{tcr}_{\mathcal{CF},Hfam}(\lambda)+\frac{q_d}{p}\Big).\end{aligned}$$

Proof. Given any adversary $\mathcal{A}$ that attacks our scheme CP-ABE in the IND-sel-CCA game, we construct an adversary $\mathcal{B}$ that attacks the Waters KEM CP-ABKEM$_{\mathrm{cpa}}$ in the IND-sel-CPA game as follows.

Commit to a Target Access Structure. The same as that of CP-ABKEM.

Set up. In return to outputting $\mathbb{A}^*$, $\mathcal{B}$ receives the public key PK$_{\mathrm{cpa}}$ for CP-ABKEM$_{\mathrm{cpa}}$. To set up a public key PK for CP-ABE, $\mathcal{B}$ *herein* needs a challenge instance: $\mathcal{B}$ queries its challenger and gets a challenge instance $(\tilde{\kappa},\psi^*_{\mathrm{cpa}})$. The rest of procedure is the same as that of CP-ABKEM, and $\mathcal{B}$ inputs PK into $\mathcal{A}$.

Phase 1. The same as that of CP-ABKEM except that $\mathcal{B}$ replies a decrypted message $\hat{m}$ to $\mathcal{A}$ for a decryption query.

Challenge. In the case that $\mathcal{A}$ submits two plaintexts (m^*_0,m^*_1) of equal length, $\mathcal{B}$ makes a challenge ciphertext CT* as follows and feeds CT* to $\mathcal{A}$.

$$\begin{aligned}&\text{Choose } b'\leftarrow\{0,1\},\text{ put } \tilde{C}^*=m^*_{b'},\tilde{\kappa};\\ &\text{Put } d^*_1=e(C'^*,g)^{\mu_1},d^*_2=e(C'^*,g)^{\mu_2};\\ &\text{Put CT}^*=(\tilde{C}^*,\psi^*=(\psi^*_{\mathrm{cpa}},d^*_1,d^*_2)).\end{aligned}$$

Phase 2. The same as in Phase 1.

Guess. In the case that $\mathcal{A}$ returns $\mathcal{A}$'s guess $\tilde{b}$, $\mathcal{B}$ returns $\tilde{b}$ as $\mathcal{B}$'s guess.

Evaluation of the Advantage of $\mathcal{B}$. A standard argument deduces a loss of tightness by a factor of 1/2. That is;

$$\begin{aligned}&\mathbf{Adv}^{\text{ind-sel-cpa}}_{\mathcal{B},\text{CP-ABKEM}_{\text{cpa}}}(\lambda,\mathcal{U})\\ \geq&\frac{1}{2}\mathbf{Adv}^{\text{ind-sel-cca}}_{\mathcal{A},\text{CP-ABE}}(\lambda,\mathcal{U})-\frac{q_{\mathrm{d}}}{p}-\mathbf{Adv}^{\mathrm{tcr}}_{\mathcal{CF},Hfam}(\lambda).\square\end{aligned}$$

5 Securing the Ostrovsky-Sahai-Waters KP-ABKEM against Chosen-Ciphertext Attacks

In this section, we describe our direct chosen-ciphertext security modification by applying it to the Ostrovsky-Sahai-Waters KP-ABE [11].

Overview of Our Modification The Ostrovsky-Sahai-Waters KP-ABE is proved to be secure in the IND-sel-CPA game [11]. We convert it into a scheme that is secure in the IND-sel-CCA game by employing the Twin Diffie-Hellman technique of Cash, Kiltz and Shoup [12] and the algebraic trick of Boneh and Boyen [13] and Kiltz [14].

In encryption, a ciphertext becomes to contain additional two elements (d_1,d_2), which function in decryption as a "check sum" to verify that a tuple is certainly a twin DH tuple.

In security proof, the Twin Diffie-Hellman Trapdoor Test does the function instead. It is noteworthy that we are unable to use the bilinear map instead because the tuple to be verified is in the target group. In addition, the algebraic trick enables to answer for adversary's decryption queries. Note also that the both technique become compatible by introducing random variables.

Key Encapsulation and Encryption. The Ostrovsky-Sahai-Waters KP-ABE can be captured as a KP-ABKEM: the blinding factor of the form $e(g,g)^{a\alpha s}$ in the Ostrovsky-Sahai-Waters KP-ABE can be considered as a random one-time key. So we call it the Ostrovsky-Sahai-Waters KP-ABKEM hereafter and denote it as KP-ABKEM$_{\mathrm{cpa}}$. Likewise, we distinguish parameters and algorithms of KP-ABKEM$_{\mathrm{cpa}}$ by the index $_{\mathrm{cpa}}$. For theoretical simplicity, we first develop a KEM KP-ABKEM.

5.1 Our Construction

Our KP-ABKEM consists of the following four PPT algorithms (Setup, Encap, KeyGen, Decap). Roughly speaking, the Ostrovsky-Sahai-Waters original scheme KP-ABKEM$_{\mathrm{cpa}}$ (the first scheme in [11]) corresponds to the case $k=1$ below excluding the "check sum" (d_1,d_2).

Setup$(\lambda,\mathcal{U})$**.** Setup takes as input the security parameter λ and the attribute universe $\mathcal{U}=\{1,\ldots,u\}$. It runs $\mathsf{Grp}(\lambda)$ to get $(p,\mathbb{G},\mathbb{G}_T,g,e)$, where $\mathbb{G}$ and $\mathbb{G}_T$ are cyclic groups of order p, $e:\mathbb{G}\to\mathbb{G}_T$ is a bilinear map and g is a generator of $\mathbb{G}$. These become public parameters. Then Setup chooses u random group elements $h_1,\ldots,h_u\in\mathbb{G}$ that are associated with the u attributes. In addition, it chooses random exponents $\alpha_k\in\mathbb{Z}_p,k=1,\ldots,4,a\in\mathbb{Z}_p$ and a hash key $\eta\in HKey(\lambda)$. The public key is published as PK $=(g,g^a,h_1,\ldots,h_u,e(g,g)^{a\alpha_1},\ldots,e(g,g)^{a\alpha_4},\eta)$. The authority sets MSK $=(\alpha_1,\ldots,\alpha_4)$ as the master secret key.

Encap$(\mathbf{PK},S)$**.** The encapsulation algorithm Encap takes as input the public key PK and a set S of attributes. Encap first chooses a random value $s\in\mathbb{Z}_p$ that is the encryption randomness. Then, a pair of a random one-time key and its encapsulation (κ,ψ) is computed as follows.

$$\begin{aligned}&\text{Put } C'=g^s;\text{For } x\in S: C_x=h^s_x\\ &\psi_{\mathrm{cpa}}=(S,C',(C_x;x\in S)),\tau\leftarrow H_\eta(\psi_{\mathrm{cpa}});\\ &\text{For } k=1\text{ to }4:\kappa_k=e(g,g)^{a\alpha_k s};d_1=\kappa_1^\tau\kappa_3,d_2=\kappa_2^\tau\kappa_4;\\ &(\kappa,\psi)=(\kappa_1,(\psi_{\mathrm{cpa}},d_1,d_2)).\end{aligned}$$

KeyGen$(\mathbf{MSK},\mathbf{PK},\mathbb{A})$**.** The key generation algorithm KeyGen takes as input the master secret key MSK, the public key PK and an LSSS access structure $\mathbb{A}=(M,\rho)$, where M is an $l\times n$ matrix and ρ is the function which maps each row index i of M to an attribute in $\mathcal{U}=\{1,\ldots,u\}$. For $k=1$ to 4, KeyGen first chooses

random values $y_{k,2},\ldots,y_{k,n} \in \mathbb{Z}_p$ and forms a vector $\vec{v_k} = (\alpha_k, y_{k,2},\ldots,y_{k,n})$. Then, for $i = 1$ to l, it calculates $\lambda_{k,i} = \vec{v_k} \cdot M_i$, where M_i denotes the i-th row vector of M, and it chooses random values $r_{k,i} \in \mathbb{Z}_p$. KeyGen generates the secret key $\mathrm{SK}_{\mathbb{A}}$ as follows.

$$\begin{aligned}&\text{For } k = 1 \text{ to } 4 : \text{For } l = 1 \text{ to } l :\\ &\quad K_{k,i} = g^{a\lambda_{k,i}} h_{\rho(i)}^{r_{k,i}}, L_{k,i} = g^{r_{k,i}}\\ &\mathrm{SK}_{\mathbb{A}} = (((K_{k,i}, L_{k,i}); i = 1,\ldots,l)k = 1,\ldots,4).\end{aligned}$$

Decap$(\mathbf{PK}, \psi, \mathbf{SK}_{\mathbb{A}})$**.** The decapsulation algorithm Decap takes as input the public key PK, an encapsulation ψ for an attribute set S and a private key $\mathrm{SK}_{\mathbb{A}}$ for an access structure $\mathbb{A} = (M, \rho)$. It first checks whether $S \in \mathbb{A}$. If the result is False, put $\hat{\kappa} = \perp$. Otherwise, let $I_S = \rho^{-1}(S) \subset \{1,\ldots,l\}$ and let $\{\omega_i \in \mathbb{Z}_p; i \in I_S\}$ be a set of linear reconstruction constants. Then, the decapsulation $\hat{\kappa}$ is computed as follows.

$$\begin{aligned}&\text{Parse } \psi \text{ into } (\psi_{\mathrm{cpa}} = (S, C', (C_x; x \in S)), d_1, d_2);\\ &\tau \leftarrow H_\eta(\psi_{\mathrm{cpa}});\\ &\text{For } k = 1 \text{ to } 4 :\\ &\quad \hat{\kappa_k} = \prod_{i \in I_S} (e(C', K_{k,i})/e(L_{k,i}, C_{\rho(i)}))^{\omega_i} = e(g,g)^{a\alpha_k s}\\ &\text{If } \hat{\kappa_1}^{\tau}\hat{\kappa_3} \neq d_1 \vee \hat{\kappa_2}^{\tau}\hat{\kappa_4} \neq d_2,\\ &\quad \text{then put } \hat{\kappa} = \perp, \text{else put } \hat{\kappa} = \hat{\kappa_1}.\end{aligned}$$

5.2 Security and Proof

Theorem 3 *If the Ostrovsky-Sahai-Waters* KP-ABKEM$_{cpa}$ *[11] is selectively secure against chosen-plaintext attacks and an employed hash function family* Hfam *has target collision resistance, then our* KP-ABKEM *is selectively secure against chosen-ciphertext attacks. More precisely, for any given PPT adversary* $\mathcal{A}$ *that attacks* KP-ABKEM *in the IND-sel-CCA game where decapsulation queries are at most* q_d *times, and for any small attribute universe* $\mathcal{U}$*, there exist a PPT adversary* $\mathcal{B}$ *that attacks* KP-ABKEM$_{cpa}$ *in the IND-sel-CPA game and a PPT target collision finder* $\mathcal{CF}$ *on* Hfam *that satisfy the following tight reduction.*

$$\begin{aligned}&\mathbf{Adv}^{ind\text{-}sel\text{-}cca}_{\mathcal{A},KP\text{-}ABKEM}(\lambda,\mathcal{U})\\ \leq &\mathbf{Adv}^{ind\text{-}sel\text{-}cpa}_{\mathcal{B},KP\text{-}ABKEM_{cpa}}(\lambda,\mathcal{U}) + \mathbf{Adv}^{tcr}_{\mathcal{CF},Hfam}(\lambda) + \frac{q_d}{p}.\end{aligned}$$

Proof. We will omit the description of the proof because the proof goes analogously to the case of CP-ABKEM in Section 4.2. □

5.3 Encryption Scheme from KEM

It is straightforward to construct our encryption scheme KP-ABE from KP-ABKEM. The IND-sel-CCA security of KP-ABE is proved based on IND-sel-CPA security of the Waters *KEM* KP-ABKEM$_{\mathrm{cpa}}$.

Setup$(\lambda, \mathcal{U})$**.** The same as Setup of KP-ABKEM.

Encrypt$(\mathbf{PK}, \mathbb{A}, m)$**.** The same as Encap of KP-ABKEM except that Encrypt multiplies m by the blinding factor κ in the group $\mathbb{G}_T$. Encrypt returns $\mathrm{CT} = (C = m\kappa, \psi = (\psi_{\mathrm{cpa}}, d_1, d_2))$.

KeyGen$(\mathbf{MSK}, \mathbf{PK}, S)$**.** The same as KeyGen of KP-ABKEM.

Decrypt$(\mathbf{PK}, \mathbf{CT}, \mathbf{SK}_S)$**.** The same as Decap of KP-ABKEM except that Decrypt divides out C by the decapsulated blinding factor $\hat{\kappa}$. Decrypt returns the result $\hat{m}$.

5.4 Security and Proof

Theorem 4 *If the Ostrovsky-Sahai-Waters* KP-ABKEM$_{cpa}$ *[11] is selectively secure against chosen-plaintext attacks and an employed hash function family* Hfam *has target collision resistance, then our* KP-ABE *is selectively secure against chosen-ciphertext attacks. More precisely, for any given PPT adversary* $\mathcal{A}$ *that attacks* KP-ABE *in the IND-sel-CCA game where decryption queries are at most* q_d *times, and for any small attribute universe* $\mathcal{U}$*, there exist a PPT adversary* $\mathcal{B}$ *that attacks* KP-ABKEM$_{cpa}$ *in the IND-sel-CPA game and a PPT target collision finder* $\mathcal{CF}$ *on* Hfam *that satisfy the following inequality.*

$$\begin{aligned}&\mathbf{Adv}^{ind\text{-}sel\text{-}cca}_{\mathcal{A},KP\text{-}ABE}(\lambda,\mathcal{U})\\ \leq &2\Big(\mathbf{Adv}^{ind\text{-}sel\text{-}cpa}_{\mathcal{B},KP\text{-}ABKEM_{cpa}}(\lambda,\mathcal{U}) + \mathbf{Adv}^{tcr}_{\mathcal{CF},Hfam}(\lambda) + \frac{q_d}{p}\Big).\end{aligned}$$

Proof. We will omit the description of the proof because the proof goes in the same way as the case of CP-ABE in Section 4.4. □

6 Efficiency Discussion

First of all, we remark that our individual modification to attain CCA security is applicable when a Diffie-Hellman tuple to be verified is in the target group of a bilinear map $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$. Especially, it is applicable even when an original CPA secure scheme is based on asymmetric pairing [19], $e : \mathbb{G}_1 \times \mathbb{G}_2 \to \mathbb{G}_T$. For example, the Type 3 version [19] of the Waters CP-ABE scheme [4] can be found in [20]. Detailed discussions and results on real implementations are found for the case of CPA-secure ABE schemes [21, 20]. We note here that the efficiency comparison below enables to guess the implementation results of CCA-secure ABE schemes via our modification.

We compare the efficiency of our CP-ABE with the original Waters CP-ABE$_{\mathrm{cpa}}$, and our KP-ABE with the original Ostrovsky-Sahai-Waters KP-ABE$_{\mathrm{cpa}}$. We also compare the efficiency of our schemes with the CCA-secure CP-ABE and KP-ABE schemes obtained by the generic transformation in [10]. Here the generic transformation [10] is considered in the case of a small attribute universe, the delegation type [10] and the Lamport one-time signature [22]. Table 1 shows these comparison. Note that a hash function is applied to generate a message digest of bit-length λ before signing by a secret key of the one-time signature. Note also, for simplicity, we evaluate the lengths and the amounts of computation below in the case that an access structure $\mathbb{A}$ is "all-AND" and an attribute map ρ is injective (i.e "single-use" that is opposed to "multi-use").

Table 1: Efficiency comparison of IND-sel-CCA secure ABEs ([10] and ours) with the original IND-sel-CPA secure ABEs [4, 11].

Scheme	L(PK)	L(SK$_S$)	L(CT)	C(Enc)	C(Dec)
Generic transform [10], CP-ABE	$+4\lambda^2(\mathbb{G})$	$+4\lambda^2(\mathbb{G})$	$+3\lambda^2$(bit)	$+2\lambda^2$exp.($\mathbb{G}$)	$+2\lambda^2$pair.(e)
Our individual modification (CP-ABE)	$+3(\mathbb{G}_T)$	$\times 4$	$+2(\mathbb{G}_T)$	$+4$exp.($\mathbb{G}_T$)	$\times 4$
Generic transform [10], KP-ABE	$+4\lambda^2(\mathbb{G})$	$+0$	$+3\lambda^2$(bit)	$+2\lambda^2$exp.($\mathbb{G}$)	$+2\lambda^2$pair.(e)
Our individual modification (KP-ABE)	$+3(\mathbb{G}_T)$	$\times 4$	$+2(\mathbb{G}_T)$	$+4$exp.($\mathbb{G}_T$)	$\times 4$

1) λ is the security parameter. (For instance, $\lambda = 224$ or 256.)
2) L(data) denotes the length of the data. C(algorithm) denotes the computational amount of the algorithm.
3) + and $\times$ mean the increment and the multiplier to the length or to the computational amount of the Waters CP-ABKEM$_{\mathrm{cpa}}$ and the Ostrovsky-Sahai-Waters KP-ABE$_{\mathrm{cpa}}$.
4) ($\mathbb{G}$), ($\mathbb{G}_T$) and (bit) mean that the lengths are evaluated in the number of elements in $\mathbb{G}$, elements in $\mathbb{G}_T$ and bits, respectively.
5) exp.($\mathbb{G}$) and pair.(e) mean the computational amount of one exponentiation in $\mathbb{G}$ and one pairing computation by the map e, respectively.

Our individual modification results in expansion of the length of a secret-key and the amount of decryption computation by a factor of four, while the length of a public-key, the length of a ciphertext and the amount of encryption computation are almost the same as those of the original CPA-secure schemes. In the case that the size of an attribute set is up to ($\frac{2}{3}$ of) the square of the security parameter λ, the amount of decryption computation of our CP-ABE and KP-ABE are smaller than those of the CP-ABE and KP-ABE obtained by the generic transformation [10], respectively.

7 Conclusion

We demonstrated direct chosen-ciphertext security modification for ABE in the standard model in the case of the Waters scheme (CP-ABKEM$_{\mathrm{cpa}}$, CP-ABE$_{\mathrm{cpa}}$) and the Ostrovsky-Sahai-Waters scheme (KP-ABKEM$_{\mathrm{cpa}}$, KP-ABE$_{\mathrm{cpa}}$). We utilized the Twin Diffie-Hellman Trapdoor Test of Cash, Kiltz and Shoup and the algebraic trick of Boneh and Boyen [13] and Kiltz [14]. Our modification worked for the setting that the Diffie-Hellman tuple to be verified in decryption was in the target group of the bilinear map. We compared the efficiency of our CCA-secure ABE schemes with the original CPA-secure ABE schemes and with the CCA-secure ABE schemes obtained by the versatile generic transformation.

Acknowledgment This work was supported by JSPS KAKENHI Grant Number JP15K00029.

References

[1] Amit Sahai and Brent Waters. Fuzzy identity-based encryption. In *Advances in Cryptology - EUROCRYPT 2005, 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Aarhus, Denmark, May 22-26, 2005, Proceedings*, pages 457–473, 2005.

[2] Vipul Goyal, Omkant Pandey, Amit Sahai, and Brent Waters. Attribute-based encryption for fine-grained access control of encrypted data. In *Proceedings of the 13th ACM Conference on Computer and Communications Security, CCS 2006, Alexandria, VA, USA, Ioctober 30 - November 3, 2006*, pages 89–98, 2006.

[3] John Bethencourt, Amit Sahai, and Brent Waters. Ciphertext-policy attribute-based encryption. In *2007 IEEE Symposium on Security and Privacy (S&P 2007), 20-23 May 2007, Oakland, California, USA*, pages 321–334, 2007.

[4] Brent Waters. Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization. In *Public Key Cryptography - PKC 2011 - 14th International Conference on Practice and Theory in Public Key Cryptography, Taormina, Italy, March 6-9, 2011. Proceedings*, pages 53–70, 2011.

[5] Allison B. Lewko, Tatsuaki Okamoto, Amit Sahai, Katsuyuki Takashima, and Brent Waters. Fully secure functional encryption: Attribute-based encryption and (hierarchical) inner product encryption. In *Advances in Cryptology - EUROCRYPT 2010, 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, French Riviera, May 30 - June 3, 2010. Proceedings*, pages 62–91, 2010.

[6] Hiroaki Anada and Seiko Arita. Short cca-secure ciphertext-policy attribute-based encryption. In *2017 IEEE International Conference on Smart Computing, SMARTCOMP 2017, Hong Kong, China, May 29-31, 2017*, pages 1–6, 2017.

[7] Ran Canetti, Shai Halevi, and Jonathan Katz. Chosen-ciphertext security from identity-based encryption. In *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings*, pages 207–222, 2004.

[8] Dan Boneh and Xavier Boyen. Efficient selective-id secure identity-based encryption without random oracles. In *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings*, pages 223–238, 2004.

[9] Xavier Boyen, Qixiang Mei, and Brent Waters. Direct chosen ciphertext security from identity-based techniques. In *Proceedings of the 12th ACM Conference on Computer and Communications Security, CCS 2005, Alexandria, VA, USA, November 7-11, 2005*, pages 320–329, 2005.

[10] Shota Yamada, Nuttapong Attrapadung, Goichiro Hanaoka, and Noboru Kunihiro. Generic constructions for chosen-ciphertext secure attribute based encryption. In *Public Key Cryptography - PKC 2011 - 14th International Conference on Practice and Theory in Public Key Cryptography, Taormina, Italy, March 6-9, 2011. Proceedings*, pages 71–89, 2011.

[11] Rafail Ostrovsky, Amit Sahai, and Brent Waters. Attribute-based encryption with non-monotonic access structures. In *Proceedings of the 2007 ACM Conference on Computer and Communications Security, CCS 2007, Alexandria, Virginia, USA, October 28-31, 2007*, pages 195–203, 2007.

[12] David Cash, Eike Kiltz, and Victor Shoup. The twin diffie-hellman problem and applications. In *Advances in Cryptology - EUROCRYPT 2008, 27th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Istanbul, Turkey, April 13-17, 2008. Proceedings*, pages 127–145, 2008.

[13] Dan Boneh and Xavier Boyen. Efficient selective-id secure identity-based encryption without random oracles. In *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings*, pages 223–238, 2004.

[14] Eike Kiltz. Chosen-ciphertext security from tag-based encryption. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, pages 581–600, 2006.

[15] M. Choudary Gorantla, Colin Boyd, and Juan Manuel González Nieto. Attribute-based authenticated key exchange. In *Information Security and Privacy - 15th Australasian Conference, ACISP 2010, Sydney, Australia, July 5-7, 2010. Proceedings*, pages 300–317, 2010.

[16] Nuttapong Attrapadung, Benoît Libert, and Elie de Panafieu. Expressive key-policy attribute-based encryption with constant-size ciphertexts. In *Public Key Cryptography - PKC 2011 - 14th International Conference on Practice and Theory in Public Key Cryptography, Taormina, Italy, March 6-9, 2011. Proceedings*, pages 90–108, 2011.

[17] Nuttapong Attrapadung. Dual system encryption via doubly selective security: Framework, fully secure functional encryption for regular languages, and more. In *Advances in Cryptology - EUROCRYPT 2014 - 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings*, pages 557–577, 2014.

[18] Moni Naor and Moti Yung. Universal one-way hash functions and their cryptographic applications. In *Proceedings of the 21st Annual ACM Symposium on Theory of Computing, May 14-17, 1989, Seattle, Washigton, USA*, pages 33–43, 1989.

[19] Steven D. Galbraith, Kenneth G. Paterson, and Nigel P. Smart. Pairings for cryptographers. *Discrete Applied Mathematics*, 156(16):3113–3121, 2008.

[20] Eric Zavattoni, Luis J. Dominguez Perez, Shigeo Mitsunari, Ana H. Sánchez-Ramírez, Tadanori Teruya, and Francisco Rodríguez-Henríquez. Software implementation of an attribute-based encryption scheme. *IEEE Trans. Computers*, 64(5):1429–1441, 2015.

[21] Ana Helena Sánchez and Francisco Rodríguez-Henríquez. NEON implementation of an attribute-based encryption scheme. In *Applied Cryptography and Network Security - 11th International Conference, ACNS 2013, Banff, AB, Canada, June 25-28, 2013. Proceedings*, pages 322–338, 2013.

[22] L. Lamport. Constructing digital signatures from a one-way function. Technical report, October 1979.

Appendix

A Proof of Lemma 2

The only one point to be complemented to the original proof (in [12]) is that even for any algorithm $\mathcal{A}$ with unbounded computational power, the statement holds. This is because, conditioning on the input fixed values (g, X_1, X_2), $\mathcal{A}$ only reduces the two-dimensional freedom $(r,s) \in \mathbb{Z}_p^2$ into the one-dimensional freedom $r \in \mathbb{Z}_p$ even if $\mathcal{A}$ correctly guesses the relation $s = rx_1 + x_2$. □

B Proof of Claim 1

For the index 3, we have implicitly set $t_3 = t_{\mathrm{cpa},3}(-\tau^*)$ and we get:

$$K_3 = g^{\mu_1} K_{\mathrm{cpa},3}^{-\tau^*} = g^{\mu_1}(g^{\alpha_1} g^{at_{\mathrm{cpa},3}})^{-\tau^*} = g^{\mu_1-\alpha_1\tau^*} g^{at_3} = g^{\alpha_3} g^{at_3},$$

$$L_3 = L_{\mathrm{cpa},3}^{-\tau^*} = (g^{t_{\mathrm{cpa},3}})^{-\tau^*} = g^{t_3},$$

$$K_{3,x} = K_{\mathrm{cpa},3,x}^{-\tau^*} = (h_x^{t_{\mathrm{cpa},3}})^{-\tau^*} = h_x^{t_3}, x \in S.$$

For the index 4, we have implicitly set $t_4 = t_{\mathrm{cpa},4}(\gamma_1\tau^*)$ and we get:

$$K_4 = g^{\mu_2-\gamma_2\tau^*} K_{\mathrm{cpa},4}^{\gamma_1\tau^*} = g^{\mu_2-\gamma_2\tau^*}(g^{\alpha_1} g^{at_{\mathrm{cpa},4}})^{\gamma_1\tau^*}$$

$$= g^{\mu_2-\gamma_2\tau^*} g^{\alpha_1\gamma_1\tau^*} g^{at_4} = g^{\mu_2-(\gamma_2-\alpha_1\gamma_1)\tau^*} g^{at_4} = g^{\mu_2-\alpha_2\tau^*} g^{at_4}$$

$$= g^{\alpha_4} g^{at_4},$$

$$L_4 = L_{\mathrm{cpa},4}^{\gamma_1\tau^*} = (g^{t_{\mathrm{cpa},4}})^{\gamma_1\tau^*} = g^{t_4},$$

$$K_{4,x} = K_{\mathrm{cpa},4,x}^{\gamma_1\tau^*} = (h_x^{t_{\mathrm{cpa},4}})^{\gamma_1\tau^*} = h_x^{t_4}, x \in S.$$

□

C Proof of Claim 2

Suppose that we are given a twin DH tuple $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$. Then, $d_i/e(C',g)^{\mu_i} = (e(g,g)^{\alpha_i})^{s(\tau-\tau^*)}, i = 1,2$. So, using the implicit relations (1), we have:

$$d_i = e(g,g)^{\alpha_i s(\tau-\tau^*)} e(g^s, g)^{\mu_i} = (e(g,g)^{\alpha_i(\tau-\tau^*)} e(g,g)^{\mu_i})^s$$

$$= (e(g,g)^{\alpha_i(\tau-\tau^*)} e(g,g)^{\alpha_i\tau^*+\alpha_{(i+2)}})^s = (e(g,g)^{\alpha_i\tau} e(g,g)^{\alpha_{(i+2)}})^s, i = 1,2.$$

This means that $(e(g,g),\ e(g,g)^{\alpha_1\tau} e(g,g)^{\alpha_3},\ e(g,g)^{\alpha_2\tau} e(g,g)^{\alpha_4}, e(C',g), d_1, d_2)$ is a twin Diffie-Hellman tuple.

The converse is also verified by the reverse calculation. □

D Proof of Claim 5

To reduce to the target collision resistance of an employed hash function family *Hfam*, we construct a PPT target collision finder $\mathcal{CF}$ that attacks *Hfam* using $\mathcal{A}$ as a subroutine. The construction is shown in Fig.2. (Note that the case Collision is defined in Fig.2.)

Note that the view of $\mathcal{A}$ in $\mathcal{CF}$ is the same as the real view of $\mathcal{A}$ until the case Collision occurs and hence the algorithm $\mathcal{A}$ works as designed.

To evaluate the probability in Claim 5, we consider the following two cases.

Case 1: the case that Abort ($\tau = \tau^*$) occurs in $\mathcal{B}$ in Phase 1. In this case, the target τ^* has not been given to $\mathcal{A}$. So $\mathcal{A}$ needs to guess τ^* to cause a collision $\tau = \tau^*$. Hence:

$$\Pr[\text{Phase 1} \wedge \Big(\bigwedge_{i=1}^{q_\mathrm{d}} \neg\textsc{Overlook}_i\Big) \wedge \textsc{Abort}] \leq \Pr[\text{Phase 1} \wedge \textsc{Collision}]. \tag{6}$$

Case 2: the case that Abort ($\tau = \tau^*$) occurs in $\mathcal{B}$ in Phase 2. In this case, if, in addition to $\tau = \tau^*$, it occurred that $\psi_{\mathrm{cpa}} = \psi^*_{\mathrm{cpa}}$ (and hence $C' = C'^*$), then it would occur that $\psi = \psi^*$. This is because the following two tuples are equal twin DH tuples by the fact that $\textsc{Overlook}_i$ never occurs:

$$(e(g,g), e(g,g)^{\alpha_1\tau} e(g,g)^{\alpha_3}, e(g,g)^{\alpha_2\tau} e(g,g)^{\alpha_4}, e(C',g), d_1, d_2),$$

$$(e(g,g), e(g,g)^{\alpha_1\tau^*} e(g,g)^{\alpha_3}, e(g,g)^{\alpha_2\tau^*} e(g,g)^{\alpha_4}, e(C'^*,g), d_1^*, d_2^*).$$

Hence both $S \in \mathbb{A}$ and $\psi = \psi^*$ would occur. This is ruled out in decapsulation query; a contradiction. So we have $\psi_{\mathrm{cpa}} \neq \psi^*_{\mathrm{cpa}}$; that is, a collision:

$$\psi_{\mathrm{cpa}} \neq \psi^*_{\mathrm{cpa}} \wedge H_\eta(\psi_{\mathrm{cpa}}) = \tau = \tau^* = H_\eta(\psi^*_{\mathrm{cpa}}).$$

Therefore, if $\textsc{Overlook}_i$ never occurs for each i, then only decapsulation queries for which $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ are certainly twin DH tuples have the chance to cause a collision $\tau = \tau^*$, as is the case in $\mathcal{CF}$. Hence we have:

$$\Pr[\text{Phase 2} \wedge \Big(\bigwedge_{i=1}^{q_\mathrm{d}} \neg\textsc{Overlook}_i\Big) \wedge \textsc{Abort}] \leq \Pr[\text{Phase 2} \wedge \textsc{Collision}]. \tag{7}$$

Given λ as input :
Set up
Initialize inner state;
Choose a polynomial size attribute universe $\mathcal{U}$ at random;
Get a target access structure $\mathbb{A}^* \leftarrow \mathcal{A}(\lambda, \mathcal{U})$;
Run $\mathbf{Setup}_{\mathrm{cpa}}(\lambda, \mathcal{U})$ to get $(p, \mathbb{G}, \mathbb{G}_T, g, e), \mathrm{PK}_{\mathrm{cpa}}, \mathrm{MSK}_{\mathrm{cpa}}$;
Get $(\kappa^*, \psi^*_{\mathrm{cpa}}) \leftarrow \mathbf{Encap}_{\mathrm{cpa}}(\mathrm{PK}_{\mathrm{cpa}}, \mathbb{A}^*)$;
Output ψ^*_{cpa};
Receive, in return, $\eta \leftarrow HKey(\lambda)$ and compute $\tau^* \leftarrow H_\eta(\psi^*_{\mathrm{cpa}})$;
Choose $\alpha_2, \alpha_3, \alpha_4 \leftarrow \mathbb{Z}_p$;
Put $\mathrm{PK} = (\mathrm{PK}_{\mathrm{cpa}}, e(g,g)^{\alpha_2}, e(g,g)^{\alpha_3}, e(g,g)^{\alpha_4}, \eta), \mathrm{MSK} = (g^{\alpha_1}, g^{\alpha_2}, g^{\alpha_3}, g^{\alpha_4})$;
Give PK to $\mathcal{A}$;
Phase 1
In the case that $\mathcal{A}$ makes a key-extraction query for $S \subset \mathcal{U}$;
Reply SK_S to $\mathcal{A}$ in the same way as **KeyGen** does using MSK;
In the case that $\mathcal{A}$ makes a decapsulation query for $(S, \psi = (\psi_{\mathrm{cpa}}, d_1, d_2))$;
Reply $\hat{\kappa}$ to $\mathcal{A}$ in the same way as **Decap** does using MSK;
If $\hat{\kappa} \neq \perp$ and $\tau = \tau^*$ (: call this case Collision)
then return ψ_{cpa} and stop;
Challenge
In the case that $\mathcal{A}$ makes a challenge instance query;
Using MSK, put $d_1^* = e(g^{\alpha_1}, C'^*)^{\tau^*} e(g^{\alpha_3}, C'^*), d_2^* = e(g^{\alpha_2}, C'^*)^{\tau^*} e(g^{\alpha_4}, C'^*)$,
$\psi^* = (\psi^*_{\mathrm{cpa}}, d_1^*, d_2^*)$;
Choose $\kappa \leftarrow KeySp(\lambda), b \leftarrow \{0,1\}$;
If $b = 1$ then put $\tilde{\kappa} = \kappa^*$ else put $\tilde{\kappa} = \kappa$;
Reply $(\tilde{\kappa}, \psi^*)$ to $\mathcal{A}$;
Phase 2
The same as in Phase 1;
Return
In the case that $\mathcal{A}$ returns its guess b^*;
Stop.

Figure 2: A PPT Collision Finder $\mathcal{CF}$ that attacks *Hfam* for the proof of Claim 5.

Taking a sum of both sides of (6) and (7), we get:

$$\Pr[\Big(\bigwedge_{i=1}^{q_{\mathrm{d}}} \neg \textsc{Overlook}_i\Big) \wedge \textsc{Abort}] \leq \Pr[\textsc{Collision}] = \mathbf{Adv}^{\mathrm{tcr}}_{\mathcal{CF}, Hfam}(\lambda).$$

□

E Proof of Claim 6

It is enough to prove that

When $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ is a twin DH tuple,
$\hat{\kappa} = \hat{Z_1}^{1/(\tau - \tau^*)} = e(g,g)^{\alpha_1 s}$ holds.

This is deduced as follows:

$$\hat{\kappa} = (d_1 / e(C', g)^{\mu_1})^{1/(\tau - \tau^*)} = ((e(g,g)^{\alpha_1})^{s(\tau - \tau^*)})^{1/(\tau - \tau^*)} = e(g,g)^{\alpha_1 s}.$$

□

F Proof of Claim 7

A direct calculation with equalities (1) shows:

$$d_i^* = e(C'^*, g)^{\mu_i} = e(g,g)^{s^*(\alpha_i \tau^* + \alpha_{(i+2)})} = e(g,g)^{\alpha_i s^* \tau^*} e(g,g)^{\alpha_{(i+2)} s^*},$$
$i = 1, 2.$

Hence $\psi^* = (\psi^*_{\mathrm{cpa}}, d_1^*, d_2^*)$ is legitimate and correctly distributed. □

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 272-274 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Cyclical Wave Bolt for Sound Waves in a Gas Stream

Vladimir Arabadzhi*

Institute of Applied Physics, Division of Geophysical Research, 603950 Russian Federation

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 02 December, 2017
Online: 23 December, 2017

Keywords :
Cyclical Wave Bolt, Sound in Gas Stream, Rotary Switch, Impeller, Bundle of Tubes, Echelon of Waveguides Bundle of Tubes, Automotive Silencer

ABSTRACT

This article is devoted to the problem of blocking the propagation of sound in a gas stream. This can be useful in the problems of pipeline acoustics and the design of automotive silencers, when it is necessary to ensure the opacity of the boundary (cross section of the pipe) for sound simultaneously with the free (ideally) flow through this boundary. The tool for solving this problem is the rapid periodic overlapping of waveguide (gas-conducting) channels in a system of parallel waveguides. The design of the proposed device assumes the following technological requirements: high mechanical rigidity of the elements, high accuracy of their manufacture and high rotor speed. Fulfillment of these conditions allows creating an effective device for blocking sound in the gas flow. Such a device may have small wave dimensions (with respect to the wavelength of sound) and a small expenditure of mechanical power to push gas through it very wide frequency range of silencing.

1. Introduction

In practice, sometimes it is necessary to place a filter in the path of the gas flow (with velocity v) in the pipe (and simultaneously sound with speed c in the tube as an acoustic waveguide). This filter must pass the flow of gas and block sound waves on its output. Such situations are encountered in problems of the acoustics of pipelines and also in the problem of designing exhaust silencers for cars with internal combustion engines. Next we will call such a device a silencer. The area of the inlet and outlet cross-section of the silencer is assumed equal S_0 , the average velocity of the flow at the inlet and outlet of the silencer is V_0 (see Figure 1). The time-average gas pressures at Q the outlet and inlet of the silencer are, respectively, p_0 and $p_0 + \Delta p$, where $\Delta p > 0$. The quality of the silencers is characterized by the following parameters:

(a) the degree

$$Q_1 = P_{out} / P_{in} \tag{1}$$

of sound pressure suppression (of course preferably $Q_1 << 1$);

(b) the frequency range

$$\omega_{\min} \le \omega \le \omega_{\max} \tag{2}$$

of sound suppression, preferably when

$$Q_2 = \omega_{\min} /(\omega_{\max} - \omega_{\min}) << 1 \tag{3}$$

(c) the wave dimensions

$$Q_3 = \ell \omega_{\min} / c \tag{4}$$

of the device (where ℓ is the geometric size of the silencer, c - is a sound speed, preferably when $Q_3 << 1$);

$$Q_4 \approx 2\Delta p / \rho V_0^2 << 1 \tag{5}$$

(d) the relative loss of power to propel the gas through the silencer (ρ - gas mass density, preferably when $Q_4 << 1$). For a silencer, this condition is necessary

$$Q_1 = P_{out} / P_{in} \tag{6}$$

Always this combination is desirable (in addition to (1))

$$Q_2 << 1,\ Q_3 << 1,\ Q_4 << 1. \tag{7}$$

Violation of the conditions (7) (or one of these conditions) is usually (for systems with parameters that are constant in time) a payment for the fulfillment of the condition (6). Widely known are the following three approaches [1] in the design of such devices: when narrow-band ($Q_2 > 1$) absorption is permissible

*Vladimir Arabadzhi: Institute of Applied Physics (RAS), 603950, Nizhny Novgorod, Ulianov st. 46, Russia, v.v.arabadzhi@appl.sci-nnov.ru

and one can be satisfied with resonant absorption of active suppression of sound; when large silencer dimensions are permissible ($Q_3 \geq 1$) and one can be satisfied by a Helmholtz resonator with a resonant frequency below the minimum sound frequency $\omega_{\min}$ in the flow at low energy costs to propel the gas; when the high energy expenditure for pushing the gas through a classical rigid plate, perforated with small holes ($Q_4 \sim 1$).

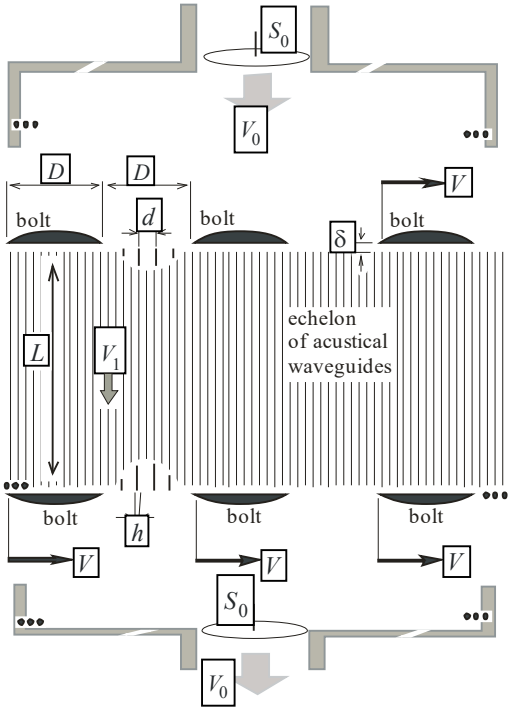


Figure 1. Two-dimensional version of cyclical wave bolt

2. Cyclical Wave Bolt

The above solutions [1] for blocking sound in a stream (with or without sound absorption) were based on systems with time-constant parameters. We now consider a system with parameters modulated in time [2]. We consider the infinite echelon (see Figure 1) of waveguides with rigid vertical plane walls, located at a characteristic distance

$$d >> h \tag{8}$$

from each other, where h is characteristic thickness of walls. Two horizontal systems of rigid thin bands (wave bolts) of the width D, separated from each other by the distance

$$D >> d \tag{9}$$

move horizontally (to the right) with the velocity V near the upper and lower edges of the echelon of above mentioned waveguides of length L. A sound wave is incident on top of this system and an excessive pressure Δp is applied to push the gas flow through this system. The gap between bolts and ends of waveguides is limited by condition $\delta < d$. We call such a parametric structure a cyclical wave bolt (CWB). From above, a sound wave falls on the echelon of waveguides, periodically blocked by wave bolts. If the conditions

$$L / c = D / V \tag{10}$$

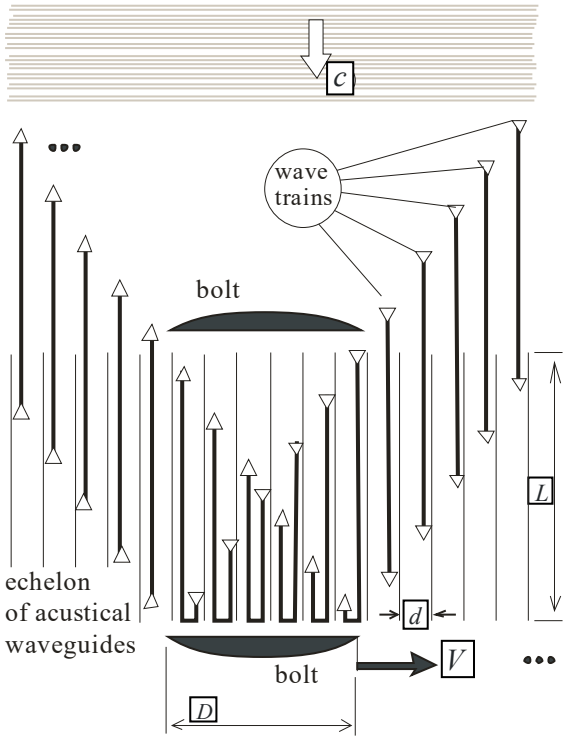


Figure 2. Space diagram of seven incoming wave trains, moving inside the echelon of waveguide channels

of synchronism are satisfied, the incident wave either reflects upward from the upper running bolts, or from the lower ones (Figure 2). Due to condition (10), each waveguide is open at both ends for a certain period of time D / V and is closed by wave bolts at both ends during a subsequent time interval of the same duration D / V. During the first interval (with open bolts), the sound wave fills this waveguide, during the subsequent interval (with closed bolts), the wave turns completely to reverse motion (due to full reflection from closed wave bolts). Thus, any perturbation propagating in a gas at a speed c cannot overcome CWB. On the other hand, for a gas flow this system is transparent. The only limitation of its transparency for the flow is the friction of the gas against the walls of the waveguides, which leads to additional energy loss. Waves, reflected not from the bolts, but from the open ends of the waveguides, do not leave the CWB. In principle, the operation of the system will be the same if the bolts are made stationary, and the train vertical walls is made to move at speed V.

3. Three-Dimensional Version of Cyclical Wave Bolt

Figure 3 shows a three dimensional cylindrical analogue (axially symmetric device) of the infinite two dimensional system described above.

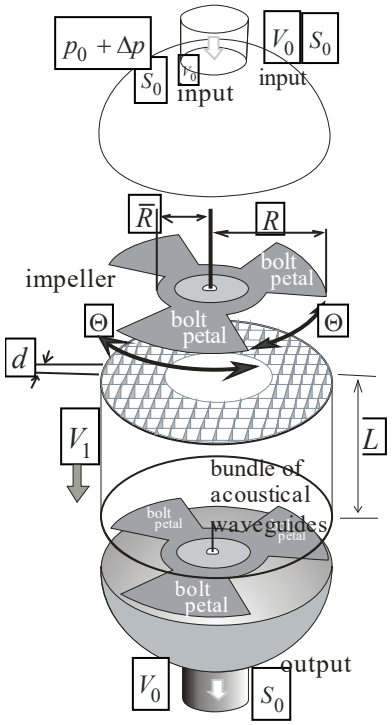


Figure 3. Three-dimensional version of silencer.

An infinite spatially periodic two dimensional bolt system is replaced by rotating (with angular frequency Ω) acoustically rigid (for sound in gas) impeller or rotary switch with N petals (bolts) with angular sectors $\Theta = \pi / N$ of the petals and the same angular sector between adjacent petals. Both ends of each waveguide of length L with a characteristic transverse dimension d (with rigid walls of thickness $h << d$) are open during time intervals $\tilde{t} + 2Tn \le t \le \tilde{t} + T + 2Tn$ and are closed at intervals $\tilde{t} + T + 2Tn \le t \le \tilde{t} + 2T + 2Tn$ of time by bolts (or petals of rotating impeller), where $\tilde{t}$ is a certain initial moment, determined by the point of placement of this waveguide and the phase of rotation of the impeller, where

$$T = \Theta / \Omega = L / c . \qquad (11)$$

The relation (11) is analog of synchronism condition (10). A bundle of tubes (waveguides) is located between the radii $\bar{R}$ and R. Any wave train that gets into the waveguide of length L in time $T = L / c$ has time to fully rotate to exit this waveguide. The maximum rigidity of structural elements and the maximum rotor speed are determined by modern state of technology level. We note that the gas velocity $V_1 = \chi V_0$ inside the waveguide tubes must ensure that the Mach number is small, i.e. $V_1 / c << 1$. By decreasing the value $\chi = S_0 / \pi(\bar{R}^2 - R^2) < 1$ we simultaneously reduce the Mach number inside the waveguides and the relative power loss Q_4. The rotor of the impeller is driven by an electric motor with a stabilized rotational frequency Ω. The frequency of incident waves in the flow is bounded from above only by the condition of one dimensionality of sound propagation in the rigid tube, i.e. its characteristic transverse scale d, i.e. $\omega_{\max} << c / 2\pi d$. Minimum cross-sectional dimension d of the tubes is limited below by the condition of a fairly low energy expenses to push gas trough the waveguides. The maximum cross-sectional dimension of the tubes is bounded above by the condition $d / V << L / c$. That the switching time of one tube is small compared with the propagation time of sound along the tube. Minimal length $L = c\Theta / \Omega$ of the tubes-waveguides is limited from below by the maximum achievable frequency Ω of the rotation of impeller. The minimum gap δ between the impeller (wave bolts, petals) and the ends of the tubes is limited from below by the manufacturing accuracy and stiffness.

Conclusions

Silencer considered can provide a low level of sound at the output, less engine power loss, a wider bands of sound suppression at smaller wave dimensions or, in other words, joint fulfillment of conditions (6), (7). The loss of power to push gas through acoustic waveguides can be reduced by shortening L and by increasing the rotational speed Ω of the impeller (until too much power is required to rotate it). The input and output of device can be interchanged, it acts symmetrically: it does not pass through itself the perturbation propagating in gas at the speed of sound.

References

[1] H. Fuchs, Applied Acoustics: Concepts, Absorbers, and Silencers for Acoustical Comfort and Noise Control: Alternative Solutions – Innovative Tools – Practical Examples, Springer Science & Business Media, 2013.

[2] V. Arabadzhi, Solutions to Problems of Controlling Long Waves with the Help of Micro-Structure Tools, Bentham Science Publishers Ltd., 2011.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 275-284 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning

Batoul Haidar*,[1], Maroun Chamoun[1], Ahmed Serhrouchni[2]

[1]Saint Joseph University, Lebanon

[2]Telecom ParisTech, France

ARTICLEINFO

Article history:
Received: 12 November, 2017
Accepted: 03 December, 2017
Online: 23 December, 2017

Keywords:
Cyberbullying
Machine Learning
Natural Language Processing
Arabic Natural Language Processing

ABSTRACT

With the abundance of Internet and electronic devices bullying has moved its place from schools and backyards into cyberspace; to be now known as Cyberbullying. Cyberbullying is affecting a lot of children around the world, especially Arab countries. Thus, concerns from cyberbullying are rising. A lot of research is ongoing with the purpose of diminishing cyberbullying. The current research efforts are focused around detection and mitigation of cyberbullying. Previously, researches dealt with the psychological effects of cyberbullying on the victim and the predator. A lot of research work proposed solutions for detecting cyberbullying in English language and a few more languages, but none till now covered cyberbullying in Arabic language. Several techniques contribute in cyberbullying detection, mainly Machine Learning (ML) and Natural Language Processing (NLP). This journal extends on a previous paper to elaborate on a solution for detecting and stopping cyberbullying. It first presents a thorough survey for the previous work done in cyberbullying detection. Then a solution that focuses on detecting cyberbullying in Arabic content is displayed and assessed.

1. Introduction

As stated in the Abstract, this work is an extension to the author's previous work [1] which was published in the EMS2016 conference.

Children and teens were exposed to physical bullying before the abundance of internet, computers and handheld devices. Nowadays, bullying is performed using cyber technology. Around 50% of the youth of America are suffering from cyberbullying [2]. As for the Arab world: 20.9% of middle-school adolescents report bullying in UAE, 31.9% in Morocco, 33.6% in Lebanon, 39.1% in Oman and 44.2% in Jordan [3].

Awareness of cyberbullying is rising in the Arab world. Arab News [4] declares that they heard of numerous cyberbullying incidents in Saudi Arabia. One of the rare reports [5] on cyberbullying states that 60% of Gulf Countries' youth openly admit the presence of cyberbullying amongst their peers. This study also states that only quarter the predators online do bully their victims offline. Which means that internet have encouraged three quarters of the predators to bully others, while they wouldn't have considered bullying face to face.

Most of the previous research dealing with cyberbullying focused on the effects of cyberbullying. It was concerned with helping the victims after cyberbullying attacks, mainly from the psychological aspect. Less work was directed towards implementing technical methods to detect and stop an ongoing cyberbullying attack, or even to prevent cyberbullying attacks while or before they happen [6]

After the Introduction, this paper furnishes a background that covers the technologies underlying cyberbullying detection in Section 2. Then in section 3 presents a survey of all existing literature in both cyberbullying detection and multilingual techniques. It concludes at the end of section 3 that there is no work done on detecting cyberbullying attacks in Arabic language. Therefore, in section 4 a system for cyberbullying detection in Arabic content is proposed. This system is designed for preventing cyberbullying attacks, by detecting and stopping them. It uses Natural Language Processing (NLP) to identify and process Arabic words. Then Machine Learning (ML) techniques are used to classify bullying content. In section 5 the results obtained from the system are displayed and analyzed. Conclusions and future works are finally stated in section 6.

2. Background

2.1. Cyberbullying

Cyberbullying is defined as the use of Internet, cell phones, video game systems, or other technologies to send or post text or

* Batoul Haidar: batoul.haidar@net.usj.edu.lb

images intended to hurt or embarrass another person or group of people [7]. Some examples of cyberbullying include sending mean or threatening messages, tricking someone into revealing personal or embarrassing information and sending it to others, sending or forwarding private messages to others, sharing explicit pictures with others without consent, starting rumors via text message or online or creating fake online profiles on websites such as Facebook, Myspace, Twitter, etc. to make fun of people [7]. There are several categories of cyberbullying as stated by [8] and [9] :

- Flaming: starting a form of online fight.
- Masquerade: where there is a bully pretending to be someone else, in order to perform malicious intents.
- Denigration: sending or posting gossip to ruin someone's reputation.
- Impersonation: Pretending to be someone else and sending or posting material to get that person in trouble or danger or to damage that person's reputation or friendships.
- Harassment: Repeatedly sending profane and cruel messages.
- Outing: Publishing someone's embarrassing information, images or secrets.
- Trickery: Talking someone into revealing secrets or embarrassing information for the sake of sharing it online.
- Exclusion: Intentionally and cruelly excluding someone from an online group.
- Cyberstalking: Repeated, intense harassment and denigration that includes threats or creates significant fear.

Bullying and cyberbullying leave mental and physical effects on both the bully (predator) and the victim. Cyberbullying is more severe than physical bullying due to the fact that it is wider, public, and the victim has nowhere to escape. Victims of cyberbullying reported emotional, concentration, and behavioral issues, as well as trouble getting along with their peers. These victims were more likely to report frequent headaches, recurrent stomach pain, and difficulty sleeping. One out of four students revealed that they felt unsafe at school. They were also more likely to be hyperactive, have conduct problems, abuse alcohol, and smoke cigarettes [10].

As stated previously, cyberbullying detection systems mainly use Machine Learning and Natural Language Processing techniques in the course of detection. Thus before dwelling in the previous literature in the area of cyberbullying detection, a thorough background for ML, NLP and other techniques collaborating in the process of cyberbullying detection is displayed.

2.2. Machine Learning

Machine Learning (ML) is defined as the ability of a computer to teach itself how to take a decision using available data and experiences [11]. Available Data is known as *Training Data*. Decisions to be taken in ML might be a classification or prediction for new objects or data. The computer classifies a new piece of data by depending on learning algorithms. When the training data is labeled, i.e. classified by human experts, the algorithms depending on these labeled data are called Supervised Learning algorithms [12].

In cyberbullying detection, there could be a corpus of data manually labeled (or classified) by people as either containing harm or not, as described in Section 3. When the training data is unlabeled, the algorithms depending on these non-labeled data are called Unsupervised Learning algorithms [12]. They teach themselves how to classify the data based on similarities and differences between data. When both supervised and unsupervised learnings are combined together by using labeled and unlabeled data, to get the most out of both ways, the algorithm is known as Semi-supervised Learning algorithm [12].

When ML is used to classify a certain object as belonging or not belonging to a certain category, the machine learner is called *Binary Classifier* [13]; for example in spam email filtering, ML algorithms are used to take decisions against incoming emails and label them as either spam or not spam. A second type is when the task given to the classifier, is to match a certain object against several classes or categories, then it is called *Multi-Class Classifier*. A third type might be predicting a value for an object and is called *Regression*, i.e. predicting a priority level for an incoming email.

There are several ML algorithms available, from which the most frequently used in relevance to the scope of research of this paper will be mentioned.

- Naive Bayes: A probabilistic supervised learning method [14] that mainly calculates the probability of an item belonging to a certain class, depending on metrics obtained from training data. Naïve Bayes algorithm was used in some cyberbullying detection research, such as in [15] and [14]. It was used for sexual predation detection.
- Nearest Neighbor Estimators: A simple estimator [16] that uses distance between data instances, in order to map a certain instance to its closest distance neighbor, thus estimating the class of this instance, this algorithm was used in [17] and [18].
- Support Vector Machine (SVM): Also a supervised algorithm. SVM is a binary classifier that assumes a clear distinction between data samples. It tries to find an optimal hyper plane that maximizes the margin between the classes [12]. SVM was used in many cyberbullying detection systems [19] , [20].
- Decision Tree: Decision tree learners use a set of labeled data, thus they are supervised learners. Decision trees classify data using a command and conquer approach. The Trees compose of leaves and arcs. Each leaf of the tree represents a classification class and each arc represents a feature inspected from training data [21]. The C4.5 algorithm is an implementation of decision trees. It was employed in cyberbullying detection by [22] and [17].

ML algorithms are widely incorporated in cyberbullying detection systems as seen in Section III, due to the huge amount of

data incorporated in social networking platforms, which makes it hard to be processed by human power, thus comes the need for a machine learner.

2.3. Natural Language Processing

Natural Language Processing (NLP) is the collection of techniques employed to make computers capable of understanding the natural unprocessed language spoken between humans by extracting grammatical structure and meaning from input [23].

NLP is a common branch of all of Linguistics, Artificial Intelligence and Computer Science [24]. NLP research started with Machine Translation in the late 1940s [25]. Then it spread to other areas of application, such as information retrieval, text summarization, question answering, information extraction, topic modeling, opinion mining [26], optical character recognition, finding words boundary, word sense disambiguation, and speech recognition [24].

According to Chandhana [27], NLP can be divided into three areas; Acoustic – Phonetic: where acoustic knowledge studies rhythm and intonation of language; i.e. how to form phonemes, the smallest unit of sounds. Phonemes and phones are aggregated into word sounds. Phonetic knowledge relates sounds to the words we recognize. Morphological – Syntactic: Morphology is lexical knowledge which studies sub words (morphemes) that would form a word. Syntactic knowledge studies the structural roles of words or collection of words to form correct sentences. Semantic - Pragmatic: Semantic knowledge deals with the meaning of words and sentences, while pragmatic knowledge deals with deriving sentence meanings from the outside world or outside the content of the document [28].

2.4. Sentiment Analysis

Sentiment analysis is a textual analysis technique. Sentiment analysis is used to define the polarity, subjectivity or features of a certain text. By polarity we mean defining whether a certain content is positive, negative or neutral [29]. Both Machine Learning and Natural Language Processing techniques are incorporated in Sentiment Analysis.

Sentiment analysis is currently used in detecting the opinions of social media users. Opinions are studied in several areas such as marketing and politics

2.5. Common Sense Reasoning/Sentic Computing

Common sense is the knowledge (usually acquired in early stages of life) concerning social, political, economic and environmental aspects of the society we live in. Common sense usually varies among different cultures and is built from layers of learning experiments we acquire throughout life [30].

Computers do not have common sense reasoning by nature, but there is a research field, known by Sentic Computing [31], that aims towards transforming computers into machines that could feel. This field of research is a multidisciplinary approach to opinion mining and sentiment analysis, which uses common sense reasoning and web semantics in order to inspect the emotions, not just the opinions from certain text. The term 'Sentic' derives from the Latin 'Sentire', the root of words like sentiment and sensations.

2.6. Performance Measures

Some evaluation metrics were adapted in Information Retrieval (IR) and then extended to other fields of computer science such as ML. These evaluation metrics are used as measures for the performance of IR and ML systems. The most widely used metrics are Recall, Precision, Area under the ROC and F-Measure.

- Recall is the proportion of returned documents (or values) which are relevant (or correct) $Rl \cap Rt$ [32] out of all relevant documents returned and not returned [33]. The metric is also known as Sensitivity of a system.

$$R=(Rl\cap Rt)/Rl \quad (1)$$

- Precision is the proportion of returned documents (or values) which are relevant (or correct) Rl∩Rt [32]. The metric is also known as Accuracy of a system.

$$P = (Rl \cap Rt)/Rt \quad (2)$$

- F-Measure, proposed by van Rijsbergen in 1979, is a weighted harmonic mean of precision and recall. It is a combination between Recall and Precision metrics, which was introduced to overcome the negative correlation between Precision and Recall [34].

$$F\beta=(1+\beta^2)PR/(\beta^2P+R) \quad (3)$$

- F1 is a special case of F- measure with β =1. β Is a parameter to control the balance between Recall and Precision where $0\leq \beta \leq \infty$. When β is set to 0, it implies giving no importance to recall, when β tends to ∞ then no importance is given to precision, and when $\beta = 1$ then Recall and Precision will be given equal importance [35].

$$F1= 2PR/(P+R) \quad (4)$$

P: Precision
R: Recall
Rt: Returned documents
Rl: Relevant documents

- ROC (Receiver Operating Characteristics) graph [36] –or area- is an analysis technique which had been used originally in medical diagnosis, to be later adopted in Machine Learning evaluation. ROC depicts the tradeoff between True Positive and False Negative rates. ROC areas are categorized roughly according to the values:

0.9-1 = excellent (A)
0.8-0.9 = good (B)
0.7-0.8 = fair (C)
0.6-0.7 = poor (D)
0.5-0.6 = fail (F)

3. Previous Work

As stated previously, the research efforts in cyberbullying covered several areas, including the detection of online bullying when it occurs; reporting it to law enforcement agencies, Internet service providers and others for the purpose of prevention and awareness;

and identifying predators with their victims. No effort was directed towards detecting cyberbullying in Arabic language. This section focuses on the previous work done in the areas of cyberbullying detection, ML, feature extraction and cross language transliteration and translation since those are the main techniques used in the implementation of cyberbullying detection systems.

3.1 Cyberbullying Detection

Most of the research done in detecting cyberbullying constituted of either a filtration software or ML techniques. A filtration software has to be employed by social networking platforms, in order to automatically delete or shade profane words [37] [38] [39]; but the filtration method is first limited by its inability for detecting subtle language harassment and second it has to be manually installed [40].

Most work other than filtration methods employs ML techniques, where old corpora of comments or conversations is crawled, whether from Facebook, Twitter, Formspring (a platform similar to Facebook, popular between teens) or even real conversations of sexual attackers [41]. These corpora are used to feed ML algorithms responsible for detecting cyberbullying attacks by building a classification rule from the training set. The obtained classification rule classifies the testing set comments. Such work was done in [22] where the authors crawled data from Formspring and used the Amazons Mechanical Turk [42] for labelling comments. Then they used the learning methods from the Waikato Environment for Knowledge Analysis (WEKA) toolkit [43] to teach and test the model for classifying comments.

The problem of detecting subtle language cyberbullying attacks was tackled by Dinakar et al [6]. They depended on commonsense reasoning in the detection of cyberbullying content. As an example of the commonsense they used: they considered comments of wearing makeup when subjected on Males might indicate the presence of harassment. They built their datasets from both YouTube and Formspring for both training and testing. They used NB, JRip, J48 and SVM for text classifications. For feature sets they used general features, such as a list of unigrams or profane words, tf-idf weighting scheme, Ortony Lexicon for negative affect, Part-of-speech tags for commonly occurring bigrams, and Label Specific Features including frequently used forms of verbal abuse.

The weighting scheme tf-idf is the product of term frequency and the inverse document frequency in the dataset. It involves multiplying term frequency (tf), that represents the number of times a term occurs in a document, by inverse document frequency (idf), which varies inversely with the number of documents to which a word is assigned [44].

Nahar, Li and Pang [40] employed the tf-idf weighting scheme for building features. In addition they built a network composing of bullies and their victims. The network was used to rank the most active predator and its target. In [45] Dinakar et al. stated that detecting profane language cyberbullying is easier than detecting sarcasm and subtle language attacks. Chayan and Shylaja of [46] enhanced the performance of the cyberbullying detection model by 4%, through looking for comments directed towards peers by using Supervised ML and Logical Regression models. However they didn't detect sarcasm comments. Dadvar et al. [47] state that incorporating user context such as the user's history as a feature for training the cyberbullying detection model increases accuracy of classification; however they didn't include sarcasm detection in their system.

SVM was also used by Yin et al. [48] for classifying posts as containing harassment or not. They used documents from CAW 2 dataset, which included posts from Kongregate, Slashdot and Myspace. For feature selection they incorporated several features;

- *Local features,* they used tf-idf.
- S*entiment features,* such as 1- grams, 2- grams and 3- grams, and they also captured second pronouns.
- *Contextual features* in which they studied both the similarity of a post to other neighboring posts and the cluster of posts neighboring around a certain harassment post.

Capturing sentiment features only didn't perform well so they compared the performance of their system by mixing features. Tf-idf performed better than n-grams and foul words, however, combining tf-idf with contextual and sentiment features achieved an additional enhancement in results in Precision, Recall and F1-measure. A similar work was done by Dadvar et al. [47], who built their feature space from Content-based, User based and Context based features. They also proved that incorporating contextual features such as gender information from the user's profile enhances cyberbullying detection.

Other research efforts were focused around social network profiles, such as [17]. They presented a methodology to detect and associate fake profiles on Twitter social networks to real users. This system had been capable of linking the owners of a fake account on Twitter to a real account for one or more students in a school class; this was a case of a real cyberbullying incident. The system was devised by collecting features from tweets then analyzing the features using various supervised ML techniques included in WEKA. Afterwards the performance among these techniques was compared on True Positive Ratio (TPR), False Positive Ratio (FPR) and Area Under ROC Curve (AUC). Bayzick, Kontostathis and Edwards [49] proposed the BULLYTRACER software which detects cyberbullying in chat rooms 58.63% of the time.

Chen et al. [50] proposed a new model for detection which they named as the Lexical Syntactic Feature-based (LSF) model; it achieved a precision of 98.24% and recall of 94.34%. Their model calculated both a post and a user's offensiveness depending on the ratio of offense appearing in a user's posts. This model detects "strong profanity" in online posts by using lexical analysis methods such as Bag of Words; and subtle language harassment to which the authors referred as "weak profanity". Then the model uses semantic analysis and NLP techniques to analyze the context of sentences by studying the grammatical relations among words. This research was an extension to the work presented in [51] for cyberbullying posts filtration.

Most of the research in cyberbullying did not give importance for the distinction between cyberbullying and cyberagression, but

for Hosseinmardi et al. [52]. They proposed a definition for cyberbullying which is the ***repetition*** of harmful actions using electronic devices over a certain period of time. They stated that most of the work previously done in detecting cyberbullying was actually focusing on detecting cyberagression. They define cyberagression as a single instance of harmful action that if repeated over time would be considered as cyberbullying. They also demonstrated that a Linear SVM classifier can significantly improve the accuracy of identifying cyberbullying to 87%. In addition they incorporated using features other than text such as images for better detection of cyberbullying.

Another study on cyber aggression was done by Nakano et al [53]. They analyzed anonymous and non-anonymous questions and answers from ask.fm. Their study shows that anonymous questions tend to be more aggressive than non-anonymous ones. They also showed that replies to anonymous questions tend to be less aggressive than replies to questions from known profiles.

Potha and Maragoudakis [54] stressed on a window of time in order to study the textual patterns of previous conversations in order of predicting the upcoming actions of a predator. They incorporated time series modelling in their research in addition to SVM for features selection. SVD (Singular Value Decomposition) [55] was used for feature reduction and DTW (Dynamic Time Warping) [56] for matching time series collections.

Fuzzy Logic and Genetic algorithms were also used in cyberbullying detection [57], where a new system was proposed using those two methods. This system's performance was compared across precision, recall and F1 measures. The system achieved better in Accuracy, F1-measure and Recall than previous fuzzy classification methods with 0.87, 0.91 and 0.98 respectively.

3.2 Arabic Language

Work related to Arabic language is scarce due to the complex morphological nature of Arabic. Arabic language is used by around 300 million Arabs around the world, mainly Muslims. It is a script language which is read and written from right to left and it constitutes of an alphabet of 28 letters. Vowels in Arabic are represented by special punctuation marks called Diacritics [58]. There are three variations for Arabic languages going on together. The Classical Arabic which is the language of the Islamic manuscripts -such as the Quran and prayers- and Arab people until Mid-20th century. The Modern Standard Arabic (MSA) which is the formal language used nowadays in schools and news, and it is known by all Arabs. Finally comes the dialects, which are accents for the Arabic language, usually used informally between people. There are around 10 dialects, one for every country - or group of countries [59], [60]. Arabic dialects imply a difference in meanings of words between different countries. We might find some words that are considered profane in one country, while good or ordinary in others, for example the word "Yetqalash" in Yemen is a compliment while in Morocco it is an offensive word [61].

Arabic language is a challenging and complex language due to its nature, where Arabic words do not include capitalization [60]. The morphologic nature of Arabic inflicts a lot of ambiguity, and the Arabic corpus is very scarce. Arabic language is ranked the 7th around the world, and its use over internet is growing vastly [59], thus arose the research interest in Arabic language fields.

An extensive search was performed on available articles and publications and no previous work for cyberbullying detection in Arabic language texts and comments was found. But some papers in the fields of ML and NLP were found. The previous work done in Arabic deals with text preprocessing or text classification.

Ghaleb Ali and Omar [62], proposed a key phrase extraction method that combines several key phrase extraction methods with ML algorithms. The output from the key phrase extraction methods is used as features to the ML algorithms. The ML algorithm in turn classifies the feature as either a key phrase or not. They compared their results by using three ML algorithms: Linear Logistic Regression [63], SVM and Linear Discriminant Analysis [64]. They have proved that SVM gives the best results in key phrases extraction among the three algorithms.

Some work had been done in Arabic named entity extraction, such as Named Entity Recognition for Arabic (NERA) [65] to identify proper names in Arabic documents. NERA used a white-list of named entities and corpora compiled from various sources; its performance was measured across recall, precision and F1. The results were satisfactory; 86.3% 89.2% 87.7% respectively for person named entities.

Filtering for spam emails written in Arabic and English was done by El-Halees [66] on pure English, pure Arabic and mixed collections of emails. Several ML techniques were used, including SVM, NB, k-Nearest Neighbor (k-NN) [67] and Neural Networks. The performance of the system was measured across all three variations and SVM was proved to be best in pure English environment. The system performed less on pure Arabic emails, due to the inflictive nature of Arabic Language. The authors also proved that stemming Arabic words enhances the performance of the classifiers, where NB performed best with 96.78% Recall and 92.42% F1-measure.

Other than emails there are also attempts for detecting spam in social networks, such as Twitter. Such work was done by El-Mawass and Alaboodi in [68]. They elaborated a system to detect spam in Arabic tweets. Their system achieved significant accuracy, precision and recall measures.

Sentiment analysis is one of the text classification categories. Sentiment analysis classifies a certain text as positive, negative or neutral [69]. Sentiment analysis was done by Hamouda [70] on Facebook comments written in Arabic. They built a corpora from 6000 comments sampled from Facebook, preprocessed this corpora, and then applied classifications to determine the sentiment behind the comment. Three classifiers were used: SVM, NB and Decision trees. The best performance was achieved by SVM with 73.4% accuracy. Another attempt for sentiment analysis was done in [71] for Arabic Tweets, their special contribution was in handling Arabizi and dialects. They incorporated NB, SVM and k-NN for classification and the best accuracy was approached by NB.

In [72], Duwairi detected sentiments from dialectical Arabic texts. Two methods were applied for detection. First, by translating dialectical words into MSA, then detecting according to MSA

lexicon. Second method was by detecting dialectical lexicon. NB and SVM classifiers were used to detect both negative and positive polarities. The results obtained showed improvement in Precision, Recall and F-measure upon translating into MSA.

Sentiment analysis was applied on Arabizi also by Duwairi et al [69]. In their system they first converted Arabizi text into Arabic by using their own rule based method. They labeled their data using their crowdsourcing tool [71] and then applied SVM and NB for classification. A comparison between SVM and NB showed that SVM outperformed NB. However better results were achieved when they first eliminated neutral entries from the dataset.

Arabic tweets in Saudi Arabia were analyzed by Alhumoud, Albuhairi and Altuwaijri [73]. They analyzed the tweets using a hybrid approach. Their hybrid approach composed of building a classifier and training it using a one-word dictionary. They compared the results obtained from both their hybrid approach and the supervised learning approach. Two classifiers were used from WEKA, NB and SVM. Their results showed the outperformance of the hybrid approach.

A significant research effort was done on stemming for Arabic language. Stemming is a text preprocessing technique. In stemming words are truncated to obtain their roots [74]. Several stemmers for Arabic are available, including rule-based stemmers such as Khoja's [75] and light stemmers. Light stemmers blindly remove letters from words –affixes and suffices – without prior knowledge of roots [76]. Stemmers are either monolingual or multilingual. Gadri and Moussaoui [77] elaborated a multilingual stemmer. Their stemmer is Language independent and it used the n-gram technique. This stemmer segments words into bigrams, then statistical measures are used to reach the best root. This stemmer was tested against English, French and Arabic. The best success rate (94%) was achieved in small Arabic Datasets. In large datasets, the best results were for English (86, 50%) and the worst for Arabic (67, 66%).

4. Proposed System

As stated previously, one of the purposes of this journal is presenting a solution for the problem of cyberbullying in both English and Arabic languages. As seen in the previous section, there is some work done for detection in English, but none in Arabic Language. Proving the hypothesis that Arabic cyberbullying can be detected was a challenge. Thus in the first stage of the system the focus was on detecting cyberbullying in Arabic language.

Since the proposed system employs ML, a dataset had to be prepared to be used for training and testing the system. Two toolkits were tested for ML, Dataiku DSS and WEKA. The decision was to use WEKA toolkit because it supports Arabic language.

4.1 Data Preparation

In order to train and test the system, a huge amount of data had to be obtained. Thus, the choice was to scrap data from both Facebook and Twitter. This choice was inflicted by the fact that those two social media portals are the most widely used by the Arab nation, especially Arab youth.

For the data acquirement phase two custom tools were built. Those tools were implemented to scrap data from social networks, one for Facebook and the latter for Twitter. Twitter Scrapper was written using PHP while Facebook scrapper incorporated python. Both scrappers connected to a mongo dB server, where all downloaded tweets and messages were stored.

In the process of getting the data from Twitter, the Twitter Scrapper searches for Tweets according to a given location and perimeter. A huge database was accumulated using this tool. In order to get the tweets from the Arab countries we focused on the center of the Middle East Region as a location, and selected an area of 10,000 kilometers radius in each run. Thus the tweets were collected from Lebanon, Syria, Gulf Area and Egypt mainly. The size of the tweets database summed up to 4.93 GB.

Collecting Facebook data was harder, due to the restrictions imposed by Facebook security and privacy measures. In Facebook the data source had to be specified beforehand. It is specified by page IDs. The aim was for pages of public figures and news agency, since those pages include more discussions and interactions between Facebook users. The tool had to be run many times to get data. Each run required that the Facebook Page ID is specified to collect the data from it. Many requests were blocked due to privacy measures on the targeted pages. The Facebook database reached 0.98GB of size, thus the decision was to keep Facebook data for validating the system in a later stage and use the tweets dB for training and testing the system at the current stage.

4.2 Data Labelling and Preprocessing

Before the system training phase, the obtained data had to be cleaned and preprocessed. WEKA was used for this purpose as mentioned above. For the system in hand, version 3.9.1 of WEKA was used since it contains specific packages that will be mentioned later.

The only features included in the first stage were text (the content of the tweet) and language. Several languages were observed in the database, mostly Arabic, English and Turkish. All the tweets of languages other than English and Arabic were discarded. The remaining tweets were separated into two datasets, one having only tweets in Arabic language and the latter English. The dataset with only Arabic language was denoted by ArabicTweets. This Dataset contained 35273 unique tweets after removing all duplicates. While the English Dataset contained 91431 tweets.

Due to the sensitivity of Arabic language and to ensure the correct labelling of sarcastic cyberbullying content, Arabic tweets were labelled manually by adding an extra attribute to the dataset, which is the "bullying". This attribute was assigned to be the class attribute, where labelling will take place and it will be used by WEKA to evaluate the performance of the system. The values of bullying instances are either "yes" or "no". "Yes" is when the tweet contains cyberbullying and "no" otherwise. The server used to preprocess the datasets was a 3.5 GHz CPU and 32GB RAM's windows 10 virtual machine.

4.3 System Training and Testing

Training and classification procedures were done several times for the purpose of reaching the best results. Two models were chosen in the first stage, Naïve Bayes and SVM. Those two models were chosen based on analysis of the previous work done in the field of ML. Researchers reached a conclusion that those are the best two algorithms for text classifications – as mention in section 2 [54, 57, 59, and 61]. In the first model the system was trained using Naïve Bayes and promising results were achieved. The precision was 90.8514 % for the overall system. However this precision is somehow tricky. The dataset contained only 2196 bullying content out of the total 35273. Thus even if in the worst case the system classifies all the tweets as not bullying then it will be still achieving a high precision. Hence the analysis of the performance was shifted towards the actual cyberbullying content which was detected (or in other words classified as "yes"). The system in the first run was capable of detecting 801 out of 2196 actual bullying instances (or tweets) which is a good result and it proves the aim of the paper: Cyberbullying in Arabic is detectable!

In training the second model; the AffectiveTweets package was used [78] and specifically the TweetToSentiStrengthFeatureVector filter. This filter depends on the SentiStrength mechanism developed by Thelwall et al [79]. SentiStrength originally works with English, but it is capable of supporting other languages. SentiStrength works on weights of tweets: positive weights (2 to 5) are given to words denoting positive feelings and negative weights for negative feelings (-2 to -5). 1 and -1 are considered neutral. Then a list of Arabic profane words was collected and weighted manually. The weight given for profane words was (-5) which is the extreme negativity in SentiStrength. Afterwards the English lexicon files used by SentiStrength were replaced with custom built Arabic files including the weighted profane words.

In the process of customizing the SentiStrength files, the phrases added contained the profanity from several Arabic countries. This gave a chance for handling dialectical difference between Arabic countries implicitly. With more effort in including more dialectical varieties of phrases all the dialects in Arabic countries would be covered.

The dataset prepared had to be preprocessed before building the training model. Preprocessing was performed using WEKA. The set of techniques used for preprocessing included applying the TweetToSentiStrengthFeatureVector filter, converting strings to word vectors and normalizing. After that the system was trained using SVM. Training and testing the model is a time-consuming process (it took around 8 hours). As a preliminary stage, the system achieved results somehow close to the results achieved by Naïve Bayes in the first model. The model was capable of detecting 710 bullying instances. Those results are promising and they show that SentiStrength can be customized to be used in cyberbullying detection. The results achieved by the system are summarized in the next section.

5. Results Obtained

Table 1 shows the division of the classified instances. How many instances were classified correctly and how many wrong in each model and class. As seen, in Naïve Bayes model 31245 non-bullying instances were classified correctly. There are also 1832 non-bullying instances classified as containing bullying. As for the bullying instances, it is seen that 801 were classified correctly while 1395 were classified as not containing bullying.

Table 1: Classification Results

	Correctly Classified		Miss Classified	
Class	***no***	***yes***	***no***	***yes***
Naïve Bayes	31245	801	1832	1395
SVM	32479	710	1923	161

Table 2: Summary OF Rates

	Class	Precision %	Recall %	TP %	FP %	F-Measure %
Naïve Bayes	***no***	94.5	95.7	95.7	69.6	95.1
	yes	36.5	30.4	30.4	4.3	33.2
	Overall	90.1	90.9	90.9	64.7	90.5
SVM	***no***	94.4	99.5	99.5	73	96.9
	yes	81.5	27	27	0.5	40.5
	Overall	93.4	94.1	94.1	67.6	92.7

Table 2 displays the results obtained by WEKA after training and testing the model. Precision, Recall, True Positive Rate, False Positive Rate and F-Measure are displayed for both models, Naïve Bayes and SVM. In each model the measures mentioned are displayed for the two classes of the "bullying" attribute which are "yes" and "no". In the third row of every model there is the weighted measure from both classes.

There is a difference between the "yes" precisions of the two models noted in Table 2. SVM achieved a much higher precision for the "yes" class -0.815 for SVM and 0.365 for NB. This difference is due to the lower number of wrong classified instances by SVM as seen in Table 2.

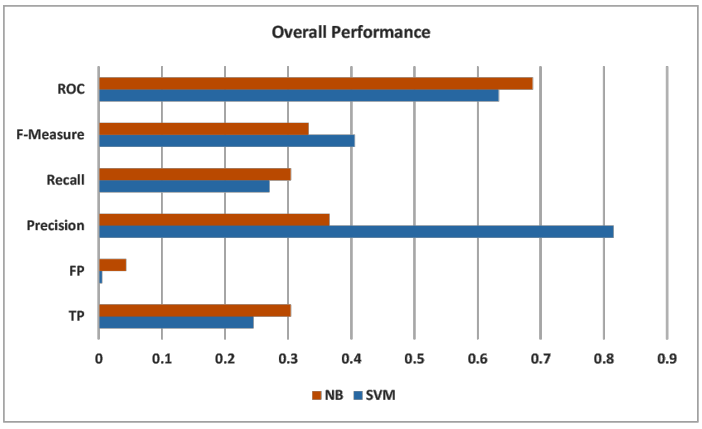


Figure 1: Difference in performance between NB and SVM

Figure 1 displays graphically the difference in performance between NB and SVM for the weighted class, or the overall per-

formance. As seen the precision achieved by SVM was much better than NB, while the other measures were somehow close, such as ROC, TP, FP, etc.

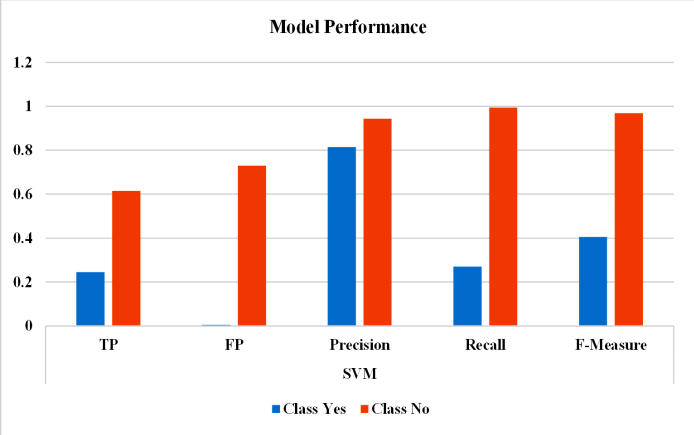


Figure 2: Comparison between the measures achieved by SVM model

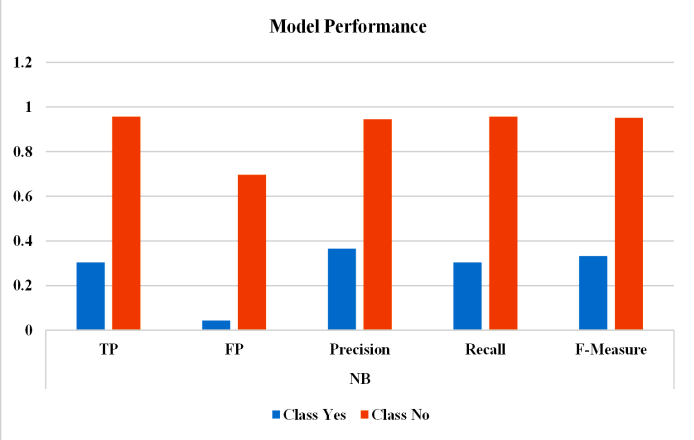


Figure 3: Results of NB Model

Figure 2 compares graphically between the measures achieved in both classes by SVM model, while Figure 3 displays the results of NB. As shown previously Table 1 and Table 2, the low number of False Positives, high precision and recall are significant.

The results obtained by this system are not perfect if compared with previous work done in English cyberbullying detection systems. But the aim of the paper was to prove that Arabic cyberbullying is detectable. Thus, the results displayed in the Tables and Figures above cannot be compared with previous work done for Arabic Language due to the negligibility of such work. Table 2 shows a recall of 30.4 for the "yes" class instances in the NB model, which means at least one third of the bullying in Arabic is detectable by the system.

The overall measures for both models give reasonable results. Although the focus was on the analysis of the "yes" class since it contains the bullying instances which are less than the "no" count. But the high precisions approached by the system denote that the model is not classifying a high number of non-bullying instances as bullying ones. Which is also an important point to consider.

6. Conclusion

The work done in this paper proves that detecting Arabic cyberbullying is possible. However, some effort is yet needed to enhance the performance of the system presented. Therefore, some future plans still need to be completed in later stages.

The first step is to enhance the performance measures achieved by the system through using hybrid training models, such as combinations of Distance Functions, NB and SVM. Treatment of Arabizi content is also in the future plans. This is a crucial point, since a lot of Arabic youth use the chat language on social networks for communication among each other. Thus the proposed system will be upgraded to handle cyberbullying when written in Arabizi.

Further plans incur training the system using deep learning methods instead of machine learning and then comparing the differences of the outcomes from the two schemes.

References

[1] B. Haidar, M. Chamoun and F. Yamout, "Cyberbullying Detection: A Survey on Multilingual Techniques," in UKSIM 2016, Pisa, Italy, 2016.

[2] K. Poels, A. DeSmet, K. Van Cleemput, S. Bastiaensens, H. Vandebosch and I. De Bourdeaudhuij, "Cyberbullying on social network sites. An experimental study into bystanders," Cyberbullying on social network sites, vol. 31, p. 259–271, 2014.

[3] S. S. Kazarian and J. Ammar, "School Bullying in the Arab World: A Review," The Arab Journal of Psychiatry , vol. 24, no. 1, pp. 37 - 45, 2013.

[4] M. Y. Ba-Isa, "Beaten up in cyberspace," 25 8 2010. [Online]. Available: http://www.arabnews.com/node/353552. [Accessed 16 9 2016].

[5] ICDL, "Cyber Safety Report: Research into the online behaviour of Arab youth and the risks they face," ICDL Arabia, 2015.

[6] K. DINAKAR, B. JONES, C. HAVASI, H. LIEBERMAN and R. PICARD, "Common Sense Reasoning for Detection, Prevention,and Mitigation of Cyberbullying," in ACM Transactions on Interactive Intelligent Systems, NY, September 2012.

[7] O. f. V. o. C. National Crime Prevention Council, "Cyberbullying Tip Sheets," National Crime Prevention Council, 2016. [Online]. Available: http://www.ncpc.org/topics/cyberbullying/cyberbullying-tip-sheets/. [Accessed 10 June 2016].

[8] N. Willard, "Educator's Guide to Cyberbullying and Cyberthreats," Center for Safe and Responsible Internet Use, 2007.

[9] N. Samaneh, A. Masrah, M. Azmi, M. S. Nurfadhilna, A. Mustapha and S. Shojaee, "13th International Conrence on Intelligent Systems Design and Applications (ISDA)," in A Review of Cyberbullying Detection . An Overview, 2013.

[10] D. Mann, "Emotional Troubles for 'Cyberbullies' and Victims," WebMD Health News, 6 July 2010. [Online]. Available: http://www.webmd.com/parenting/news/20100706/emotional-troubles-for-cyberbullies-and-victims. [Accessed 24 August 2015].

[11] T. M. Mitchell, "The Discipline of Machine Learning," CMU-ML-06-108, Pittsburgh, July 2006.

[12] P. Kulkarni, Reinforcement And Systemic Machine Learning For Decision Making, New Jersey: IEEE, WILEY, 2012.

[13] P. FLACH, MACHINE LEARNING The Art and Science of Algorithms that Make Sense of Data, Cambridge University Press, 2012.

[14] D. Vilariño, C. Esteban, D. Pinto, I. Olmos and S. León, "Information Retrieval and Classification based Approaches for the Sexual Predator Identification," Faculty of Computer Science, Mexico.

[15] H. José María Gómez and A. A. Caurcel Diaz, "Combining Predation Heuristics and Chat-Like Features in Sexual Predator Identification," 2012.

[16] A. S. a. S. Vishwanathan, Introduction to Machine Learning, Cambridge: Cambridge University Press, 2008.

[17] I. Santos, P. G. Bringas, P. Gal´an-Garc´ıa and J. Gaviria de la Puerta, "Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying," DeustoTech Computing, University of Deusto, 2013.

[18] I.-S. Kang, . C.-K. Kim, . S.-J. Kang and S.-H. Na, IR-based k-Nearest Neighbor Approach for Identifying Abnormal Chat Users, 2012.

[19] C. M. a. G. Hirst, Identifying Sexual Predators by SVM Classification with Lexical and Behavioral Features, 2012.

[20] D. E. L. a. A. B. Javier Parapar, "A learning-based approach for the identification of sexual predators in chat logs," 2012.

[21] Ron Kohavi and R. Quinlan, "Decision Tree Discovery," 1999.

[22] K. Reynolds, "Using Machine Learning to Detect Cyberbullying," 2012.

[23] S. Ahmad, "Tutorial on Natural Language Processing," Artificial Intelligence (810:161) Fall 2007.

[24] V. Gupta, "A Survey of Natural Language Processing Techniques," vol. 5, 01 Jan 2014.

[25] B. MANARIS, "Natural Language Processing: A Human–Computer Interaction Perspective," vol. 47, no. pp. 1-66, 1998..

[26] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research," IEEE ComputatIonal IntEllIgEnCE magazIne, May 2014.

[27] C. Surabhi.M, "Natural Language Processing Future," in International Conference on Optical Imaging Sensor and Security, Coimbatore, Tamil Nadu, India, July 2-3, 2013.

[28] G. G. Chowdhury, "Natural Language Processing," Annual Review of Information Science and Technology, vol. 37, no. 0066-4200, pp. 51-89, 2003.

[29] S. Dinakar, P. Andhale and M. Rege, "Sentiment Analysis of Social Network Content," in IEEE 16th International Conference on Information Reuse and Integration, 2015.

[30] E. Cambria, Application of Common Sense Computing for the Development of a Novel Knowledge-Based Opinion Mining Engine, University of Stirling, Scotland, UK, 2011.

[31] M. Grassi, E. Cambria, A. Hussain and F. Piazza, "Sentic Web: A New Paradigm for Managing Social Media Affective Information," Cogn Comput (2011) 3:480–489.

[32] W. E. Webber, Measurement in Information Retrieval Evaluation (Doctor of Philosophy), The University of Melbourne, September 2010.

[33] C. J. v. RIJSBERGEN, INFORMATION RETRIEVAL, University of Glasgow.

[34] N. Chinchor, "MUC-4 EVALUATION METRICS," in Fourth Message Understanding Conference, 1992.

[35] Y. Sasaki, "The truth of the F-measure," University of Manchester, 26th October, 2007.

[36] T. Fawcett, "An introduction to ROC analysis," in Pattern Recognition Letters 27 (2006) 861–874.

[37] WatchGuard, "Stop Cyber-Bullying in its Tracks - Protect Schools and the Workplace," WatchGuard Technologies, 2011.

[38] "https://blog.barracuda.com/2015/02/16/3-ways-the-barracuda-web-filter-can-protect-your-classroom-from-cyberbullying/".

[39] "Internet Monitoring and Web Filtering Solutions," PEARL SOFTWARE, 2015. [Online]. Available: http://www.pearlsoftware.com/solutions/cyber-bullying-in-schools.html. [Accessed 2 June 2016].

[40] V. Nahar, X. Li and C. Pang, "An Effective Approach for Cyberbullying Detection," in Communications in Information Science and Management Engineering, May 2013.

[41] "Perverted Justice," Perverted Justice Foundation, [Online]. Available: http://www.perverted-justice.com/.

[42] "Amazon Mechanical Turk," 15 August 2014. [Online]. Available: http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanical-TurkGettingStartedGuide/ SvcIntro.html. [Accessed 2 June 2016].

[43] S. Garner, "Weka: The waikato environment for knowledge analysis," New Zealand, 1995.

[44] "tf-idf: A single Page Tutorial," [Online]. Available: http://www.tfidf.com. [Accessed 13 May 2016].

[45] K. Dinakar , R. Reichart and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," Cambridge, 2011.

[46] V. S. Chavan and Shylaja S S , "Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network," in International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015.

[47] M. Dadvar, D. Trieschnigg, R. Ordelman and F. De Jong, "Improving cyberbullying detection with user context," 2013.

[48] D. Yin, Z. Xue, L. Hong, B. D. Davidson, A. Kontostathis and L. Edwards, "Detection of Harassment on Web 2.0," Madrid, Spain, April 21, 2009.

[49] J. Bayzick, A. Kontostathis and L. Edwards, "Detecting the Presence of Cyberbullying Using Computer Software," Koblenz, Germany, June 14-17, 2011.

[50] Y. Chen, S. Zhu, Y. Zhou and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," 2012.

[51] Z. Xu and S. Zhu, "Filtering Offensive Language in Online Communities using Grammatical Relations," Redmond, Washington, US, July 13-14, 2010.

[52] H. Hosseinmardi, S. Arredondo Mattson, R. IbnRafiq, R. Han, Q. Lv and S. Mishra, "Detection of Cyberbullying Incidents on the Instagram Social Network," 2015.

[53] T. Nakano, S. Tatsuya, Y. Okaie and M. J. Moore, "Analysis of Cyber Aggression and Cyber-bullying in Social Networking," in IEEE Tenth International Conference on Semantic Computing, 2016.

[54] N. Potha and M. Maragoudakis, "Cyberbullying Detection using Time Series Modeling," 2014.

[55] K. Baker, "Singular Value Decomposition Tutorial," 2013.

[56] M. Muller, "Dynamic Time Warping," in Information Retrieval for Music and Motion, Springer, 2007, pp. 69 - 84.

[57] B. Nandhinia and J. Sheebab , "Online Social Network Bullying Detection Using Intelligence Techniques," 2015.

[58] M. A. Attia, Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation, Doctor of Philosophy in the Faculty of Humanities, 2008.

[59] K. Darwish and W. Magdy, "Arabic Information Retrieval," vol. 7, no. 4, 2013.

[60] A. FARGHALY and K. Shaalan, "Arabic Natural Language Processing:Challenges and Solutions," vol. 8, December 2009.

[61] "12 Arabic Swear Words and Their Meanings You Didn't Know," [Online]. Available: http://scoopempire.com/swear-words-meanings-around-middle-east/#.V0fdjPl96M9. [Accessed 2 June 2016].

[62] N. Ghaleb Ali and N. Omar, "Arabic Keyphrases Extraction Using a Hybrid of Statistical and Machine Learning," in International Conference on Information Technology and Multimedia (ICIMU), Putrajaya, Malaysia, 2014.

[63] T. Haifley, "Linear Logistic Regression: An Introduction," IEEE, 2002.

[64] G. J. McLACHLAN, "Discriminant Analysis and Statistical Pattern Recognition," Wiley InterScience, New Jersey, 2004.

[65] K. Shaalan and H. Raza, "Arabic Named Entity Recognition from Diverse Text Types," Berlin Heidelberg, GoTAL 2008.

[66] A. El-Halees, "Filtering Spam E-Mail from Mixed Arabic and English Messages: A Comparison of Machine Learning Techniques," The International Arab Journal of Information Technology, vol. 6, no. 1, 2009.

[67] T. M. COVER and P. E. HART, "Nearest Neighbor Pattern Classification," IEEE TRANSACTIONS ON INFORMATION THEORY, vol. 13, no. 1, 1967.

[68] N. EI-Mawass and S. Alaboodi, "Detecting Arabic Spammers and Content Polluters on Twitter," in Sixth International Conference on Digital Information Processing and Communications (ICDIPC), 2016.

[69] R. M. Duwairi, M. Alfaqeh, M. Wardat and A. Alrabadi, "Sentiment Analysis for Arabizi Text," in 7th International Conference on Information and Communication Systems (ICICS), 2016.

[70] A. E.-D. A. Hamouda and F. E.-z. El-taher, "Sentiment Analyzer for Arabic Comments System," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 4, no. 3, 2013.

[71] R. M. Duwairi, R. Marji, N. Sha'ban and S. Rushaidat, "Sentiment Analysis in Arabic Tweets," in 5th International Conference on Information and Communication Systems (ICICS), 2014.

[72] R. M. Duwairi, "Sentiment Analysis for Dialectical Arabic," in 6th International Conference on Information and Communication Systems (ICICS), 2015.

[73] S. Alhumoud, T. Albuhairi and M. Altuwaijri, "Arabic sentiment analysis using WEKA a hybrid learning approach," in 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Lisbon, Purtogal, 2015.

[74] A. Al-Zyoud and W. A. Al-Rabayah, "Arabic Stemming Techniques: Comparisons and New Vision," in Proceedings of the 8th IEEE GCC Conference and Exhibition, Muscat, Oman, 2015.

[75] S. Khoja and R. Garside, "Stemming arabic text," Computing Department, Lancaster University, Lancaster, UK, 1999.

[76] L. S. Larkey, L. Ballesteros and M. E. Connell, "Light Stemming for Arabic Information Retrieval," in Arabic Computational Morphology, book chapter, , , Springer, 2007.

[77] S. Gadri and A. Moussaoui, "Information Retrieval: A New Multilingual Stemmer Based on a Statistical Approach," in 3rd International Conference on Control, Engineering & Information Technology (CEIT), 2015.

[78] S. M. Mohammad and F. Bravo-Marquez, "Emotion Intensities in Tweets," in Joint Conference on Lexical and Computational Semantics (*Sem), Vancouver, Canada, August 2017.

[79] M. B. K. P. G. C. D. &. K. A. Thelwall, "Sentiment strength detection in short informal text," Journal of the American Society for Information Science and Technology, vol. 61, no. 12, p. 2544–2558, 2010.

[80] Hewlett-Packard Development Company. L.P., 2013. [Online]. Available: http://www.autonomy.com/html/power/idol-10.5/index.html. [Accessed 2 June 2016].

[81] "Arabic chat alphabet," 23 May 2016. [Online]. Available: https://en.wikipedia.org/wiki/Arabic_chat_alphabet. [Accessed 2 June 2016].

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 285-289 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Applying Machine Learning and High Performance Computing to Water Quality Assessment and Prediction

Ruijian Zhang*[1], Deren Li[2]

[1]*Purdue University Northwest, Computer Science, 46323, USA*

[2]*Indiana University-Purdue University Indianapolis, Computer Science, 46202, USA*

ARTICLE INFO

Article history:
Received: 13 November, 2017
Accepted: 03 December, 2017
Online: 23 December, 2017

Keywords:
Machine Learning
Classification
Decision Tree
High Performance Computing
Water Quality Assessment and Prediction

ABSTRACT

Water quality assessment and prediction is a more and more important issue. Traditional ways either take lots of time or they can only do assessments. In this research, by applying machine learning algorithm to a long period time of water attributes' data; we can generate a decision tree so that it can predict the future day's water quality in an easy and efficient way. The idea is to combine the traditional ways and the computer algorithms together. Using machine learning algorithms, the assessment of water quality will be far more efficient, and by generating the decision tree, the prediction will be quite accurate. The drawback of the machine learning modeling is that the execution takes quite long time, especially when we employ a better accuracy but more time-consuming algorithm in clustering. Therefore, we applied the high performance computing (HPC) System to deal with this problem. Up to now, the pilot experiments have achieved very promising preliminary results. The visualized water quality assessment and prediction obtained from this project would be published in an interactive website so that the public and the environmental managers could use the information for their decision making.

1. Introduction

This paper is an extension of work of "Applying Parallel Programming and High Performance Computing to Speed up Data Mining Processing", originally presented in the 16th IEEE international Conference on Computer and Information Science, 2017 [1]. Water quality is an important issue. Not only can the drinking water directly influence our body as about 70% of the human body is made up of water, the natural water is also an important part as people like to have some entertainment activities near lakes or rivers. Nowadays the prediction of river or lake water qualities is more and more important. The traditional way to assess water quality is analyzing the water attributes' value. If the value is in a certain range, the quality of water can be classified. This is an easy and fast way, but it can only be used for assessment. To predict water quality, scientists need to use the chart to show the changes of an attribute in a period of time. This is a statistic way and can predict water quality in a relatively accurate way. The problem is it takes lots of time to analyze the water attributes and it can only predict a long period of time's water quality. As the results are analyzed by people, increasing water attributes for analyzing will greatly increase the workload [2, 3, 4, 5, 6, 7].

In this research, we use machine learning algorithm on water quality prediction. By applying machine learning algorithms, the increasing water attributes and cases are no longer a big problem. Compared with the traditional way, using machine learning can deal with a great number of data in a short time. It can also slightly reduce the human error on the analysis phase. The idea of implementation is not very complex. First, we use the K-MEANS algorithm to classify the data. The K-MEANS algorithm is a very basic, easy and important clustering algorithm. In this research, we specify K is 5, which means we want to classify the water data into five clusters to represent five water qualities from good, relatively good, medium, relatively bad, and bad. By analyzing each cluster's center; we will know which cluster represents what water quality. After that, we use the C5.0 algorithm to generate a decision tree [8]. In this research, we use five or seven water attributes for classification.

*Ruijian Zhang, 2200 169th Street, Hammond IN USA 46323, Phone: (219)9892706, zhangrj@pnw.edu

2. Significance

As water quality is more and more important, not only scientists but also the general publics are interested in the future natural water quality. Thus, the request to predict natural water quality like weather forecast is getting stronger. Unlike weather forecast which can use satellites to get, we can't use several satellites to cover a large area to collect water data. The data needs to be collected by observations in different places. It also takes far more time to analyze the data so that people can make a water quality prediction. The United States is a country with many rivers and lakes. The government has already established many water observations to monitor the water system of the country [9]. Collecting data is not so difficult. Making predictions in an efficient way is the key problem.

For this research, the most significant part is it completely uses the data from currently operational observations. This means the hardware cost of prediction is greatly reduced. As long as an observation station can provide enough data, this method can be used to predict water quality. By using machine learning technology, we can deal with a great amount of data in a short time, which can make water quality analyze very quickly. Although these kinds of water quality analysis may not be as accurate as biology analysis, it is accurate enough for water quality forecast to the generate public [10]. Compared with the traditional way for water quality research, the human factors in this research is relatively low. This can slightly reduce the possibility of human error.

3. Design and Implementation

There are four steps in this research: data selection, data preprocessing, classification, and assessment and prediction. The first two steps are preparing a suitable data for classifying. Step three is to make the classifications. It will be implemented by the K-MEANS algorithm. Step four will generate different decision trees for assessment and predicting by using the C5.0 algorithm.

Data selection is very important. This step is the root of this research. As this is a machine learning research, a "good" data definitely can greatly affect the results. The "good" here means it needs to have enough cases, enough attributes, and less missing data. Enough cases can ensure the number of training data. The accuracy of classifications is based on the number of training data. The more cases we have, the better the classifications will be.

The data we used was from the USGS website. USGS keeps the daily water data from 10472 observation sites all over the United States. Many of the sites keep about four attributes of the water. After checking all the sites keeping at least seven attributes, we finally choose the site numbered 08070200. This site is in San Jacinto River in Texas and the data is still updating till now. It keeps 7 attributes of the water, and has few missing values. In this research, we choose the data between 11/22/2005 and 09/08/2012 which has 2483 records [9].

The second step is data preprocessing. We made the data more meaningful for use. It included resetting the extreme data, and normalizing the data. To deal with the extreme data, we put the entire data of one attribute on a scatter chart and find the extreme value. Fig. 1 is an example of original "discharge" value shown in a chart. The x-axis represents different cases and the y-axis is value. It is very clear that the maximum value of discharge is about 8000. If we use these values directly and normalize them, the problem is most normalized value will be 0. This is definitely not good for classification as it cannot reflect the data distribution. To solve this problem, the idea is set all values greater than an "extreme value" to be equal to this "extreme value". In Fig. 2, we use the same data in Fig. 1 but set all values greater than 700 to be equal to 700.

Fig. 2 looks better than Fig. 1. It now can reflect the changes and distribution of the discharge. A problem here is how to set the "extreme value". In Fig. 2, we can see that there is not too much difference if we set the extreme value to 600 compared with 700, as most value are under 200. The key point is we do not want to change too much value as changing data is a kind of human factor. Making the charts have better data distribution with the least changes is the role of the extreme value. In this attribute, totally 136 values have been affected by setting the extreme value to 700, which is about 5% of the total 2483 values.

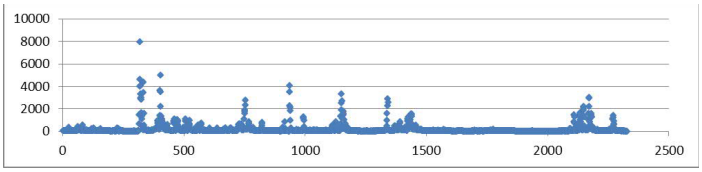


Figure 1, the original scatter chart of the discharge in this research. All values are either the original values or calculated values.

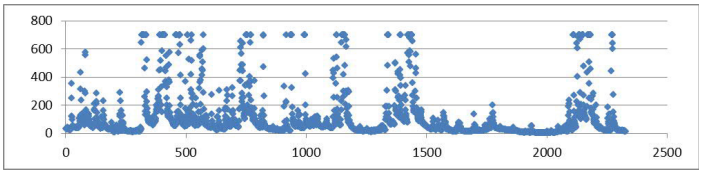


Figure 2, the new discharge chart which set all values greater than 700 to be equal to 700.

The third step of this research is classification. We want to classify the data into five clusters (classifications). Each cluster represents a certain water quality range from very bad, bad, medium, good, to very good. The method for classification is to implement the K-MEANS algorithm. The K-MEANS algorithm is a very basic but important algorithm for classification. It treats every node in an N-dimensional space. Each dimension represents an attribute. In this research, we used seven attributes, so each record was a node in 7-dimension spaces. We first gave five initial centers. Then based on the distance of each node and each center, we assigned every node a group. After that, we recalculated the distance of every node inside one group to move the center. So we repeated the iteration of assigning groups and moving centers until the distance between the new center and the previous center was smaller than a "quit distance" or the new center moved to a previous center node.

The last step of this research was assessment and prediction. Before we could make the assessment and prediction, the very first thing was to analyze which group represents what kind of water quality. This can be done by analyzing the center node's attributes. We needed to look up some references and some criteria of water. This was not easy but also not hard. The results of classification actually were the results for water quality assessment. We already classified the data into five groups. We could make a water quality assessment by generate a decision tree.

The main topic was how to make prediction. Before we could make the decision tree, we needed the training data. The format of the training data we used to generate the decision tree was "attributes" plus "classification results". If the training data was "day 1's attributes" plus "day 1's classification result", it was used to generate a decision tree for assessment. So, we "shifted" the classification results by 1day. The training data would be like "day 1's attributes" plus "day 2's classification result". This data can generate a decision tree for prediction. Fig. 3 shows how to "shift" the data. In this way, we lose one day's case as the last day only has attributes. This is a basic way for prediction. It uses previous 1day's attributes for prediction. An improved way is using more than 1 day's attributes. We first created an additional attribute which made the first line look like "day 1's attributes" plus "day 1's attributes" plus "day 2's classification result". Then we shifted it to "day 1's attributes" plus "day 2's attributes" plus "day 3's classification result". This was using previous 2 days' attributes to predict the third day's water quality. In our research, we used up to previous 8 day's attributes to predict the ninth day's water quality to see if the results have any improvements. This does not mean the more the better. In fact, adding more day's attributes is the same with adding water attributes. It can increase the data interference, which will increase the difficulty to generate the decision tree and reduce the accuracy of prediction.

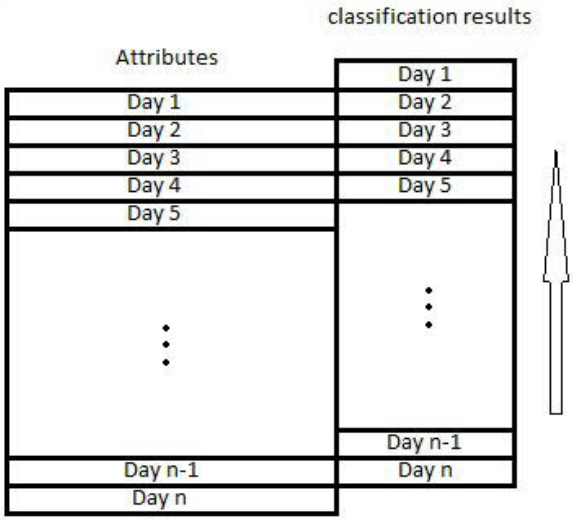


Figure 3, the idea of how to "shift" the classification result.

After preparing the training data, we used See5.0 to generate a decision tree. See5.0 is a software tool which implements the C5.0 algorithm. It can deal with a great amount of data in a short time and generate the decision trees by the classic method, the winnow method, and the boost method. It generates a decision tree and gives the importance of every attribute. Most times all attributes are used. If one attribute has not been used by the C5.0 algorithm, very often this attribute may have a linear relationship with some other attributes [9].

4. Result and Analysis

The results analysis of this research is based on the prediction accuracy. Among the 2483 records, the even records are used as training data to generate decision tree, the odd records are used as testing data. We tested the decision tree by testing data to get the prediction error rate.

The accuracy for assessment was quite good. Using the winnow method, the prediction accuracy was 93.5% by using seven attributes and 93.6% by using five attributes. Dissolved oxygen, turbidity and PH were mainly used attributes for assessment. The usage of temperature was relatively low, which shows the way we normalized the temperature reasonable. Using boost method, the prediction accuracy was 96.1% by using 7 attributes and 95% by using five attributes. The results of boost method show increasing attributes can increase the prediction accuracy.

The prediction has 32 results into 2 groups. One group used five attributes to predict, another group used seven attributes to predict. In this way, we can know if increasing attributes can improve the prediction accuracy. Both the winnow and the boost methods are used for test. One group have 8 tests differed by using different number of day's attributes.

Table 1 is the result for group 1. It shows the error rate for prediction the next day's water quality by using five attributes to classify the data into five classifications. Table 2 is the result for group 2. It shows the error rate for prediction the next day's water quality by using seven attributes to classify the data into five classifications. All the results above are the test error rate assigned by See5.0. Based on these results, we can change the data from error rate to prediction accuracy as an alternative.

Table 1, the error rate of predict the next day's water quality by using 5 attributes and 5 classifications. W stand for winnow method and B stand for Boost method.

Predict 2nd Day		Predict 3rd Day		Predict 4th Day		Predict 5th Day	
W	B	W	B	W	B	W	B
19.1%	16.5%	19.4%	17.1%	19.6%	16.9%	20.7%	17.0%
Predict 6th Day		Predict 7th Day		Predict 8th Day		Predict 9th Day	
W	B	W	B	W	B	W	B
22.8%	16.1%	21.2%	17.0%	21.4%	17.0%	22.6%	16.9%

Table 2, the error rate of predict the next day's water quality by using 7 attributes and 5 classifications. W stand for winnow method and B stand for Boost method.

Predict 2nd Day		Predict 3rd Day		Predict 4th Day		Predict 5th Day	
W	B	W	B	W	B	W	B
19.6%	18.1%	21.2%	16.8%	20.2%	16.9%	20.3%	16.9%
Predict 6th Day		Predict 7th Day		Predict 8th Day		Predict 9th Day	
W	B	W	B	W	B	W	B
21.5%	16.6%	21.1%	16.0%	22.1%	17.7%	20.7%	16.3%

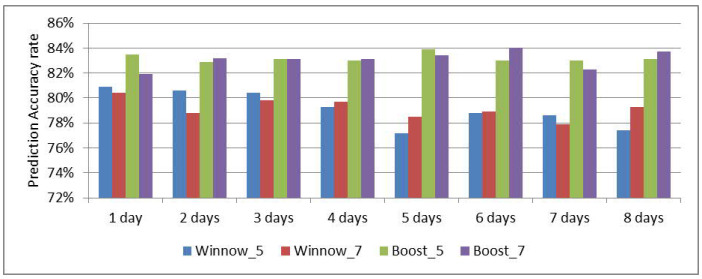


Figure 4, Prediction accuracy of two different methods and using different number of attributes on predict the next day's water quality to 5 classifications.

Figure. 4 shows the prediction accuracy for each method using different attributes. It proves the great performance of boost method. We can see using seven attributes has better performance. The winnow method seems to have no great difference. The boost method, however, shows the advantage of using seven attributes. Out of eight tests, five tests show using seven attributes can reach a better performance than using five attributes. Also using the boost method reaches the highest prediction accuracy, which is 84% by using previous six day's attributes to predict the next day's water quality.

This figure also shows some interesting results. As we mentioned before, increasing more day's attributes may not improve the prediction accuracy. Actually, these eight tests are all used to predict the next day's water quality. With the increase number of attributes, the prediction accuracy of the winnow method, generally speaking, is reducing. This is because data interference. No matter we increase the number of single day's attributes or we increase the number of days we use for prediction; they have same effect for the winnow method. When the attributes' number is greatly increased, the difficulty to generate the decision tree also increasing tremendously. The boost method, however, doesn't take too much effect from data interference. As each decision tree make a slightly improvement based on the previous decision tree's error, the "vote" mechanism can greatly reduce the data interference and has a better performance.

5. High Performance Computing Speed up the Processing

In this project, we were running machine learning program on Purdue University's High Performance Cluster – *Falcon*, which has 8 nodes, each with 4 CPUs, and/or *Miner*, which has 512 nodes, each with 4 CPUs. This high performance computing (HPC) system is supported by the Northwest Indiana Computational Grid (NWICG) program funded by the U.S. Department of Energy grant ($4.9 million). Taking advantage of the NWICG's HPC system, it is practical for us to use complexity to exchange the water quality prediction accuracy. HPC dramatically decreases the running time of the enumeration based algorithm, which is time consuming but more accurate.

Facilitated perfectly by the current existing Northwest Indiana Computational Grid, the project has applied the HPC system which dramatically decreases the running time of the enumeration based algorithm, so that makes this very time-consuming algorithm for water quality prediction becoming practical.

Up to now, the experiments of this innovative approach obtained very promising results. The sixteen-processor HPC system achieved more accurate water quality prediction and reduced the execution time by more than ten times. The experiment results showed applying machine learning models on HPC system is a very encourage direction for Northwest Indiana's water quality assessment and prediction.

In light of almost 3000 cases, the optimal enumerative algorithm is very time-consuming (running in days) when executing on a single machine. In order to improve the performance, we applied parallel programming on the enumerative algorithm. The program run on multi-processors in parallelism. The parallel computing assigned one process as the master process and others as slave process. The master process read in the data, broadcasted the data and other information to the slave processes. The parallel programming split the clustering calculation into threads for slaves. Each slave process worked on calculations for clustering. Finally, the master processor received the results from each slave process and chose the global optimized clustering [11, 12, 13].

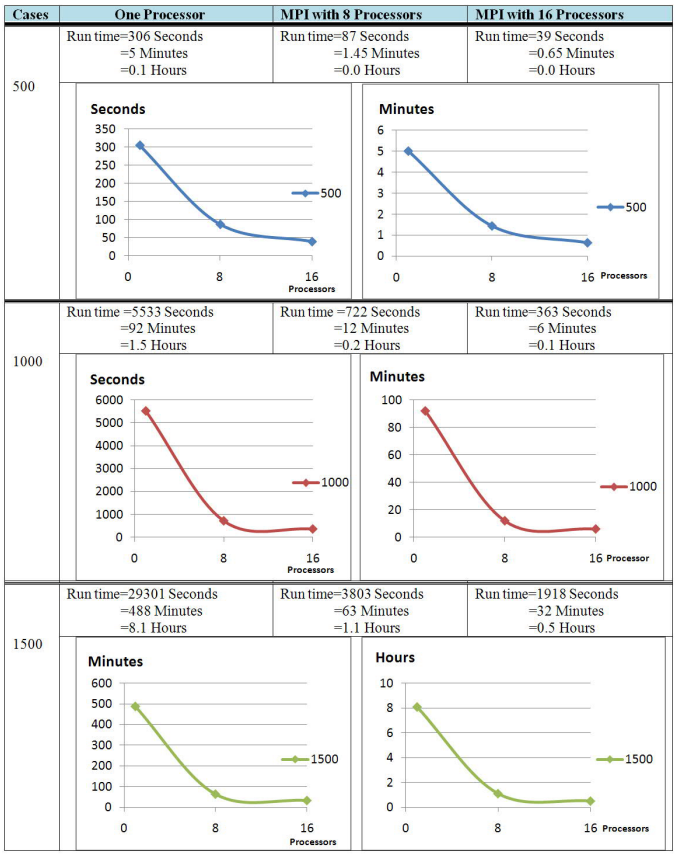


Figure 5, Experiment results with MPI on the HPC

Our experiments compared the running time of the enumerative algorithm using a single processor with that using HPC with 8 processors and 16 processors. The number of the cases was 500, 1000, 1500, 2000, 2500, and 2700 respectively. Figure 5 and Figure 6 below show the execution time of our preliminary experiments. It exposed the critical role of HPC for the time consuming enumerative algorithm to process large amount of data. For example, for 1500 cases, running on a single processor, it took more than eight hours to finish the clustering; whereas running on HPC with 16 processors, it only took about 32 minutes. This improvement of execution time is critical, because it is important to have a reasonable amount of indicative and representative data samples (cases) in order to discover patterns in data samples using machine learning technologies. It is not unusual to have a data set of 1500 cases by this approach.

The results of these experiments produced a solid fundamental for future developments in our water quality assessment and prediction project. We created decision tree employed C5.0 for the water quality assessment and prediction. Using the produced decision trees to predict unseen cases, the prediction accuracies reach 82 percent, two percent better than applying k-means algorithm and about the same improvement than applying the

mechanistic simulation models. It is anticipated that if we use more attributes of the water quality in the future, the accuracy rate could be further improved.

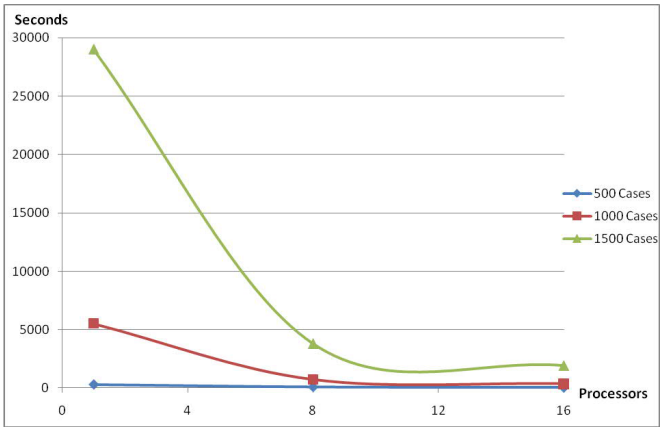


Figure 6, Execution time in seconds for comparison of various processors in HPC

6. Summary and Conclusion

In this research, by applying machine learning algorithm to a long period time of water attributes' data; we can generate a decision tree so that it can predict the future day's water quality in an easy and efficient way. The idea is to combine the traditional ways and the computer algorithms together. Using machine learning algorithms, the assessment of water quality will be far more efficient, and by generating the decision tree, the prediction will be quite accurate. The drawback of the machine learning modeling is that the execution takes quite long time, especially when we employ a better accuracy but more time-consuming algorithm in clustering. Therefore, we applied the HPC System to deal with this problem. Up to now, the pilot experiments have achieved very promising preliminary results. The visualized water quality assessment and prediction obtained from this project would be published in an interactive website so that the public and the environmental managers could use the information for their decision making.

References

1. Ruijian Zhang, "Applying Parallel Programming and High Performance Computing to Speed up Data Mining Processing", Proceedings of the 16th IEEE international Conference on Computer and Information Science, May 2017.
2. Faruk, D. "A hybrid neural network and ARIMA model for water quality time series prediction" Eng. Appl. Aritf. Intel., 23, 586-594, 2009. https://doi.org/10.1016/j.engappai.2009.09.015
3. L. M. He, Z. L. He, "Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California", Water. Res., 42, 2563-2573, 2008. https://doi.org/10.1016/j.watres.2008.01.002
4. W. Labiosa, J. Leckie, R. Shachter, D. Freyberg, J. Rytuba, "Incorporating Uncertainty in Watershed Management Decision-Making: A Mercury TMDL Case Study" in Watershed Management Conference, Williamsburg, Virginia, United States, 2005. https://doi.org/10.1061/40763(178)125
5. A. Said, "The Implementation of a Bayesian Network for Watershed Management Decisions" Int. Ser. Prog. Wat. Res., 20, 591–605. 2005. https://doi.org/10.1007/s11269-006-3088-z
6. S. Shrestha, F. Kazama, "Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan" Environ. Modell. Softw., 22, 464-475, 2006. https://doi.org/10.1016/j.envsoft.2006.02.001
7. C. A. Stow, M. E. Borsuk, K. H. Reckhow, "Nitrogen TMDL development in the Neuse River watershed: An imperative for adaptive management", Water Resources Update, 122, 16–26, 2002.
8. See 5.0 [Computer software]. Australia: Rulequest Research, 2012.
9. USGS Water-Quality Data for the Nation. USGS Water-Quality Data for the Nation from U.S. Geological Survey, U.S Department of the Interior, 2012. http://waterdata.usgs.gov/nwis/qw
10. S. Tong, W. Chen, "Modeling the relationship between land use and surface water quality". Environ. Manage., 66, 377-393, 2002. https://doi.org/10.1006/jema.2002.0593
11. C. Severance, K. Dowd, High performance computing, 2nd Edition, 1998.
12. G. Stellner, "Cocheck: Checkpointing and Process Migration for MPI", Proceedings of International Conference on Parallel Processing, Hawaii, USA, 1996. https://doi.org/10.1109/IPPS.1996.508106
13. P.L. Vaughan, A. Skjellum, D.S. Reese, F.C. Cheng, "Migrating from PVM to MPI, Part I: The Unify System", Frontiers of Massively Parallel Computation, McLean, Virginia, United States, 1995. https://doi.org/10.1109/FMPC.1995.380476

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 290-301 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Flexible lengthening-shortening arm mechanism for fishery resource management

Yoshiki Iwamochi[1], Motoki Takagi[2], Tasuku Miyoshi[1,3*]

[1]*Department of Mechanical Engineering, Iwate University, Morioka-city, Iwate 020-8551, Japan*

[2]*Department of Information and Electronic Engineering, Teikyo University, Utsunomiya-City, Tochigi 320-8551, Japan*

[3]*Soft-Path Science and Engineering Research Center (SPERC), Faculty of Science and Engineering, Iwate University, Morioka-city, Iwate 020-8551, Japan*

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 16 December, 2017
Online: 23 December, 2017

Keywords :
Underwater robot arm
Flexible arm
Sliding screw mechanism
Obstacle avoidance
Path-tracing algorithm

ABSTRACT

The goal of this study was to use underwater robots instead of a diver's observations to monitor and record the condition of an obstructed seabed in a shallow area. It is difficult to investigate marine resources that exist in deep water shaded by rock due to large and/or small rocks on the seabed. To solve these problems, we newly constructed a flexible lengthening-shortening arm with a small camera unit for an underwater robot to assist in the management of fishery resources. In this paper, we describe the concept and configuration of the newly developed arm mechanism using a sliding screw mechanism to overcome obstacles by changing arm posture in a two-dimensional plane, and we demonstrate the experimental results of a path-tracing controller for the rear links. The results were that the maximum deviations between the target path and the tracing path were less than 4.0% of the total width of the arm mechanism. These results suggest that the newly developed path-tracing algorithm is effective for our flexible lengthening-shortening arm mechanism.

1. Introduction

This report is an extension of work that was originally presented at the 2017 IEEE International Conference on Mechatronics and Automation (ICMA) [1]. Extended points are as follows; (a) the geometric configurations and control parameters for an unique slider-screw mechanism of our system in Section 2.2.3, (b) waterproof design for the slider-screw mechanism in Section 2.2.4, (c) an operational test on land to clarify the validity of the flexible lengthening-shortening arm mechanism in Section 3.1, (d) an explanation of the details of the path-tracing algorithm in Section 4.2.3 and Appendix, and (e) improvements of the movement error occurred in previous work [1].

Underwater robots, such as autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs), were developed to reduce the big burden on fishermen in managing fishery resources and/or the cost of submarine operations and risk to human life (e.g.; AUV for work in shallow waters and coastal areas [2], reduce the burden on divers [3], achieving the core sampling at shallow seabed [4], and achieving the video transect method instead of divers [5]). Generally, underwater robots are equipped with cameras to monitor and record the condition of the seabed rather than relying on a diver's observations. In the case of shallow sea areas, such as surf zones, there are many rocks and marine algae. Although these areas are useful as growth environments of fishery resources, it is assumed that the view of the camera is hindered and/or the robots cannot approach the target objects closely. In addition, it is difficult to investigate rock shadows, which cannot be penetrated by underwater robots, and the shellfish living deep within them.

There is a possibility that a survey could not be achieved using current robot arms (e.g.; dual arm operation [6] and tele-operated manipulators installed to ROV [7]). Instead of using robot arms, developing snake-type robots is one solution. Snake robots can search for and investigate targets, avoiding multiple obstacles by their serpentine motion in narrow environments where humans cannot intrude. Previously, many researchers had been developing

*Corresponding Author: Tasuku Miyoshi, Department of Mechanical Engineering, Iwate University, Morioka-city, Iwate 020-8551, Japan, Email: tmiyoshi@iwate-u.ac.jp

snake-type robots for investigating unknown environments and searching for disaster sites (e.g.; review of snake-type robots [8], tutorial of snake-type robot [9], and carrying out the inspection work [10]). Also, an amphibious snake-type robot, which moves in a serpentine motion for generating the forward propulsion in water, has been studied [11]. With this unique motion, however, it is necessary for the robot to keep thrusting in order to stay in place in shallow areas where wave disturbances occur from all directions. Without any propulsion force for forward movement by a serpentine motion, a mobile robot equipped snake-type arm mechanisms had been studied (e.g.; surveying robot equipped with snake-type arm on wheel robot [12], assembling an aircraft using a snake-type arm mounted on a traditional industrial robot arm [13], and linear motion mechanism mounted on base of snake-type arm [14]). When entering the camera at the tip of the arm into the narrow space, the snake-type arm changes its attitude so as to avoid obstacles and the intrusion operation (forward approach action) is in charge of the base unit. In order to investigate a narrow area in the shallow sea using AUV or ROV equipped an arm with snake-type mechanism, it is necessary to take forward motion while maintaining posture and position against the tidal current near the seabed. Therefore, we assumed that a snake-type arm, which had a forward movement mechanism and installed to AUV and/or ROV, is effective for investigating narrow areas in shallow sea. Hence, the purposes of this study were: 1) to construct a new robot arm that utilizes the features of a snake-type robot's obstacle avoidance, and 2) to evaluate the autonomous path-tracing algorithm to achieve obstacle avoidance at the part of the rear-link mechanism. In this paper, Section 2 outlines our newly developed arm mechanism and the results of the operation test. In Section 3, we propose a path-tracing algorithm of the rear-link mechanism required for obstacle avoidance and show the effectiveness of this algorithm experimentally.

2. Flexible lengthening-shortening arm mechanism (FLSAM)

2.1. Design specifications

We selected the following design specifications for the newly developed flexible lengthening-shortening arm mechanism (FLSAM):

1. The FLSAM can avoid obstacles,

2. The FLSAM can reach into rock shadows and penetralia,

3. The FLSAM can approach target objects closely, and

4. The FLSAM can obtain photos and movies for further analysis.

To satisfy the required specifications of the rotational and translational degrees of freedom, the FLSAM's design considers the following points:

A1. To avoid obstacles, the FLSAM should be equipped with rotational degrees of freedom,

A2. To reach into rock shadows and go near the target, the FLSAM should be able to expand its arm length by translational degrees of freedom, and

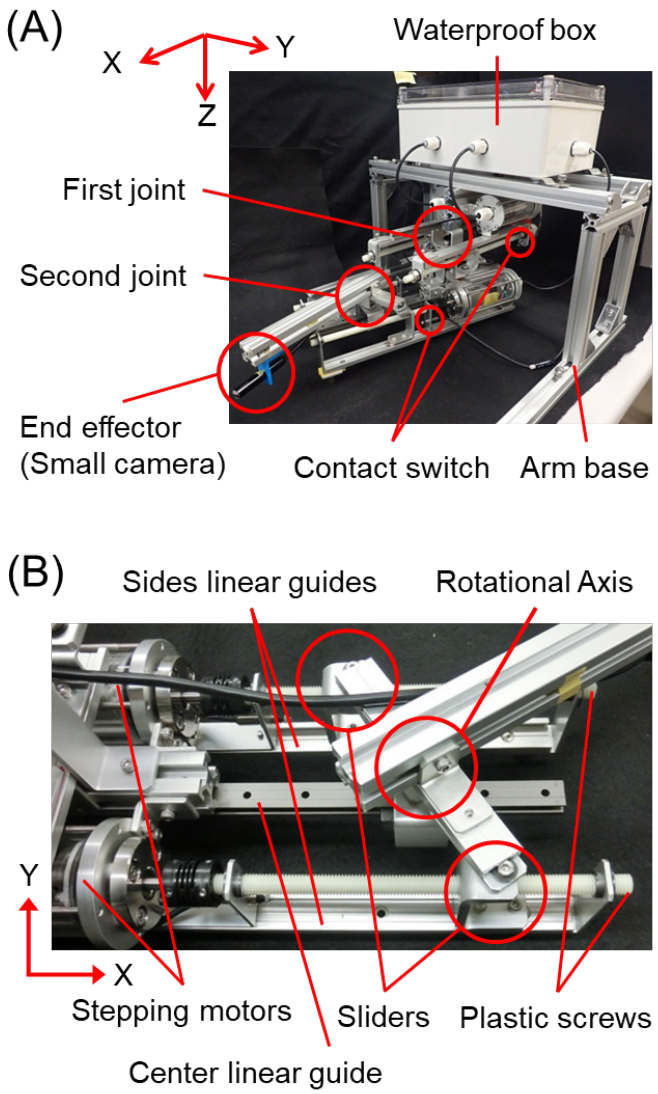


Figure 1: (A) Overview of the newly developed flexible lengthening-shortening arm mechanism, "FLSAM"; (B) mechanical details around the second joint of the FLSAM

A3. To obtain photos and movies, a small video camera should be installed on the FLSAM.

2.2. Configuration of the FLSAM

2.2.1. Overview of the FLSAM

Figure 1A shows an overview of the FLSAM, and the specifications are denoted in Table 1. Each joint consists of two stepping motors (PKP213D05A, ORIENTAL MOTOR Co., Ltd.), two plastic screws (MB-RENY-M8-300, MISUMI Corporation), and three linear guides (for both sides: NK-02-17-1-220-LLYZ, igus k.k.; for the center: TK-04-12-1-250, MISUMI Corporation). Each linear guide was in parallel, and sliders were installed on both sides of the linear guides. To achieve rotational degrees of freedom, the rotational axis was set at the center linear guide. As to the detailed structure of the joint, the appearance of the second joint is shown in Figure 1B. Contact switches (PTP5M3CB-L, MISUMI Corporation) were installed to detect the origin of the

Table 1: Specifications of the FLSAM

Total height [mm]	190
Shortened length [mm]	290
Maximum lengthened condition [mm]	560
Total width [mm]	140
First joint stroke [mm]	90
Second joint stroke [mm]	180
Range of motion [deg]	Plus/minus 45
Total weight [kg]	3.9
Sensors	Small camera Contact switch

sliders and to prevent an overrun. A small waterproof camera (WPMISC7M, THANKO Co., Ltd.) was equipped to record photos and movies from the end effector. Electric circuits with an AVR microcomputer (Arduino Uno R3, Arduino AG) were placed in a waterproof box (SPCM13-18-10-T, MISUMI Corporation).

2.2.2. *Sliding screw mechanism to achieve rotational and translational degrees of freedom of the FLSAM*

When constructing an arm with the simultaneous motion of the translational and rotational degrees of freedom, it is natural to couple the translational and rotational mechanisms. However, the arm mechanism increases in size due to the increased number of joints when the translational and rotational joints are arranged separately. Therefore, we adopted a sliding screw mechanism to achieve the translational and rotational degrees of freedom in one joint. The sliding screw mechanism could be expected to reduce the size and number of joints. Moreover, it could be said of this mechanism that (1) power consumption is low because it is not necessary to maintain the posture, and (2) it can operate in water without a lubricant [15].

Figure 2A–C shows the schematic of the sliding screw mechanism installed in the FLSAM. We arranged two sliding screws for one joint, and the slider part moves forward and backward with respect to the rotational volume of the first and second actuators. The FLSAM installed two sliding screw mechanisms; in total, four degrees of freedom were achieved. The changes in the stroke length of each sliding screw mechanism were defined by the number P_k [pulse] and frequency A_k [pulse/s] of pulses' input to the stepping motor, as

$$\begin{bmatrix} P_k \\ A_k \end{bmatrix} = \frac{R}{L_e} \begin{bmatrix} X_k \\ V_k \end{bmatrix} \quad (1),$$

where, k is the number of the slider (1 and 2 for the first joint, 3 and 4 for the second joint), X_k [mm] is the stroke length of each slider, V_k [mm/s] is the moving velocity of each slider, L_e [mm] is the lead of the screw, and R [pulse] is the number of pulses per

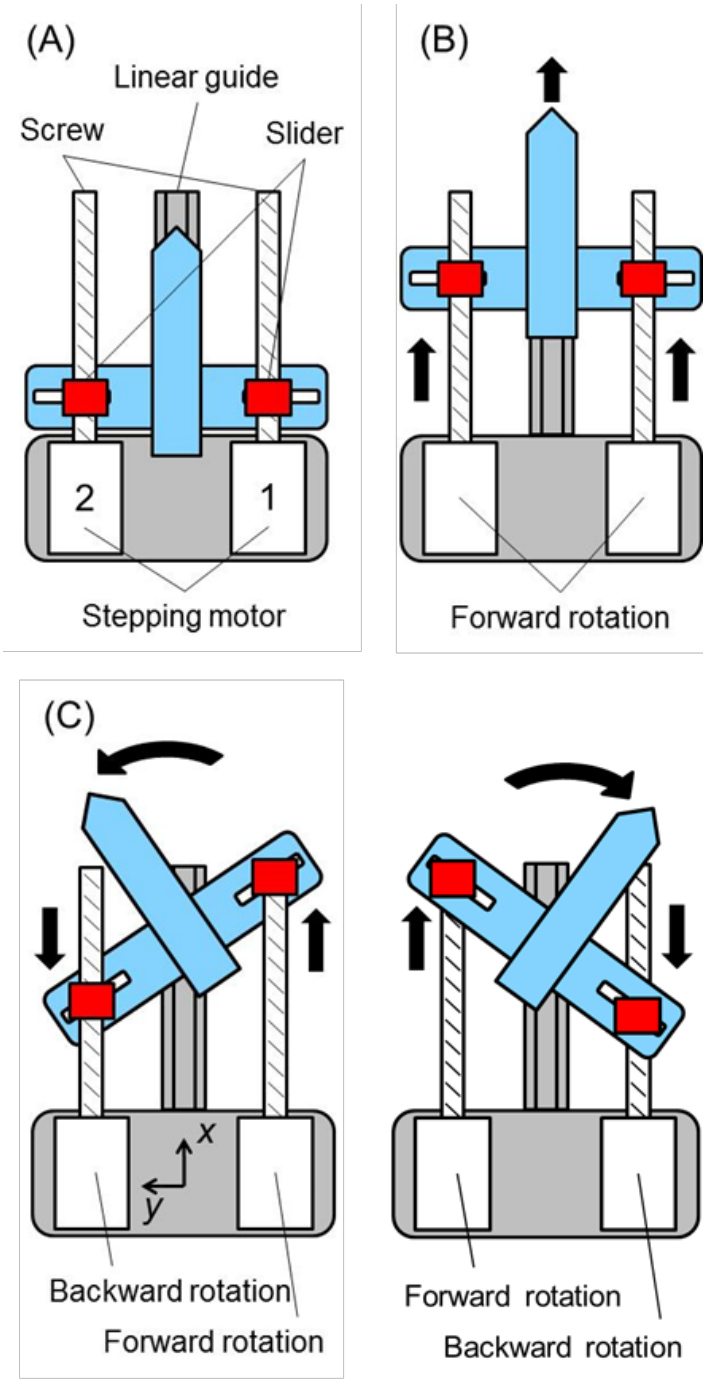


Figure 2: Schematic figure of the sliding screw mechanism installed in the FLSAM; (A) initial condition, (B) lengthened condition, (C) flexible condition

revolution of the stepping motor. In the FLSAM, the L_e is set at 1.25, and the R is 200.

2.2.3. *Geometric configuration of the sliding screw mechanism*

Figure 3 is the schematic of the geometric constraints and the definition of the control parameters by the sliding screw mechanism for the FLSAM. The $O_{i-1}X_{i-1}Y_{i-1}$ joint coordinate system was set at the base of the sliding screw mechanism. The

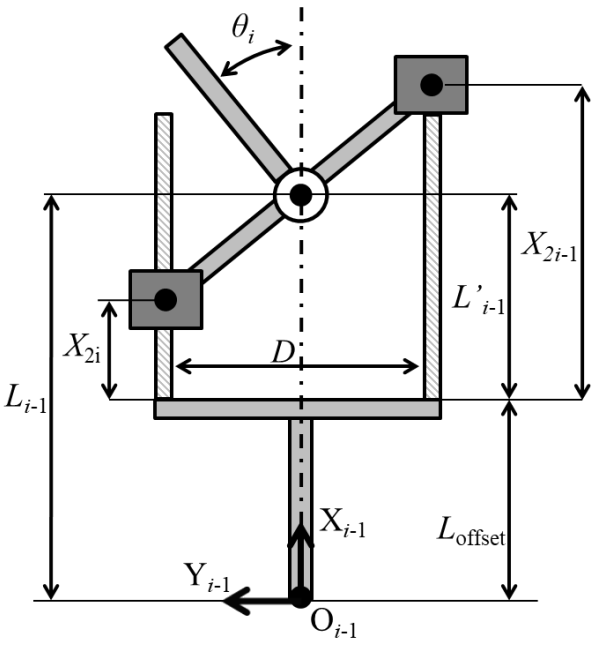


Figure 3: Schematic figure for the parameters of the sliding screw mechanism

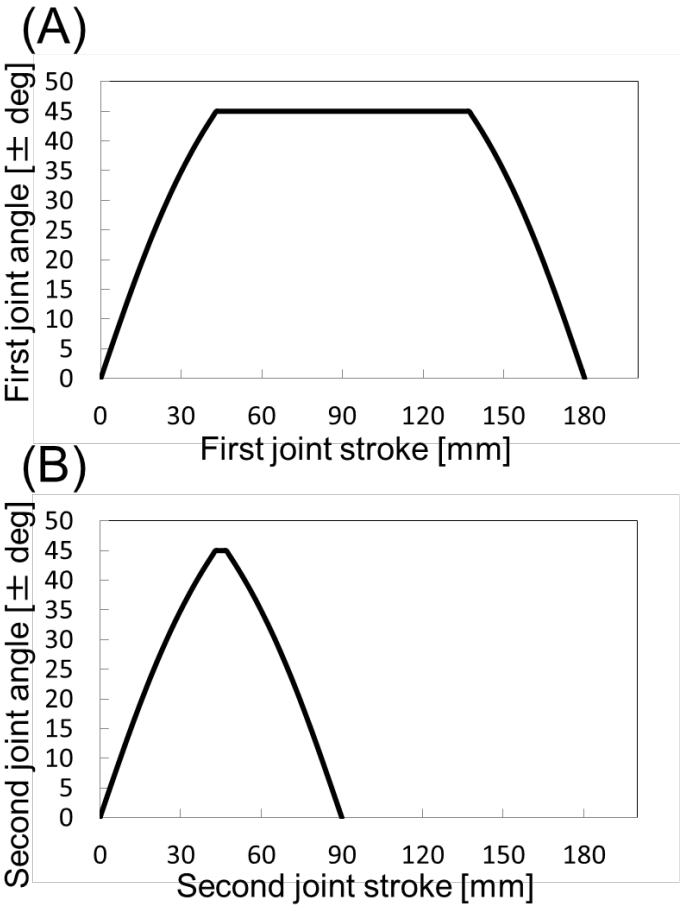


Figure 4: Geometric relationships between the joint range of motion and the joint stroke length; (A) first joint, (B) second joint

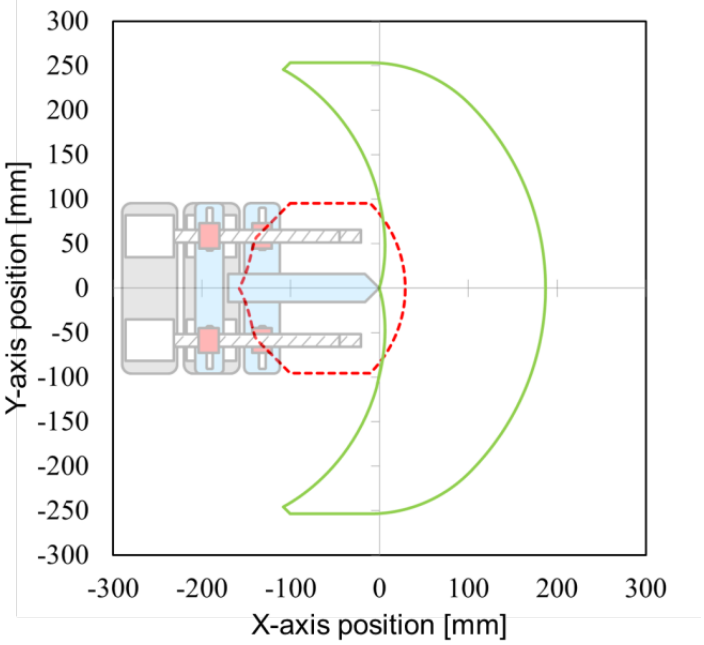


Figure 5: Schematic figure of the range of motion of the FLSAM; red broken line denotes the range of motion for maintaining the maximum flexibility, green line denotes the maximum range of motion

link length, L_{i-1}, and the joint angle, θ_i, were controlled by right- and left-moving sliders. The L'_{i-1}, L_{i-1}, and θ_i were calculated as

$$L'_{i-1} = \left(X_{2i-1} + X_{2i}\right)/2 \quad (2),$$

$$L_{i-1} = \left(X_{2i-1} + X_{2i}\right)/2 + L_{\text{offset}} \quad (3),$$

$$\theta_i = \tan^{-1}\left(\left(X_{2i} - X_{2i-1}\right)/D\right) \quad (4),$$

where L'_{i-1} (i=1, 2) is the link length while lengthening, shortening, and/or rotating. X_{2i-1} is the slider position at the side of the positive in the y-axis, X_{2i} is the slider position at the side of the negative in the y-axis, and D is the length between the two sliders. The position of the end effector of the FLSAM was operated by combinations of the L_{i-1} and θ_i. The movement speed of both sides of the slider is determined so that the operation time is the same according to the L_{i-1} and θ_i, since only the slider on one side works as a singular point.

The sliding screw mechanism has a unique characteristic, in which the maximum range of motion of a joint angle varies with the changes in the link length, as shown in Figure 4. Both maximum joint ranges of motion were set at plus/minus 45 degrees; however, the maximum joint range of motion is 0 degrees when the left and right sliders are at the minimum and/or the maximum position of the link length simultaneously. In addition, these maximum joint ranges of motion were independent parameters, so that it is possible to take the most flexible posture from 43 mm to 47 mm in the first joint and from 43 mm to 137 mm in the second joint. Figure 5 shows a range of motion of FLSAM. In this study, the range of motion was set at the inside area depicted by red line, which maintain the maximum flexibility of FLSAM (joint ranges of motion was at ± 45 deg). When

operating beyond this range, the maximum joint ranges of motion become small and the flexibility of the FLSAM decreases.

2.2.4. *Waterproof structure of the sliding screw mechanism*

The main parts of the FLSAM (sliding screws, sliders, guide rails, and a small camera) are highly water resistant. Cable entry and exit of the waterproof box equipped with an electrical system were waterproofed by cable grounding. Four stepping motors were waterproofed using custom-designed cylindrical housings, as shown in Figure 6. Basically, the housing is made watertight by careful fitting of the acrylic tube, aluminum lid, and O ring. The rotary shaft was waterproofed by combining stainless steel bushes, resin bushes, and a sealing material.

2.2.5. *System configuration of the FLSAM*

Figure 7 shows the system configuration of the FLSAM. The FLSAM's control program was constructed using commercialized software, LabVIEW 12.0.1 (National Instruments Japan Corp.). All control signals from the controller were communicated to Arduino via a serial port, and Arduino then calculated the control signals for the motor drivers (SFE-ROB-12779, SWITCH SCIENCE). Contact switch signals are the only feedback signals to determine the initial and/or the maximum lengthened condition of the FLSAM.

3. Operational tests of the FLSAM

3.1. *Operational test on land*

In order to check whether lengthening-shortening and flexion motions by a sliding screw mechanism of design requirement for image acquisition by a small camera are possible, a ground motion test was carried out. As targets for image acquisition, models of aquatic organisms (turtle, shark, dolphin, and seahorse) and an obstacle were placed randomly. The seahorse model was placed in the shadow of an obstacle that would be a blind spot from the initial posture of the FLSAM. One of the objects of this investigation is abalone, and the catch size is defined as more than 90 mm. For this reason, an imaging target model in this operational test was also the same size of abalone. Each mode size is shown in Table 2. Hence, the FLSAM avoided obstacles due to the lengthening-shortening and flexion motion of the arm and eventually was able to discover the seahorse model. The amount of the stroke length and velocity of each slider of the FLSAM was given as a numerical input from a PC, and the operator checked the posture of the FLSAM and the camera image visually. Figure 8 shows a typical image acquisition and a birds-eye view of the FLSAM operation. In the process of finally photographing seahorses, images were acquired, in order, of the turtle, shark, dolphin, and then the seahorse in this arrangement. When the base unit is stable, FLSAM is not affected by posture change due to disturbance in water because of the characteristics of the sliding screw mechanism. Therefore, when the FLSAM achieves the ideal motion on land, the FLSAM can perform the same operation even in water as on land. It was confirmed that the FLSAM can achieve lengthening-shortening and flexion motions as demanded of the design. Also, image acquisition with a small camera was achieved.

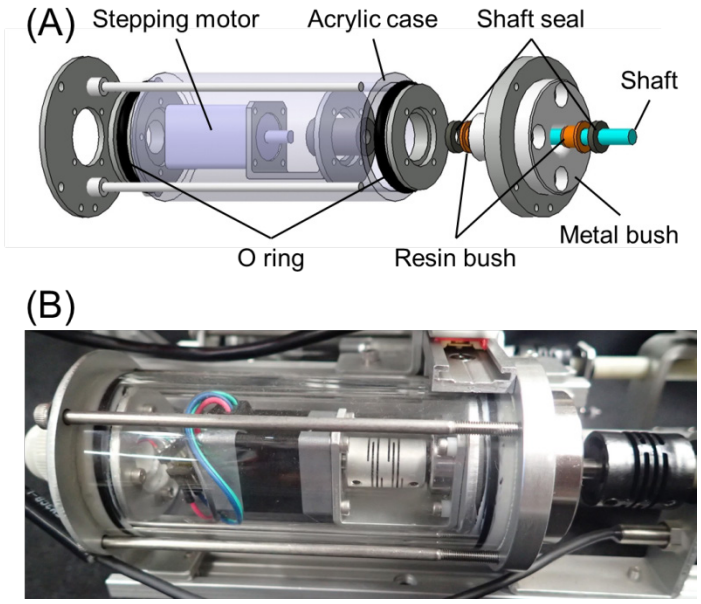


Figure 6: Waterproof motor housing; (A) exploded view of the designed housing, (B) photo of the manufactured housing

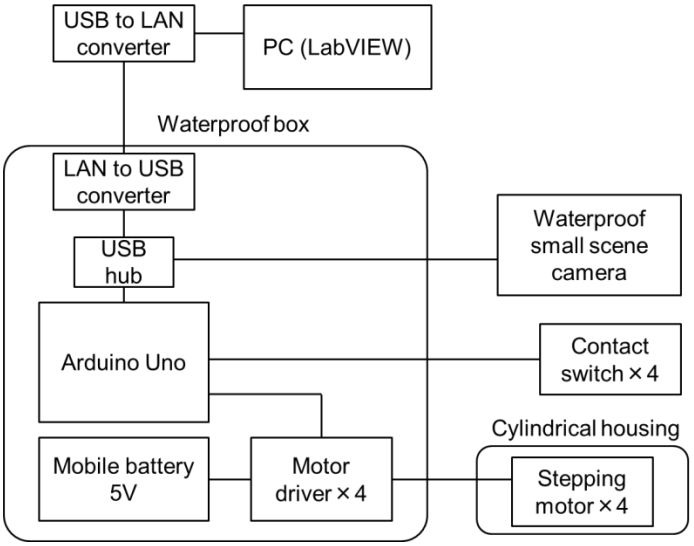


Figure 7: System configuration of the FLSAM

Table 2: Each mode size

Models of aquatic organisms	High [mm]	Length [mm]	Width [mm]
Turtle	40	90	70
Shark	60	170	90
Dolphin	50	110	40
Seahorse	60	40	30

3.2. Operational tests in water

Although it was confirmed that FLSAM could operate without any problems during the ground test, FLSAM was developed to be used in water. Therefore, it was necessary to conduct verification tests to confirm whether it can function properly underwater. Hence, this experiment was planned (1) to check the underwater

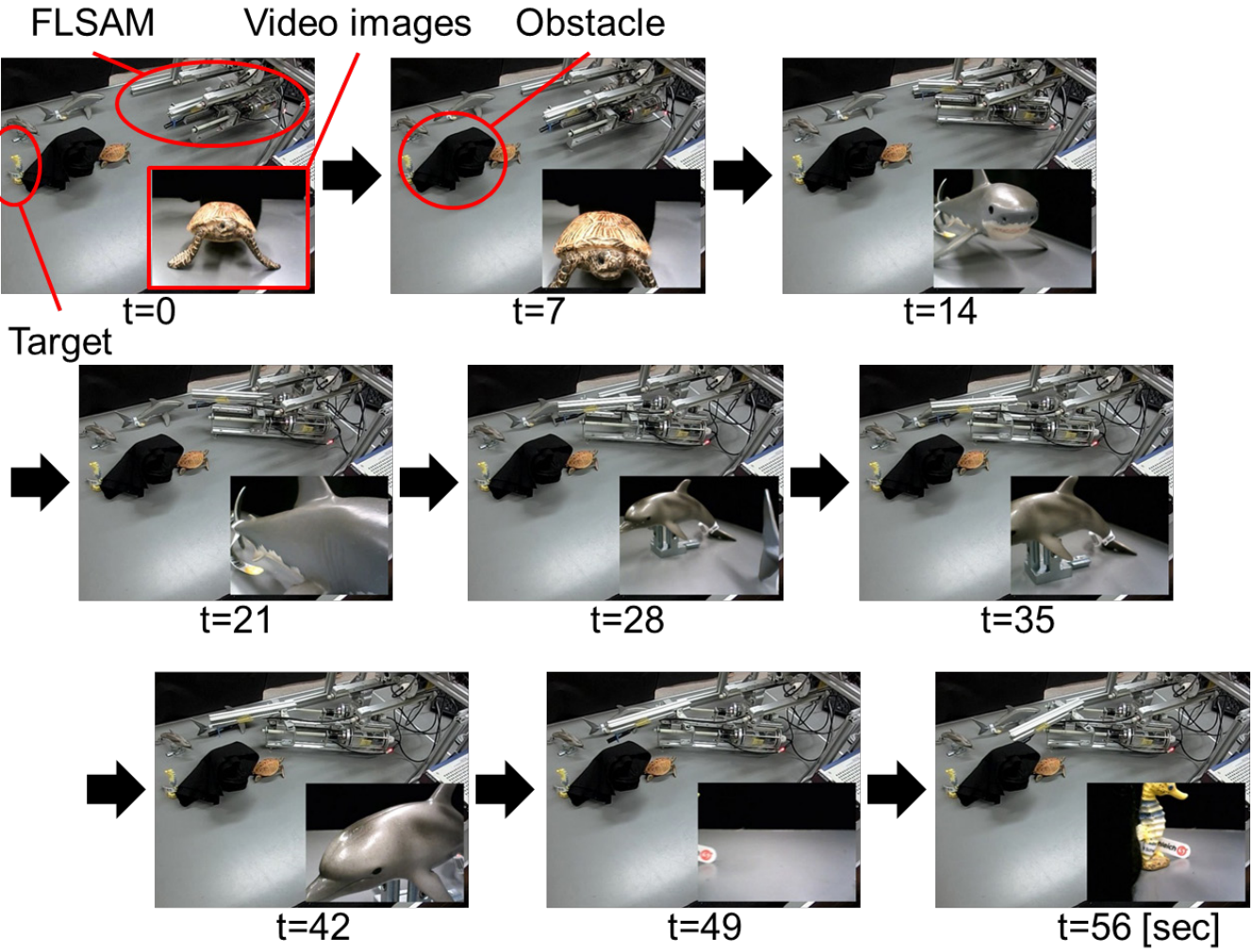


Figure 8: Time series of the state of the operational test conducted on land

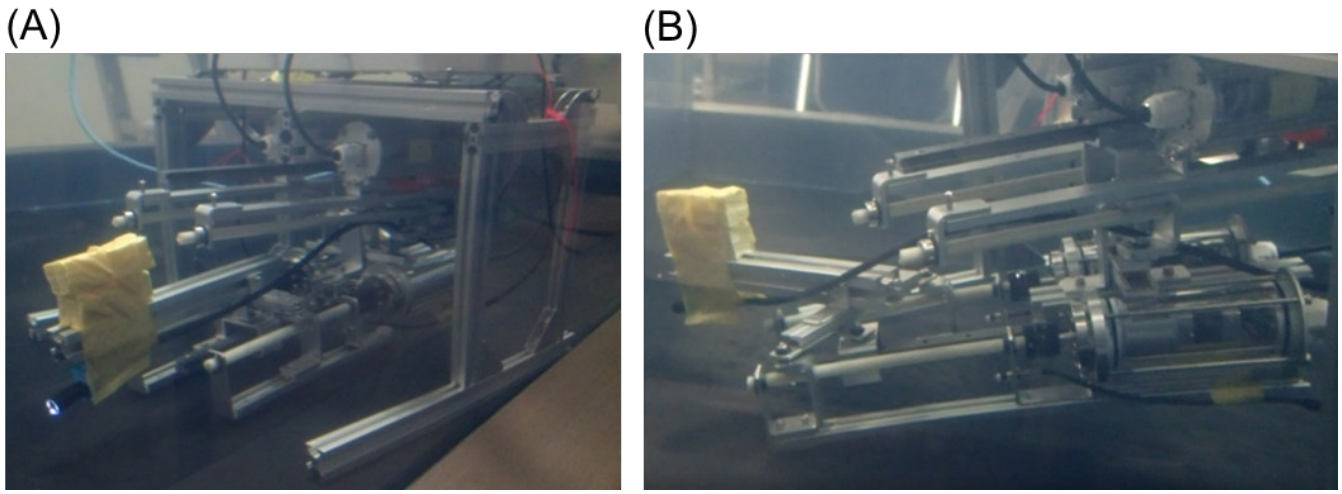


Figure 9: Typical states of the operational test in water; (A) the initial condition, (B) left-side flexed condition

motion, (2) to obtain camera images, and (3) to check the waterproof performance of the FLSAM.

The experiment was carried out in static water conditions (the water tank was 1.57 m deep × 1.2 m wide × 1.27 m high). The FLSAM was placed on the floor of the water tank in its initial condition, and then any arbitrary lengthening/shortening and/or flexion motion was operated manually. To make sure that the sliding screw mechanism of the FLSAM functions properly in water, lengthening-shortening and flexion motions were tested multiple times with reciprocated motions. Further, in order to ascertain whether there is a possibility of water leak into the motor housing due to vibration of the motor and/or the rotation of the shaft, the operation speed is changed while the underwater test continues. Also, video images were checked on the PC simultaneously to determine whether it is possible to acquire images via the small underwater camera installed. The operation of FLSAM was confirmed from the outside of the water tank.

From outside the water tank, it was confirmed that, for 30 minutes, the FLSAM performed every operation on the floor of the water tank. In addition, no water leaked into the housing. Figure 9 shows the typical condition during the operation tests of the FLSAM in water: (A) is the initial condition, and (B) is in the left-side flexed condition. In water, the FLSAM achieved the desired behavior commanded by the operator. However, as the arm lengthened, the FLSAM unexpectedly became tilted, and the small camera at the tip of the arm contacted the floor. This was due to the movement of the center of gravity that occurred when the arm was lengthened. Therefore, it is suggested that attitude control by the support base is required to acquire a stable image using the FLSAM.

4. Autonomous path-tracing experiment for rear-link mechanisms

4.1. Purpose of the path-tracing experiment

Path planning and tracing are key issues for investigating the shadows of rocks. In the case of the FLSAM, path planning and its tracing algorithm should also be installed to achieve obstacle avoidance and to reach into the depths of rock shadows to enable surveying. Methods of controlling a flexible robot arm with a retractable mechanism have been studied (e.g.; series studies of path-tracking control of a moray-type robot arm by same group [16–19], expands and/or contracts arm mechanism [20], and path tracking control of underwater snake-type robot [21]); however, it is impossible to plan a path in environments where the camera cannot obtain data due to blind sports and/or deep rock shadows.

Previous studies have developed methods of manipulating the tip of the arm using a camera image to intrude into an unknown environment (e.g.; surveying robot to avoid obstacles based on screw principle [22], double headed snake-type robot [23], and remotely operated snake-type robot [24]). The FLSAM that can enter unknown environments, such as rock shadows in water, is considered to be effective. Therefore, we adapted to control the FLSAM manually using only the video data obtained from the small installed camera. Figure 10 shows the schematic figure entering the rock shadows by changing the posture of the FLSAM. The operator decides the traveling direction of the tip of the arm using the obtained video data and then manipulates the traveling direction forward/backward and left/right so that only

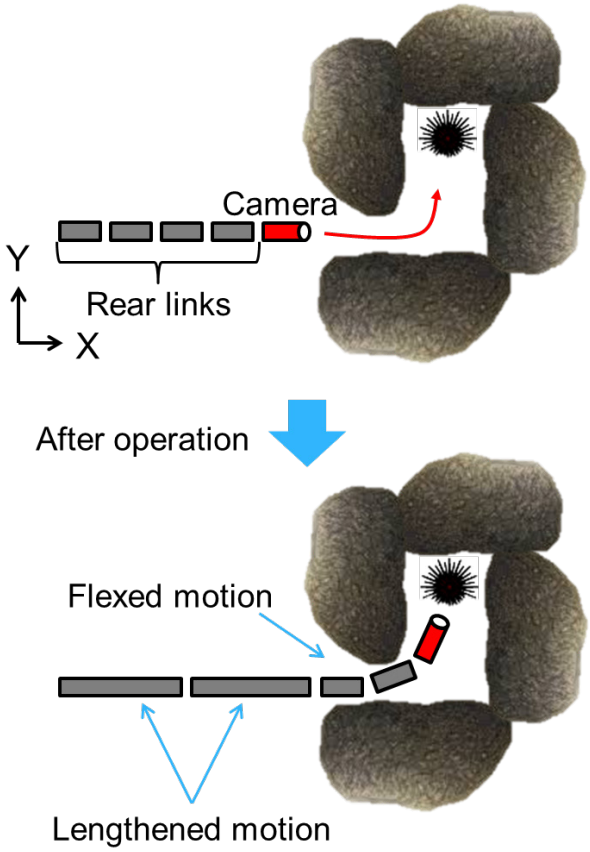


Figure 10: Schematic figure of entering the rock shadow by changing the posture of the FLSAM

the tip of the arm should be controlled manually. This means that the rear links would autonomously follow the movement path of the arm tip. Hence, the purpose of this experiment was to evaluate the accuracy of path tracing of the rear links using the FLSAM's control path-tracing algorithm.

4.2. Path tracing in the rear-link parts of the FLSAM

4.2.1. Definition of the rear-link in FLSAM

FLSAM has two sliding screw mechanism joints. That means that if the second link of the FLSAM equipped with the small camera is defined as the arm tip, the rear link is only the first link. However, it is difficult to evaluate the validity of the path-tracing algorithm by measuring the movement of one link. Hence, as shown in Figure 11AB, the virtual arm tip is defined as existing at the tip of the second link of the FLSAM, and the first link and the second link of the FLSAM are set as the rear links.

4.2.2. Target path defined by the Spline interpolation curve

The first and second rear links of the FLSAM autonomously trace the tip's moving path. In this research, we prepared the target path under the assumption that the tip's moving path is generated

by the movement of the virtual tip, and the first and second links of FLSAM trace this virtual path. For the target path, a cubic Spline interpolation curve capable of creating a two-dimensional curve in space was used. The details are described in Appendix-1; however, the first derivatives of the start point of the Spline curve are set to 0 in the boundary condition (A5). With respect to this definition, the slope of the Spline curve at the start point becomes 0. Therefore, the Spline curves of the target path and the X axis are continuous at the origin of the global coordinate system, and discontinuity of the arm joint angle can be avoided.

4.2.3. *Development of a path-tracing algorithm in the rear-links of the FLSAM*

Figure 11C shows a schematic of path tracing in a rear link of the FLSAM. In this study, the rear links of the FLSAM trace the Spline curve created as the target path to evaluate the validity of the path-tracing algorithm. In order to trace the target path, it is necessary to calculate each joint angle as output with the target path and each link length as input. Thus, as shown in Figure 11C, links L_0 and L_1 should lengthen and/or shorten to control the joint angles for tracing the target path.

At first, we defined each joint coordinate and link length as follows: the origin of the arm base as $J_0=[(x_0, y_0) = (0, 0)]$, the position of the first joint as $J_1=[(x_1, y_1) = (x_0 + L_0, 0)]$, the position of the later joints as $J_{j+1} = [x_{j+1}, y_{j+1}]$ (j = 1 to n); n is the number of total joints, the position of the arm tip is Jc (= J_{n+1}) = $[x_{n+1}, y_{n+1}]$, and the link lengths are L_j. J_0 and J_1 are determined by the initial arrangement and the link length of L_0. The outline of the procedure for deriving each joint position from each link length is as follows:

(1) let J_{j+1} be the point of intersection with the circle C_j of radius L_j centered on J_j and the Spline curve,

(2) let J_{j+2} be the point of intersection with the circle C_{j+1} of radius L_{j+1} centered on J_{j+1} derived in procedure (1) and the Spline curve,

(3) then repeat procedure (2) until Jc.

Here, we will describe procedures (1) and (2) in greater details. Where the C_j intersects with the interpolation interval, $S_i(x)$ is determined by the following conditional expression, since the Spline curve has a number of ith interpolation intervals:

$$(x_i - x_j)^2 + (y_i - y_j)^2 < L_j^2 < (x_{i+1} - x_j)^2 + (y_{i+1} - y_j)^2 \quad (5).$$

Then, solve the following simultaneous equations to find x_{j+1}:

$$C_j : (x_{j+1} - x_j)^2 + (y_{j+1} - y_j)^2 = L_j^2 \quad (6),$$

$$S_i(x_{j+1}) = a_i x_{j+1}^3 + b_i x_{j+1}^2 + c_i x_{j+1} + d_i (= y_{j+1}) \quad (7),$$

where x_{j+1} satisfies $x_i < x_{j+1} < x_{i+1}$. By repeating these procedures, each joint position is derived. Finally, each joint angle is derived by the following formula:

$$\theta_j = \tan^{-1} \frac{y_{j+1} - y_j}{x_{j+1} - x_j} - \theta_{j-1} \quad (8).$$

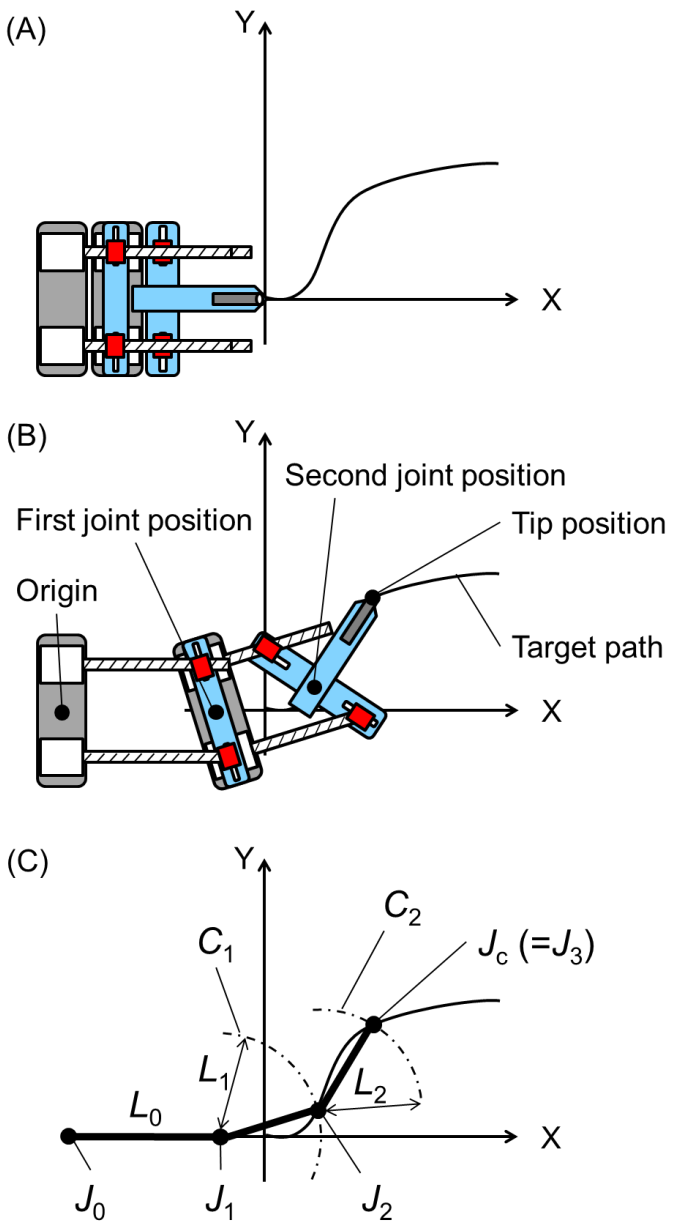


Figure 11: Schematic figure of the definition of the rear links in the FLSAM; (A) state of the arm tip on the ordinate of the global coordinate system, (B) definition of the rear links in FLSAM, (C) definition of each joint coordinate and link length

In this study, using each link length, L_j, and joint angle, θ_j, as input values, the path-tracing algorithm for the rear links calculates the output value as the number of pulses input to each stepping motor of the sliding screw mechanism. Thus, in the case of the FLSAM, $J_0=[(x_0, y_0) = (0, 0)]$ is the base position, the position of the first joint is $J_1=[(x_1, y_1) = (x_0 + L_0, 0)]$, the position of the second joint is $J_2 = [x_2, y_2]$, and the position of the arm tip is Jc (= J_3) = $[x_3, y_3]$, since the number of joints is two.

4.2.4. *Simulation of a path-tracing algorithm in the rear links of the FLSAM*

Computational simulation was performed to evaluate whether the proposed rear-link path-tracing algorithm works precisely enough to trace the target path (an arbitrary Spline curve). This simulation was done with a custom-designed graphical user

interface using LabVIEW 12.0.1. Figure 12 shows the simulation results of the trajectories of the tip and the second joint calculated by using each link length and joint angle output from the path-tracing algorithm. Thus, we confirmed that each joint is moving on the Spline curve completely, and that the path-tracing algorithm of the rear link of the FLSAM is correct.

4.3. Methods

The accuracy of the FLSAM motion was measured using a real-time three-dimensional motion capture system (VENUS3D, Nobby Tech., Ltd.). The conceptual motion pattern for path tracing in the FLSAM is shown in Figure 12, which shows that the initial posture was determined, in which the origin of the global coordinate system of the FLSAM was harmonized with the position of the arm tip. The tip of the end effector traced a predetermined target path, and then the rear links followed the tip motion, as shown in Figure 11B. For each slider of the FLSAM, the maximum moving speed was set to 5 mm/s. In addition, the L_{i-1} (i=1, 2) was lengthened by a step of 2 mm. Three target paths (a figure "S" pattern, an exponential figure pattern, and a sigmoid curve figure pattern) and three mirrored target paths were defined before the experiment and were started from the origin of the global coordinate system. Each pattern was followed in ten trials. The motion trajectory of the arm tip and the second joint were measured to compare the accuracy of the target paths and the motion of the FLSAM. All experiments were conducted on land.

4.4. Results and Discussion

Figure 13 shows the results of all ten trials of tracing (A) the figure "S" and its mirrored patterns, (B) the exponential figure and its mirrored patterns, and (C) the sigmoid curve figure and its mirrored patterns. Time series of the position of the tip in X and Y axis are shown in Fig. 14. It was observed that all ten path trials are highly reproducible, the tip of the FLSAM traced the target paths, and the tracking results and errors were also symmetrical in mirrored paths. These results confirm the validity of the path-tracing algorithm. The maximum and the root mean square of the deviations were (2.25 mm, 1.7 mm) for pattern (A), (1.86 mm, 0.66 mm) for pattern (B), and (4.7 mm, 1.48 mm) for pattern (C), respectively. All of these deviations were under 4.0% (2.8 mm) of the total width of the FLSAM (140 mm). Although the same tendencies had found in the previous studies (e.g.; series studies of path tracing of a snake-type robot using screw driving theorem by same group [25-26]) that the deviations are approximately 4.0%, it is suggested that the path-tracing algorithm for the FLSAM while path tracing is effective, and minimizing the deviations is a task for future studies. In addition, the movement error of 4% (2.8 mm) is below 4% or less of the survey target size of 90 mm or more. In this respect, the validity of the path-tracing algorithm of the rear links of the FLSAM proposed in this paper was confirmed.

We would like to construct an algorithm to achieve three-dimensional motion by changing the arrangement and the number of sliding screw mechanisms of the FLSAM. Further studies should be needed at least the following points; (1) to improve the accuracy of path tracing, since the tracing deviations would accumulate if the FLSAM would be expanded to increase the number of joints, making it possible for a large error to occur at the tip position and (2) to generate a following path of rear links from the movement path of the tip instead of the Spline curve used in this study.

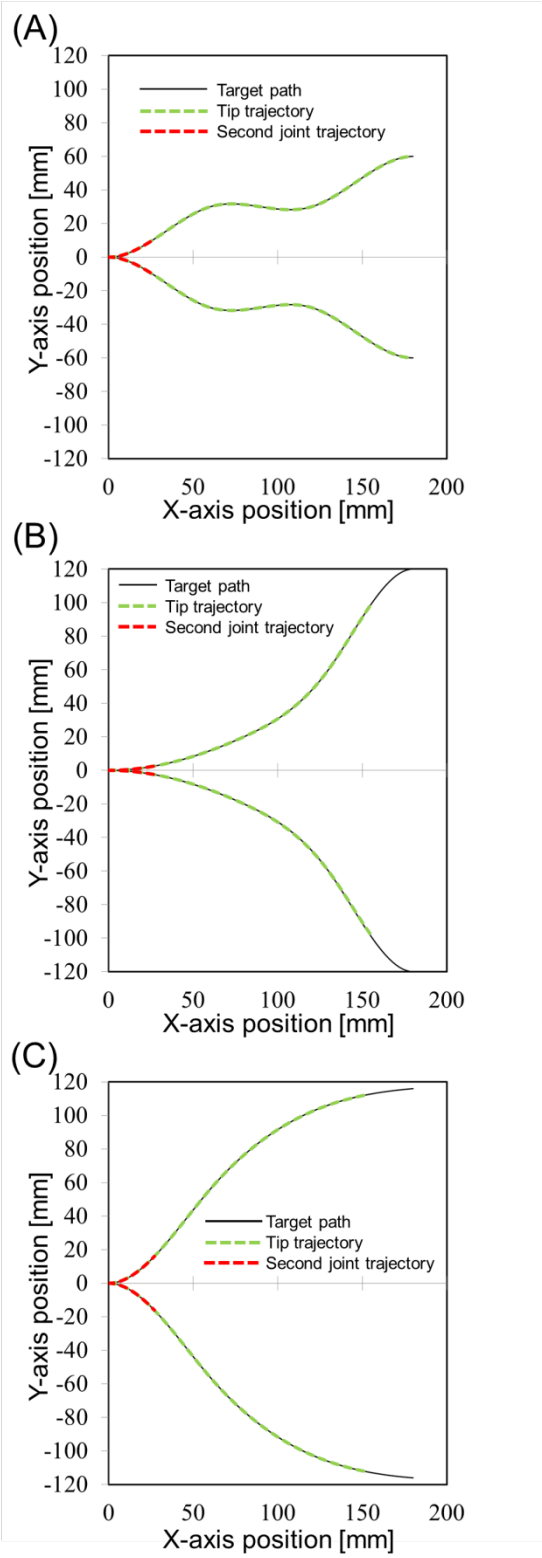


Figure 12: Simulation results of the path tracing algorithm; (A) figure "S" and its mirrored pattern, (B) the exponential figure and its mirrored patterns, and (C) the sigmoid curve figure and its mirrored patterns. Black lines are the target paths, green lines are the tip trajectories, and red lines are the trajectories of the second joint position.

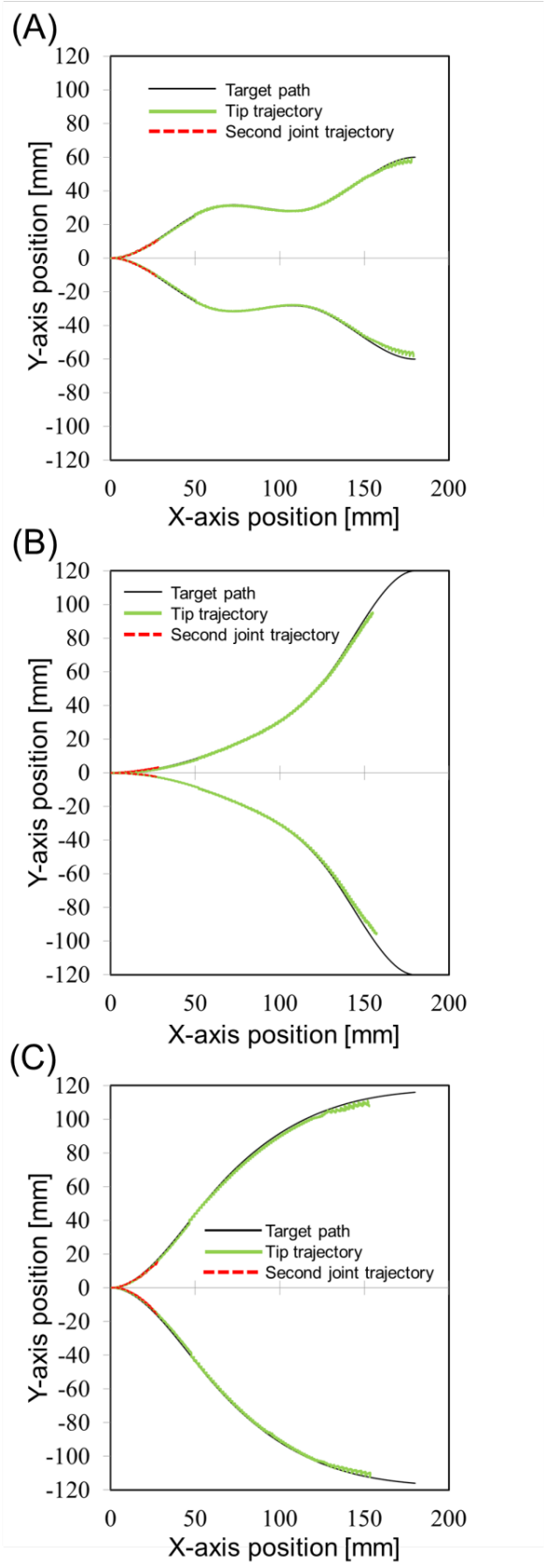


Figure 13: Results of path tracing by the FLSAM; (A) figure "S" and its mirrored pattern, (B) the exponential figure and its mirrored patterns, and (C) the sigmoid curve figure and its mirrored patterns. Black lines are the target paths, green lines are the tip trajectories, and red lines are the trajectories of the second joint position.

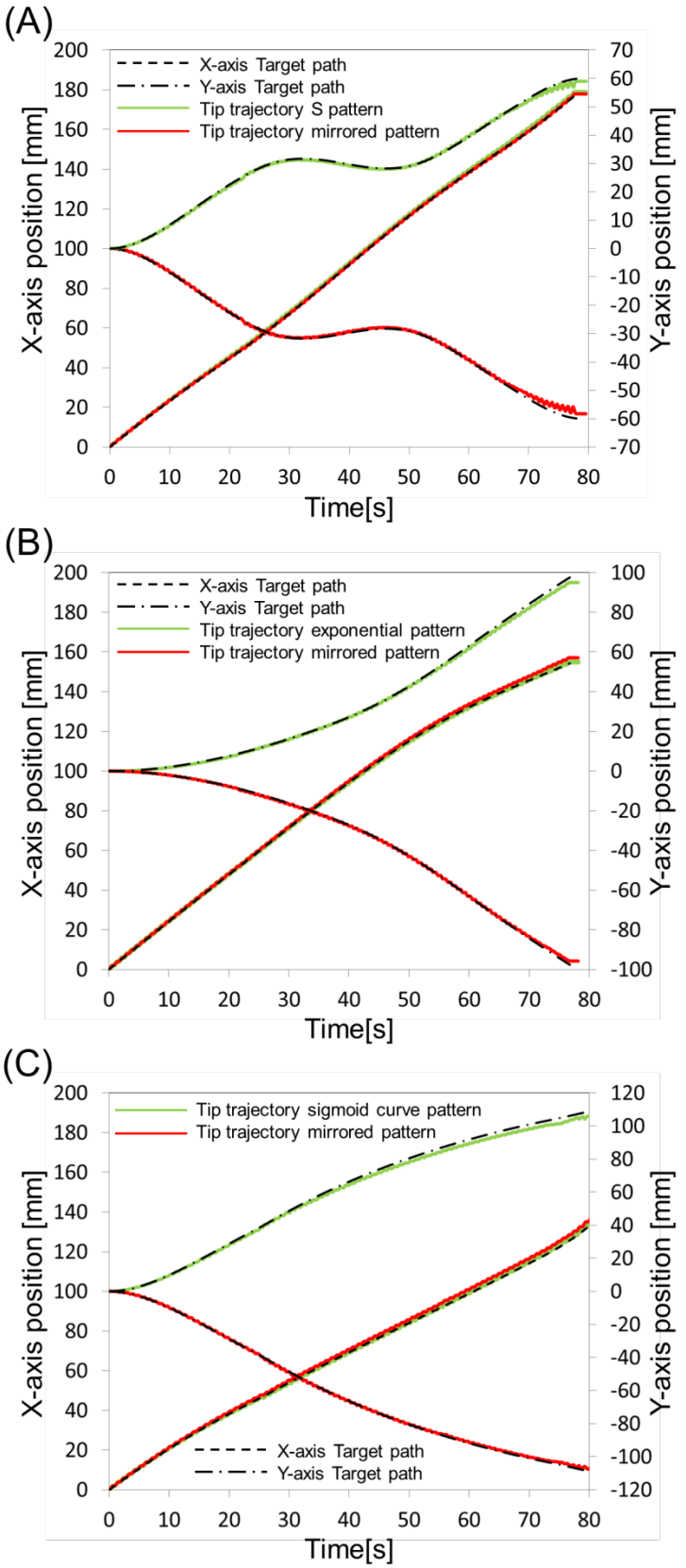


Figure 14: Time series of the position of the tip in X and Y axis; (A) figure "S" and its mirrored pattern, (B) the exponential figure and its mirrored patterns, and (C) the sigmoid curve figure and its mirrored patterns.

5. Conclusion

In order to realize a flexible underwater robot arm, we adopted a sliding screw mechanism and developed a prototype device, a "FLSAM." Path tracing in the rear-link mechanisms and underwater operation tests were executed, and the results suggested that the sliding screw mechanism is sufficient for construction of a flexible underwater robot arm. However, to obtain stable video data, further modification, such as using an autonomous postural control base, is necessary. Further study is also needed to improve the path-tracing control performance using kinematic and dynamic analysis of the FLSAM.

Appendix-1: Calculation of the cubic Spline interpolation curve

For the target path, a cubic Spline interpolation curve capable of creating a two-dimensional curve in space is used. The *n*th Spline interpolation is continuous to the (*n*-1)th derivative function, and oscillations between the representative points hardly occur. In addition, it is effective for generating a target path simulating an obstacle-avoidance route, since it is possible to produce a complicated shape in accordance with the given interpolation points. In this study, a cubic Spline curve was calculated by setting the first interpolation point as the origin of the global coordinate system and sequentially setting the next interpolation point at an arbitrary position in two-dimensional space.

When (*n*+1) points are given, there are *n* interpolation divisions, and the polynomials exist in each division as follows:

$$S_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i \qquad \text{(A1)} ,$$

where $S_i(x)$ is the cubic Spline curve function in the *i*th division; a_i, b_i, c_i, and d_i are the coefficients in *i*th division; and *i* (*i*=1 to *N*) equals the number of deviations. The coefficients of the polynomial (a_i, b_i, c_i, and d_i) are derived using the continuous condition (A2–4) at the interpolation point and the two boundary conditions (A5–6), as follows:

$$S_i(x_{i+1}) = S_{i+1}(x_{i+1}) \qquad \text{(A2)} ,$$

$$S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}) \qquad \text{(A3)} ,$$

$$S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}) \qquad \text{(A4)} ,$$

$$S_1'(x_1) = 0 \qquad \text{(A5)} ,$$

$$S_N'(x_{N+1}) = const \qquad \text{(A6)} ,$$

where the ith interpolation point coordinate is taken as (x_i, y_i), and $S_N'(x_{N+1}) = 0$ in this study. In the boundary condition (A5), the first derivatives of the start point of the Spline curve are set to 0, so that the Spline curves of the target path the and X axis are continuous at the origin of the global coordinate system.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 17K08029.

References

[1] Y. Iwamochi, M. Takagi and T. Miyoshi "Flexible Lengthening-Shortening Arm Mechanism for an Underwater Vehicle", Proceedings of 2017 IEEE International Conference on Mechatronics and Automation, MP1-6(6), 2017. http://doi.org/10.1109/ICMA.2017.8015921

[2] J. Pyo. H. Cho, H. Joe, T. Ura and S. Yu, "Development of hovering type AUV "Cyclops" and its performance evaluation using image mosaicking", Ocean Engineering, vol.109, pp.517-530, 2015. https://doi.org/10.1016/j.oceaneng.2015.09.023

[3] S. Kawamura, "Underwater robot development for manipulation task and their uses in Biwa lake", IFAC-PapersOnLine, vol.48, No.2, pp.14-19, 2015. https://doi.org/10.1016/j.ifacol.2015.06.003

[4] P. Gao. B. Zgeng, J. Liang and C. Ren, "Development of an Underwater Robot for Sediment Soil Sampling", Proceedings of the 3rd International Conference on Mechatronics and Industrial Informatics, pp.6-11, 2015. http://doi.org/10.2991/icmii-15.2015.2

[5] K. Satoh, M. Takagi, H. Mori and T. Miyoshi, "Development of an Underwater Robot for Researching of Underwater Environmental", The Japan Society of Mechanical Engineers, The Proceedings of JSME annual Conference on Robotics and Mechatronics (Robomec), Session ID: 1A2-16b5, 2016. http://doi.org/10.1299/jsmermd.2016.1A2-16b5

[6] N. Sakagami, M. Shibata, H. Hashizume and Y. Hagiwara, "Development of a Human-Sized ROV with Dual-Arm", The proceedings of the Oceans 2010 IEEE, 2010. http://doi.org/10.1109/OCEANSSYD.2010.5603897

[7] H. Shim, B.H. Jun, P.M. Lee, H. Beak and J. Lee, "Workspace control system of underwater tele-operated manipulators on an ROV", Ocean Engineering, vol.37, No.11-12, pp.1036-1047, 2010. https://doi.org/10.1016/j.oceaneng.2010.03.017

[8] P. Liljeback, O. Stavdahl, K. Pettersen and J. T. Gravdahl, "A review on modelling, implementation, and control of snake robots", Robotics and Autonomous Systems, vol.60, No.1, pp.29-40, 2012. https://doi.org/10.1016/j.robot.2011.08.010

[9] S. Hirose ans H. Yamada, "Snake-like robots [Tutorial]", IEEE Robotics & Automation Magazine, vol.16, No.1, pp.88-98, 2010. http://doi.org/10.1109/MRA.2009.932130

[10] G. Granosik, J. Borenstein and M. G. Hansen, Serpentine Robots for Industrial Inspection and Surveillance, Industrial Robotics: Programming, Simulation and Applications, 2006. http://doi.org/10.5772/4921

[11] S. Stromsoyen, "Propulsion Methods for Under Water Snake Robots - Investigation and Simulation Using Foil for Propulsion of a Snake Robot", Master thesis, Norwegian University of Science and Technology, 2015.

[12] A. Wolf, H. B. Brown, R. Casciola, A. Costa, M. Schwerin, E. Shamas and H. Choset, "A Mobile Hyper Redundant Mechanism for Search and Rescue Tasks", The proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2003. http://doi.org/10.1109/IROS.2003.1249309

[13] R. Buckingham, V. Chitrakaran, R. Conkie and G. Ferguson et al, "Snake-Arm Robots: A New Approach to Aircraft Assembly", SAE Technical Paper, 2007-01-3870, 2007. https://doi.org/10.4271/2007-01-3870.

[14] D. Palmer, S. C. Guzman and D. Axinte, "Real-time method for tip following navigation of continuum snake arm robots", Robotics and Autonomous Systems, vol.62, No.10, pp.1478-1485, 2014. https://doi.org/10.1016/j.robot.2014.05.013

[15] H. Nakagawa and K. Yoshimura, "Super-multi Joint Manipulator Composed of Linking 3 DOF Joint Unit", The Japan Society of Mechanical Engineers, The proceedings of the JSME annual meeting, Session ID: 834, May 2005. http://doi.org/10.1299/jsmemecjo.2005.5.0_7

[16] S. Hirose and S. Ma, "A moray drive of multi-joint manipulator", Journal of the Robotics Society of Japan, vol. 8, No.1, pp.9-16, 1990. http://doi.org/10.7210/jrsj.8.9

[17] S. Ma, S. Hirose and K. Yokoshima, "2 DOF moray drive for multijoint manipulators", Journal of the Robotics Society of Japan, vol.14, No.3, pp.436-443, 1996. http://doi.org/10.7210/jrsj.14.436

[18] S. Ma and I. Kobayashi, "Path-tracking control of a moray-type robot arm with consideration of actuator's actuation characteristics", Journal of the Robotics Society of Japan, vol.18, No.1, pp.135-141, 2000. http://doi.org/10.7210/jrsj.18.135

[19] S. Ma and Y. Ohmameuda, "Obstacle avoidance control of moray-type robot arm", Transactions of the Japan Society of Mechanical Engineers Series C, vol.68, No.670, pp.1791-1797, 2002. http://doi.org/10.1299/kikaic.68.1791

[20] Y. Cao, V.J. Modi, C.W.de Silva, M. Chen and A.K.Misra, "Trajectory tracking experiments using a novel manipulator", Acta Astronautica, vol.52, pp.523-540, 2003. https://doi.org/10.1016/S0094-5765(02)00123-6

[21] E. Kelasidi, P. Liljeback, K. Y. Pettersen and J. Tommy Gravdahl, "Integral line-of-sight guidance for path following control of underwater snake robots: Theory and Experiments", IEEE Transactions on Robotics, vol.33, No.3, pp.610-628, 2017. http://doi.org/10.1109/TRO.2017.2651119

[22] M. Hara, S. Satomura and F. Matsuno et al, "Development and Test of Fundamental Action of Snake-Like Robot with Driving Mechanism Based on Screw Principle", The Japan Society of Mechanical Engineers, The Proceedings of JSME annual Conference on Robotics and Mechatronics (Robomec), Session ID: 2P2-D27, 2006. http://doi.org/10.1299/jsmermd.2006._2P2-D27_1

[23] T. Kamegawa and F. Matsuno, "Development of a remote-controlled double headed snake-like rescue robot KOHGA", Journal of the Robotics Society of Japan, vol.25, No.7, pp.1074-1081, 2007. http://doi.org/10.7210/jrsj.25.1074

[24] K. Ito, R. Murai and K. Nakamichi, "Development of snake-like rescue robot designed for ease of use ", Journal of the Robotics Society of Japan, vol.27, No.4, pp.419-426, 2009. http://doi.org/10.7210/jrsj.27.419

[25] M. Hara, S. Satomura, H. Fukushima, T. Kamegawa, H. Igarashi and F. Matsuno, "Control of a Snake-like Robot Using the Screw Drive Mechanism", Proceedings of 2007 IEEE International Conference on Robotics and Automation, Roma Italy, 2007. http://doi.org/10.1109/ROBOT.2007.364074

[26] H. Fukushima, M. Tanaka, T. Kamegawa and F. Matsuno, "Front-Unit-Following Control of a Snake-like Robot Using Screw Drive Mechanism", Journal of the Robotics Society of Japan, vol.28, No.6, pp.707-714, 2010. http://doi.org/10.7210/jrsj.28.707

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 297-309 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

The Use of LMS AMESim in the Fault Diagnosis of a Commercial PEM Fuel Cell System

Reem Izzeldin Salim[*,1], Hassan Noura[2], Abbas Fardoun[3]

[1]*Electrical Engineering Department, United Arab Emirates University, Al Ain, United Arab Emirates*

[2]*Faculty of Engineering, Islamic University of Lebanon, Wardanyeh, Lebanon*

[3]*Electrical Engineering Department, Al Maaref University, Beirut, Lebanon*

ARTICLE INFO

Article history:
Received: 30 October, 2017
Accepted: 18 January, 2018
Online: 30 January, 2018

Keywords:
Proton Exchange Membrane Fuel Cells
Modeling
Simulation
AMESim
Fault Diagnosis

ABSTRACT

The world's pollution rates have been increasing exponentially due to the many reckless lifestyle practices of human beings as well as their choices of contaminating power sources. Eventually, this lead to a worldwide awareness on the risks of those power sources, and in turn, a movement towards the exploration and deployment of several green technologies emerged.

Proton Exchange Membrane Fuel cells (PEMFCs) are one of those green technologies. However, in order to be able to successfully and efficiently deploy PEMFC systems, a solid fault diagnosis scheme is needed. The development of accurate model based fault diagnosis schemes has been imposing a lot of challenge and difficulty on researchers due to the high complexity of the PEMFC system. Furthermore, confidentiality issues with the manufacturer can also impose further constraints on the model development of a commercial PEMFC system. In this work, an approach to develop an accurate PEMFC system model despite the lack of crucial system information is presented through the use of Siemens LMS AMESim software. The developed model is then used to develop a fault diagnosis scheme that is able to detect and isolate five system faults.

1. Introduction

Proton Exchange Membrane Fuel Cells are complex multi-physics systems (chemical, electrical, fluidic, thermal, and mechanical phenomena are inter-acting with one another). This makes the modeling and fault diagnosis of PEMFC systems a very difficult task. Furthermore, when the modeling is based on experimental testing and experimental data, some limitations are usually faced. For example, many physical system data may be absent or difficult to obtain with the system's supplied data acquisition. Furthermore, due to warranty issues, the addition of sensors might be difficult since only limited access to the systems is allowed. Likewise, other specific parameters might be unobtainable due to manufacturer confidentiality issues.

This paper is an extension of work originally presented in the 7th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO) [1]. It presents a novel approach, in which the Siemens software, LMS AMESim 14, is used as an alternative modeling tool to model a PEMFC system when such limitations are faced. The LMS AMESim software can serve as an excellent and attractive simulation and modeling platform for different complex physical systems such as automobile and aviation systems as well as power generation systems and transmission lines. Moreover, it has outstanding simulation capabilities that makes it an excellent platform for fault simulation, and fault diagnosis studies. The graphical user interface comprehensive library it contains makes it very easy to compile a complete system model from different system components. Instead of writing all the modeling equations which could be time consuming and is prone to modeling errors, a user

[*]Reem I. Salim, Electrical Engineering Dept., UAEU, 15551 Al Ain, UAE, Tel: +971507649740, Email: reemsalim@gmail.com

can use a ready accurate model better and focus on the parameter identification step in order to find the model that matches the pursued system performance measures. Furthermore, a researcher can easily change the scripts of the components used to better suit their system's design, as well as create new libraries with more specific components.

The ElectraGen™ 3 kW PEMFC system shown in Figure 1, is an actual commercial system used in practice in many sectors, especially in telecommunication companies. It uses hydrogen gas supplied from pressurized hydrogen cylinders as the anode fuel, and atmospheric air supplied by a compressor and humidified through a built-in humidity exchanger as the cathode fuel. The ElectraGen™ system is an outdoor air cooled system, and it is only operable at ambient temperatures ranging from –40°C to 50°C. It contains a total of 38 cells and can produce up to 3 kW of unregulated DC output power and has a rated voltage of 48V.



Figure 1: The ElectraGen™ PEMFC system and the 3 kW Load.

The module contains the ElectraGen™ 3 kW Fuel Cell stack with integrated microprocessor controller and safety features, a hydrogen pressure gauge which gives an indication of fuel level and the 3kW load which consists of 30 lamps, 100W each (see Figure 1). The system is connected through data acquisition to a GUI based on LabVIEW to monitor and log the different system variables (stack current, stack voltage, external voltage, individual cell voltages, cabinet temperature, cathode air temperature, coolant temperature, exhaust temperature, and hydrogen pressure). The ElectraGen™ system is installed outdoor and the experimental data sets collected from the system were taken at extreme environmental conditions in the summer at noon with the ambient temperature ranging from 48°C to 52°C. However, due to warranty issues, limited access to the systems was allowed and limited data was obtainable from the data acquisitions. Moreover, due to confidentiality issues with the manufacturer, several physical parameters were unobtainable such as active cell area, membrane length, volumes of cathode and anode chambers, mass of stack, etc. This made it very difficult to model the system and develop an accurate fault diagnosis scheme. Therefore LMS AMESim was used as an alternative modeling platform since all the common PEMFC modeling equations available in literature [2, 3] are already embedded in its library components.

Furthermore, it is convenient to mention here that the simulated model will not be an exact match for the ElectraGen™ system since not enough system readings were available to match the model to. However, it is the aim of this work to develop an AMESim model that is as realistic as possible by matching all the known features of the actual physical system (the ElectraGen™ system) to their equivalents in the AMESim model as best as possible. This model can then be used in the fault diagnosis study.

Section 2 presents the ElectraGen™ modeling and validation results using the Siemens LMS AMESim Software. In section 3, the simulation of five different system faults is presented, and in section 4 two residual generation techniques are evaluated and then outperforming technique is used to develop a fault diagnosis scheme in AMESim for the ElectraGen™ system. The fault diagnosis scheme is then evaluated and concluding remarks are finally given in section 5.

2. AMESim Modeling of the ElectraGen™ 3 kW System

The LMS AMESim software contains several embedded parameter identification tools including Genetic Algorithms (GA). However, in order to be able to use such tools efficiently, several software licenses are needed. To further explain this, GA is a population based mechanism that is known to be successful because it performs parallel evaluations, and when only one AMESim license is available, using GA becomes impractical. As an example, if the GA has 20 individuals in its population, and with a preset maximum number of generations of 500, this is equivalent to 10,000 runs of the software. With one system license, the software will be unable to perform parallel evaluations. Thus, if a single run of the software takes 1 minute, then the parameter identification process using GA and one system license will take 10,000 minutes (almost seven days).

As a result, matching the parameters of the AMESim model to the actual system performance of the ElectraGen™ system was done through trial and error. The LMS AMESim model developed for the ElectraGen™ PEMFC system is presented in Figure 2.

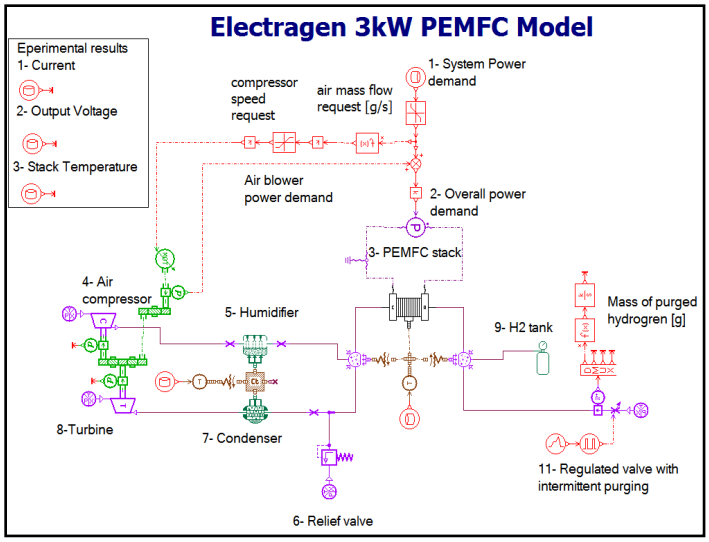


Figure 2: ElectraGen™ 3 kW system's model in AMESim.

2.1. AMESim Model Parameter Identification

The actual power demand data collected experimentally from the ElectraGen™ system was used as the input to the system in

Figure 2. After several trial and error attempts to match the stack voltage, stack current, and stack temperature values as best as possible to the actual experimental data collected from the system; the best obtained match is depicted in the following figures: Figure 3 gives the power demand input of the modeling data set, and Figures 4, 5 and 6 presents a comparison between the actual experimental current, stack voltage and stack temperature respectively to those resulting from the AMESim simulation model.

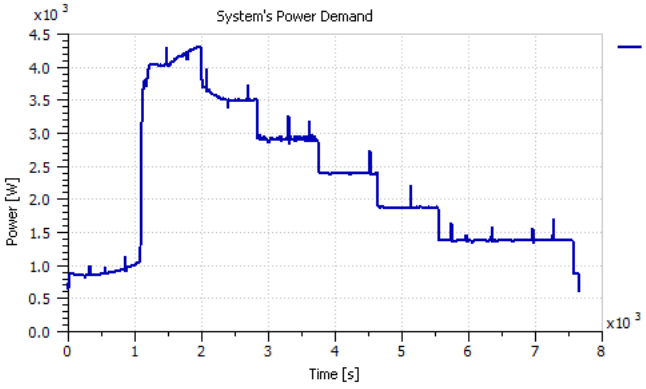


Figure 3: Power demand of the AMESim modeling data set.

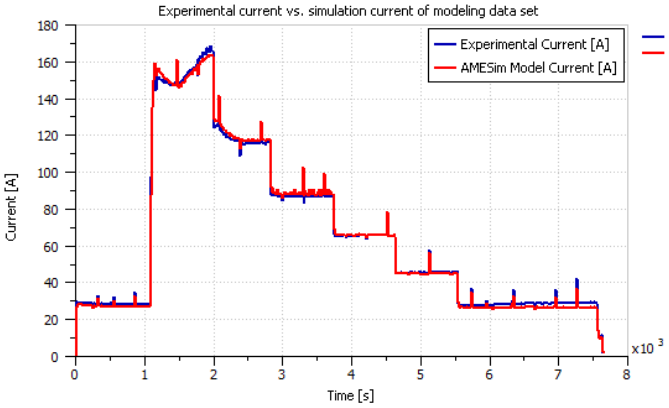


Figure 4: Experimental versus AMESim model's resulting current of the modeling data set.

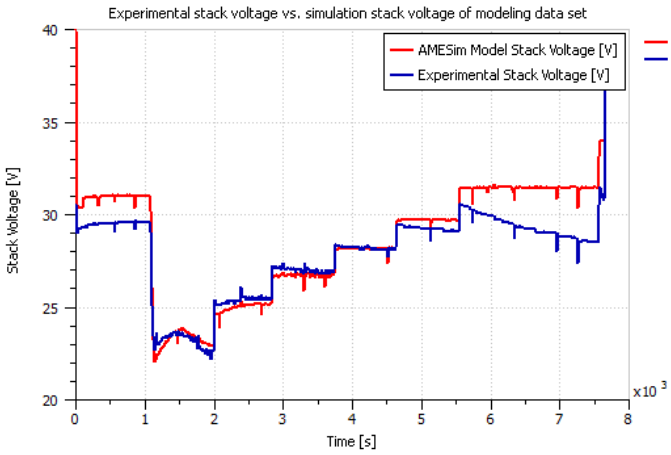


Figure 5: Experimental versus AMESim model's resulting stack voltage of the modeling data set.

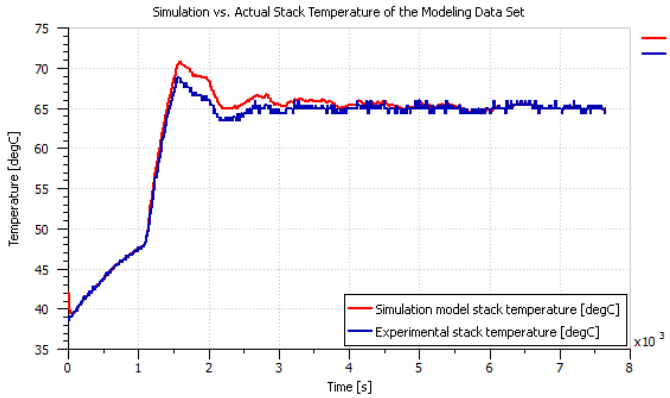


Figure 6: Experimental versus AMESim model's resulting stack temperature of the modeling data set.

2.2. AMESim Model Validation

Two other ElectraGen™ experimental data sets were used to validate the obtained AMESim model of Figure 2. Figure 7 presents the power demand of the first validation example and Figures 8 and 9 compare the actual experimental current and stack voltage to those obtained from the LMS AMESim model respectively. Similarly, the power demand of the second validation example is given in Figure 10, and a comparison between the actual and simulation current is presented in Figure 11, whereas a comparison between the actual and simulation stack voltage is presented in Figure 12.

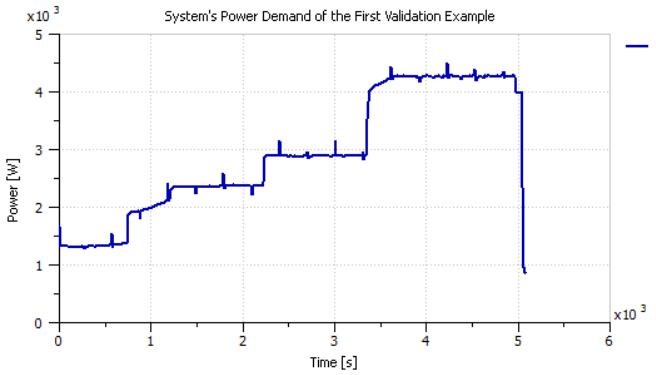


Figure 7: Power demand of the first validation example.

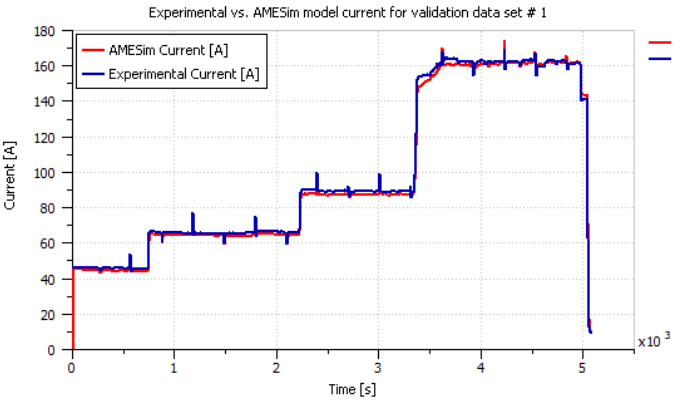


Figure 8: Actual versus simulation current of the first validation example.

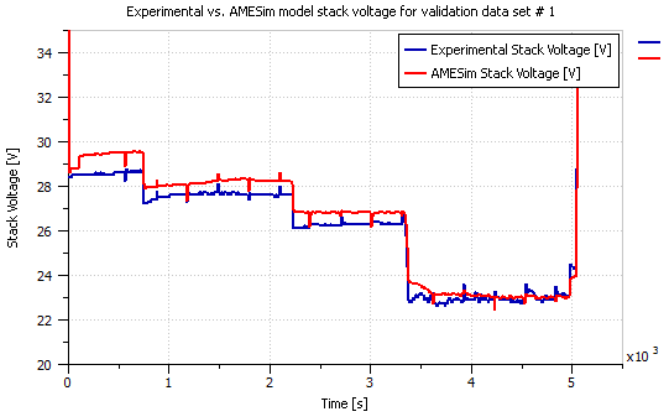


Figure 9: Actual versus simulation stack voltage of the first validation example.

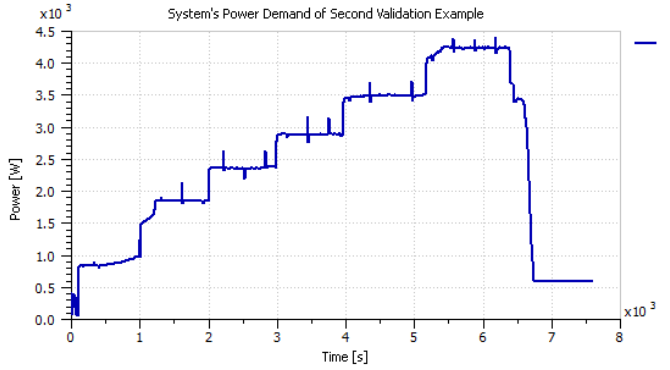


Figure 10: Power demand of the second validation example.

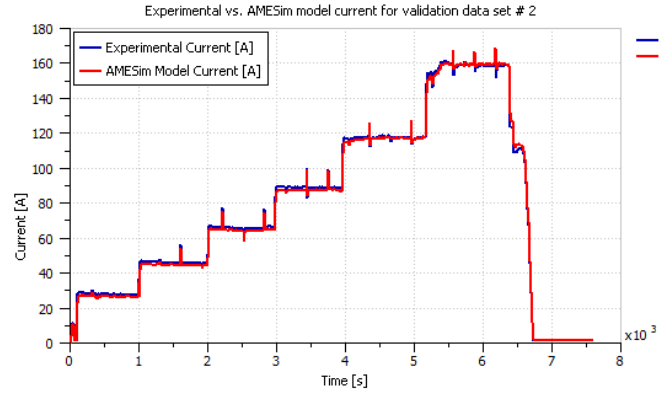


Figure 11: Actual versus simulation current of the second validation example.

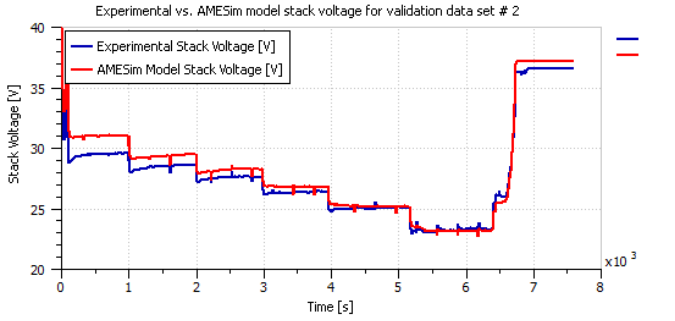


Figure 12: Actual versus simulation stack voltage of the second validation example.

Obviously, a perfect representation of the ElectraGen™ system is unobtainable through trial and error alone. Nevertheless, after validation, the AMESim obtained model proved to give a proper representation of the system's performance and was therefore used in the following fault diagnosis study.

Moreover, if resources (multiple licenses of the software) were available, a much more accurate representation of the system would have been achievable through the use of Genetic Algorithms in the parameter identification approach in AMESim, which would have in turn lead to a better fault diagnosis study.

3. Fault Simulation Using AMESim Model

Several faults were induced in the AMESim model and the system's performance measures towards those faults were recorded. The induced faults are:

1. Drying
2. Flooding
3. Air leakage
4. Hydrogen Leakage
5. Cooling System Failure

The above faults were simulated using different techniques.

3.1. Drying

Zawodzinski et al. were the first to describe the water content in the membrane by (λ) in [4] in order to estimate its state of humidity. As presented in (1), λ represents the ratio between the number of water molecules in the membrane to the number of ($SO_3^- H^+$) charge sites in the Nafion layer of the membrane [5].

$$\lambda = \frac{H_2O}{SO_3^- H^+} \quad (1)$$

The AMESim PEMFC stack module automatically calculates the water content λ in the membrane. Therefore, when simulating drying or flooding, λ would give an indication towards the membrane's state of health.

However, λ cannot be used in the fault diagnosis study because this parameter is not readily available in commercial fuel cells. Furthermore, the PEMFC undergoes drying condition when the water content in the membrane λ drops below 4 [6].

Therefore in order to simulate drying, the humidity level of the input air was set to be 0% and the target humidity level of the humidifier was dropped to 10% only. The water content in a healthy stack and that of a drying stack are compared in Figure 13. The water content of the stack that is undergoing drying is obviously well below 4. Figure 14 on the other hand depicts the difference between the polarization curve of a healthy stack and a drying stack. It is noticed that drying results in a significant voltage.

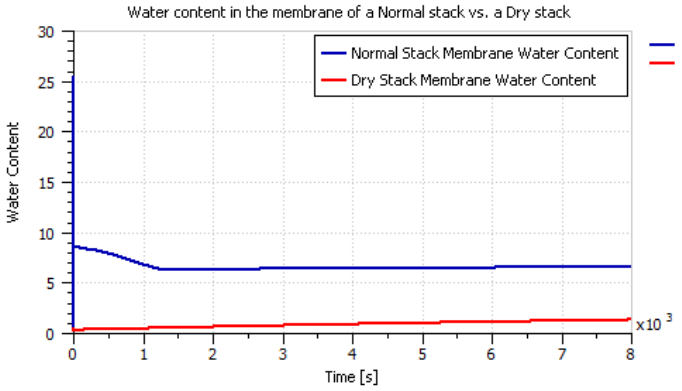


Figure 13: The water content in the membrane (λ) of a normal stack versus a dry stack.

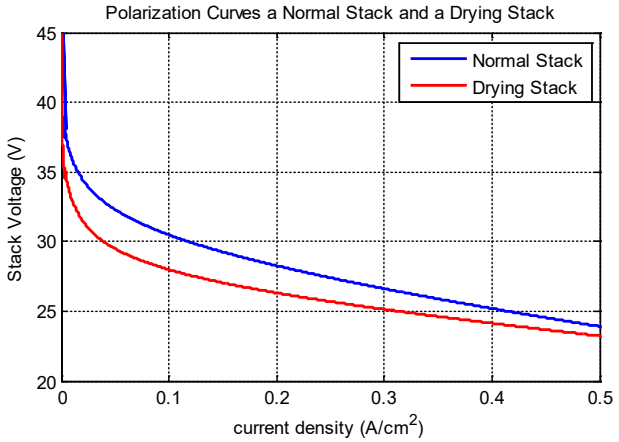


Figure 14: Normal stack's polarization curve in comparison to a dry stack's polarization curve.

3.2. Flooding

Flooding was simulated by dropping the stack temperature to 25°C while increasing the humidifier's target humidity level to 100%. Flooding affected both the voltage and current profiles of the stack as depicted in the polarization curve comparison given in Figure 15. Furthermore, Figure 16 shows the water content λ of the simulated flooded stack.

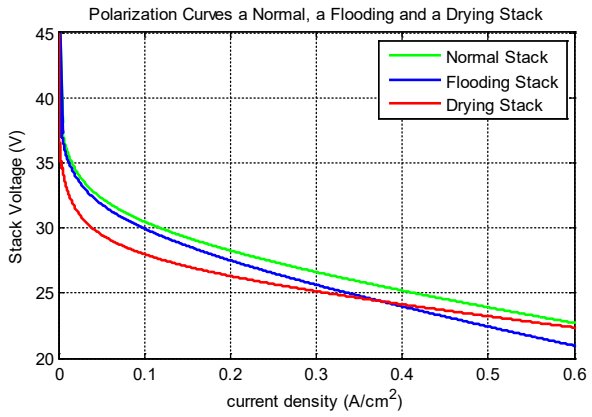


Figure 15: Comparison of pressure drop in a normal stack, a flooding stack and a drying stack.

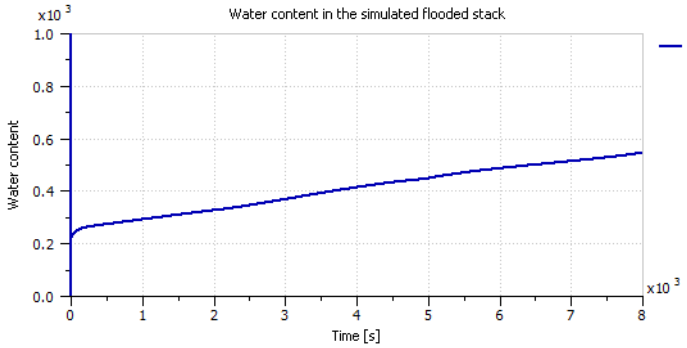


Figure 16: The water content in the membrane (λ) of the simulated flooding stack

Note that both flooding and drying result in similar effects to the stack's polarization curve. However, flooding seemed to also cause a distinctive effect on the cathode pressure drop. As the stack gets flooded with water, the pressure drop across the cathode increases. Nonetheless, in order to clearly see this effect on cathode pressure, the stack had to be fed with a step power demand profile such as that of Figure 17. The cathode pressure of both the healthy and flooded stacks with respect to the power demand input of Figure 16 are compared in Figure 17. The flooded stack showed a steeper drop in pressure when compared to the healthy stack.

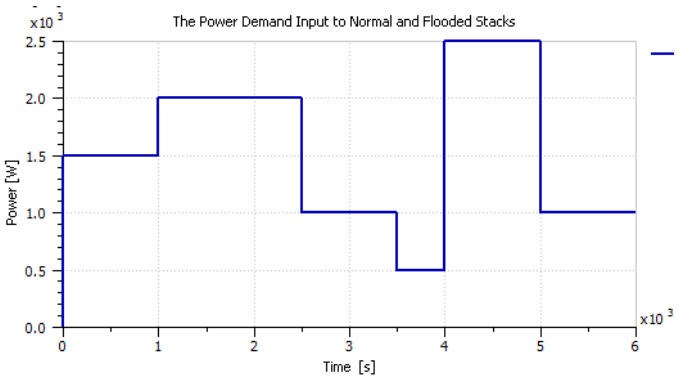


Figure 17: Power demand input to normal stack and flooding stacks.

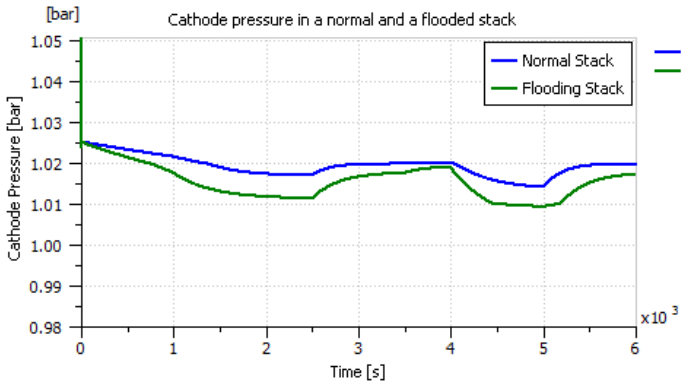


Figure 18: Comparison of pressure drop in a normal stack and a flooding stack.

3.3. Air leakage

In order to simulate air leakage, a relief valve was added right after the humidifier as depicted in Figure 19, and was set to leak air at a rate of around 10 g/s. Figure 20 shows the amount of leakage introduced in g/s. Furthermore, Figure 21 compares the input air flow rate to the PEMFC stack with and without air leakage.

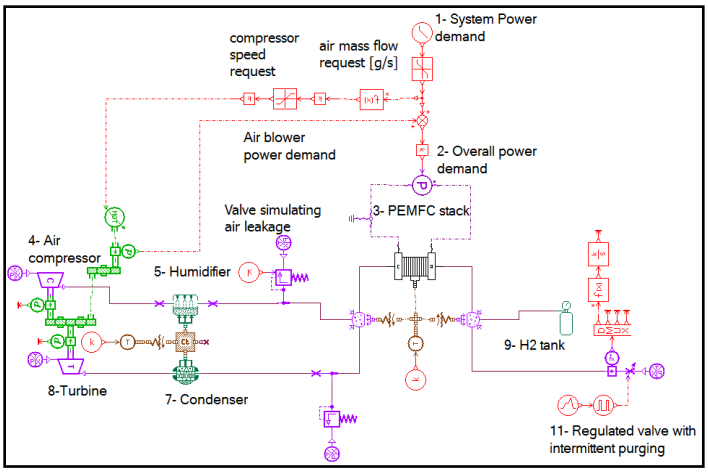


Figure 19: The relief valve added to the AMESim model to simulate air leakage.

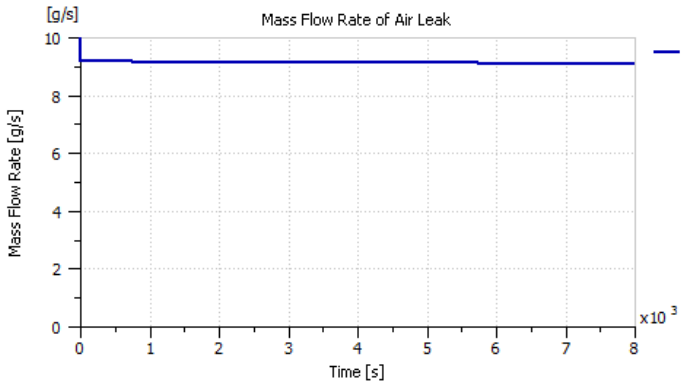


Figure 20: The flow rate of the air leakage induced to the PEMFC system.

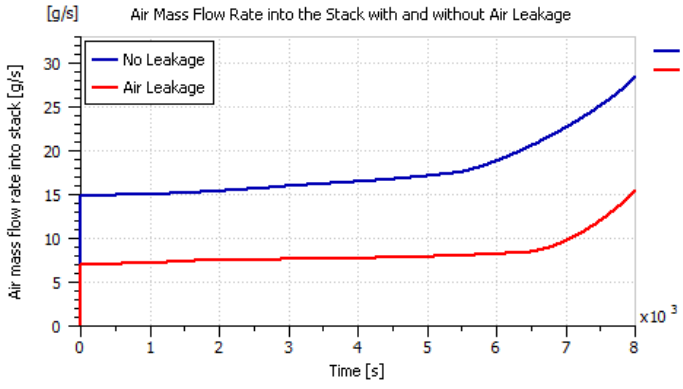


Figure 21: Input air flow rate to the stack with and without air leakage.

It can be deduced from the polarization curve comparison depicted in Figure 22 that the air leakage had no noticeable effect on the stack's current or voltage. However, it was noticed to impose a significant effect on the cathode pressure. To better see this effect, both the healthy and air leaking stacks were fed with the step power demand of Figure 17. Figure 23 shows the air leakage effect on the cathode pressure drop when compared to a non-leaking stack.

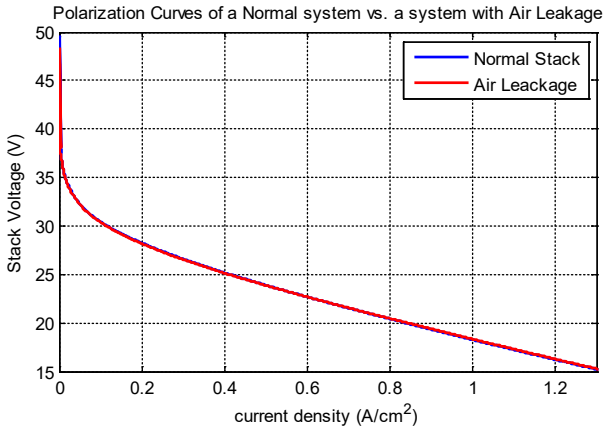


Figure 22: Polarization curves of a normal system vs. a system undergoing air leakage.

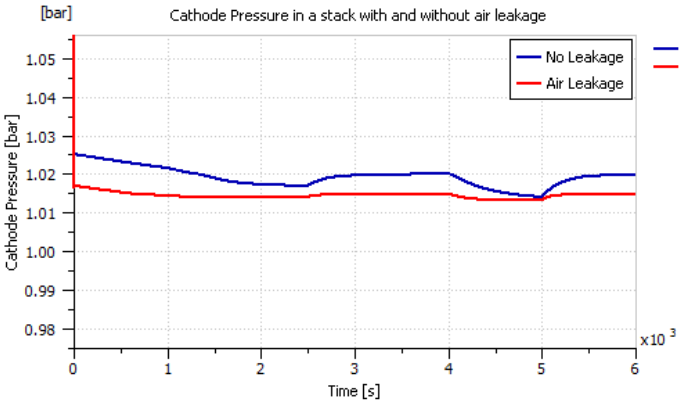


Figure 23: Pressure drop in the cathode of a normal system vs. a system undergoing air leakage.

3.4. Hydrogen leakage

Similar to air leakage, hydrogen leakage was also simulated through the addition of a relief valve right after the hydrogen canister as depicted in Figure 24 in order to leak hydrogen at a rate of 10 g/s. Figure 25 shows the amount of hydrogen leakage introduced in g/s.

Hydrogen leakage was found to result in a slight effect on stack's polarization curve as shown in the comparison of Figure 26. Furthermore, hydrogen leakage was also found to affect the anode pressure as shown in Figure 27, but had no significant effect on the cathode pressure as seen in Figure 28.

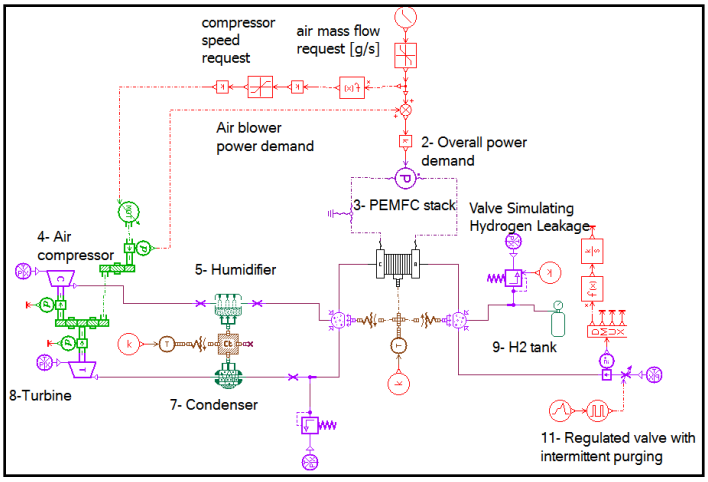


Figure 24: The relief valve added to the AMESim model to simulate Hydrogen leakage.

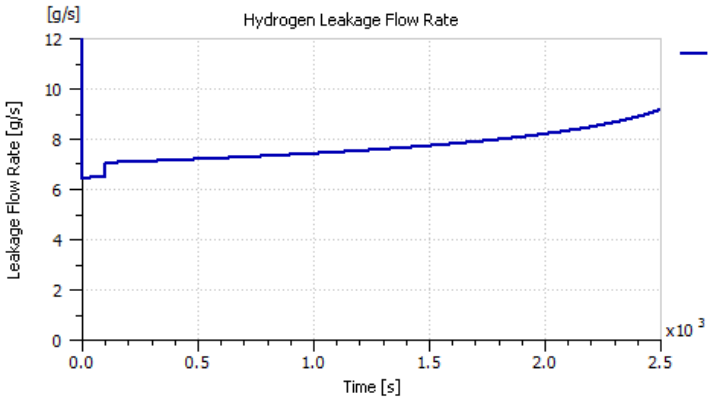


Figure 25: The flow rate of the Hydrogen leakage induced to the PEMFC system.

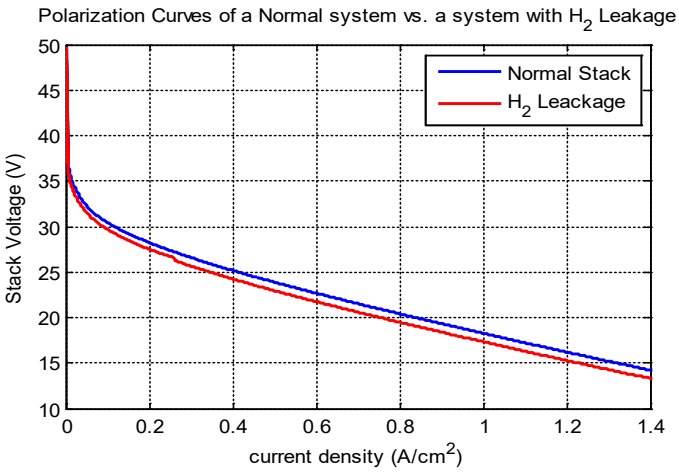


Figure 26: Polarization curves of a normal system vs. a system undergoing Hydrogen leakage.

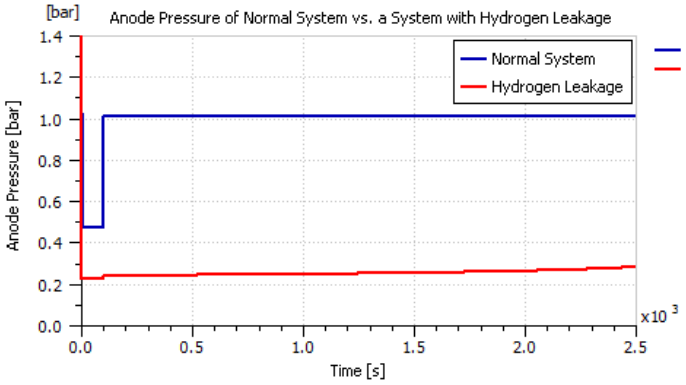


Figure 27: Anode pressure of a normal system vs. a system undergoing Hydrogen leakage.

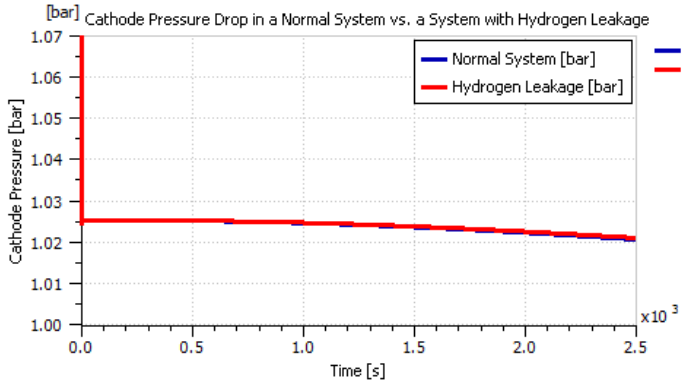


Figure 28: Pressure drop in the cathode of a normal system vs. a system undergoing Hydrogen leakage.

3.5. Cooling Failure

The stack temperature of the ElectraGen™ system should be maintained at a temperature value between 60°C to 65°C for maximum efficiency through air cooling. However, it is also convenient to mention that the stack temperature should never reach any value above 75°C in order to avoid damage of the membrane. Thus, it is important to flag a cooling system failure as soon as the stack temperature reaches 75°C or higher. Hence, the stack's temperature was increased to 75°C in order to simulate cooling failure. The polarization curve was found to undergo a significant effect with the increase in stack temperature as shown in Figure 29.

Note that the 75°C stack temperature results in a slight improvement in the system's polarization curve because the resistive components in the activation and ohmic voltage drop will decrease with the increase in stack temperature. However, at such an elevated temperature, drying of the stack will be inevitable. Furthermore, comprehensive literature review [7] revealed that operating the PEMFC at higher stack temperatures is a common stressor for almost all health degradation mechanisms. Thus, this slight improvement in the polarization curve is worthless since it will significantly shorten the PEMFCs lifespan.

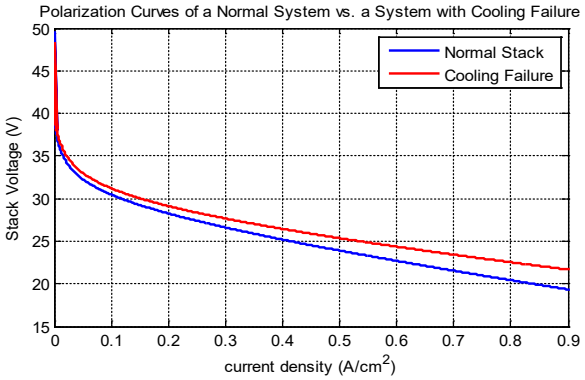


Figure 29: Polarization curve of a normal system vs. a system undergoing cooling failure.

Nonetheless, the stack temperature should be the main parameter used to detect the presence of cooling failure regardless of the effects on the voltage and current profiles. As soon the stack's temperature reading reaches 75°C is a cooling system failure fault should be flagged.

Note from Figure 29 that there is an improvement in the voltage performance of the fuel cell with the increase of stack temperature. This is expected since the rate of chemical reactions increase with the temperature causing this increase in voltage. However, it should also be noted that at this temperature value of 75°C, drying of the membrane is inevitable. Furthermore, high stack temperature is a well-known stressor for PEMFCs [8 – 11]. Thus, operating the system at such elevated stack temperatures will significantly reduce its lifespan, which makes this small voltage improvement at high stack temperature values worthless.

4. Fault Diagnosis

In this work, the model based fault diagnosis approach is based on the real-time comparison between the actual system performance and the performance predicted by the developed AMESim model. Any predicted discrepancies will be analyzed to determine the type of system fault occurring at the moment.

4.1. Residual generation

It can be concluded from the previous section that in order to detect discrepancies between the actual system and its developed model that can help in the fault detection and isolation process, five residuals (see Figure 30) should be generated based on the following five system variables: stack voltage (V_{stack}), current (I), stack temperature (T_{stack}), cathode pressure ($P_{Cathode}$) and anode pressure (P_{anode}).

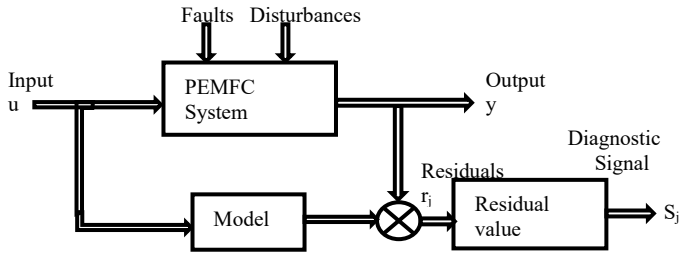


Figure 30: Residual generation diagram [12].

Several forms exist in literature for the calculation of residuals. In the simplest form given in (2), the absolute value of the residual (r_j) is compared to a relative threshold value (τ_j). The diagnostic signal (S_j) is then set to be "0" if the absolute residual is less than or equal to the threshold to indicate that no discrepancy is detected, and "1" if the absolute residual is greater than the set threshold which indicates that a discrepancy is detected [12]

$$S_j = \begin{cases} 0 & if \quad |r_j| \le \tau_j \\ 1 & if \quad |r_j| > \tau_j \end{cases} \tag{2}$$

Note that basing the fault diagnosis approach on the residuals calculated in (2) makes the approach prone to diagnosis errors. This is because the residuals of (2) are highly sensitive to instantaneous changes in the system such as the electromagnetic disturbance pulses which act on the system's measured signals as well as the occurrence of actual system faults.

Therefore, the calculation of those residuals can be slightly altered to decrease their sensitivity to disturbances. Thus, instead of calculating instantaneous residuals, they can be calculated over a moving window of time and judging the average of all residuals in that window as explained by (3), where N_r represents the number of residuals in the window [13]:

$$S_j = \begin{cases} 0 & if \quad \tilde{r}_j(N_r) = \frac{1}{N_r}\left|\sum_{n=0}^{N_r-1} r_{j,k-n}\right| \le \tau_j \\ 1 & if \quad \tilde{r}_j(N_r) = \frac{1}{N_r}\left|\sum_{n=0}^{N_r-1} r_{j,k-n}\right| > \tau_j \end{cases} \tag{3}$$

The threshold value for each residual is given in Table 1, and it was set based on the accuracy of the actual system sensors installed in the ElectraGen™ PEMFC system.

Table 1: Accuracy of actual sensors used to measure system variables

System variable	Sensor accuracy
Stack Voltage (V_{stack})	± 0.5 V
Current (I)	± 1 A
Stack Temperature (T_{stack})	± 3 °C
Cathode Pressure ($P_{cathode}$)	± 1 mbar
Anode Pressure (P_{anode})	± 1 mbar

4.2. Sensitivity assessment

Both the residual calculation techniques of (2) and (3) were implemented in LMS AMESim to study and assess their effectiveness and test their sensitivity towards electromagnetic disturbance pulses. Thus, a noisy step power demand signal with disturbance pulses was generated to help with the assessment. The system estimator however was fed with a clean power demand signal to test the diagnostic signals sensitivity toward those disturbance pulses.

The residuals that were calculated by the system are: r_1 which is based on the stack temperature (T_{stack}), r_2 which is based on the stack voltage (V_{stack}), r_3 which is based on the current (I), r_4 which is based on the anode pressure (P_{anode}) and r_5 which is based on the

cathode pressure ($P_{cathode}$). On the other hand, S_i is the respective diagnostic signal r_i.

The faults being detected by the proposed fault diagnosis approach are: (f_1 = Drying, f_2 = Flooding, f_3 = Air Leakage, f_4 = Hydrogen Leakage, and f_5 = Cooling Failure). The results of section III are summarized in Table 2 to help in the fault discrimination process, where 0 indicates that no discrepancies are detected between the predicted value and the actual measured value, and 1 indicates the positive detection of discrepancies whereas X indicates the fault exists whether the diagnostic signal detects discrepancies in the system variable or not. Note that the cooling failure that f_5, is set to be flagged as soon as the stack temperature reaches a value of 75°C (i.e. based on S_1) regardless of the effects on the stack voltage and current (i.e. S_2 and S_3).

Table 2: The effect of faults on diagnostic signals

fault	S_1	S_2	S_3	S_4	S_5
f_1	0	1	1	0	0
f_2	0	1	1	0	1
f_3	0	0	0	0	1
f_4	0	1	1	1	0
f_5	1	X	X	0	0

Figure 31 shows the fault diagnosis system built based on the instantaneous diagnostic signals calculation of (2). Figure 32 shows the noisy power demand signal fed to the actual system being diagnosed and the clean power demand signal fed to the system estimator. The diagnostic signals calculated based on the instantaneous residuals of (2) are shown in Figure 33.

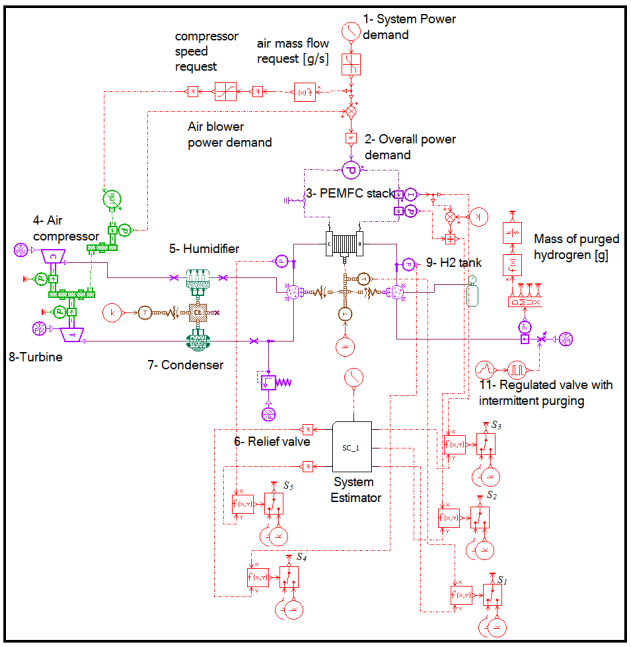


Figure 31: Fault diagnosis system based on instantaneous diagnostic signals.

It is obvious from Figure 33 that the diagnosis approach based on the instantaneous residuals in (2) is impractical and highly sensitive to electromagnetic disturbance pulses, which can therefore result in the detection of a faults when no fault actually exists. For instance, at around 150 s, both S_2 and S_3 diagnostic signals were triggered, which from Table 2 , will indicate the presence of the first fault f_1 (Drying).

Afterwards, the diagnostic signal of (3) was also implemented in LMS AMESim in order to be assessed based on three different moving windows of time (5 s, 10 s and 15 s). The same power signals of Figure 32 were used in the assessment to evaluate its sensitivity towards electromagnetic disturbance pulses. The fault diagnosis system built based on the diagnostic signals calculation of (3) is presented in Figure 34.

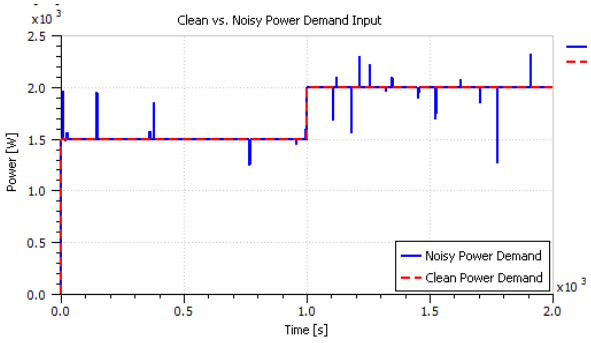


Figure 32: Clean and noisy power demand signals.

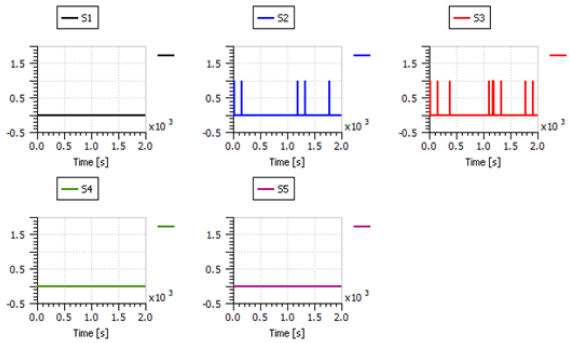


Figure 33: The effect of the pulsating noise on the five diagnostic signals.

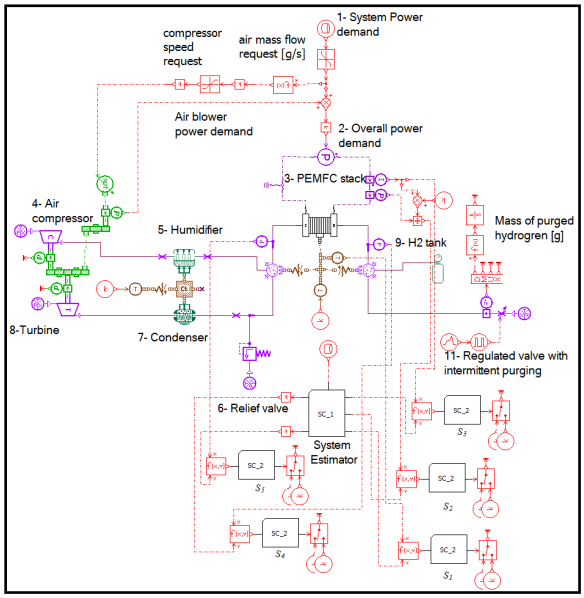


Figure 34: Fault diagnosis system based on diagnostic signals calculated over a window of time.

Note that a super-component was added to this model to help calculate the diagnostic signal over three different windows of time equal to 5 s, 10 s and 15 s. The contents of this super-component for the 5 s moving window of time is presented in Figure 35, whereas the super-component for the 10 s moving window of time is presented in Figure 36, and finally the super-component for the 15 s moving window of time is presented in Figure 37.

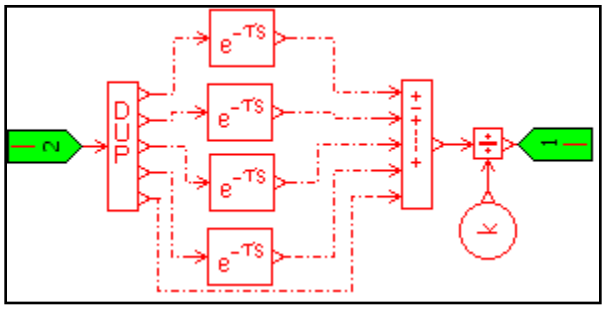


Figure 35: Super-component calculating diagnostic signal over a 5 *s* time window

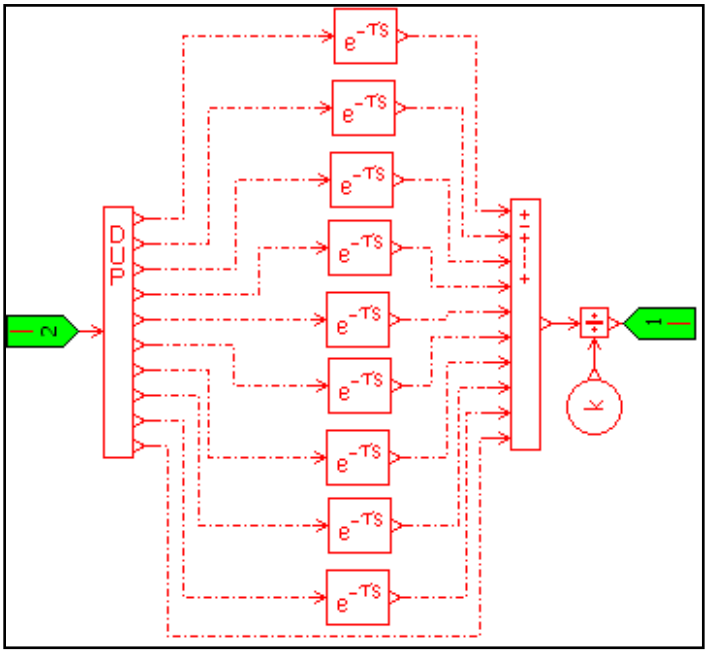


Figure 36: Super-component calculating diagnostic signal over a 10 *s* time window

Furthermore, the effect of the electromagnetic disturbance pulses on the calculation of the five diagnostic signals evaluated using a 5 s moving window of time is presented in Figure 38. Moreover, the effect of the disturbance pulses on the diagnostic signals utilizing a 10 s moving window is presented in Figure 39, and the effect of the disturbance pulses on the diagnostic signals utilizing a 15 s moving window is presented in Figure 40.

Figures 38, 39 and 40 prove that this diagnostic approach of (3) outperforms that of (2). However, it is convenient to mention here that most commercial PEMFC systems commonly have a minimum sampling time of 1 s. Therefore, basing the diagnostic signal on a 5 s window of time can prove to be impractical as seen in Figure 38. Furthermore, the diagnostic signals calculated based on both the 10 s and the 15 s moving windows of time proved to be effective and insensitive to electromagnetic disturbance pulses as seen from Figures 39 and 40. Nevertheless, using a 10 s moving window of time is preferred over the 15 s moving window of time in order to avoid delays in detecting faults as well as saving memory.

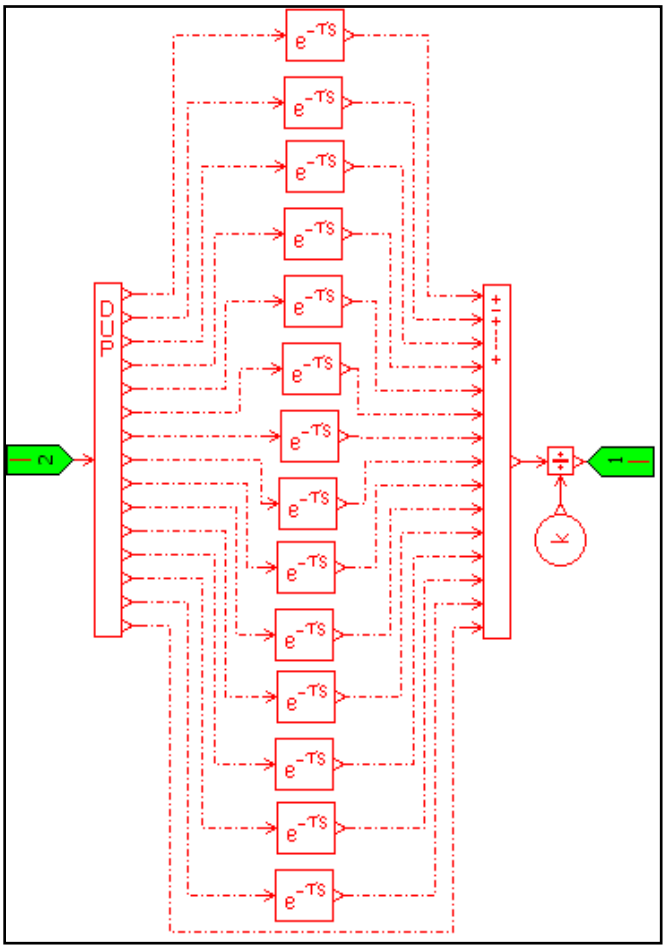


Figure 37: Super-component calculating diagnostic signal over a 15 *s* time window

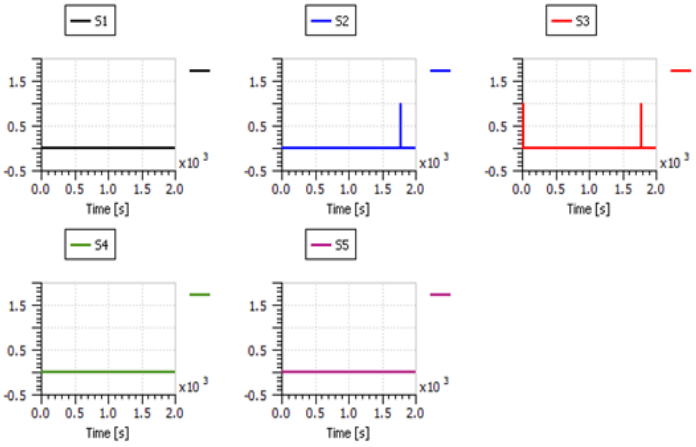


Figure 38: The effect of the disturbance pulses on the five diagnostic signals calculated using a 5 s moving window.

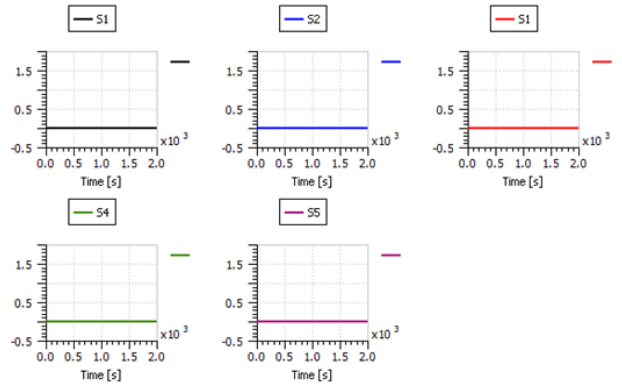


Figure 39: The effect of the disturbance pulses on the five diagnostic signals calculated using a 10 s moving window.

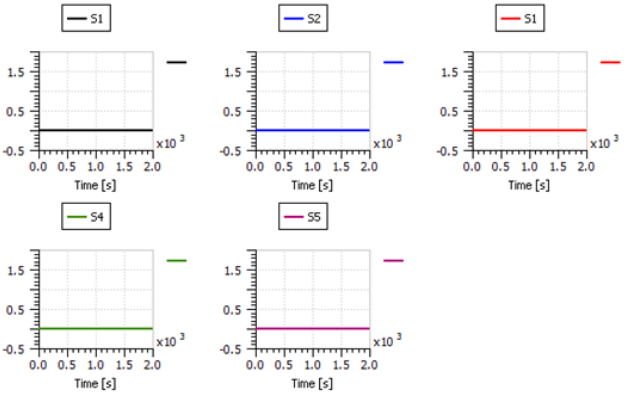


Figure 40: The effect of the disturbance pulses on the five diagnostic signals calculated using a 15 s moving window.

4.3. Fault diagnosis results

To test the proposed fault diagnosis scheme designed using diagnostic signals calculated over a 10 s moving window of time; the five system faults were induced at different times to evaluate the system's ability to detect and isolate the five system faults.

The first fault to be simulated was membrane drying. Thus, the relative humidity target of the humidifier was dropped to 10% and the effect of this action was immediately captured by the diagnostic signals S_2 and S_3 as seen in Figure 41, which – from Table 2 – clearly indicates the presence fault f_1 (drying of the membrane).

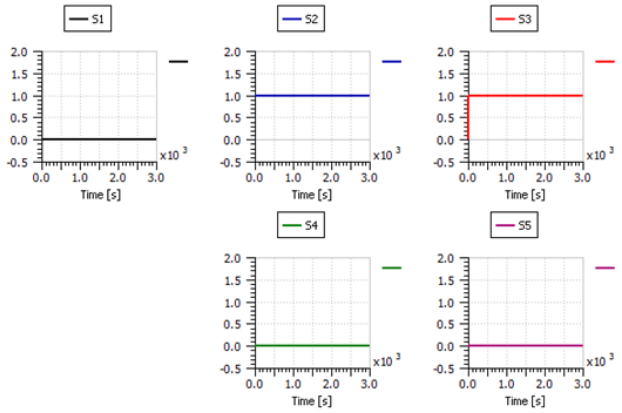


Figure 41: The diagnostic signals successfully detecting drying fault.

The second fault to be tested was membrane flooding. Therefore, the relative humidity target of the humidifier was raised to 100% while reducing the stack temperature to 25°C. Again, this action was clearly reflected on the diagnostic signals as seen from Figure 42, and the three diagnostic Signals S_2, S_3 and S_5 were triggered. However, it was noticed that the effect on the diagnostic signal S_5 was not as fast as that on diagnostic signals S_2 and S_3. This is expected since the increase in pressure drop needs some time to take effect. From Table 2, triggering S_2, S_3 and S_5 at the same time indicates the presence fault f_2 (flooding of the membrane).

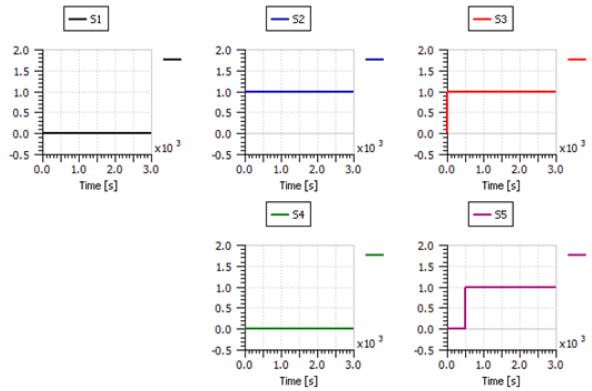


Figure 42: The diagnostic signals successfully detecting flooding fault.

The third fault to be tested was the air leakage in the supply manifold, which was simulated through the addition of a relief valve after the humidifier as shown in Figure 43. This relief valve was set to open with a flow rate of 9 g/s after 500 s of operation. Only one diagnostic signal was affected by this action at around 500 s as depicted in Figure 44. From Table 2, it can be deduced that triggering only S_5 indicates the occurrence of fault f_3 (air leakage).

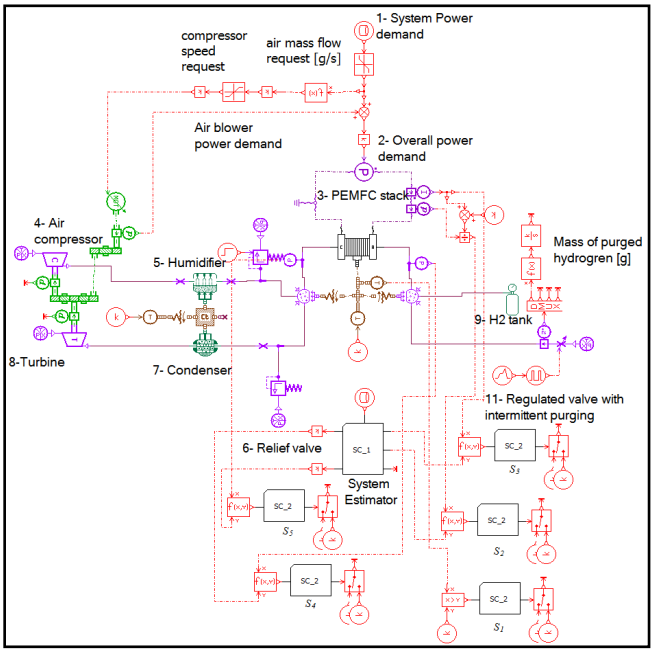


Figure 43: The modified diagnosis model with a relief valve to simulate air leakage.

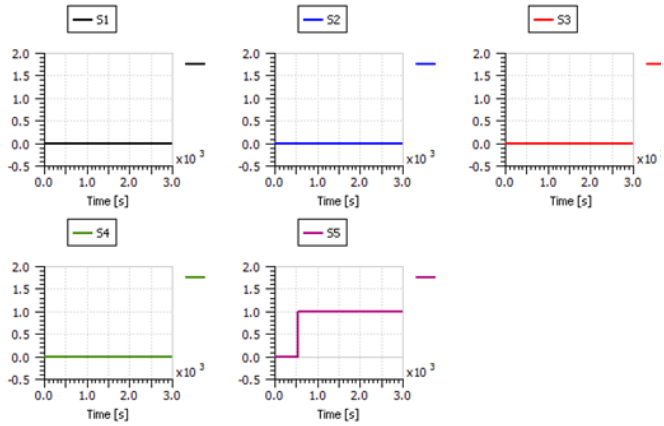


Figure 44: The diagnostic signals successfully detecting air leakage fault.

The next fault to be simulated was hydrogen leakage. As shown in Figure 45, the leakage was simulated through the addition of a relief valve right after the hydrogen supply canister. This valve was set to open with a flow rate of 8 g/s after 500 seconds of operation. The amount of hydrogen leakage can be seen in Figure 46.

From Figure 47, it is noticed that three diagnostic signals were triggered at around 500 s, which are S_2, S_3 and S_4. This clearly indicates the occurrence of fault f_4 (hydrogen leakage) as it can be deduced from Table 2.

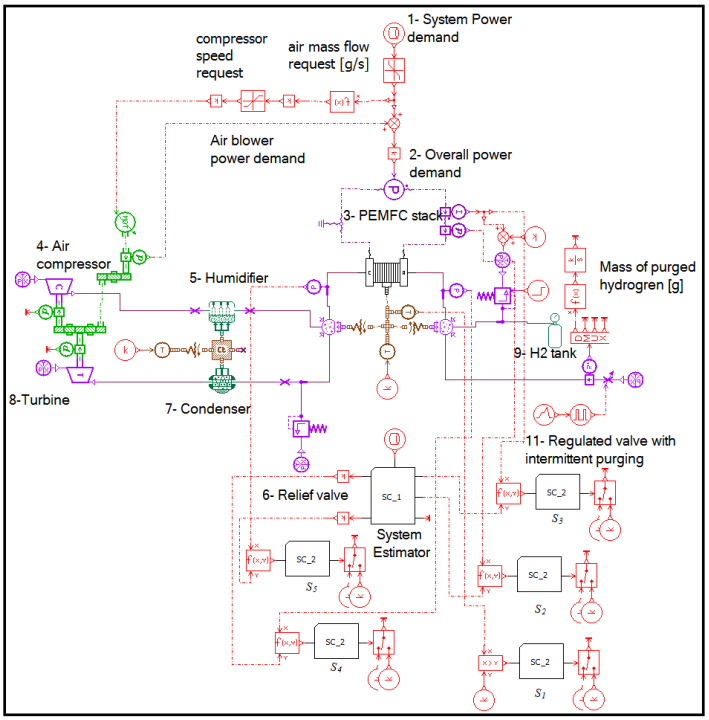


Figure 45: The modified diagnosis model with a relief valve to simulate hydrogen leakage.

Finally, the cooling system failure was tested by forcing the stack temperature to rise to 75°C after 1500 s of operation. The three diagnostic signals S_1, S_2 and S_3 were triggered by this action at around 1500 s of operation as shown in Figure 48. As it can be deduced from Table 2, and as previously explained in section III, S_1 alone is enough to flag fault f_5 (cooling failure) regardless of the effects on other diagnostic signals.

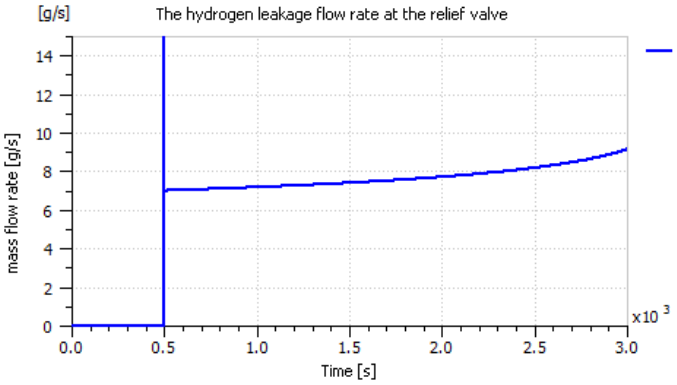


Figure 46: The hydrogen leakage flow rate at the relief valve.

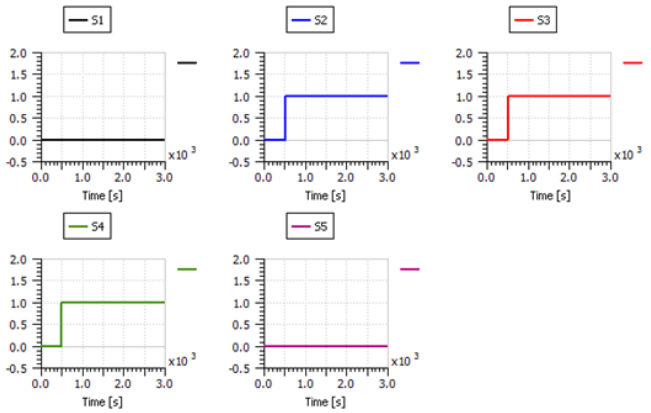


Figure 47: The diagnostic signals successfully detecting hydrogen leakage fault

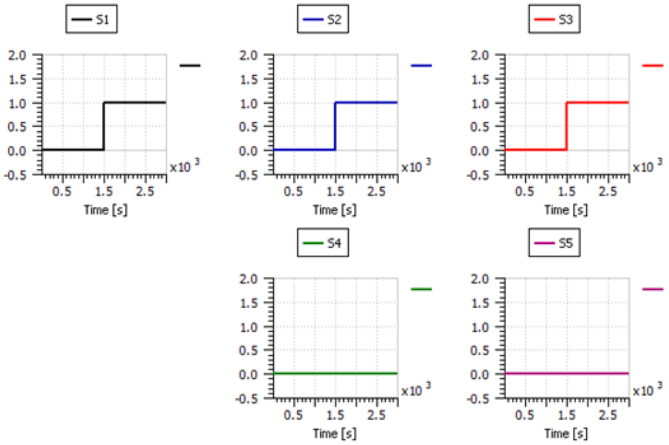


Figure 48: The diagnostic signals successfully detecting cooling system failure fault.

5. Conclusion

Fuel cells are extremely attractive clean power generation systems with the capability of someday replacing fossil fuels in the areas of power generation and transportation, while helping clean the environment by significantly lowering the world's pollution rates. However, to turn this green technology dream into reality, an accurate model that can effectively predict the fuel cell's performance in different conditions is desired. Such model can

then be used to study, simulate, and monitor the behavior of PEMFCs to detect any potential faults that can affect their performance.

Moreover, the complexity of the PEMFC model makes it very difficult and mathematically demanding to try and identify the modeling parameters. Furthermore, other limitations such as the absence of some parameters and confidentiality issues with the manufacturer can also limit the researchers' ability to develop an accurate fault diagnosis oriented model for a commercial PEMFC system. The Siemens software LMS AMESim 14.2 was used in this work as a solution to overcome such limitations.

A diagnosis oriented model of the ElectraGen™ PEMFC system was developed in LMS AMESim and five system faults (drying of the membrane, flooding of the membrane, air leakage, hydrogen leakage in the supply manifold, and cooling failure) were simulated to analyze their effect on different system parameters.

Diagnostic signals based on two different residual generation techniques were also assessed in this work, and the outperforming technique was implemented in the proposed diagnosis scheme. This diagnosis scheme was then tested in LMS AMESim against the five system faults under study and it was found to be very successful in both fault detection and discrimination.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

This paper is part of a joint research program funded by United Arab Emirates University (UAEU) and Japan Cooperation Center, Petroleum (JCCP).

References

[1] R. I. Salim, H. Noura and A. Fardoun, "Fault diagnosis of a commercial PEM fuel cell system using LMS AMESim," in 7th International Conference on Modeling, Simulation and Applied Optimization, Sharjah, UAE, April, 2017. DOI: 10.1109/ICMSAO.2017.7934890

[2] M. Nehrir and C. Wang, Modeling and Control of Fuel Cells: Distributed Generations Applications, IEEE Press Series on Power Engineering, Wiley, 2009.

[3] EG&G Services, Inc., Fuel Cell Handbook, 7th ed., Science Applications International Corporation, DOE, Office of Fossil Energy, National Energy Technology Laboratory, 2004.

[4] T. A. Zawodzinski, M. Neeman, L. O. Sillerud and S. Gottesfeld, "Determination of water diffusion coefficients in perfluorosulfonate ionomeric membranes," J. Phys. Chem., vol. 95, no. 15, p. 6040–6044, 1991. DOI: 10.1021/j100168a060

[5] T. Springer, T. Zawodzinski and S. Gottesfeld, "Polymer electrolyte fuel cell model," J. Electrochem. Soc., vol. 138, no. 8, p. 2334–2342, 1991. DOI: 10.1149/1.2085971

[6] P. Khanungkhid and P. Piumsomboon, "200W PEM fuel cell stack with online model-based monitoring system," Eng. J., vol. 18, no. 4, p. 13 – 26, Oct. 2014. https://doi.org/10.4186/ej.2014.18.4.13

[7] R. I. Salim, H. Noura and A. Fardoun, "A Review on fault diagnosis tools of the proton exchange membrane fuel cell," 2013 Conference on Control and Fault-Tolerant Systems (SysTol), pp. 686 - 693, Nice, France, Oct. 9-11, 2013. DOI: 10.1109/SysTol.2013.6693877

[8] J. Zhang, C. Song and J. Zhang, "Accelerated lifetime Testing for Proton Exchange Membrane Fuel Cells Using Extremely High Temperature and Unusually High Load," J. Fuel Cell Sci. Technol., vol. 8, no. 5, p. 051006, 2011. DOI: 10.1115/1.4003977

[9] C. A. Wilkie, J. R.Thomsen and M. L. Mittleman, "Interaction of poly (methyl methacrylate) and nafions," J. Appl. Polym. Sci., vol. 42, no. 4, p. 901–909, 1991. DOI:10.1002/app.1991.070420404

[10] S. R. Samms, S. Wasmus and R. F. Savinell, "Thermal stability of nafion in simulated fuel cell environments," J. Electrochem. Soc., vol. 143, no. 5, p. 1498–1504, 1996. DOI: 10.1149/1.1836669

[11] N. Ramaswamy, N. Hakim and S. Mukerjee, "Degradation mechanism study of perfluorinated proton exchange membrane under fuel cell operating conditions," Electrochimica Acta, vol. 53, no. 8, p. 3279–3295, 2008. https://doi.org/10.1016/j.electacta.2007.11.010

[12] J. Korbicz, J. Koscielny, Z. Kowalczuk and W. Cholewa, Fault Diagnosis. Models, Artificial Intelligence, Applications, Springer, 2004.

[13] J.-H. Wee, "Applications of proton exchange membrane fuel cell systems," Renewable and Sustainable Energy Reviews, vol. 11, no. 8, p. 1720–1738, 2007. https://doi.org/10.1016/j.rser.2006.01.005

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 310-316 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Parametric Study of Micro Strip Patch Antenna Using Different Feeding Techniques for Wireless and Medical Applications

Debajyoti Chatterjee, Anjan Kumar Kundu*

Department of Radio Physics and Electronics, University of Calcutta, Kolkata-700009, India

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords:
Micro Strip Patch Antenna
Feeding Mechanism
Return Loss
VSWR
Bandwidth
Dielectric Substrate

ABSTRACT

Due to the increasing bandwidth requirement of modern wireless communication systems, developing antenna having wider bandwidth have been receiving significant attention in the recent years. In this paper, a comparative analysis of the contacting feed (micro strip line and coaxial probe) and non-contacting feed mechanism (both Aperture Coupling and Proximity Coupling) in micro strip patch antenna has been done. In case of contacting feed, RF power is fed directly to the radiating patch using a connecting element such as a micro strip line whereas in case of non-contacting feeds, electromagnetic field coupling is done to transfer power between the micro strip line and the radiating patch. As per the latest research, ultra wide band technology is used in the frequency range from 3 GHz to 10 GHz. We have analyzed and compared the return loss and corresponding bandwidth of these four types of antenna at 5.853GHz so that this antenna may be used in medical as well as wireless applications.

1. Introduction

Micro strip patch antennas are considered as an indispensable tool in today's research oriented activities. The design and manufacturing cost of micro strip antenna is very cheap because of its 2D geometrical structure [1]. The patch antenna is a popular resonant antenna used for microwave wireless communications that require semispherical coverage. Some patch antennas avoid using a dielectric substrate and suspend a metal patch in the air above a ground plane using dielectric spacers; the resulting structure provides increased bandwidth. With an increase in frequency, the input impedance moves to the clockwise direction on the Smith chart [2]. Wireless and Medical applications requires small, low-cost, low profile antennas which has Omni directional radiation pattern in horizontal planes. Micro Strip patch antenna meets all requirements. Figure 1 shows a typical structure of a rectangular micro strip antenna.

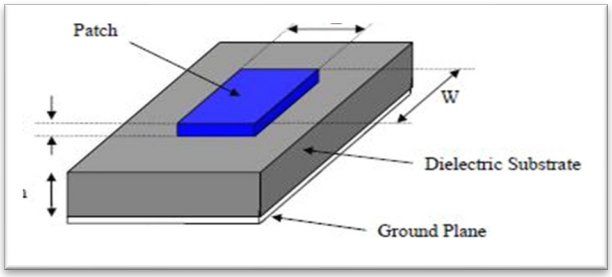


Fig 1: Micro Strip Patch Antenna

2. Feeding Mechanism

Suitable feeding technique plays an important role for antenna efficiency and better impedance matching. The feeding techniques used in the micro strip antenna are given below:

*Dr. Anjan Kumar Kundu, Institute of Radio Physics & Electronics,92, A.P.C Road, Kolkata-700009, +91-9874191584 (M), anjankumarkundu@gmail.com

2.1. Contacting Feed:

In this method, contacting element such as micro strip line or coaxial line is used to help the patch so that it can be fed directly to RF power. The most commonly used contacting feed methods are [3]:

1. Line Feed
2. Co-axial probe Feed

- **Micro Strip Line Feed:**

Micro strip line feed is one of the easier methods to fabricate as it is just a conducting strip connecting to the patch and therefore can be considered as extension of patch.[3-4] It is simple to model and easy to match by controlling the inset position. However the disadvantage of this method is that as substrate thickness increases, surface wave and spurious feed radiation increases which limit the bandwidth.

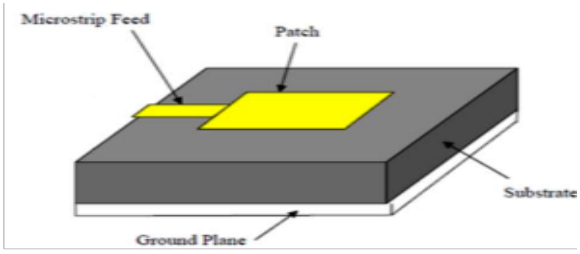


Fig 2: Micro strip Line feeding

- **Coaxial Probe Feed:**

Coaxial feeding is feeding method in which that the inner conductor of the coaxial is attached to the radiation patch of the antenna while the outer conductor is connected to the ground plane.

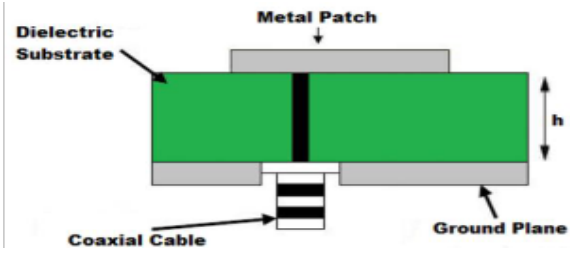


Fig 3: Coaxial Probe feeding

2.2. Non-Contacting Feed:

In this method, the RF power is transferred to the path from the feed line through electromagnetic coupling instead of feeding directly. The commonly used non-contacting feed methods are:

1. Aperture Coupled feed
2. Proximity Coupled feed

- **Aperture Coupling Feed:**

In the aperture coupled feed technique, the radiating patch and the micro strip feed line are separated by the ground plane. Coupling between the patch and the feed line is made through a slot or an aperture in the ground plane.

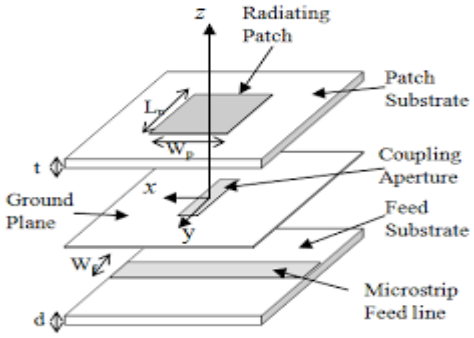


Fig 4: Aperture Coupled feeding

- **Proximity Coupling Feed:**

In this method, two dielectric substrates are placed such that the feed line lies in between the two substrates and the radiating patch is placed at the top of the upper substrate.

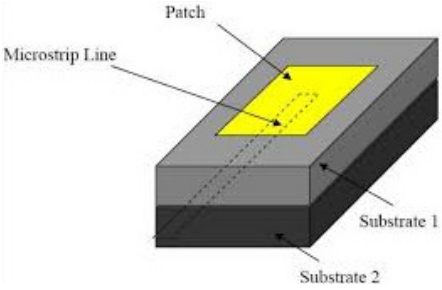


Fig 5: Proximity Coupled feeding

The two dielectric substrates can be selected independently to optimize both micro strip guided waves and patch radiating waves. The ranges of operating thickness of the substrate have a big effect on the resonant frequency and bandwidth of the antenna. Bandwidth of the micro strip antenna will increase with increasing substrate thickness. However, certain limits must not be exceeded, and otherwise the antenna will stop resonating. Therefore, [4], the measures for selecting a substrate may include the following:

(a) Surface-wave excitation.

(b) Dispersion of the dielectric constant and loss tangent of the substrate.

(c) Anisotropy in the substrate.

(d) Cost Effective

3. Study of Antenna Designing Parameters

There are three essential parameters for design of a rectangular micro strip Patch Antenna. Firstly, the resonant frequency (f_0) of the antenna must be selected appropriately. The frequency range for ultra wide band applications is from 3.1GHz to 10.6 GHz and the design antenna must be able to operate within this frequency range.

The second important parameter of good antenna design is dielectric substrate ($\in_r$). A thick dielectric substrate having low dielectric constant is desirable. This provides better efficiency, larger bandwidth and better radiation. FR-4 Epoxy which has a dielectric constant of 4.4 for lower substrate and RT –Duroid for upper substrate having dielectric constant of 2.2 and loss tangent equal to 0.009 can be used for new antenna design [5-6]. The other antenna parameters to be considered for design are length of the patch L, width W, height of dielectric substrate h and Loss Tangent. The patch length is considered around $L_g/2$ (L_g - Length of the ground plane) to initiate the radiation. The antenna is typically fed at the diverging edge on the dimension W because it offers sensible Polarization. The antenna parameters can be calculated by the transmission line method as exemplified below [7]:

3.1 Width of the Patch (mm):

Having specified the height of the patch antenna, the first step in the design procedure is to determine the width of the patch. This can be calculated using the following equation:

$$W = \frac{c}{2f_0\sqrt{\in r + 1}/2} \quad (1)$$

Where,

c = Speed of light in free-space (3×10^8 m/s)

f_0= Resonating frequency

3.2 Calculation of Effective Dielectric Constant (ϵ_{reff}):

This is calculated using the following equation:

$$\epsilon_{reff} = (\in r + 1)/2 + (\in r - 1)/2\left[1 + 12\ \frac{h}{w}\right]^{-0.5} \quad (2)$$

Where, h= height of the patch (mm)

W= width of the patch (mm)

3.3 Calculation of Effective Length (L_{eff}):

This is calculated using the following equation:

$$L_{eff} = \frac{c}{2f_0\sqrt{\in reff}} \quad (3)$$

3.4 Calculation of Length Extension (Δ_L):

The length extension is calculated using the following equation;

$$\Delta L = \frac{0.412h\ (\in reff + 0.3)(w/h + 0.264)}{(\in reff - 0.258)(w/h + 0.8)} \quad (4)$$

Where,

L= Patch length extension (mm)

h = height (mm)

W = width of the patch (mm)

3.5 Calculation of Actual Length:

The actual length of the patch antenna is calculated using the following equation;

$$L = Leff - 2\Delta L \quad (5)$$

4. Design and Simulation of Micro Strip Patch Antenna for Different Feeding Methods

4.1. Micro Strip Line Feed:

The antenna is designed at resonating frequency 5.853 GHz of WLAN. It is designed using transmission line model [8]. This section describes the design of rectangular patch antenna satisfying the given specifications:

Table 1: Design specification of Aperture Coupled Patch

Frequency	5.853 GHz
Antenna Dimension	29mm x 41mm
Dielectric Constant (RT Duroid)	2.2

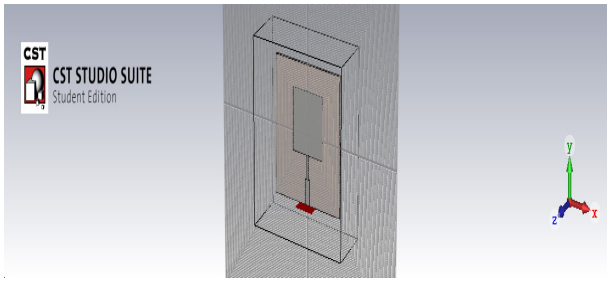


Fig 6: Designed Structures on CST Studio

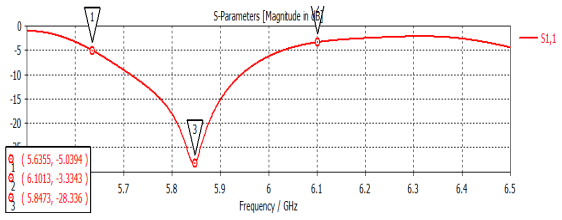


Fig 7: Return Loss S_{11} of simulated antenna at 5.853 GHz

Return Loss:

Fig 7 shows the S11 parameters (Return Loss = - 28dB) for the proposed antenna. The designed antenna resonates at 5.853 GHz.

Smith Chart:

The Smith Chart plot (Fig. 8) represents that how the antenna impedance varies with frequency and gives impedance of 50.4 ohms. For proper matching, the locus must be large so that it connects with the center point of the smith chart.

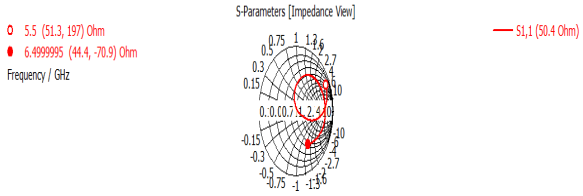


Fig 8: Smith Chart of simulated antenna at 5.853 GHz

Calculation of VSWR:

Fig. 9 shows the VSWR plot against frequency that numerically describes how well the antenna matches with the transmission line it is connected to.

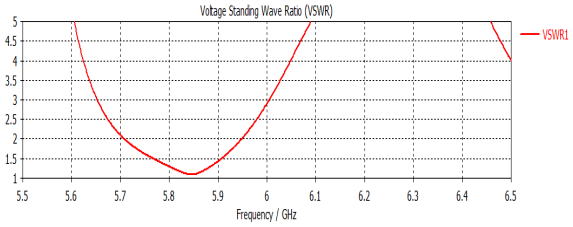


Fig 9: VSWR of simulated antenna at 5.853 GHz

4.2. Coaxial Probe Feeding:

The antenna is designed at resonating frequency 5.853 GHz of WLAN. It is designed using transmission line model. This section describes the design of rectangular patch antenna satisfying the given specifications:

Return Loss:

Fig 11 shows the S_{11} parameters (Return Loss = -37dB) for the proposed antenna. The designed antenna resonates at 5.853 GHz.

Table 2: Design specification of Proximity Coupled Patch

Frequency	5.853 GHz
Antenna Dimension	29mm x 41mm
Dielectric Constant (RT Duroid)	2.2

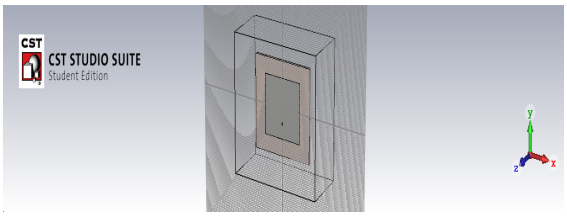


Fig 10: Designed Structure on CST Studio

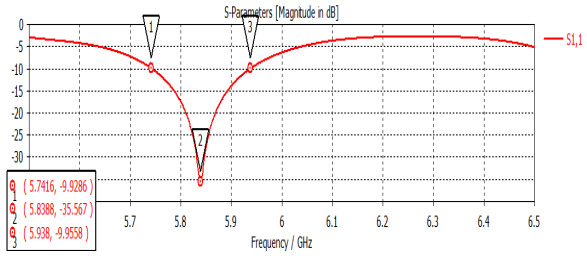


Fig 11: Return Loss S_{11} of simulated antenna at 5.853 GHz

Smith Chart:

The Smith Chart plot (Fig. 12) represents that how the antenna impedance varies with frequency and gives impedance of 50.21 ohms. For proper matching, the locus must be large enough that it passes through the center of the smith chart.

As it can be seen from Fig. 12, the circle cuts the resistive part at 0.5021, thus matching at 50.21 ohm.

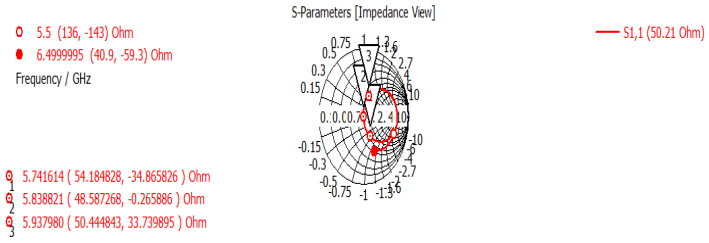


Fig 12: Smith Chart of simulated antenna at 5.853 GHz

Calculation of VSWR:

Fig. 13 shows the VSWR plot against frequency that numerically describes how well the antenna matches with the transmission line it is connected to.

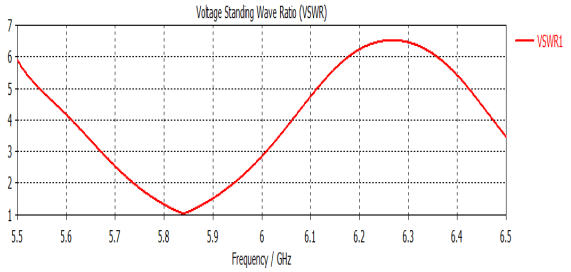


Fig 13: VSWR of simulated antenna at 5.853 GHz

4.3. Aperture Coupled Feeding:

The antenna is designed at resonating frequency 5.853 GHz of WLAN. It is designed using transmission line model. This section describes the design of rectangular patch antenna satisfying the given specifications:

Table 3: Design specification of Aperture Coupled Patch

Frequency	5.853 GHz
Antenna Dimension	29mm x 41mm
Lower Dielectric Constant (FR4)	4.4
Upper Dielectric Constant (RT Duroid)	2.2

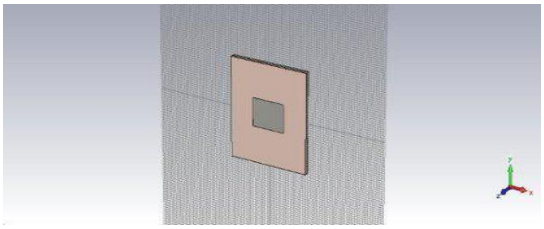


Fig 14: Designed Structures on CST Studio

Return Loss:

Fig 15 shows the S_{11} parameters (Return Loss = - 33dB) for the proposed antenna. The designed antenna resonates at 5.853 GHz.

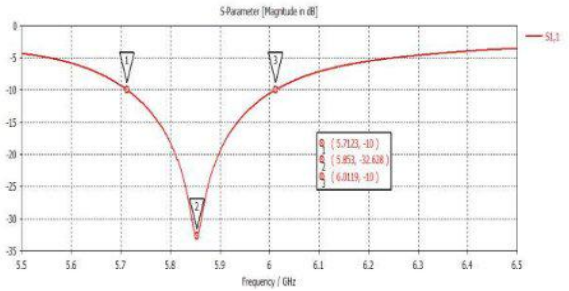


Fig 15: Return Loss S_{11} of simulated antenna at 5.853 GHz

Smith Chart:

The Smith Chart plot (Fig. 16) represents that how the antenna impedance varies with frequency and gives impedance of 49.93 ohms. For proper matching, the locus must be large so that it connects with the center point of the smith chart.

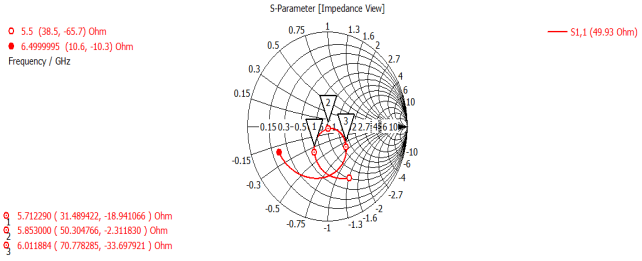


Fig 16: Smith Chart of simulated antenna at 5.853 GHz

Calculation of VSWR:

Fig. 17 shows the VSWR plot against frequency that numerically describes how well the antenna matches with the transmission line it is connected to.

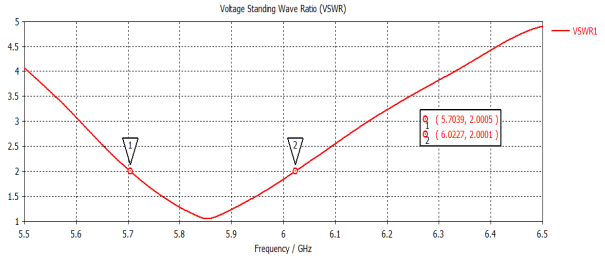


Fig 17: VSWR of simulated antenna at 5.853 GHz

4.4. Proximity Coupled feeding:

The antenna is designed at resonating frequency 5.853 GHz of WLAN. It is designed using transmission line model. This section describes the design of rectangular patch antenna satisfying the given specifications:

Table 4: Design specification of Proximity Coupled Patch

Frequency	5.853 GHz
Antenna Dimension	29mm x 41mm
Lower Dielectric Constant (FR4)	4.4
Upper Dielectric Constant (RT Duroid)	2.2

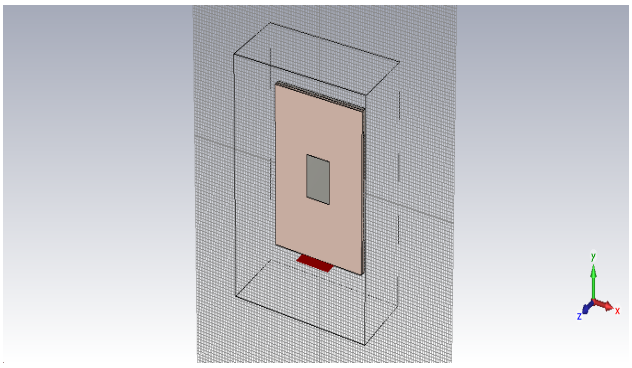


Fig 18: Designed Structure on CST Studio

Return Loss:

Fig 19 shows the S_{11} parameters (Return Loss = -47dB) for the proposed antenna. The designed antenna resonates at 5.853 GHz. More negative the return loss, higher the directivity and gain of the proposed antenna in particular direction.

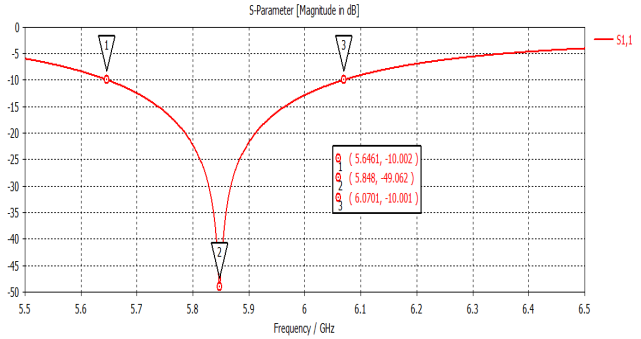


Fig 19: Return Loss S_{11} of simulated antenna at 5.853 GHz

1) Smith Chart:

The Smith Chart plot (Fig. 19) represents that how the antenna impedance varies with frequency and gives impedance of 49.97 ohms. For proper matching, the locus must be large enough that it passes through the center of the smith chart.

As it can be seen from Fig. 19, the circle cuts the resistive part at 0.4997, thus matching at 49.97 ohm.

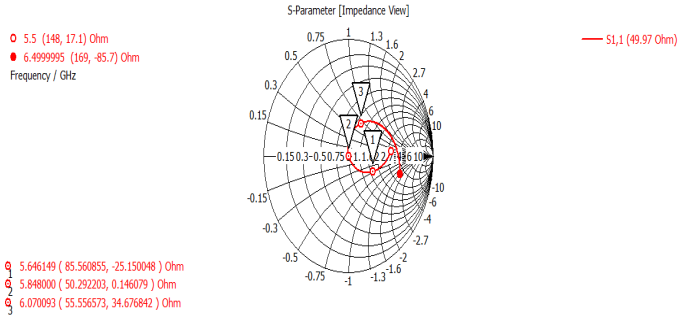


Fig 6.3: Smith Chart of simulated antenna at 5.853 GHz

3) Calculation of VSWR:

Fig. 20 shows the VSWR plot against frequency that numerically describes how well the antenna matches with the transmission line it is connected to.

5. Comparative Study of Important Parameters

Characteristics	Line	Coaxial	Aperture	Proximity
Return Loss (S_{11}) (dB)	-28	-37	-33	-47
Bandwidth (MHz)	250	210	320	475
VSWR	1.08	1.13	1.04	1.16
Impedance (ohm)	50.4	50.21	49.93	49.97

Table E : Comparison of Different Feeding Techniques

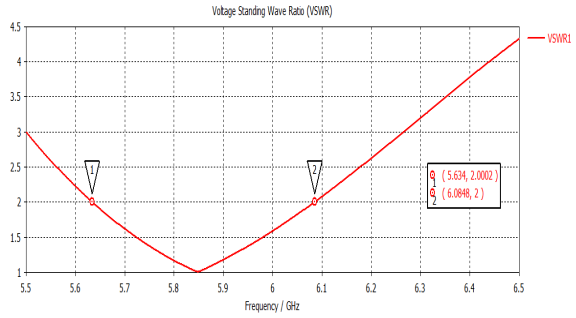


Fig 20: VSWR of simulated antenna at 5.853 GHz

Conclusion

Finally, the optimum result of all four feeding techniques of rectangular patch antenna on FR4 and RT-duroid substrate for Wi-max applications has been investigated. A comparison is made between feeding techniques [9-10] in terms of bandwidth, return loss, VSWR and impedance matching. So, we can see that selection of the feeding technique for a micro strip patch antenna is an important decision because it affects the bandwidth and other parameters also. A micro strip patch antenna excited by different excitation techniques gives different bandwidth, different gain, different efficiency etc. The maximum bandwidth can be achieved by proximity coupling. Proximity coupling gives the best impedance matching and radiation efficiency. Coaxial feeding technique gives the least bandwidth. We can also conclude that by changing the feed point where matching is perfect, the high return loss can be achieved at the resonant frequency. Various micro strip patch antennas with each different feeding technique are presented. The various parameters like return loss, smith chart, and VSWR are plotted for each antenna.

Conflict of Interest

The authors declare no conflict of interest.

Future Scope of Work

The designed patch antenna is intended for wireless applications. This paper mainly deals with the simulation of its characteristics for different feeding techniques which must have to be verified with the fabricated one. The size of this antenna ultimately restricts its usage mainly in biotelemetry in the field of medical applications. So, to extend its usage as an implantable one, its reduced size with similar characteristics has to be investigated. Using bio-compatible ceramic resins instead of conventional one to reduce its size has a great future scope of this work. Also, this design has to be compared with other miniaturized antenna techniques to ensure proper selection of implantable body antenna.

Acknowledgment

The authors of this paper would like to thank University Grant Commission (UGC) for giving the opportunity to work in

Modern Biology Group B: “Image and Imaging” under UPE-II. We also acknowledge all the contribution of corresponding IEEE authors and most importantly the publishers of related books and journals which gave immense support and inspiration in preparing this manuscript. Above all, the extreme mental support and source of inspiration from all the family members and friends are widely acknowledged.

References

[1] C.A.Balanis, Antenna Theory Analysis And Design,Second Edition,John Wiley & Sons. Ramesh Garg,Prakash Bhartie,inder Bahl,Apisak Illipiboon ,Microstrip Antenna Design Handbook, pp.1-68,253-316 Artec House Inc.Norwood,MA

[2] Ramesh Garg,Prakash Bhartie,inder Bahl,Apisak Illipiboon, Microstrip Antenna Design Handbook,pp.168,253-316 Artec House Inc.Norwood.

[3] Amit kumar Jaspreet kaur Rajinder singh,(2013), Performance analysis of different feeding technique,vol 3 issue 3.

[4] K. Praveen Kumar, K. Sanjeeva Rao, T. Sumanth, N. Mohana Rao, R. Anil Kumar, Y.Harish,(2013) Effect of Feeding Techniques on the Radiation Characteristics of Patch Antenna: Design and Analysis International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 2.

[5] Devan Bhalla And Krishan Bansal,(2013) Design of a Rectangular Microstrip Patch Antenna Using Inset Feed Technique IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: 2278-2834,p- ISSN: 22788735. Volume 7, Issue 4 PP 08-1

[6] Brajlata Chauhan, Sandeep vijay, S C Gupta(2013) Comparative analysis of Microstrip Patch Antenna using different substrate and observe effect of changing parameter at 5.4 GHz, Conference on Advances in Communication and Control Systems

[7] Fouzi Harrou, Abdelwahab Tassadit (2010), Analysis and Synthesis of Rectangular Microstrip Antenna ,Journal of Modelling and Simulation of Systems vol 1Issue 1 pp. 34-39.

[8] Rajesh Kumar Vishwakarma, Sanjay Tiwari,(2011) Aperture Coupled Microstrip Antenna for Dual-Band ,Wireless Engineering and Technology vol 2,93-101

[9] John R. Ojha Marc Peters and Igor Mini,(2010) Patch Antennas and Microstrip Lines, microwave and millimeter wavetechnologies modern uwb antennas and equipment ISBN: 978-953-7619-67-1.

[10] Hemant Kumar Varshney, Mukesh Kumar, A.K.Jaiswal, Rohini Saxena and Anil Kumar (2014) Design Characterization of Rectangular Microstrip Patch Antenna for Wi-Fi Application, Vol.4, No.2, E-ISSN 2277 – 4106, P-ISSN 2347 – 5161.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 317-321 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Simulation and FPGA Implementation of a Ring Oscillator Sensor for Complex System Design

Aziz Oukaira[*,1], Idir Mellal[1], Ouafaa Ettahri[1], Mohamed Tabaa[2], Ahmed Lakhssassi[1]

[1]*University of Quebec in Outaouais, Computer Engineering Department, Gatineau, (PQ), 18X 3X7, Canada*

[2]*Moroccan School of Engineering Sciences, LPRI Laboratory, Casablanca, 20250, Morocco*

ARTICLEINFO

Article history:
Received: 10 November 2017
Accepted: 22 January 2018
Online: 30 January 2018

Keywords:
RO
FPGA
VHDL
Complex System Design
DE1

ABSTRACT

This paper, presents the design of a temperature sensor based on RO (Ring Oscillator) in order to make a thermal study for the detection and localization of thermal peaks in a complex system. In this work, a simulation and FPGA implementation of a fully digital temperature sensor features a number of exact inverters that can be dynamically inserted. Before the transition to the implementation in FPGA board, the use of VHDL code is necessary to describe the exact number of inverters that form a single ring oscillator, in order to verify and validate the results obtained. This paper offers a solution to thermally induced stress and local overheating in complex system design which has been a major concern for the designers during the design of integrated circuit. In this paper a DE1 FPGA board cyclone V family 5CSEMA5F31C6 is used for the implementation.

1. Introduction

Decreasing feature sizes and increasing power and package contact densities are making thermal issues extremely important in complex system design. The main purpose of using a temperature sensor is to provide thermal monitoring for predicting local overheating or thermally induced stress. In this article, an intelligent sensor is used for thermal monitoring which is almost an ideal sensor due to its low power consumption. The main idea of this work is to simulate and implement a fully digital temperature sensor that can be dynamically inserted, operated and eliminated from the circuit once the test is done.

The intelligent sensor used in the design is actually based on ring oscillator. One of the important questions in the field of thermal issues of VLSI systems and micro-systems is how to perform the thermal monitoring, in order to indicate the overheating situations, without complicated control circuits.

Traditional approach consists of placement of many sensors everywhere on the chip, and then their output can be read simultaneously and compared with the reference voltage recognized as the overheating level. These techniques, though helpful to reduce the overall power consumption, may cause significant on-chip thermal gradients and local hot spots due to different clock/power gating activities and varying voltage scaling. It has been reported in [1] that temperature variations of 30 °C can occur in a high-performance microprocessor design.

The magnitude of thermal gradients and associated thermo-mechanical stress is expected to increase further as VLSI complex designs move into nanometer processes and multi-GHz frequencies. After the actual temperature is read, the value will be returned through the configuration ports of the FPGA. Then, the sensor will be removed from the chip. This type of oscillator used here is based on the switching time of an inverter. By connecting an odd number of inverters, a naturally oscillating signal is obtained at the output of each inverter of the chain. The oscillation frequency is directly related to the number of inverters. The more inverters there are, the lower the frequency is. To control the oscillation, an inverter can be replaced by a NAND gate, in order to stop or activate the oscillating system. Reference [2], [3], [4], [5] and [6] presents various techniques for varying the frequency of oscillation of the structure. These suggestions based on the voltage control of the delay generated by each cell, on the current control of the rise and falling time of each cell or on the controlled variation of the number of cells.

The advantage of this structure lies in the fact that it can be implemented both with an even number and an odd number of cells. This offers a greater tolerance on the constellation in phase of the signals available at output, but in the case of an implementation with an odd number of cells gives a sinusoidal signal in the output which makes the ring oscillator the perfect solution used ever to

*Corresponding Author: Aziz Oukaira; Email: ouka02@uqo.ca

give more information during the thermal monitoring. In fact, a ring oscillator consists of a feedback loop that includes a necessary odd number of inverters to produce the displacement of the phase which maintains the oscillation the total period is twice the sum of the delays of all the elements that make up the loop. Reversals can be made using the LUT (look-up table) of the configurable logic blocks (CLBs) or the programmable inverters included in the FPGA blocks. In any case, it is useful to insert an external signal to open the loop, as well as an output register to prevent variations in the frequency due to different loads. Thus, the sensor can give instantaneous temperatures.

The future of detectors based on the ring oscillator method is to help designers of more complex integrated circuits to optimize the management of thermal dynamics on the chip [7], [8] and [9]. In this paper in particular, the sensor used allows the detection of thermal peaks. The interests of a sensor based on this method are multiple [10] and [11]. Among the interests of the use of this sensor based on ring oscillator method, is that it can be easily integrated on a chip and can be dynamically inserted or removed from the design at any time due to its small size.

The rest of our work is organized as follows. Section 2, gives a description of the methodology used. Section 3, shows the implementation of the ring oscillator on the FPGA board to verify and validate the results.

2. Description of the methodology used

The new methodology adopted is to validate the control temperature of complex microsystems on a chip based of five inverters forming a ring oscillator [12]. This methodology has given very encouraging results for thermal monitoring in more complex integrated circuits. These simulations and FPGA implementation will then be generalized in high density microsystems, for the development of an integrated thermo mechanical stress control unit using our proposal described in this paper. The proposed ring oscillator depends on the temperature and its frequency changes accordingly. At a given temperature, the oscillator will exhibit a fixed frequency of oscillation.

2.1. Material and geometry of the complex design in COMSOL

We have made the simulation in COMSOL tool appointment with different materials and hardware layers and well on their ranking on the semiconductor. In this figure, we clearly see the reconciliation of our complex circuit layers. The complex model contains 36 Radial Board (RB) in which each RB contains 12 adapter board detector module (ABDM) and each ABDM contains 2 ASICs. Power of 0.6 W is dissipated in a complex circuit of the ASIC (4.68 mm x 5.97 mm) [13]. This amount of power is applied for 9 seconds to visualize the evolution and distribution of heat around the ASIC. Figure 1 shows the ASIC and its support modeled in COMSOL, the ASIC transmits event data through low voltage differential signaling links. To solve the thermal diffusion equations, the Dirichlet boundary conditions (DBC) at 298.15 °K are applied around the daughter board.

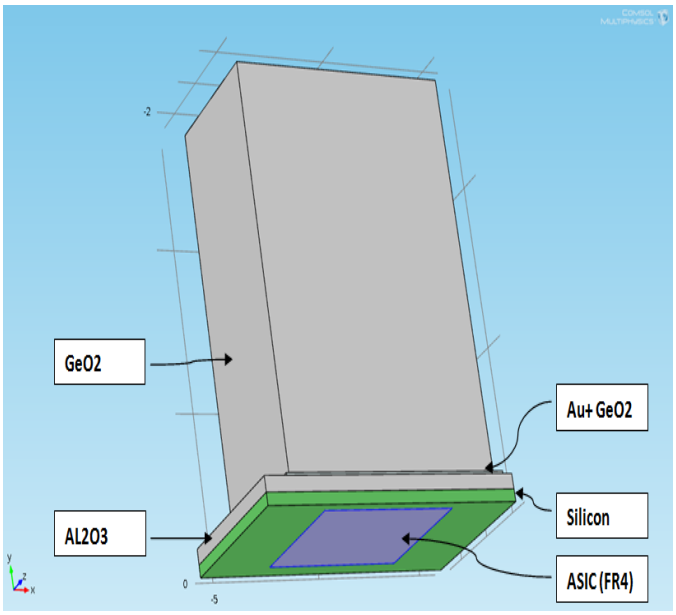


Figure 1 The ASIC modeled by the COMSOL tool.

This structure thus represents a continuous domain, by the method of finite elements consists first of a geometric discretization. The structure is subdivided into sub domains of the simple geometric form called finite element and defined not on the whole of the structure, but for each of its elements.

2.2. Thermal distribution of the complex design in COMSOL

It is very interesting to have a simulation environment that includes the ability to add different physical phenomena to the model studied. In this part, we will present the simulation results from heat sources represented the ASIC complex module in COMSOL tool. As you can see in Figure 2 shows the thermal behavior of our model complex design.

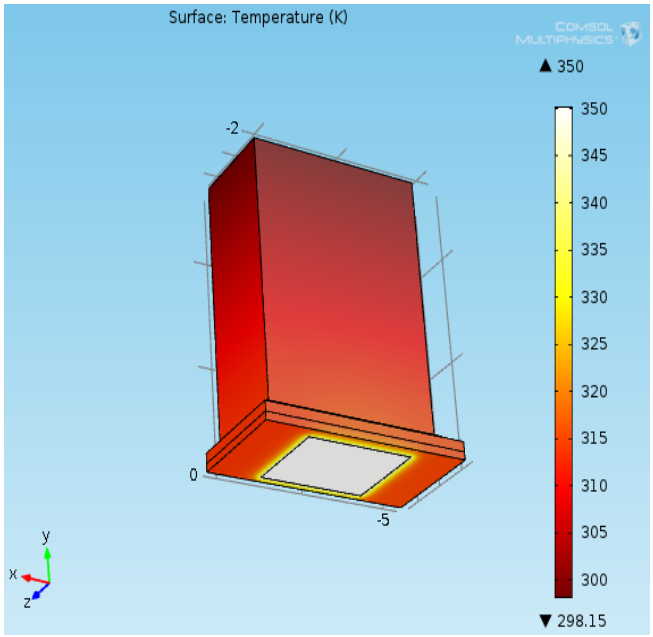


Figure 2 Thermal distribution of the ASIC in COMSOL tool.

This simulation of the ASIC complex module in COMSOL gives a good idea of the behavior and the thermal diffusion of heat sources in our complex system and shows the temperature up to 350 °K. The use of COMSOL tool gives us an idea about the thermal diffusion around the ASIC. According to another study done by [14] and the following Figure 3 presents the values of the temperature as a function of the frequencies of a ring oscillator.

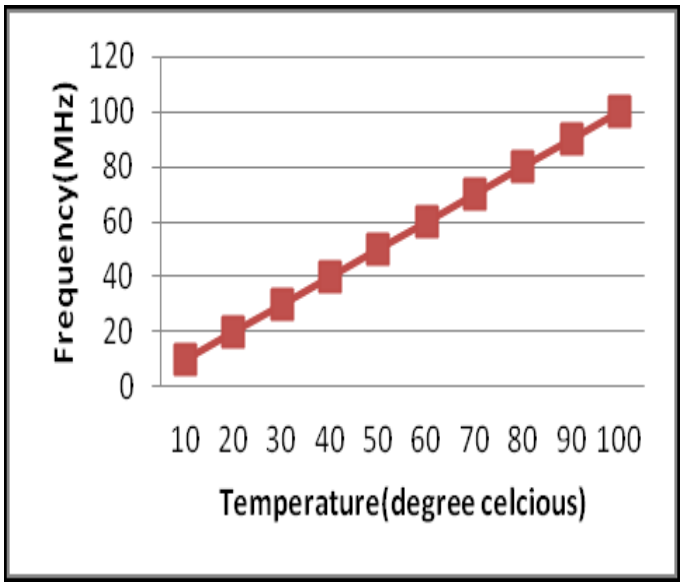


Figure 3 Température of (RO) ring oscillator showing frequency of oscillation linear to temperature.

this research which is done by [14] and the presentation of the results in ghraph form of integrated thermal sensors to monitor the temperature between 10 C to 100 C, according to our simulation under the tool COMSOL Fig. 3 and the study which is done by [12], we can deduce that temperature 77 ° C has our complex module corresponds to approximately 78 MHz (see Fig. 3). This perfectly corresponds to our simulation which is made for the ring oscillator with 5 inverters.

Now we need to validate these theoretical findings find by implementing a 5-inverter based sensor on an FPGA board that allows for different simulations, in [2] explains that more information can be found on the location of the thermal peak at the same frequency and the same temperature of the ring oscillators knowing that this type of ring oscillator sensor can only determine the necessary information if it receives sinusoidal signals, therefore a sinusoidal signal as an input.

3. Experimental implementation and results

The main purpose of this section is the implementation and validation of a single ring oscillator composed of five inverters. VHDL code is used to describe the RO module to facilitate the development of its architecture for its implementation in complex system design. This architecture is modeled in high-level language and simulated to assess its performance and finally implemented on FPGA. The simulation results are validated by using the software Modelsim under Quartus Prime, which allows simulating the behavior of the system in time. Our design flow will be divided into three main parts: simulation, synthesis, and implementation of the VHDL code on FPGA. A description of each part will be presented in the next paragraphs.

3.1. Creation and simulation of the VHDL code

This part, presents the description of the single ring oscillator based of five inverters using a VHDL code editor. The code editor used is Modelsim. Figure 4 shows the top-level module of a single ring oscillator.

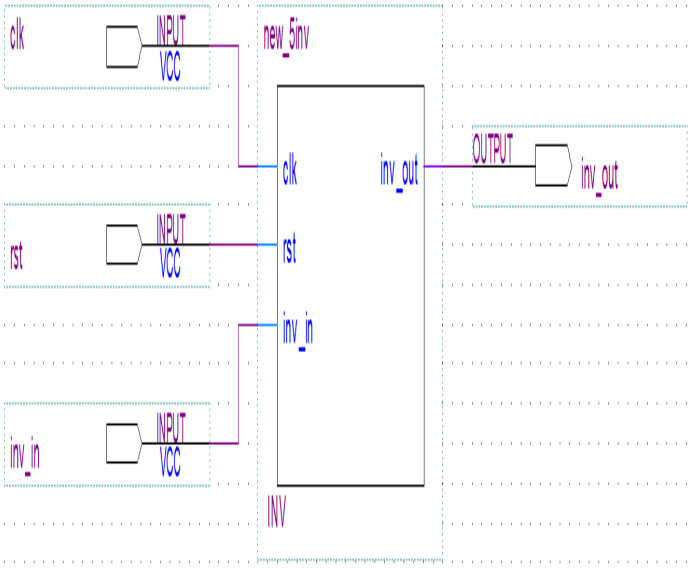


Figure 4 Top level of the ring oscillator composed of five inverters.

This structure in Figure 4 shows the top level of the ring oscillator composed of five inverters, this facilitates the simulation of the logic circuits thereafter. After having followed it, the modelization of a logic scheme of a ring oscillator based on five inverters with Quartus Prime and comes as the next step then the simulation step using ModelSim.

After generating the two .vhd files (the primary file system and the "Test Bench" file) with the "System Generator" the role of the Quartus Prime Navigator comes in order to synthesize the design and generate the RTL files as shown in Figure 5.

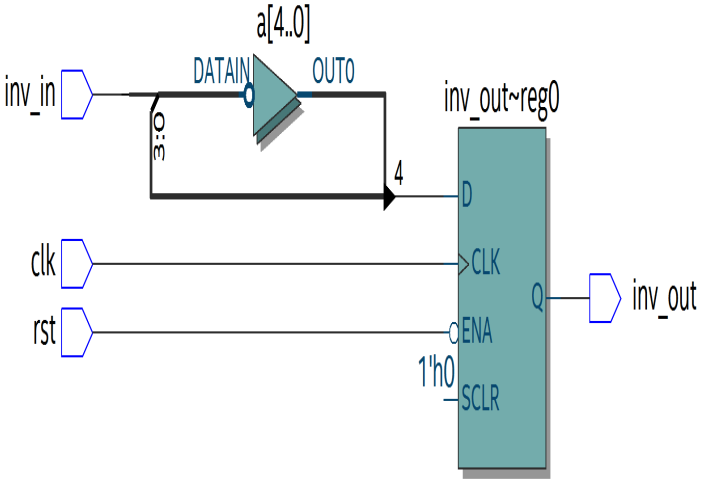


Figure 5 Structure of ring oscillator based five inverters in Quartus Prime tool.

The structure of the single ring oscillator based of five inverters, after synthesis with Quartus Prime from Altera is shown in (Figure

5). The VHDL code implemented was validated. In this part the simulation will be run with the same conditions used in [11], [12] and [14], to validate the experimental results. This Figure 6 shows the results found by the simulation using the Modelsim tool.

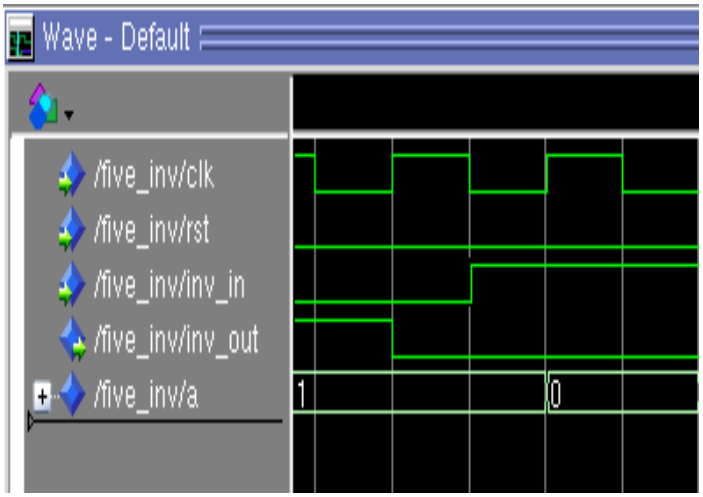


Figure 6 Display résulte of simulation the VHDL code.

As you can see in Figure 6, the signals of simulations are explaining the role of inverters for example for the logical value zero the output is the logical value 1 and the same thing for the second case. The results found validate the VHDL code since it is the correct operation of a single ring oscillator based on five inverters. In this paper, the GDS (gradient Direction Sensor) method for thermal peak detection is used, simulated and verified with a VHDL code and a 'test bench' at the laboratory LIMA the results found meet the initial specifications.

3.2. Implementation and downloading of the VHDL code

Once compiled after the assignment of the pins, the program is ready to be downloaded on the card DE1 cyclone V family and 5CSEMA5F31C6 as a device. This Figure 7 shows that the VHDL code is downloaded successfully on the card.

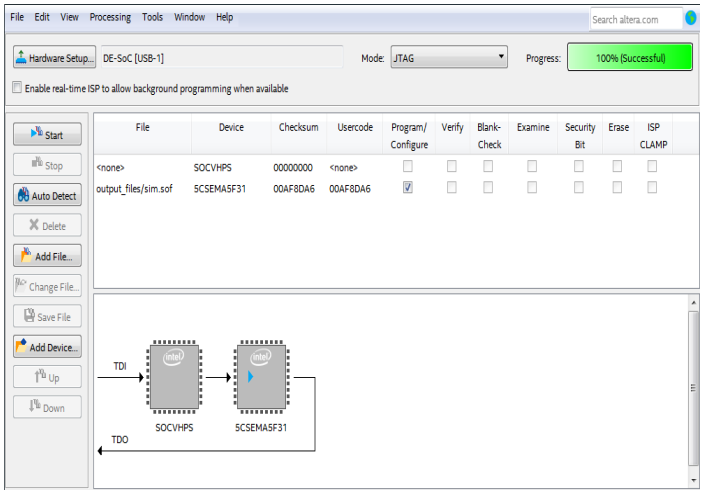


Figure 7 Downloading the code of DE1 Altera cyclone V.

After the download, the program was running and then the outputs were shown on the display of the card. The clock is at 50 MHz, so the outputs should change with a frequency of 50 MHz and the following Figure 8 shows the two values (1 and 0) after implementation on the LCD.

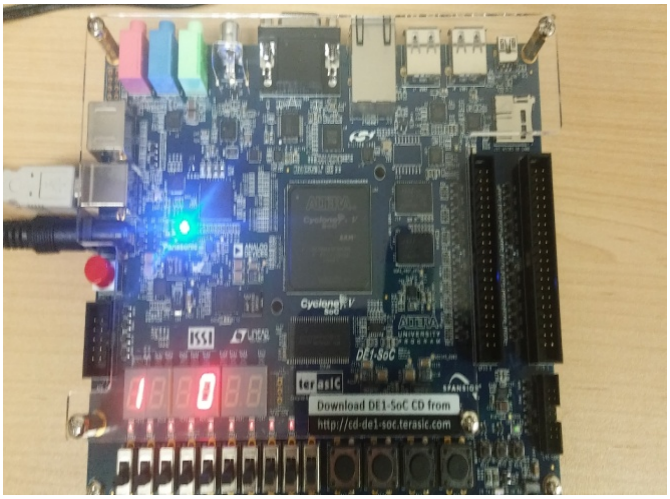


Figure 8 Validation of simulation results on the LCD after the implementation of DE1 Altera cyclone V.

This validation of the simulation results in Figure 8 shows that the value displayed on the LCD matches the results found before. The simulation and implementation on FPGA board DE1 can be applied in any kind of environment to get improved performance to control temperature of complex system design on a chip based of five inverters that will form a single ring oscillator with respect to the conventional schematic; it is also able to keep the temperature constant at the desired value regardless of changes in the load or environment. Thus, the overshooting problem can be solved up to great extent. One of the important issues in the field of electronics is overheating problems especially when it comes to integrated and complex systems and microsystems, but the mean question is how to perform thermal monitoring, to indicate overheating situations, without control.

This type of sensors all over the chip, and then their output can be shown simultaneously and be compared to the reference voltage recognized as the level of overheating. The idea of the proposed method is to validate the results predict the local temperature and gradient along the given distance in some places only on the monitored surface and evaluates obtained several real-time information in a short area in order to predict the temperature of the heat source. These peaks found are essential when monitoring the thermal matrix to avoid a critical induced thermo-mechanical stress. In addition, in most cases, overheating occurs in only one location.

4. Conclusion

The main objective of this paper is to simulate and implement a temperature sensor based on ring oscillator to make a thermal study at the junction. For this paper, we presented an experimental study for the implementation of a fully digital temperature to be dynamically inserted, operated and removed from the circuit after the test. Thus, the main advantage of this type of sensor is the analysis of the temperature during operation of the different blocks of a complex circuit implemented on FPGA. This will be useful for the integrated circuit designer because it offers a solution to thermally induced stress and local overheating in complex system

design which has been a major concern for the designers during the design of integrated circuits.

References

[1] P. Gronowski et al., "High performance microprocessor design," IEEE J. Solid-State Circuits, vol. 33, pp. 676– 686, 1998.

[2] Oukaira, A, Ettahri, O and Lakhssassi, A "Modeling and FPGA implementation of a thermal peak detection unit for complex system design", (IJACSA) International Journal of Advanced Computer Science and Applications, vol.8, no.6, pp.307-312, 2017.

[3] O. Slattery, D. O' Mahoney, E. Sheehan, and F. Waldron, "Sources of Variation in Piezoresistive Stress Sensor Measurements," IEEE transactions on components and packaging technologies, vol.27, no.1, pp.81-86, 2004.

[4] M. Banu, "100 khz – 1 ghz Nmos Variable-frequency Oscillator With Analog and Digital control," ISSCC, Solid-State Circuits Conference. Digest of Technical Papers, vol.31, pp.20-21, 1998.

[5] N. Retdian, , S. Takagi, and N. Fujii, "Voltage controlled ring oscillator with wide tuning range and fast voltage swing," IEEE Asia-Pacific Conference, ASIC Proceedings, pp.201-204, 2002.

[6] Oukaira, A, Pal, N, Ettahri, O, Kengne, E and Lakhssassi, A "Simulation and FPGA Implementation of Thermal Convection Equation for Complex System Design", (IREA) International Journal on Engineering Applications, vol.2, no.6, pp.307-312, 2016.

[7] Pinel, Stephane, et al, "Thermal modeling and management in ultrathin chip stack technology", IEEE Transactions on Components and Packaging Technologies, vol.25, no.2, pp.244–253, 2002.

[8] Oukaira, A, Lakhssassi, A, Fontaine, R, & Lecomte, R, "Thermal Model Development for LabPET II Scanner Adapter Board Detector Module", Proceedings of the COMSOL Conference, pp. 1-5, 2015.

[9] Oukaira, A, Fontaine, R, Lecomte, R, & Lakhssassi, A, "Thermal cooling system development for LabPET II scanners by forced convection flow", In New Circuits and Systems Conference (NEWCAS), 15th IEEE International, pp. 289-292, 2017.

[10] Rahmanikia, Navid, et al, "Performance evaluation metrics for ring-oscillator-based temperature sensors on FPGAs: A quality factor", Integration, the VLSI Journal, vol.57, pp.81–100, 2017.

[11] Rahmanikia, Navid, et al, "Exploring Efficiency of Ring Oscillator-Based Temperature Sensor Networks on FPGAs", Proceedings of ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp.264–264, 2015.

[12] Oukaira, A, Mellal, I, Ettahri, O, Kengne, E and Lakhssassi, A "Thermal Management and Monitoring Based on Embedded Ring Oscillator Network Sensors for Complex System Design", (IJCEIT) International Journal of Computer Engineering and Information Technology, vol.9, no.7, 2017.

[13] Oukaira, A, Taheri, S, Nour, M, & Lakhssassi, A, "Simulation and Validation of Thermal Stability for Complex System Design High Power Dissipation", In the 5th IEEE International Conference on Smart Energy Grid Engineering (SEGE), pp. 229-233, 2017.

[14] Suman, S, & Singh, B. P. "Ring oscillator-based CMOS temperature sensor design", International Journal of Scientific & Technology Research, vol.1, no.4, pp.76–81, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 322-328 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Dynamic Reallocation Based Window Access Scheme for Enhancing QoS of Vehicular Ad-hoc Networks (VANETs)

Md. Amirul Islam*, Hossen Asiful Mustafa

Institute of Information and Communication Technology,

Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh

ARTICLE INFO

Article history:
Received: 14 November, 2017
Accepted: 18 January, 2018
Online: 30 January, 2018

Keywords:
VANET
V2I communications
MAC
TDMA

ABSTRACT

This article proposes a new MAC scheme for Vehicle-to-Infrastructure (V2I) communications that dynamically reallocates unused TDMA slots. By maintaining a balanced waiting time, the proposed TDMA based scheduling scheme allocates TDMA slots in a rational way to minimize merging and one-hop neighboring collision. The proposed scheme ensures dynamic reallocation of unused slots by using "time slot reassignment" mechanism. The scheme has been simulated in VEINS framework of OMNET++ network simulator and its performance has been compared with other conventional protocols. Experimental results show that our scheme performs better than existing schemes in terms of successfully transmitted data packets.

1. Introduction

In recent years, Vehicular Ad-hoc Networks (VANETs) technology has drawn interest of many researchers as a common platform for inter vehicle communication on highways or in urban environments [1-5]. As a special type of Mobile Ad-hoc Network (MANET), VANET provides communications among vehicles and between vehicle, and infrastructure via Road Side Units (RSUs). The relevance of VANET has been confirmed by the development of a specific IEEE standard, 802.11p, to support VANETs [2].

Different from other ad-hoc networks, VANET has unique characteristics of high node mobility, dynamic topology changes and strict delay constraints. The applications devised for VANETs can be divided into the following three services: safety services, traffic management and user-oriented services. Among these services, safety services usually require bounded transmission delays as well as low access delays. As the IEEE 802.11p standard does not provide a reliable broadcast mechanism with bounded communication delay [2, 6], it is not sufficient for VANET applications which are primarily envisioned to improve road safety.

As opposed to IEEE 802.11p, Time Division Multiple Access (TDMA) is inherently a collision free scheme with bounded access delay. Several protocols have been proposed in VANETs using TDMA to provide fairness and to reduce interference among vehicles. By concatenating or rescheduling time slots based on access priority, TDMA based schemes can also assign bandwidth resources to different vehicles on-demand. However, for TDMA based Medium Access Control (MAC) protocol, an efficient slot management is very crucial to ensure fairness as well as reusability of unused slots. VANETs, which have very dynamic topology, are very much prone to allocating a time slot in overlapping areas. In this case, two types of collision may occur: access collision between vehicles trying to access the same available time slots, and merging collisions between vehicles using the same time slots. Some researchers worked on collision issues and some worked on fairness; however, a comprehensive protocol to ensure fairness and reallocation of unused slots is still missing.

In this work, we propose a new TDMA based MAC scheme, Dynamic Reallocation based Window Access MAC (DRWA-MAC), to achieve fair slot allocation as well as reallocation of unused slots dynamically. Our scheme allocates time slots to the registered vehicles evenly; then it monitors the usage of the assigned slots to ensure the dynamic reallocation of unused slots in a fair way. Simulation results in OMNET++ show that our proposed scheme outperforms several existing schemes. This work

*Corresponding Author: Md. Amirul Islam & Email: amirulctg@yahoo.com

is an extension of our previous work originally presented and published in ECCE conference [1].

2. Related Work

VeMAC [7, 8] is a TDMA based scheme proposed for inter-vehicle communication in which vehicles in opposite directions (Left, Right) and roadside units (RSUs) are assigned with time slots in the same TDMA time frame by using logical segmentation. Though VeMAC can make use of the seven DSRC channels and decrease the rates of merging, and access collisions, it is burdened with the overhead of the size of the transmission control frame and cannot ensure reusability of an unused slot.

In ACFM [9], each RSU maintains a dynamic slot assignment cycle for vehicles in its coverage adaptively. Under the scenario of light traffic, ACFM controls the excessive increase of unassigned slots by shrinking slots assignment cycle frame by frame. When there is a mass of vehicles on roads, ACFM provides more available slots by expanding cycle. However, to avoid interference between adjacent segments, ACFM requires two orthogonal frequencies to ensure that the same frequency is not used for a distance of two hops.

TC-MAC [6] is a TDMA Cluster-based MAC protocol for VANET. It subdivides slots of control channel (CCH) into mini-slots to broadcast beacons or safety messages. Here, in each frame, each vehicle gets allocation on a time slot in service channel (SCH) and competes for a mini-slot on the CCH. A vehicle uses its mini-slot to inform the other vehicles of its transmission during time slot on the SCH. Cluster members can use their time slots on the service channels to exchange non-safety data in unicast or multicast communication mode. TC-MAC protocol was designed for simple highway traffic in which all the vehicles are moving in the same direction and it has high collision in bidirectional traffic and in urban scenarios due to the merging collision problem. TC-MAC also suffers from inter-cluster interference problem when two or more clusters are in close proximity.

PTMAC [10] is a prediction-based TDMA MAC protocol that predicts encounter collisions and effectively reduces the number of collisions while maintaining high slot utilization. Here, immediate vehicles within a two-hop neighborhood detects a potential collision if two vehicles occupies the same slot and the protocol removes the collision by asking one of them to change its current slot. As PTMAC makes TDMA slot assignment on a contention based approach, it is unable to ensure fairness.

FAWAC-MAC [1] is a TDMA based scheme that uses capture effect to send extra data in free or unused slots. Here, all vehicles, which are not the owner of the current slot, check whether its own slot is within a predefined window size. If not, they transmit their packets in low power while the original owner of the slot transmits packets in high power. Using capture effect, the RSU can distinguish the power levels. In the presence of a high power packet, all low power packets get discarded; while in the absence, low power transmission succeeds. When vehicle density is very high, FAWAC-MAC has a chance of low power packet collision in free slots and if the coverage area is large, then it may also be suffered by near-far effect.

FAWAT-MAC [1] uses time slot reassignment mechanism to transmit data in unused slots. Here, if an RSU detects no ongoing transmission within one fourth of the beginning time of a slot, it selects the slot as unused and invites another vehicle to transmit data in that slot. Upon reception of the invitation packet, the specific vehicle transmits data for the remaining time of the slot in a collision free manner. The slot allocation process of FAWAT-MAC can cause merging and one-hop neighboring collision for boundary vehicles. In this protocol, the waiting time for some of the high speed boundary vehicles can be so high that by the time they got their assigned slot, they may have left the corresponding RSU.

3. Proposed System

In this work, we propose Dynamic Reallocation based Window Access MAC (DRWA-MAC), an extension of the FAWAT-MAC protocol [1]. Our proposed protocol operates as a TDMA based MAC protocol in a centralized topology where RSUs control distribution and reallocation of slots.

3.1. Assumptions

We made the following assumptions regarding VANET's context in which DRWA-MAC operates:

- Each vehicle and RSU has a unique ID.
- Vehicles and RSUs are equipped with GPS and are perfectly time synchronized.
- At the beginning of each frame, the RSU transmits registration beacon and then, each vehicle, which receives that message, broadcasts a registration response message with its own ID, location and data size. For a certain amount of time, the whole registration response system is a contention based process.
- Each newly joined vehicle that does not have a slot and wants to get a new slot shall listen to the channel for one frame. Then, after receiving the registration beacon, they can register to the RSU.

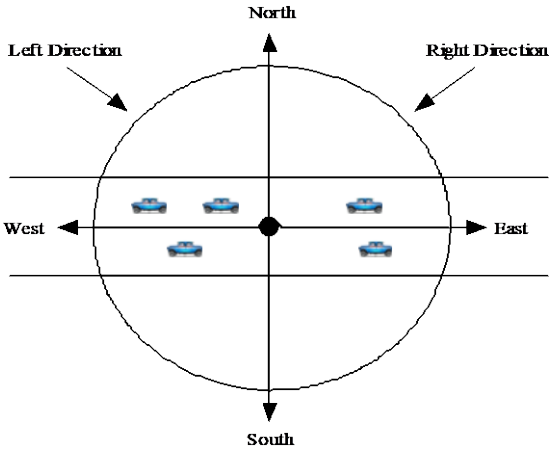


Figure 1: Right and left direction of vehicle movement as defined in the protocol

3.2. System Design

In our proposed DRWA-MAC protocol, we consider RSU based centralized VANET with a set of vehicles moving in lanes consisting of two-way traffic. In a two-way traffic system, we

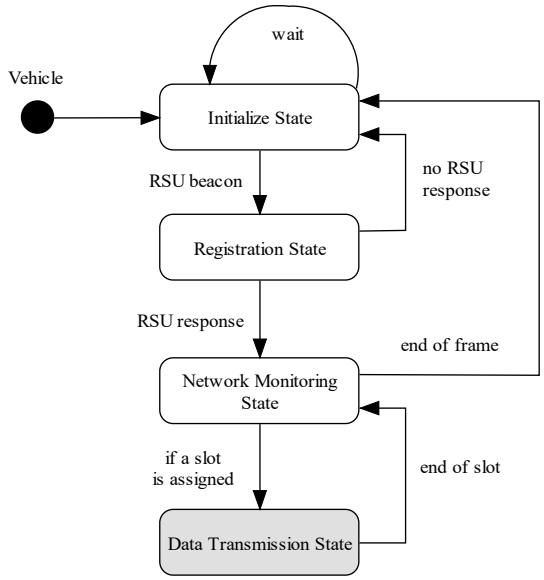


Figure 2: Vehicle state diagram in our scheme where a vehicle can be in one of the four states: initialize, registration, network monitoring, and data transmission

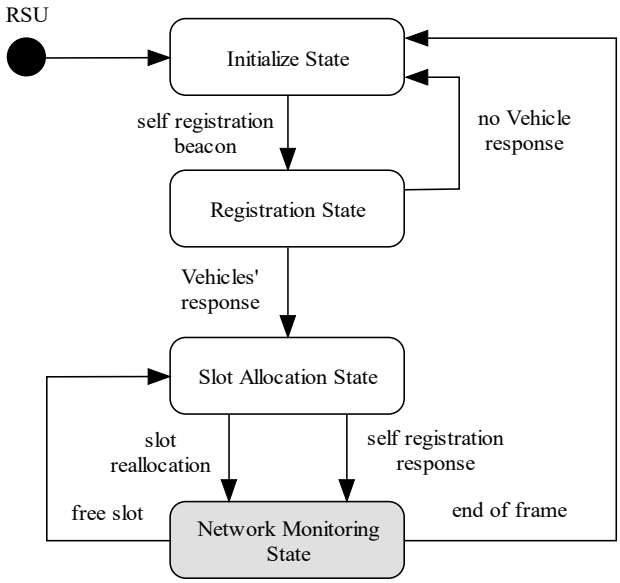


Figure 3: RSU state diagram in our scheme where the RSU can be in one of the four states: initialize, registration, slot allocation, and network monitoring

define a vehicle as positioned in left direction of a RSU's broadcasting area if it is placed in the west half and right direction, if it is placed in the east half, as shown in Figure 1.

Our scheme consists of three processes: (i) a contention based registration process, (ii) a fair slot distribution process and finally, (iii) a network monitoring process for reallocating unused slots. Considering the effect of these processes, the states of a vehicle and RSU in our proposed scheme is depicted in Figure 2 and Figure 3 respectively.

As shown in Figure 2, upon reception of RSU's registration beacon, a vehicle moves into the *registration state* from its *initialize state* and sends back a registration response. Then, after receiving RSU's response, the vehicle moves into the *network monitoring state*; once a slot is assigned, the vehicle moves to the *data transmission state* where it transmits data in the slot assigned to it. At the end of the slot, the vehicle moves back to the *network monitoring state* and waits for additional slots from the RSU. When data transmission is complete, the vehicle moves back to the *initialize state*.

As shown in Figure 3, after transmitting registration beacon, an RSU enters into the *registration state* where it receives response from vehicles for a short amount of time. After receiving registration response(s) from the vehicles, the RSU enters into the *slot allocation state* where it allocates available slot to the vehicles; then, it transmits a registration response message and moves to the *network monitoring state* where the RSU monitors network for free slots. If any free slot is detected, the RSU moves back to the *slot allocation* state; otherwise, it moves to the *initialize* state at the end of the TDMA frame.

4. Details of the DRWA-MAC Protocol

DRWA-MAC consists of three processes: (i) a contention based registration process, (ii) a fair slot distribution process, and (iii) a network monitoring process. In this section, we discuss the details of these processes and also, present our slot distribution algorithm.

4.1. Registration Process

At the beginning of the TDMA frame, RSU transmits a registration beacon for all the vehicles in its broadcasting area. Upon reception of the registration beacon, each vehicle sends back a registration response message to the RSU using a contention based approach. To accommodate the vehicles response, this contention based procedure continues for a certain predefined amount of time. Within registration response message, vehicles send back their ID, location and data size.

4.2. Slot Distribution Process

After registration process, RSU's usually have N number of participating vehicles that need to be assigned with S number of slots. As the registration response from each vehicle consists of location information, RSU can differentiate the L number of vehicles on its left and the remaining (N-L) number of vehicles on its right direction in a road with two-way traffic. Our protocol divides the available S slots into $S/2$ ODD, and S/2 EVEN sets. Then, the $S/2$ ODD slots are distributed among the L left directional vehicles and S/2 EVEN slots to (N-L) right directional vehicles. Our protocol uses a repeated round robin approach for channel assignment. In case the number of vehicles is less than the number of slots, i.e., S>N, one or more vehicle may get multiple slots in the same frame. An example is shown in Figure 4 where there are 8 slots in a frame with two left direction vehicles (*V[1]*, *V[3]*), and two right direction vehicle (*V[2]*, *V[4]*); a repeated round robin distribution will allocate slot 1, and 5 to *V[1]*, slot 2, and 6 to *V[2]*, slot 3, and 7 to *V[3]* and slot 4, and 8 to *V[4]*. On the other hand, if $L < S/2$ and $(N–L) > S/2$, one or more left directional vehicles will get multiple slots but $(N–L–S/2)$ right

Slot Numbers

1	2	3	4	5	6	7	8
V[1]	*V[2]*	*V[3]*	*V[4]*	*V[1]*	*V[2]*	*V[3]*	*V[4]*

Figure 4: Example of slot allocation in our proposed scheme using repeated round robin distribution where the index represents vehicle ID. In this example, there are 4 vehicles where each vehicle received 2 slots each

directional vehicles will not get any slot allocation. Our slot distribution algorithm is shown in Algorithm 1.

Algorithm 1: TDMA Slot Distribution

```
1:  foreach available slot s in frame f
2:      calculate even slot es and odd slot os
3:  foreach registered vehicle v
4:      If vehicle v is right directional and es>0
5:          assign an  even slot es to vehicle v
6:          es = es - 1;
7:      else if os>0
8:          assign an odd slot os to vehicle v
9:          os = os – 1;
10: s = s – v;
```

4.3. Transmission Monitoring Process

After slot assignment, vehicles usually begin transmission in their assigned slots. Even though all the vehicles register for TDMA slot, some of them may not use all of their assigned slot(s); this could be due to early completion of data transmission, passing over RSU's broadcasting area, detection of the safety message that is already broadcasted, etc. To avoid wastage, for each slot, our protocol employs a monitoring process at the RSU end to check whether an existing transmission is going on or not; if not, then it ensures reallocation. Here, an RSU detects transmission status by monitoring the first one fourth of the beginning time of a slot; if no data is transmitted during this period, the RSU mask the slot as a free and broadcasts a very light NULL packet similar to a becon signal. Inside the NULL packet, RSU sends the ID of the vehicle having maximum data size; this is possible because at the beginning of the frame, each vehicle registers with its data size. Upon reception of the NULL packet, the specific vehicle transmits data for the remaining time of the slot and hence, reallocation of unused slots is achieved in a collision free manner.

Figure 5 shows operation of the proposed DRWA-MAC in detail. Here, *V[1]*, and *V[3]* are left directional vehicles and *V[2]*, *V[4]* are right directional vehicles. *V[1]*, *V[2], V[3], and V[4]* have 24MB, 22MB, 4MB, and 4MB of data to transmit respectively. These vehicles V[1], V[2], V[3], V[4] are assigned with slots (*1*,*5*), (*3*,*7*) and (*2*,*6*), (*4*,*8*) respectively by our slot allocation algorithm. Considering max data size of each slot as *4* MB, at slot *7*, RSU will detect that there is no ongoing transmission in channel as *V[3]* has no data left to send and therefore, will select

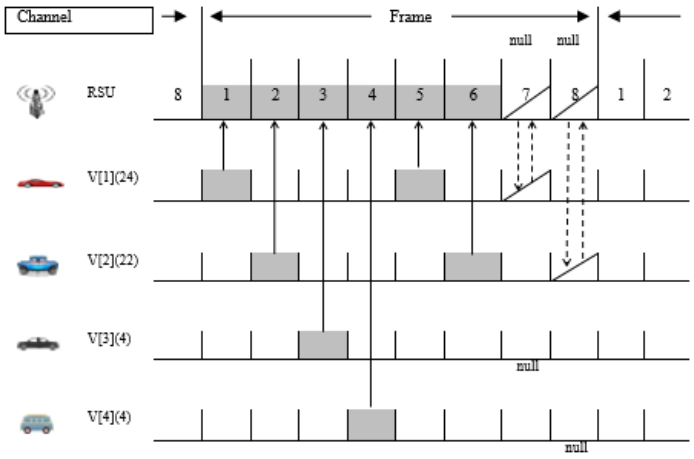


Figure 5: Operation of proposed DRWA-MAC with four vehicles and 8 slots in each frame where 4MB of data can be transmitted in each slot. Initial slot assignment is as shown in Figure 4

V[1] as it has *16* MB data remaining for transmission (out of 24 MB, V[1] has transmitted 8 MB in slot 1 and 5). So, *V[1]* will be able to send 3 MB of data in the remaining 3/4 of the slot and will have 13 MB data left for transmission. At slot *8,* RSU again will detect that there is no ongoing transmission in the channel as *V[4]* has no data left to transmit and will select *V[2]* as it has 14 MB data remaining while *V[1]* has 13 MB of data remaining to transmit.

5. Effectiveness

In our protocol, the use of "time slot reassignment" technique makes use of most of the wasted slots and hence, increases overall throughput of the TDMA network. As our protocol works in RSU based centralized topology, it is free of access collision. Since the proposed protocol uses RSU's guidance based "time slot reassignment" technique where a single vehicle is selected by RSU to send message in free slots, it is also free from hidden node collision.

Instead of straight forward segmentation like VeMAC, our protocol divides and allocates slots in ODD-EVEN basis and it balances waiting time of the partitioned vehicles. This distribution process also avoids or minimizes merging collision and one-hop neighboring collision. To support our argument, an example scenario is depicted in Figure 6. In this scenario, there are 2 RSUs, RSU[0], and RSU[1]. We show that vehicle *V[4]*, *V[6]* of *RSU[0]* and *V[5]*, *V[7]* of *RSU[1]* are one hop neighboring collision free as *RSU[0]* assigns EVEN slots to *V[4]*, and *V[6]* and on the other hand, *RSU[1]* assigns ODD slots to the vehicles *v[5],* and *v[7]*. Additionally, if we consider the position of vehicle *V[5]* of *RSU[1]* and *V[1]* of *RSU[0],* we can claim that our protocol also minimizes merging collision because both *V[5]* and *V[1]* are placed in left/west half of their respective RSU and each is assigned with ODD slots. By the time *V[5]* enters into the coverage area of *RSU[0]*, *V[1]* may leave and the slot freed by *V[1]* can be reassigned to *V[5]* so that *V[5]* can transmit data in the running frame.

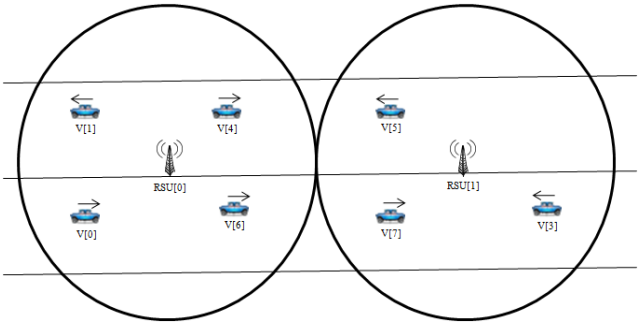


Figure 6: Example scenario of proposed DRWA-MAC protocol which depicts that the protocol is free from one-hop neighboring collision and merging collision

6. Simulation Results and Analysis

In our simulation experiments, we have compared the performance of our proposed DRWA-MAC protocol with ACFM, VeMAC, PTMAC, FAWAC-MAC and FAWAT-MAC in terms of number of successfully transmitted data packet, average delay, and data loss ratio on overall road traffic network.

6.1. Simulation Environment

We have implemented DRWA-MAC protocol on the VEINS [11] framework of OMNET++ [12] network simulator. The vehicle movement and related realistic road traffic scenario is generated using SUMO [13] urban mobility traffic generator.

In our scenarios, we considered a 2000m×1000m rectangular two way road traffic network with ten junctions as shown in Figure7, where each junction contains three lanes. We have covered our road network segment with three non-overlapping RSUs as shown in Figure 8.

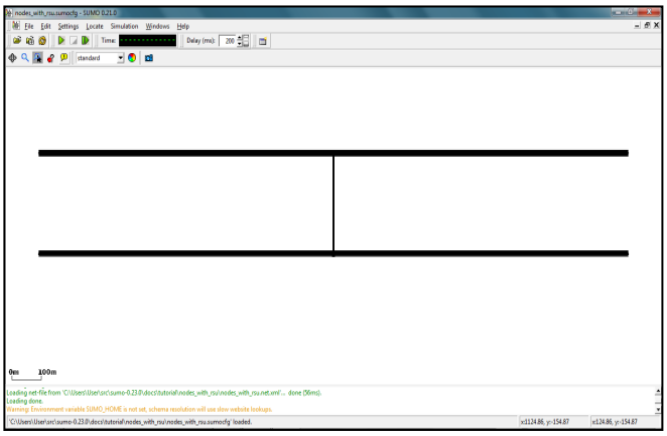


Figure 7: 2000m×1000m rectangular traffic network with ten junctions and 3×3 lane road in SUMO



Figure8: OMNET++ view of the road network with RSUs

Table 1: Simulation Parameters

Parameter	Value
Number of lanes	3+3
Number of scenario	5
Number of vehicles	22, 41, 61, 83 and 102
TxPower	20mW
TxDatasize	10Mbps
Number of slots per frame	256
Slot time	50ms
Registration time	2s
Simulation time	2 frames, on 258s to 288s

For comparison, we have devised five scenarios with 22, 41, 61, 83, and 102 vehicles and measured the performance of the mentioned protocols in cases where 10%, 15% and 20% of the registered vehicles were not using their assigned slots. Configuration of the simulation is listed in Table 1.

6.2. Simulation Result

In this section, we show performance comparison of DRWA-MAC with the aforementioned protocols.

Figure 9 shows performance comparison in terms of successfully transmitted data packets when 10% of the vehicles are not using their assigned slots. As shown in Figure 9, initially at light traffic, only FAWAC-MAC performed better than DRWA-MAC. However, the performance of FAWAC-MAC gradually decreased with the increase of traffic. In high traffic scenarios, the performance of VeMAC, ACFM, PTMAC and FAWAT-MAC degraded noticeably. This is because these protocols suffer from merging collision and one-hop neighboring collision. On the other hand, DRWA-MAC minimized these collisions and outperformed all the other schemes when network density is higher.

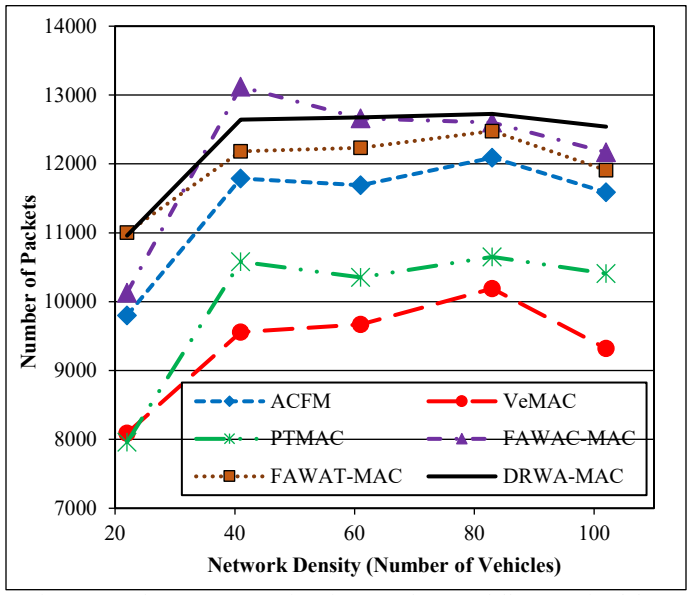


Figure 9: Performance comparison in term of successfully transmitted data packets when 10% of the vehicles are not using their assigned slots

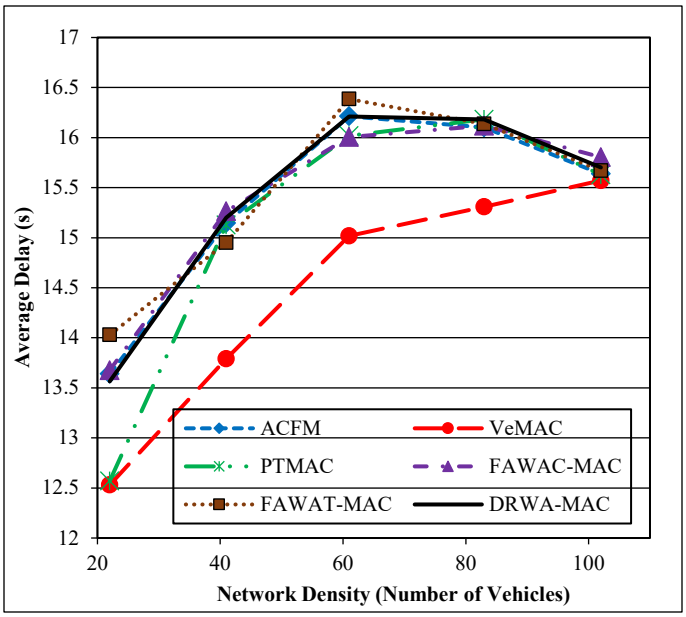


Figure10: Performance comparison in terms of average delay when 10% of the vehicles are not using their assigned slots

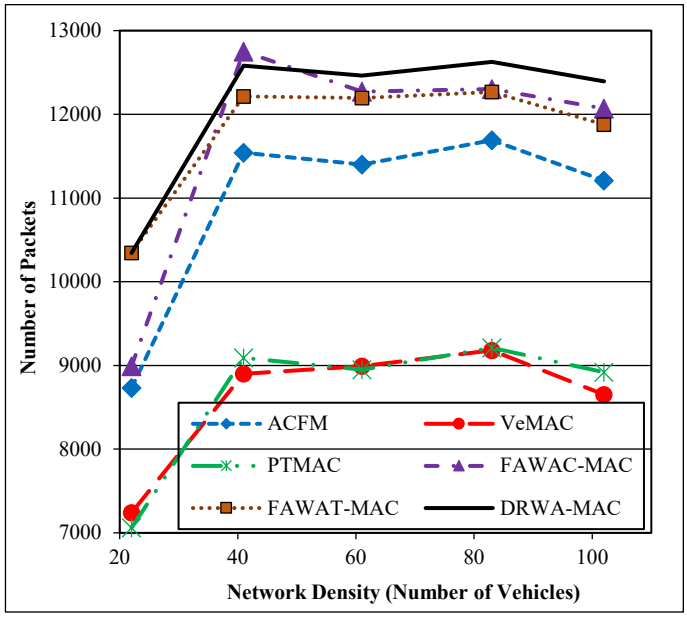


Figure 12: Performance comparison in terms of successfully transmitted data packets when 15% of the vehicles are not using their assigned slots

Figure 10 shows performance comparison in terms of average delay when 10% of the vehicles are not using their assigned slots. Though for a single channel, the overall throughput of VeMAC is not optimum, it has the most favorable average delay. On the other hand, our proposed DRWA-MAC performed better by maintaining almost the same average delay as ACFM, PTMAC, FAWAC-MAC and FAWAT-MAC.

Figure 11 shows performance comparison in terms of data loss ratio where 10% of the registered vehicles are not transmitting data in the assigned slots. Because of inherent huge low power transmission, as shown in Figure 11, FAWAC-MAC is having high data loss, though those low power packets were bearing opportunistic data. VeMAC has the lowest data loss ratio while DRWA-MAC has similar data loss ratio compared to other protocols.

In Figure 12 and 13, we show performance comparisons in terms of successfully transmitted data packets for cases where 15% and 20% of the registered vehicles are not transmitting data in their assigned slots respectively. With the increase of free slots, the performance of VeMAC, ACFM and PTMAC decreased as these protocols do not have any process to ensure reusability of slots. However, as FAWAC-MAC, FAWAT-MAC and DRWA-MAC

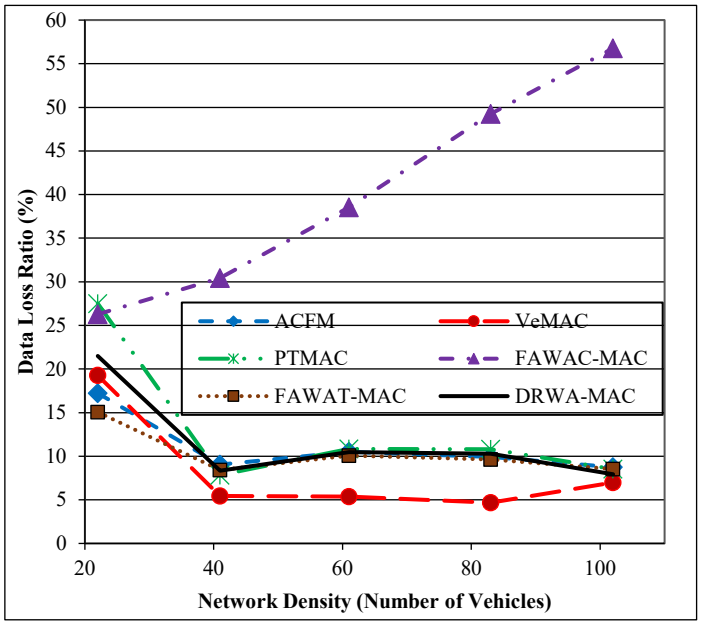


Figure11: Performance comparison in terms of data loss ratio when 10% of the vehicles are not using the slots assigned to them

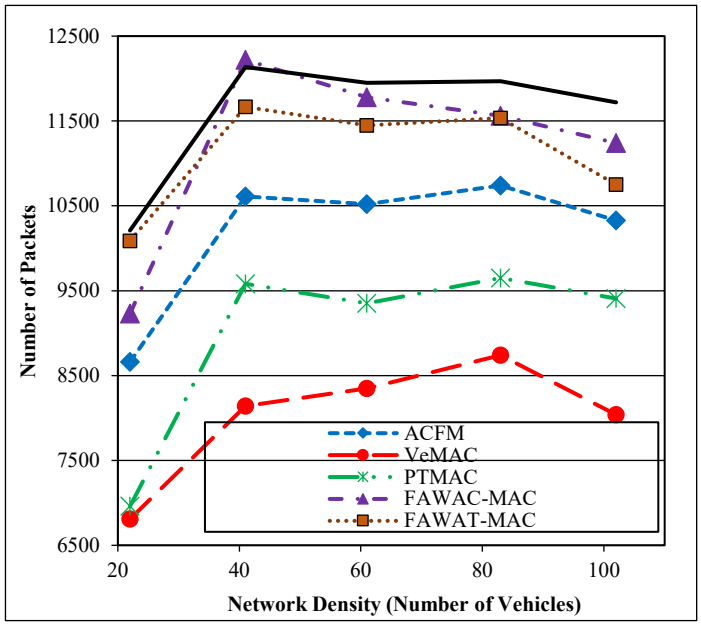


Figure 13: Performance comparison in terms of successfully transmitted data packets when 20% of the vehicles are not using their assigned slots

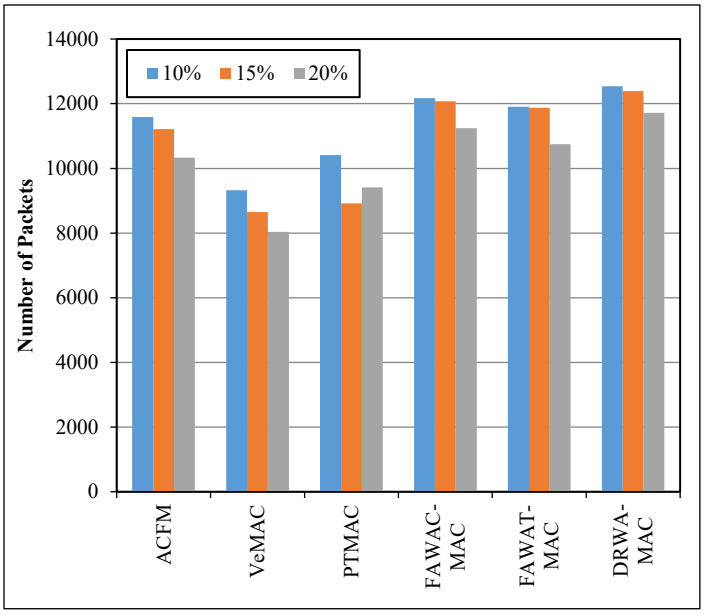


Figure 14: Bar diagram of successfully transmitted data packets by grouping scenarios where 10%, 15% and 20% of the vehicles are not using their assigned slots

can unsure reusability of free slots, their performance were not affected significantly. From Figure 9, 12 and 13, we can conclude that though FAWAT-MAC and DRWA-MAC improves the amount of successful data transmission by ensuring reallocation of free slots even though one fourth of the data capacity of those free slots degraded due to the detection delay of the RSU.

Figure 14 shows a bar diagram of successfully transmitted data packets for 102 cars in the network by grouping scenarios where 10%, 15% and 20% of the vehicles are not using their assigned slots. This diagram is also shows that with the increase of free slots, the performance of VeMAC, ACFM and PTMAC decreased, but the performance of FAWAC-MAC, FAWAT-MAC and DRWA-MAC were not affected significantly. As shown in the figure, DRWA-MAC outperforms all other schemes in terms of successfully transmitted data packet in all the experimental scenarios.

7. Conclusion

This article proposes a dynamic reallocation based window access scheme, DRWA-MAC, for centralized TDMA based VANETs. The proposed scheme maintains a balanced waiting time for vehicles by dividing TDMA slots into ODD slots and EVEN slots and then, by evenly distributing all TDMA slots among the registered vehicles in a rational way. Our slot allocation algorithm ensures fairness in slot allocation among the vehicles as well as reduces merging collision and one-hop neighboring collision. By dynamically monitoring and reallocating unused slots using “time slot reassignment” mechanism, the proposed DRWA-MAC achieved significantly higher throughput compared to other conventional TDMA based centralized protocols. Simulation results show the effectiveness of the proposed protocol in terms of successfully transmitted data packets, average delay and data loss ratio on overall road traffic network in different scenarios.

References

[1] Md. Amirul Islam, Shafika Showkat Moni and Mohammad Shah Alam, “A Fixed Assignment Based Window Access Scheme for Enhancing QoS of Vehicular Adhoc Networks” in International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16-18, 2017.

[2] Mohamed Hadded, Paul Muhlethaler, Anis Laouiti, Rachid Zagrouba, Leila Azouz Saidane, "TDMA-based MAC Protocols for Vehicular Ad Hoc Networks A Survey, Qualitative Analysis and Open Research Issues" inIEEE Communication Surveys and Tutorials, Volume: 17, Issue: 4, February 2015.

[3] Laouiti, Anis, Qayyum, Amir, Mohamad Saad, Mohamad Naufal, "Vehicular Ad-hoc Networks for Smart Cities" in First International Workshop(Advances in Intelligent Systems and Computing), 2014.

[4] M.J. Booysen, S. Zeadally, G.-J. Van Rooyen, "Survey of media access control protocols for vehicular ad hoc networks"in IET Communication, 2011, Vol. 5, Iss. 11, pp. 1619–1631.

[5] Georgios Karagiannis, Onur Altintas, Eylem Ekici, Geert Heijenk, Boangoat Jarupan, Kenneth Lin, and Timothy Weil, "Vehicular Networking: A Survey and Tutorial on Requirements, Architectures, Challenges, Standards and Solutions" in IEEE Communications Surveys & Tutorials, Volume: 13, Issue: 4, 2011.

[6] Mohammad S. Almalag, Stephan Olariu, Michele C. Weigle, "TDMA Cluster-based MAC for VANETs (TC-MAC)" in IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), June 2012.

[7] H. Omar, W. Zhuang, and L. Li, “VeMAC: A Novel Multichannel MAC Protocol for Vehicular Ad Hoc Networks” in IEEE Conf. Computer Communications Workshops (INFOCOM WORKSHOPS), April 2011, pp. 413–418.

[8] H. Omar, W. Zhuang, and L. Li, “VeMAC: A TDMA-Based MAC Protocol for Reliable Broadcast in VANETs” in IEEE Transactions on Mobile Computing, Volume: 12, Issue: 9, September 2013, pp. 1724 - 1736

[9] Weijie Guo, Liusheng Huang, Long Chen, Hongli Xu, Jietao Xie, "An Adaptive Collision-Free MAC Protocol Based on TDMA for Inter-Vehicular Communication" in International Conference on Wireless Communications & Signal Processing (WCSP), 2012.

[10] Xiaoxiao Jiang and David H.C. Du, "PTMAC: A Prediction-based TDMA MAC Protocol for Reducing Packet Collisions in VANET" in IEEE Transactions on Vehicular Technology (Volume: PP, Issue: 99), 19 January 2016.

[11] Christoph Sommer, “Veins - The open source vehicular network simulation framework”, Web Link “http://veins.car2x.org/”.

[12] András Varga, “OMNeT++, Discrete Event Simulator”, Web Link “https://omnetpp.org/”.

[13] DLR - Institute of Transportation Systems, “SUMO, Simulation of Urban MObility”, Web Link “http://sumo-sim.org/”.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 342-347 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Signal-Based Metamodels for Predictive Reliability Analysis and Virtual Testing

Veit Bayer[*,1], Stephanie Kunath[1], Roland Niemeier[1], Jürgen Horwege[2]

[1]*Dynardo GmbH, Weimar, Germany*

[2]*Airbus Operations GmbH, Bremen, Germany*

ARTICLEINFO

Article history:
Received: 29 November, 2017
Accepted: 19 January, 2018
Online: 02 February, 2018

Keywords:
Metamodel
Surrogate model
Dynamics
Random field
Aircraft
High lift
Digital twin
Virtual test

ABSTRACT

In the current industrial development, an increasing number of sensors is applied for monitoring of any kind of appliances and machines. A predictive mathematical model allows for realistic assessment of the health state of the appliance, indication of service requirements, as well as control of the appliance, serving as so-called digital twin on a control device.

Completely analogous modeling can be used for virtual testing, i.e. a (partial) substitution of physical experiments. The software-in-the-loop model provides realistic feedback to the physical specimen on the test rig and helps to increase the representativeness of the experiment and to reduce costs.

In a joint research project partially sponsored by the German Federal Ministry of Economics and Energy, the authors developed an approach for meta-modeling of dynamic systems. While the modeling process is fed by results from sophisticated simulations, or even test results as input data, the resulting model can be used for fast stochastic analyses as well as software-in-the-loop in dynamic real-time experiments. The approach was verified on tests of an aircraft high lift system.

1 Motivation and Objective

The prediction of test results is especially important for very expensive tests like in aerospace industrial applications, e.g. testing of high lift systems. As a matter of costs and manageability, the test specimen usually represents only a partial system. A realistic feedback of reactions from the remaining system by software driven actuators increases the representativeness of test results. Sophisticated procedures for realistic modeling of physical systems, e.g. algorithms for multi-body dynamics simulation, are available, but they require significant computation time and lack the ability of real-time performance. Therefore, there is a need to reduce computational cost maintaining high fidelity modeling.

Purely mathematical models can be established which are fast to compute, yet maintain accurate results. The authors' approach, which has been presented initially in [1], yields meta-models that approximate the dynamic response of the tested object. Models are built on the basis of physical simulations that represent the test results accurately, but require a computation time that would be prohibitive as for the applications indicated above. The results of real experiments can be used as input data as well. The proposed meta-models, having short response times, enable fast robustness analysis to assess the influences of uncertain parameters such as damping or friction and can be used as software-in-the-loop even in real-time dynamic experiments [2, 3], where often fast but simplified physics-based models are applied [4]. The requirements for development of the methodology, which will be presented in the following sections, are summarized as: speed in performance, flexibility to change of parameters, versatility and accuracy.

The combination of physical testing and virtual testing serves several purposes: evaluation of the test rig without risk of damage; substitution of parts of the tested system —or augmentation of the existing test—; more realistic and more complex tests and reduction of costs.

[*]Veit Bayer; Dynardo, Steubenstr. 25, D-99423 Weimar, Germany; Ph. +493643900847; E-mail veit.bayer@dynardo.de

https://dx.doi.org/10.25046/aj030141

Other applications of this approach can be thought of: fast meta-models can be implemented as digital twin on data acquisition and control devices for online monitoring of a system; in the development of a product, fast but accurate models are important for stochastic analyses such as predictive reliability assessment.

2 Methodology

Basis of the methodology proposed here is Dynardo's algorithm *Metamodel of Optimal Prognosis* (MOP). The procedure fits the best available model to given data points and avoids so-called over-fitting. It first starts a filtering of parameters by statistical criteria. Input parameters with negligible influence on the observed results are canceled from the data set, thus the dimension of the problem can be reduced effectively. Several model approaches, such as polynomials, moving least squares, kriging, are built up and tested by cross validation. The resulting *Coefficient of Prognosis* (CoP) is the complimentary value to the sum of squared residuals over the variation of result data. The CoP gives information, what amount of data variation is explainable by the meta-model, based on independent test data. The MOP result is the chosen model out of the set of available models with the largest CoP. The MOP is also used for sensitivity analysis [5, 6]. Conditional variances are computed by holding systematically one parameter at fixed values, indicating the relative contribution of this parameter to the total variation of the response.

The MOP deals with scalar response quantities only. For the intended application, results are time series, e.g. from a multi-body dynamics simulation. Hence a representation of the time series by scalar values has to be found, while avoiding to adopt the total set of discrete time steps. The approach makes use of a specific topic in probability theory, namely the random fields methodology [7, 8]. A random field is a quantity defined on a spatial domain, where the value at any point of observation is a random variable. Here, the domain is time instead of space. The training data for the model are produced by first sampling input values, either by design of experiments or quasi-random sampling, then computing the results for each input data set. The time-dependent results are then interpreted as random process or 1-D random field.

The key to a parametrization of the random process is the eigenvalue decomposition of the covariance matrix $\mathbf{C}_{XX}$ of the discretized time series $\mathbf{X}$,

$$\mathbf{\Psi}^T \mathbf{C}_{XX} \mathbf{\Psi} = \mathrm{diag}\{\lambda_i\}, \tag{1}$$

wherein $\mathbf{\Psi}$ is the matrix of eigenvectors, and $\mathrm{diag}\{\lambda_i\}$ holds the eigenvalues of the covariance matrix. From this, the so-called *spectral representation* of the random field can be derived [9]. Assuming that X are normal distributed with zero mean values (which can be subtracted for the analysis and added later again for synthesis of time series), new random variables Y are defined as

$$\mathbf{Y} = \mathbf{\Psi}^T \mathbf{X}, \tag{2}$$

which are normal distributed, independent, zero-mean and with a standard deviation given by

$$\sigma_{Y_i} = \sqrt{\lambda_i}. \tag{3}$$

For synthesis of the original time series $\mathbf{X}$, one makes use of the Karhunen–Loève series expansion

$$\mathbf{X} = \mathbf{\Psi} \mathbf{Y}. \tag{4}$$

A typical property of eigenvalue solvers is, that eigenvalues are stored in descending order. Since the eigenvalue of order i defines the amount of variation contributed by parameter Y_i to the total variation of the data, this gives a criterion for truncation of the series expansion and therefore a drastic reduction of the dimension [10, 11].

Summarizing, for generation of a time series $\mathbf{X}$ one needs the modal base $\mathbf{\Psi}$ of the covariance matrix, which we may call the set of "shape functions" in the following, and the respective "amplitudes" Y_i. Figure 1 shall illustrate the series expansion of (4) with an added mean value signal.

The shape functions are an unchangeable property of the data. Properties of the parameters Y_i have to be determined such that, the series expansion optimally represents the physical time series $\mathbf{X}$. For this purpose, MOP is applied to a training data set, from which the corresponding sample of amplitudes is calculated. This leads to the *Field-Metamodel of Optimal Prognosis* (F-MOP). Figure 2 shows the flow of analysis. Using these dynamic meta-models, we obtain a simplified and reduced parametric of the dynamic signal based on a statistical meta-model. The user does not need to find a parametrization himself.

In analogy to the sensitivity analysis of scalar data [5, 6], the Coefficient of Prognosis of the F-MOP can be plotted against the time axis, called F-CoP here. Moreover, sensitivity measures over time can be computed, such that it is possible to assess the model quality and the relative influence of input parameters on the response, locally within the entire observed time range. The procedure is realized by connecting the functionalities of the programs offered by Dynardo, optiSLang and Statistics on Structures (SoS).

3 Application Example: Aircraft High Lift System

The methodology which is described in the previous section is validated by the example of an aircraft high lift system. Instead of real experiments, virtual tests using a detailed model serve for generating the training data. The simulation model in MSC Adams/Flex comprises the inboard flap, outboard flap, transmission and actuators as well as the test rig. With that, model wing position, flap positions, loadings as well as

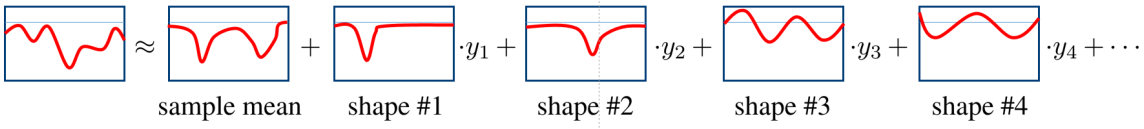


Figure 1: Illustration of Karhunen–Loève series expansion of a signal.

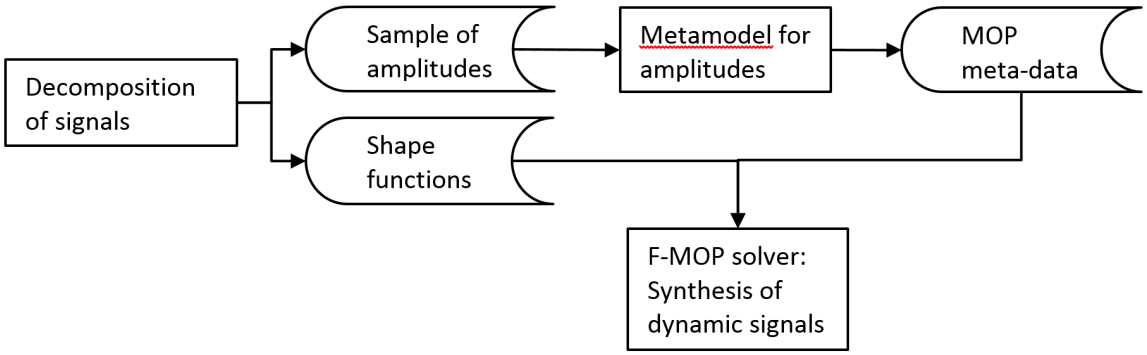


Figure 2: Overview of the approach of decomposing signals and generating meta-models for dynamic signals.

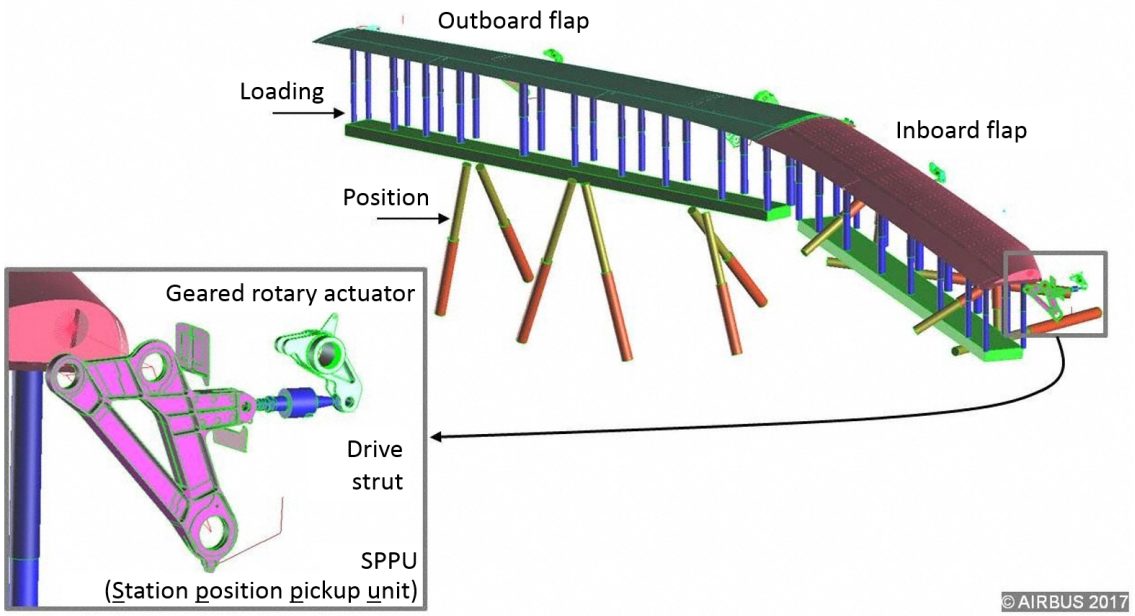


Figure 3: Adams/Flex model of the high lift system on a test rig.

actuator action and backlash can be simulated. Since friction is taken into account in the simulation, the model is actually non-linear. Figure 3 visualizes the tested system.

For the virtual test, wing and flap positions and loadings are given. The test case is a simulated rupture of two actuators simultaneously, one at the inboard flap, one outboard. As uncertain parameters, the stiffness of all actuators, backlash of all actuators, damping and friction parameters are sampled as input to the meta-modeling procedure. 200 parameter sets are simulated using MSC Adams/Flex, yielding signals over time of the dynamic responses. The responses observed are all actuator moments, drive strut forces and angles at the station position pickup units (SPPU). For these responses, dynamic meta-models shall be built.

4 Discussion of Results

Figure 4 –top– shows 4 typical signals picked out of the whole sample, here of the moment at actuator #2. Below, the prediction capability expressed as F-CoP is plotted. It can be observed that, the Coefficient of Prognosis of the model F-CoP [Total] is able to reach nearly 100% at the peaks of the signal, while its value drops where the oscillation passes the stationary value. At such points, scatter in data is rather numerical noise than information and cannot be explained by the meta-model. The relative influences of input parameters can be observed, too (largest: F-CoP[DV_Fric_Mue_PP_dyn] for Coulomb friction parameter).

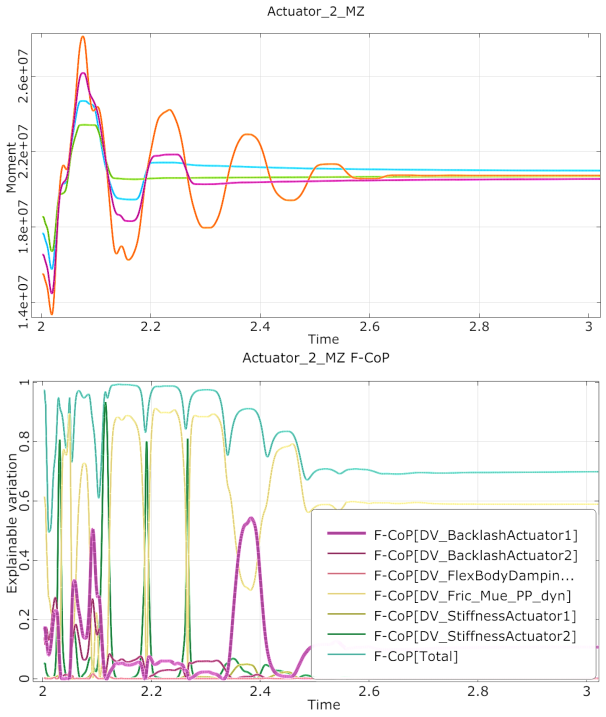


Figure 4: Top—Sample time series of moments at actuator #2. Bottom—Prediction capability of the model for the time dependent signal (F-CoP [Total]) and the impact of each input parameter on the total variation of the signal.

Another signal, the angle at SPPU #2, is examined in the same way, results are displayed in Figure 5. There is a constant shift for each sample. The F-CoP[Total] values are near 100% throughout the entire time range. The model quality is considered excellent by means of the statistical F-CoP criterion.

Finally, the original simulated time series and the series synthesized with help of the field meta-model shall be compared. Again, a few time series are arbitrarily picked from the whole sample. In Figure 6, the signals of the moment at actuator #2 are compared, and Figure 7 shows the comparison for the angles at the station position pickup unit #2. The left plots in both figures are the original simulation results obtained by MSC Adams/Flex, which serve as reference. The right plots are the signals which were synthesized by the field meta-models using the same input parameters as for the original simulations. The coincidence of original signals and meta-model results is excellent, particularly in the peaks. When the signals tend to be damped out, some artificial oscillations in the field meta-model results can be observed. These will vanish if the series expansion (refeq:KLseries) is truncated at a later position, thus more shape functions will be taken into account.

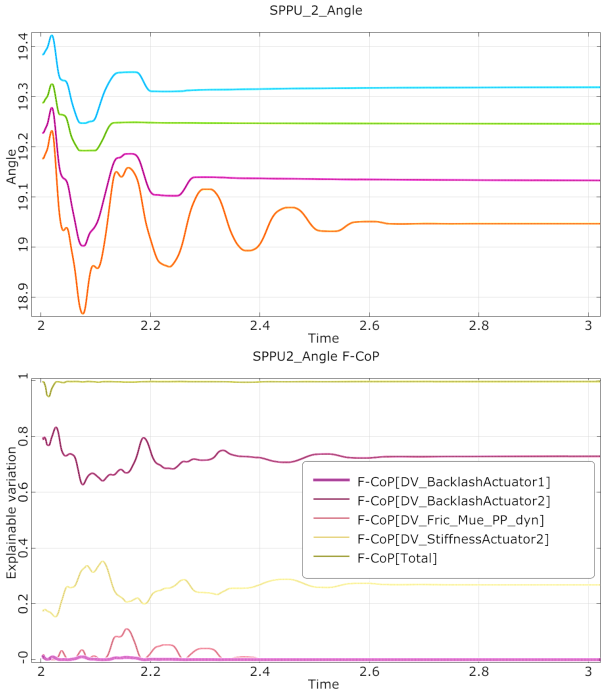


Figure 5: Top—Sample time series of angles at SPPU #2. Bottom—Prediction capability of the model for the time dependent signal (F-CoP [Total]) and the impact of each input parameter on the total variation of the signal.

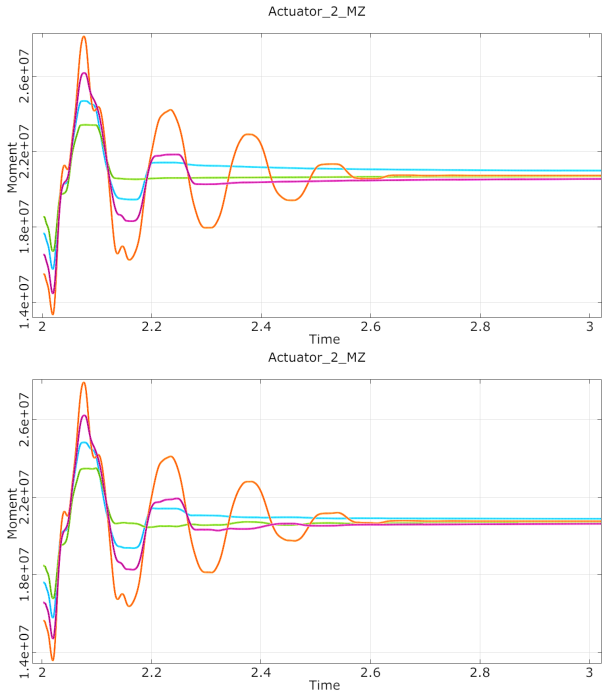


Figure 6: Sample time series of moments at actuator #2. Top—original results obtained from Adams/Flex. Bottom—synthesized signals by F-MOP.

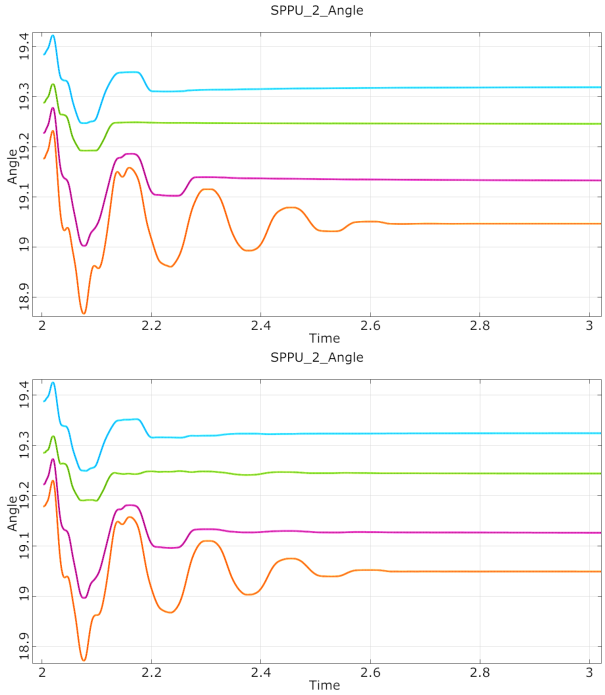


Figure 7: Sample time series of angles at SPPU #2. Top—original results obtained from Adams/Flex. Bottom—synthesized signals by F-MOP.

The proposed procedure has been applied also to other applications, e.g. for the parameter identification of non-linear material models in finite element analysis. The "signal" here is the load–displacement curve of a test specimen. In previous approaches [12], the signal was discretized into few equidistant steps, then meta-models were established by MOP for the single values on the curve. The squared residuals at these discrete points were used as criteria in a fast optimization procedure using MOP. For longer signals this approach is infeasible, moreover, it requires manual interference to identify characteristic points or relevant ranges. With the proposed approach, this is not necessary anymore. The field meta-model is valid for all points according to the raster of the original data, and one easily sees in which regions the parameters have what amount of influence.

5 Summary and Outlook

A new procedure is proposed in this article for meta-modeling of dynamic (time series) signals. It is based on the decomposition of a sample set of signals into a series consisting of shape functions scaled by amplitudes for each summand. The algorithm Metamodel of Optimal Prognosis is then applied to find the best fitting model for each amplitude. The Coefficient of Prognosis, which does an assessment of models with the statistical method of cross validation, is used as the selection criterion. It is possible to assess the model prediction capability even locally along the time axis. Moreover, sensitivity measures reveal the relative influence of parameters (as input to the dynamic analysis or experiment), also locally along the time axis.

The procedure has been applied to an aircraft high lift system. The training data were generated by virtual experiments, i.e. multi-body dynamics simulations. The surrogate models proved to be very accurate, by the CoP criterion as well as by direct comparison of the reference to the model time series. Unlike the high-fidelity simulation, the meta-models are very fast to compute, allowing e.g. for fast stochastic analyses in the scope of predictive reliability analysis. The models were also successfully implemented as software-in-the-loop into an experimental environment for real-time dynamic tests at the Airbus site.

Further developments are planned which will focus on more detailed model assessment and model improvement, locally in the time and parameter space. The range of application for this procedure is not limited to the above mentioned. Signals can be any xy-data, so fast surrogate models can be obtained, e.g., also for spectral analyses both in the low and high frequency domains, stress-strain curves of a material law etc. The field meta-models can also serve as digital twins which, fed by sensor data, monitor or control electronic appliances or machines.

Acknowledgments This work was partially financed by the German Federal Ministry of Economics and Energy, call LUFO 5.1, grant number 20Y1301E, which is gratefully acknowledged by the authors.

The authors also express their gratitude for the contributions by Thomas Töpsch, Tobias Ulmer, Achim Lenz, Airbus Operations, Bremen; Sönke Klostermann, Mario Cappitelli, Airbus Group Innovations, Hamburg; Viktor Lebsak, MSC Software, München and Michael Neumann, P3 group, Hamburg.

References

[1] S. Kunath, V. Bayer, R. Niemeier, "Predictive reliability with signal based meta-models" in 18^{th} International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE), Dresden Germany, 2017.

[2] A. Forrester, A. Sóbester, A. Keane, Engineering Design via Surrogate Modelling, Wiley 2008.

[3] D. Vogt, "Effektive Visualisierung Stochastischer Simulationen", Der Andere Verlag 2010.

[4] V. Bayer, U.E. Dorka, U. Fllekrug, J. Gschwilm, "On real-time pseudo-dynamic sub-structure testing: algorithm, numerical and experimental results" Aerospace Science and Technology, **9**(3), 223–232, 2005.

[5] T. Most, J. Will, "Metamodel of Optimal Prognosis – an automatic approach for variable reduction and optimal metamodel selection" in Proc. Weimar Optimization and Stochastic Days 5.0, Weimar Germany 2008.

[6] T. Most, J. Will, "Sensitivity analysis using the Metamodel of Optimal Prognosis" in Proc. Weimar Optimization and Stochastic Days 8.0, Weimar Germany, 2011.

[7] E. Vanmarcke, Random Fields: Analysis and Synthesis, MIT Press 1983.

[8] C. Bucher, Computational Analysis of Randomness in Structural Mechanics, Structures and Infrastructures Book Series Vol. 3, CRC Press 2009.

[9] R. Ghanem, P.D. Spanos, Stochastic Finite Elements – a Spectral Approach, Springer 1991.

[10] C.E. Brenner, "Ein Beitrag zur Zuverlässigkeitsanalyse von Strukturen unter Berücksichtigung von Systemuntersuchungen mit Hilfe der Methode der Stochastischen Finite Elemente", Ph.D. (Dr. tech.) Thesis, Leopold Franzens Universität Innsbruck, 1995.

[11] V. Bayer, J. Will, "Random Fields in Robustness and Reliability Assessment of Structural Parts" in 15. VDI Kongress Berechnung und Simulation im Fahrzeugbau SIMVEC (in German with English abstract), Baden-Baden Germany 2010.

[12] T. Most, "Effiziente Parameteridentifikation fr numerische Simulationsmodelle", in NAFEMS Konferenz: Berechnung und Simulation, Bamberg Germany, 2014 (in German)

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 348-351 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

LabVIEW-based data acquisition system for Diode I-V Characterization

Nor Shaida Mohd Saufi*, Nurul Syafiqah Yap Abdullah, Mohd Ikhwan Hadi Yaacob

Faculty of Science Mathematics, Universiti Pendidikan Sultan Idris, 35900, Malaysia

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 07 January, 2018
Online: 02 February, 2018

Keywords:
LabVIEW
Data acquisition system
I-V Characteristics
Diode

ABSTRACT

This work describes computerized data acquisition system for current-voltage (I-V) characteristics of diodes. The proposed system consist of a personal computer pre-installed with LabVIEW, two units of power supply, an interface board, current and voltage sensors. Two devices under test (DUTs) are selected namely 1N4007 silicon and 1N34 germanium diodes. The current flow through the circuit and voltage across the diode is measured and I-V characteristic is successfully plotted. When data acquisition interrupted by the user, real time I-V curve with least-square fitting is plotted. The LabVIEW was employed primarily to display data from sensors and storing collected data for further post-measurement processing. Estimated forward voltage drop from plotted I-V curve was then compared with the published value from manufacturer's datasheet. Comparison revealed that developed system has been successfully characterized both DUTs base on the forward voltage drop obtained from the plotted I-V curve. Furthermore, proposed system architecture offers extra flexibility where stored data can be manipulated with minimum programming efforts. The application of the proposed system can also be extended for in-situ device characterization in various circuits.

1. Introduction

Rapid advancement in computerized measurement system around the world has seen the introduction of fast, simple, user-friendly and reliable solutions found in various applications. In recent years, rapid progress of the virtual instrument platform such as LabVIEW graphical programming language has simplified programming effort and produce simple yet highly-productive real-time computerized measurement system.

This paper is an extension of work originally presented in 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE) [1]. In this report, we discussed in detail a system to determine I-V characteristics curve of germanium diode.

Proposed computerized I-V characterization system employed LabVIEW as the acquisition tool to obtain the data, analyze and display the result. Ease of use is one of our priority in this project. However, the complete functionality of LabVIEW is a plus factor that also satisfied the requirement of this project [2]. Virtual Instruments (VIs) in LabVIEW is utilized to construct a Graphical User Interface (GUI) as the time saving is the fair trade-off for the full low level control of other conventional programming languages such as C/C++, Visual Basic, and Matlab. Its graphical nature makes it ideal for measurement and data acquisition [3,4].

Simplifying data acquisition process and decreasing test time without scarifying the accuracy are primary goals in this project. Semiconductor diode is used in all sorts of electrical and electronic system. The electrical characterization of diode has been an important subject for more than half a century [5]. Semiconductor diode is an important device that exhibit non-linear relationship of I-V characteristics. In order to visualize the relationship, it is common to plot current versus voltage on graph. However, it is much more convenient to display the complete curve in an automated computer-based acquisition system.

2. System Description

This paper describes a LabVEW-based measurement method to plot I-V characteristics of semiconductor diodes. Figure 1 shows a block diagram of the main components of the measurement system that comprised mainly a personal computer with LabVIEW software, voltage and current sensors, interface board to be used between the sensors and computer, power supplies, and electronic circuit. In this research, LabVIEW 2012 was used in Windows 7 operating system. More precise sensors were used to measure voltage and current. The current sensor used

*Nor Shaida Mohd Saufi, Email: norshaidasaufi@gmail.com

was the Phidget CE-IZ02-32MS2-0.5 DC Current Sensor 0-1A powered by an external 12V power supply. The voltage sensor is Phidget Precision voltage sensor 1135 which measure differential voltage between input terminal and its output terminal. The current and voltage sensor was interfaced to personal computer through Phidget interface board 8/8/8 through Universal Serial Bus (USB) cable. Basically, the main components of the system and circuit used in this project is similar as described in [1] but we have used more sophisticated and better performance voltage and current sensors in terms of its measuring range and accuracy.

3. LabVIEW software

3.1. Block diagram

There are two windows for user to works on which are block diagram and front panel. LabVIEW is a programming environment which in block diagram, blocks representing functions, icons representing variables and lines representing data flows pass between different functional nodes [4]. LabVIEW is compatible with most of other hardware such as Phidget devices therefore they have their own blocks representing function. The communication between Phidget devices and LabVIEW software has been simplified using robust Application Program Interface (API) library which is accessible online for free. Figure 1 shows the block diagram of formula node in LabVIEW that was used to convert analog sensor value of voltage and current sensors to its corresponding value in volts and miliAmperes.

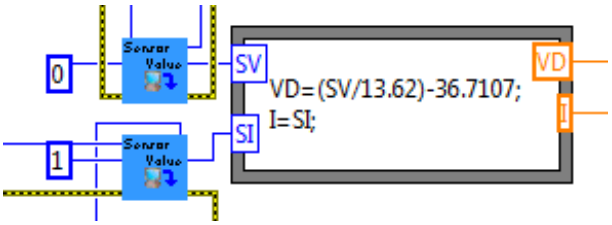


Figure 1. Formula node to write formula used to convert analog sensor value to voltage and current values

Particularly, Lev-Mar least square fitting method is expounded in the system which fit best the non-linear relation of diode. Figure 2(a) shows block diagram in LabVIEW for diode when user switch Curve fitting slide to Non-Linear and in Figure 2(b) is the block diagram when user switch slide to Linear. The saved data file in .csv can be accessed and plotted in Compare Graph when user clicked 'file path' and choose file location. The block diagram of the flow pass is shown in Figure 3.

Figure 3 shows block diagram that its function to read data from spreadsheet or .csv data file to be plot and compare in graph. The functional block can read and display multiple plot of germanium and silicon diode.

3.2. Graphical User Interface (GUI)

Front panel window is used as Graphical User Interface (GUI) of the system for user to interact with the program. Controls such as push button START, STOP, SAVE, PLOT and etc. as shown in Figure 4 are for user action to make the data displayed in indicator fields such as voltage, current value and also graphs. Before start the acquisition, user must make sure to check what is the electronic device to be tested. As we performed the measurement for I-V characteristics of diode, the Curve Fitting slide must be set to Non Linear. Next, when user clicked RUN button to start performing the acquisition, the LED of RATIOMETRIC STATE button is appeared green when the interface board is properly connected to computer and appeared red when it is not connected.

The voltage and current value appeared automatically and plotted on the I-V Characteristic graph. The fitting of I-V plot is plotted in real time until user push the STOP button to abort the execution. The graph provided a cursor for user to hover over the plotted line to find forward voltage drop of the diode as shown in Figure 4. The GUI gave user option in Curve fitting slide button to choose between linear and non-linear regression depends on what the characteristics of device the user testing.

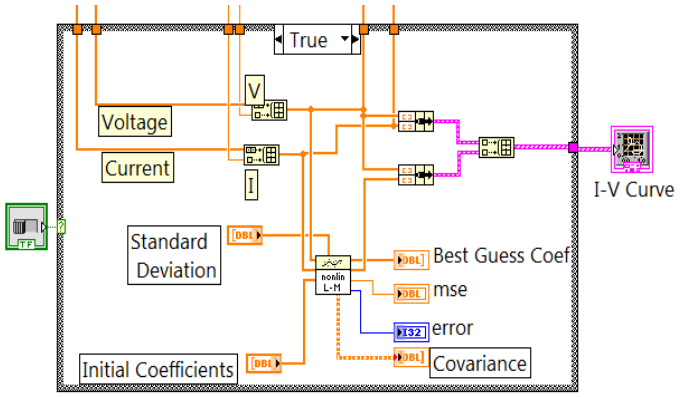


Figure 2(a). Block diagram of linear curve fitting

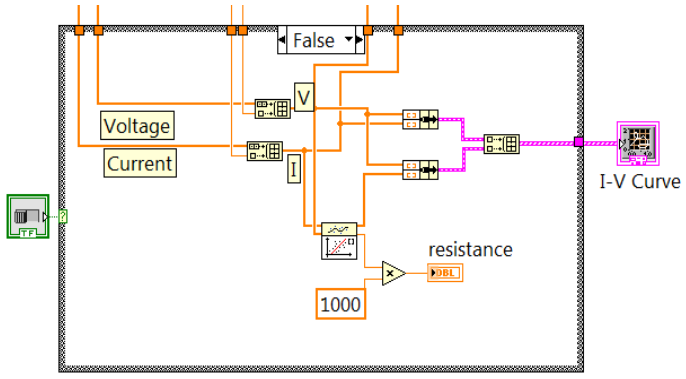


Figure 2(b). Block diagram of non-linear curve fitting

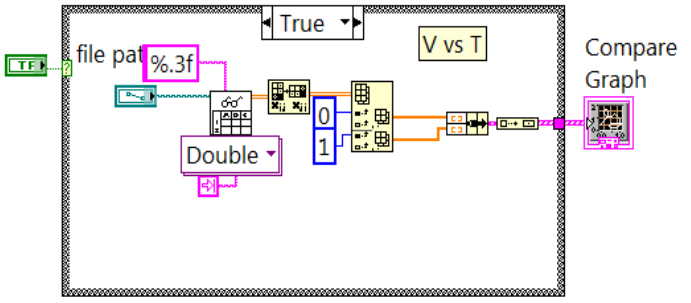


Figure 3. Functional block to read and plot graph from data file

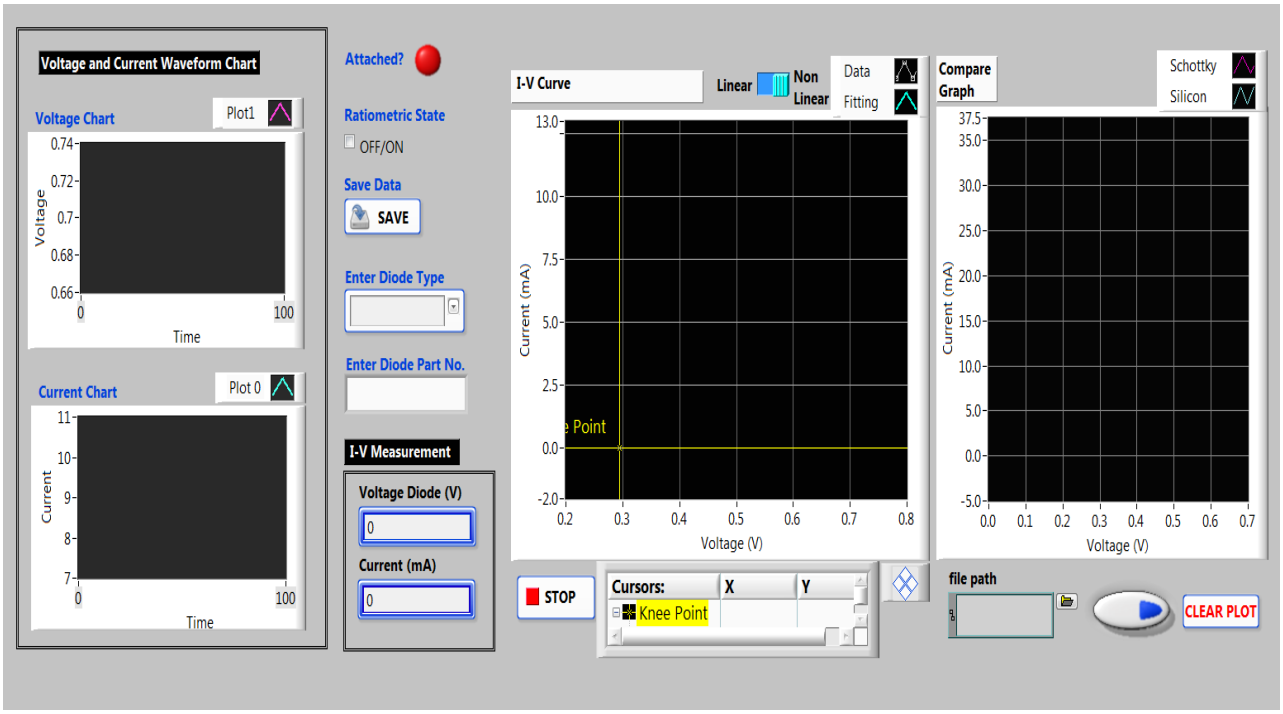


Figure 4. Graphical User Interface (GUI) developed in LabVIEW software

4. Analysis & Discussion

Recently in [1], silicon and Schottky diode has been studied which the system has shown the non-linear relationship of I-V characteristics. Upon studying the I-V graph, forward voltage drop of both diodes has been clarified which 0.7 volts for silicon and 0.4 volts for Schottky diode.

In this paper, the system has been tested on same silicon diode and characteristic of germanium diode is explored. The I-V characteristics graph plotted in Figure 5 can be seen that after the forward voltage drop, the current increases very rapidly where the device starts to conduct for both DUT. Below this voltage, current is less than 1% of maximum rated value of diode current [6]. Rated peak forward current for 1N34 and 1N4007 diode for most brand are 150mA and 30A [6,7]. Voltage increase slightly due to small internal dynamic resistance as the current change after the forward voltage drop is larger than the voltage change. The 1N34 germanium diode has much less defined result. The forward voltage drop is 0.4V which it is still shown low forward voltage drop. This germanium type of diode have low forward voltage drop which is fast switching speed and found uses in television and crystal radio application.

The silicon diode always expected to have value of forward voltage drop which is approximately 0.7 volts. The forward voltage drop was measured as 0.7 volts which this has support the theory behind its design. A germanium diode such as 1N34 have typically 0.3 volts forward voltage drop or low forward voltage drop means they are much more efficient compared to silicon diode. However, silicon diodes are resistant to heat, and better in terms of processing and stability. The results of this experiment appear to satisfy the theory behind diodes.

5. Conclusion

In this paper, a system for determination of current-voltage characteristics of diode has been developed and implemented. Based on LabVIEW software platform, this system measure current flow through the circuit and voltage across the diode to plot I-V characteristics of diode tested. From I-V plotted, forward voltage drop of silicon and germanium diode has been successfully determined. The possible research direction may include determination of voltage-current characteristics of any other two terminal components such as resistor or transistor. The ideality factor of diode can also be estimated by taking the natural log of current versus voltage plot.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors want to thank to the Ministry of Higher Education (MoHE), Malaysia, through the RAGS/1/2015/STO/UPSI/03/1 Research Acculturation Grant Scheme (RAGS), Universiti Pendidikan Sultan Idris (UPSI) for the financial support provided to this work.

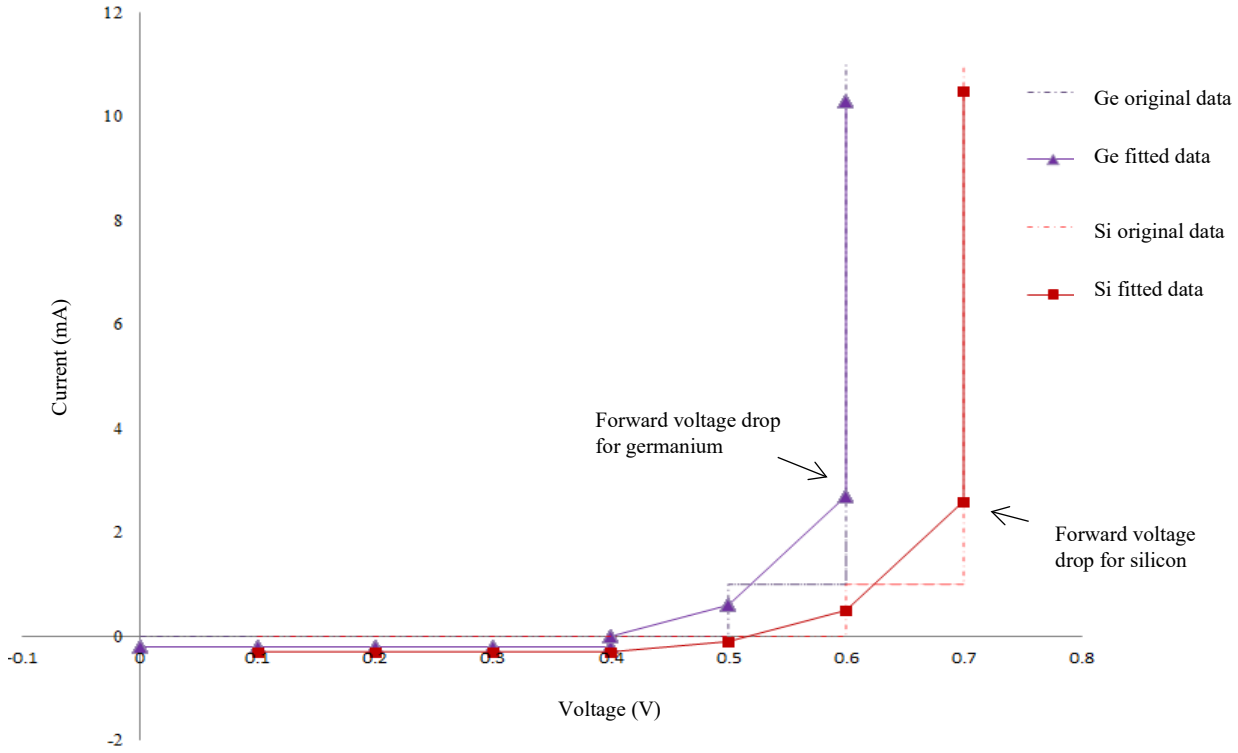


Figure 5. I-V characteristics graph of silicon and germanium diode

References

[1] N.s., M. S., Abdullah, N., & Yaacob, M. (2016). "Automated measurement system for diode I–V characterization," in 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE),pp. 498-501,2016.

[2] Sumathi, S., Surekha, P., & Surekha, P. (2007). LabVIEW based advanced instrumentation systems (Vol. 728). Berlin: Springer.

[3] J. Travis, J. King, LabVIEW for Everyone: Graphical Programming Made Easy and Fun, 2006.

[4] Bitter, R., Mohiuddin, T., & Nawrocki, M. (2006). LabVIEW: Advanced programming techniques. Crc Press.

[5] W. Shockley, "The theory of p-n junctions in semiconductors and p-n junction transistors," Bell Syst. Tech. J., vol. 28, p. 435, 1949.

[6] Bakshi, A. G. U. (2009). Basic Electronics. Technical Publication

[7] Semtech, "Point Contact Germanium Diode," 1N34A datasheet, n.d. {Retrieved from http://pdf1.alldatasheet.com/datasheet-pdf/view/42335/SEMTECH/1N34.html

[8] Vishay, "General Purpose Plastic Rectifier," 1N4001 thru 1N4007 datasheet, 2017 [Revised Feb. 2017].

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 352-356 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Innovative design with learning reflexiveness for developing the Hamiltonian circuit learning games

Meng-Chien Yang *, Hsuan-Yu, Chiang

[1] *Department of Computer Sciences and Communication Engineering, Providence University, Taichung, 43301, Taiwan ROC*

ARTICLE INFO

Article history:
Received: 30 November, 2017
Accepted: 07 January, 2018
Online: 02 February, 2018

Keywords :
Game-based learning,
Graph theories,
Learning reflexiveness,
Hamiltonian circuit

ABSTRACT

In this study, we use a new proposed framework to develop the Hamiltonian circuit learning games for college students. The framework is for enhancing learners' activities with learning reflexiveness. The design of these games is based on this framework to achieve the targeted learning outcomes. In recent years, the game-based learning is a very popular research topic. The Hamiltonian circuit is an important concepts for learning many computer science and electric engineering topics, such as IC design routing algorithm. The developed games use guiding rules to enable students to learn the Hamiltonian circuit in complicate graph problem. After the game, the learners are given a reviewing test which using the animation film for explaining the knowledge. This design concept is different from the previous studies. Through this new design, the outcome gets the better learning results under the effect of reflection. The students will have a deeper impression on the subject, and through self-learning and active thinking, in the game will have a deeper experience.

1. Introduction

The study of the paper is the research of the game-based learning technology. The game-based learning is an emerging research topics of the learning technology. Developing a video game as a learning tool might include many different tasks and considerations. Basically, how the game can be effective for learning is the main concern problem. This problem can arouse many challenging research problems such as the learning model of the game, the foci of the game, the players' attitude of the game and the game development. Many studies focused on the game developments such as the development of a new gaming technology, the usefulness of the gaming tool proposed by the game company and the development process of the learning game. Other studies might focus on the effectiveness of the game such as the learning process, the learners' attitudes. Moreover, some complicate topics of the game industry such as the immersive experience of the players brought research interests. Traditionally, the game for learning the specific subject is regarded as the serious game. Many different serious games have developed for different purpose of training and learning. Although the serious games have been used as the tool for education, most studies of the serious games remained on development of the functions of the game. Then how these games are effective and helpful for learning? This could be another challenging research topic. Our research goal is to find how to develop the effective learning method for game-based learning.

In this paper, a new framework for developing the learning game is proposed. Using advanced information technology, the design of digital games is booming in many applicable developments. Because of the unique interactive interesting and attractive features in digital games, it can arouse the students ' interest in learning, and improve the willingness to learn a specific knowledge. However, students are less likely to think about the contents after the game, but keep playing down. Usually, it is until the final online evaluation begins to help the students to recall the contents of the previous. The process of the final review is important for the design. This will not only can immediately check whether the students have learned the knowledge, but also can let students have the chance through self-initiative to get the targeted knowledge. Based on the above consideration, this study proposes a new method for developing the game for learning by using the Role-Playing Game (RPG), chapters of textbooks and story development.

*Meng-Chien Yang, Department of Computer Sciences and Communication Engineering, Providence University. Email: mcyang2@pu.edu.tw

In the beginning, we design the guiding rules to lead students to play the game. Suppose that there is a specific topic prepared for learning. The students are told to play the game linked to this topic. After the game, the learners are given a review test for enhancing the learning results. Meanwhile, we using a set of animation films to assist the learners. These films can be traced back to the pre-test which is given before the students play the game. Then, after the review test students will watch the related film with the content of the learning topic. This will strengthen the concept of knowledge link with the students in this topic.

In this paper, we use the above framework to design the game for learning an important subject: Hamiltonian circuit of the graph theory in discrete mathematics. The Hamiltonian circuit problem is to find an existing loop to visiting each vertex once on the graph. The Hamiltonian circuit problem is an important problem and can be applied to many different technological applications. To develop a game for learning the Hamiltonian circuit problem is an important problem both in computer science education and mathematical education.

In recent years, developing math games have become the popular research topic. The importance of mathematical games in mathematics education cannot be ignored. A practical approach of using the game for teaching math in the country junior high school was reported [1]. In addition, the author pointed out that the mathematical game teaching method has the following advantages and functions:

(1) to stimulate students' interest and motivation to learn mathematics [2];

(2) to help students from the specific experience to establish basic mathematical concepts and skills;

(3) to cultivate students to solve the problem of flexible and deep reasoning thinking ability

In another paper [3], the experimental results and infield tests showed that the existence of mathematical games have considerable potential. Our paper follows these previous studies and will continue to explore the learning changes brought by this new design approach. Based on the hypothesis of the above research project, we hope our study of developing the game of the Hamiltonian circuit problem can help the students learning this problem. In our study, we had invited students who taking the course of the graph theory to play the game. The testing results showed our proposed framework can develop the effectiveness game to help student learning.

The remaining of this paper is organized as the follows. Section 2 discuss related studies of the game-based learning. Section 3 describe the design of our proposed framework. Section 4 is the description of the game development and system modules of the proposed game for the Hamiltonian circuit problem. Section 5 is the report of experimental results and discussion. The conclusions are put in Section 6.

2. Studies of Serious Game and Game-Based Learning

The research projects related to our study are the studies of the serious game. Based on the definition of the famous game design textbook, the serious games are those games for education and study, simulation, health and growth [4]. Serious game gives the players fun experiences and other explorations. Serious game design is built in the "fun" under the premise of the players in the process of the game, while achieving a sense of accomplishment, knowledge and training, and further use of the nature of media and entertainment to convey the effect of information and education [5][6]. A famous study argued that the practice of the serious game has the meaning of learning and is a systematic learning, so that students can learn from the experience of the game [7].

There are new studies that "video games" will be a very influential tool for learning [8][9].Game-based learning is not a regular learning style but a revolution for education. It can increase the learner's motivation to learn and help learners to learn effectively to improve learning motivation and effectiveness. Learners playing the game, easily immerse into the game and the learning scenarios because of game challenges, level design or script story into the immersive state. Therefore, game learning can follow the experience of "heart flow" [10], a process to obtain the knowledge, so that learners immersed in the game. The theory of flow suggests that when learners are integrated into the task, they produce a state of mind that is isolated from the outside world. If students are in such a state of mind, students will have the highest motivation to learn.

The study of the serious game development could have many different aspects. Recently a reviewing paper [11] described some essential features of the serious games for education. In this paper, the authors stated that two major focus of the serious game are learners' attitude and game core mechanics. Another important study about how to use the serious game for collegiate computer science students is to use the game with course materials [12]. The development of the serious game might need the guidelines for design the game. A recent study [6] proposed the guidelines and stated the key factors of the serious game development. Based on the above studies, several factor could influence the design and implementation of the learning game. First, the connection between the learning materials and the learning game should be tight. Second, the development of the game should follow a guideline that can make the game to become a useful learning materials. Then, the third, the learning outcome should be evaluated with different and various considerations and assumptions. Based on these assumptions, our study focus on designing a new framework for developing the learning game. In this framework, we developed a new learning outcome called learning reflexiveness which can help the learners to enhance the learners' attitudes in the game-based learning process. The framework can help designer to develop the serious learning games.

3. Methodology and Framework of the Game Design

The design of our framework was initiated with the concepts of the flow simulation proposed by Csikszentmihalyi [10], [13]. When human are involve the activities with flow based on the theory proposed by Csikszentmihalyi, people can concentrate on these activities and have the immediate feedback. The flow simulation had been used the development of the playground and video games. The activities with flow can bring self-consciousness of awarding the related information and context. Based on this theoretical finding, we attempt developing the learning process with the game. We describe our design by the question what is a

good educational learning game? This will provide scaffolding for our proposed framework.

A good educational game can motivate the learners through the design of the presumptive learning situations. The presumptive learning situations are a set of conditional factors for identifying the initial condition of each learning level. The learning situation and RPG game scenarios are the main settings of the game. Meanwhile, in the domain of the game, the story narration is a factor that attracts the learners from being immersed in it. We create stories and game rules for the game. The learners follow the guideline proposed by [14] in the symbol of the code, through the code to show the navigation keys to read the story. The strategy of game design for each level should be in a gradually better mode that can help the learner to immerse into the story of the game. Following previous study, the game prepares easy, moderate and difficulty challenges to enhance the learners' self-esteem [15]. The design is to design a learning situation during learning low-challenge tasks to learning high-challenge tasks. We developed the setting of the game by adopting the concept of flow in different working environments.

We had study the game development for several years. We analyzed and found the following three motivational factors that help learners to play the games [16].

1. Challenge

Each game level can be design with a particularly challenge. The challenge will entice learners to continue the game. The design of challenges is to describe a clear game goal, then an uncertain outcome, and a related feedback. The different challenges can be design with different learning materials.

2. Curiosity

The interface on the screen and sound control can cause learners interest. The setting of the game world can be designed as an environment that the players can explore new items.

3. Fantasy

Fantasy environment is an important factor. It stimulates the intrinsic motivation of learners in two ways, one is to satisfy the learner's inner emotion, and the second is to meet the learners' cognitive needs, and to promote the learners to grasp the relationships in the game.

With these three factors, we add a new factor into the game development. The new factor is the reflexiveness of the learner. Once the learner is playing the game, the outcome and the challenges can help the learner to have the learning attitude to adjust the learning paths toward the meaningful goal. The reflexiveness for learning had been mentioned in the studies of mathematics education.

3.1. The Importance of Reflexiveness for Mathematics Learning

Reflexiveness mathematics learning refers to the fact that it is an effective way to learn mathematics through the reflection of the process of math learning activities. It is not only a general review or repetition of mathematics, but a deep understanding of mathematics activities involved in the knowledge, methods, ideas, and strategies and so on. The general operational mathematics is based on "learning knowledge" as the main purpose, concerned about the current academic performance; reflexiveness mathematics is based on "learning to learn" for the purpose of focusing on the current academic performance and students' future development [17].

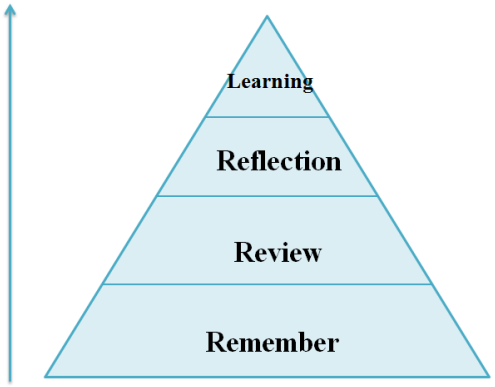


Figure 1: Reflectiveness of four learning levels

3.2. The reflexiveness in the proposed learning

Based on above discussions of the reflexiveness in mathematics learning, the reflexiveness can be applied to the other learning subject. The reflexiveness in the game-based learning can be illustrated as four different levels shown on Figure 1. The bottom level is the learning for remembering. This can be the basic of reflexiveness and are used as the fundamentals for accumulating knowledge. The second level is the learning with the reviewing content. The function of the second level is to enhance the effectiveness of the learning. The third is reflection which are the main level of our design. This level is to find the correct ways of learning the targeted subjects or courses. The final level is the effective learning process that the learners can successfully learn this subject.

We used the proposed framework to develop a game for learning the Hamiltonian circuit problem. The Hamiltonian circuit is an important graph in the graphic theory. The Hamiltonian circuit problem can be taught in the collegiate courses such as discrete mathematics, graph theory, combinatorial mathematics. The Hamiltonian circuit problem is an important concept in many computer science applications like logic design, routing problem. In next section, we describe our game that following the proposed framework.

4. Game Design and Game Play

In this section, the system design of the game is illustrated. At the beginning, the whole structure the game is to design with an RPG game structure combined with the fairy tale. Figure 2 shows the main menu of the game. The game is described as several chapters with different concepts of the Hamilton circuit. At the beginning of each chapter the learners will receive guidance rules. The guiding rules are put in a short RPG game and with a short introduction video clip. When the learner plays the RPG game, he can get the game rules. Using these rules, the learners can play and find the Hamilton circuit in a given graph (Figure 3). If the learner can successfully find the Hamilton circuit in the graph, the learner

can be guided to complete the chapter with more similar graphs. When the learners finish the chapter of the game, there is a chapter reviewing quiz which helps the learners to recall the process of gameplay. The learners can get an immediate feedback here and understand the correct answers which are the key terms of this chapter.



Figure 2 Main menu of the game

The reflexiveness setting in the whole scenario of the gameplay is a cyclic framework shown in Figure 4. In this framework, the learner can find the learning content in four different styles. The guide rules and the RPG game are used to attract the learner’ s attention. The game process is the main challenge and help the learner to understand the Hamilton circuit. The review test can help the learner to enhance the learning results. The teaching video can keep the learner’s interest and go to next chapter which is more difficult Hamiltonian graph.

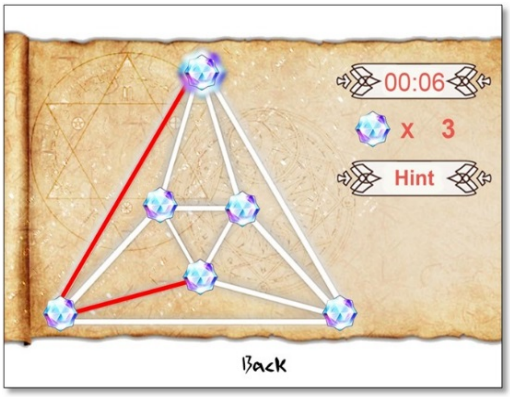


Figure 3: Gameplay for finding the Hamiltonian circuit

In addition, the reflective learning will affect the learners in two factors of each cycle. First, in the review of the test (Figure 5), learners need to recall the game experiences for answering these questions correctly. Second, the animated films with chapter concept guide, learners can link the game experience and answer process. An interesting design here is that the outcome of the story will vary depending on the score of the learner's game, which gives the learner the motivation to repeat the game.

Reflective learning should help the learners to get attention to the guidance and encouragement in the learning process. When the learners get the guide rules after the game, and in the review of the test to reflect on the game process and guide the rules of the tips, organize the answer to answer questions. When learners watch the teaching film for the second time of reflective activities, the learners will have reflective experience based on the pasting activities.

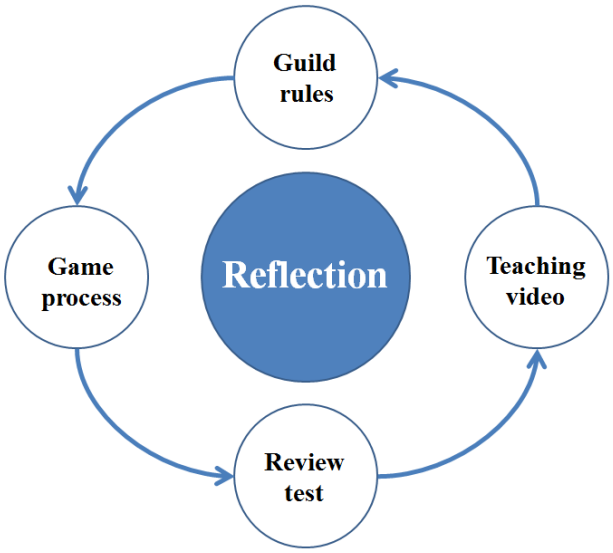


Figure 4: Game cycle that includes reflective learning.

The game is implemented with this system design. In the game, six cycles for the easy to complicate knowledge of the Hamilton circuit are implemented. After the review test, the learners can watch an animated film as the enhancement for learning reflection.

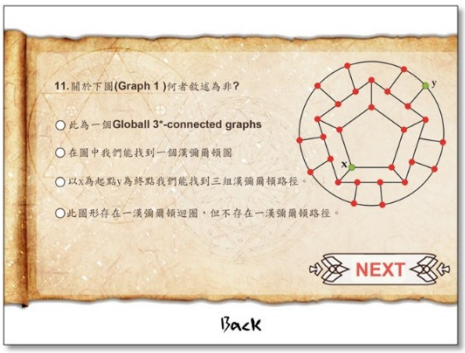


Figure 5: Screen shot of the online reviewing test

5. Implementation and Evaluation

The learning game was implemented and evaluated by a user–based evaluation and an expert-based working through evaluation. The approach was similar to a previous study of the educational serious game development [12]. A group of computer science major college students who taking the graph theory course were invited to participate the user-based evaluation. Our game was installed in a tablet computer and each participant was given 10 minutes to play the first three chapters of the game and then answered the quiz after the game. If the student could finish these three chapters successfully, the student would be given a big reviewing test and were asked to finish the remaining three chapters of the game.

The total number of students participating this experiment were 34. When the students finished the first three chapters of the

game, they were given a writing test for evaluating the basic concepts of the Hamiltonian circuit. When they finish the final chapter, the sixth chapter, the students were given a writing test with the complex knowledge of Hamiltonian circuits.

The experimental results showed 19 students had finished the first three chapters of the game and passed the basic test. Then, 15 students had finished the whole six chapters of the game and passed the advance test. Based on the evaluating, the developed game can be useful as a tool for teaching Hamiltonian circuit.

The second evaluation process is the expert walk-through test. We invited 8 computer major graduate students who had already understood the Hamiltonian circuit. When these students played the game, they were asked to find the shortcoming of the usability of the interface and evaluating the gameplay. Most students agreed that the game can help the beginner to learn the Hamiltonian circuit. Several mentioned the advanced part, the last three chapter should be improved with more illustration figures to help the learners play the games. Based on the suggestions, we are certain that the proposed framework can foster a comprehensive and effective learning.

6. Conclusion

In this paper, a framework for developing the learning games for advance computer science concept is proposed. In this framework, the learning process is arranged as a cycle with playing game, answering the review questions, watching the animation clips and being given the guiding rules. The framework is developed based on the concept of the flow and the fundamentals of the game theory. In addition, the reflexive review of the mathematical education is put in the framework and we coin a new learning attitude, reflexiveness. The reflexiveness design in the learning game can help the learners achieve the learning goal via playing.

We used this framework for developing the game for learning the comprehensive knowledge of Hamiltonian circuit. The results of evaluation showed that the game can help these participants to learn the Hamiltonian circuit. The game can be a useful teaching tool for courses related to graph theory and computer science. The game can be used for other purpose such that training course for novices to learn this concept at the IC design company.

The project has two successful results. First, we have develop a framework that can be useful as a guideline for other research studies who focus on developing the serious games for certain subject. Second, the game for teaching the Hamiltonian circuit can be useful as the teaching tool. In the next studies, we will explore and find more evidences that proposed framework is effective for the developing the serious learning game.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research was supported in part by Providence University 2017 grant for graduate student.

References

[1] C. W. Rau, Game-Based Teaching for Elementary Math Courses.Wu-Nan Culture Publish Co. 1996.

[2] E.R. Lai, Motivation: A Literature Review. Pearson. Research Report. Pearson's publications 2011.

[3] C. W. Lai, & P. J. Lin, Immersion of Mathematic Games for high school math function courses. Master Thesis, National Hsinchu Educational University, 2011.

[4] E. Adams, Fundamental of Game Design 3rd, New Riders, 2015.

[5] A. Singhal, & E. M. Rogers, "A theoretical agenda for entertainment—education." *Communication theory*, *12*(2), 117-135, 2002.

[6] E., Tsekleves, J. Cosmas, & A. Aggoun, "Benefits, barriers and guideline recommendations for the implementation of serious games in education for stakeholders and policymakers." *British Journal of Educational Technology*, *47*(1), 164-183, 2016.

[7] A. Derryberry, "Serious Games: Online Games for Learning," Adobe White Paper 2007.

[8] K. Humphrey, "The application of a serious, non-digital escape game learning experience in higher education," *Sport & Exercise Psychology Review*, *13*(2), 48-54, 2017.

[9] F. W. Kron, C. L. Gjerde, A. Sen, M. D. Fetters, "Medical student attitudes toward video games and related new media technologies in medical education." *BMC Medical Education*, *10*(1), 50, 1-11, 2010.

[10] M. Csikszentmihalyi , Creativity: Flow and the Psychology of Discovery and Invention, New York: Harper Perennial, 1996

[11] P. Lameras, S. Arnab, I. Dunwell, C. Stewart, S. Clarke, & P. Petridis, "Essential features of serious games design in higher education: Linking learning attributes to game mechanics." *British Journal of Educational Technology*, *48*(4), 972-994, 2017.

[12] A. Schäfer, J. Holz, T., Leonhardt, U. Schroeder, P. Brauner, & M. Ziefle, "From boring to scoring – a collaborative serious game for learning and practicing mathematical logic for computer science education." *Computer Science Education*, *23*(2), 87-111, 2013.

[13] J. Nakamura, & M. Csikszentmihalyi, "The concept of flow." In *Flow and the foundations of positive psychology* (pp. 239-263). Springer Netherlands, 2014.

[14] M. P. Chen," Learner's attitude and evaluation for programming language courses", Journal of National Normal University, 52, 1-21, 2007.

[15] T. W. Malone, & M. R. Lepper, "Making learning fun: A taxonomy of intrinsic motivations for learning." *Aptitude, learning, and instruction*, *3*, 223-253, 1987.

[16] M. Papastergious, "Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation", Computers and Education, 52(1), 1-12.2009.

[17] R. B. Tu, "Discussion of Reflexiveness for Mathematic Education." Journal of Mathematic Education, 4(3), 2000.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 357-365 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Theoretical Investigation of Combined Use of PSO, Tabu Search and Lagrangian Relaxation methods to solve the Unit Commitment Problem

Sahbi Marrouchi*, Nesrine Amor, Moez Ben Hessine, Souad Chebbi

Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE), National Superior School of Engineers of Tunis (ENSIT), University of Tunis, 5 Street Taha Hussein Montfleurie, 1008 Tunis, Tunisia

ARTICLE INFO

Article history:
Received: 30 November, 2017
Accepted: 15 January, 2018
Online: 02 February, 2018

Keywords:
Unit commitment
Optimization methods
Scheduling
TS-PSO-LR

ABSTRACT

Solving the Unit Commitment problem (UCP) optimizes the combination of production units operations and determines the appropriate operational scheduling of each production units to satisfy the expected consumption which varies from one day to one month. Besides, each production unit is conducted to constraints that render this problem complex, combinatorial and nonlinear. In this paper, we proposed a new strategy based on the combination three optimization methods: Tabu search, Particle swarm optimization and Lagrangian relaxation methods in order to develop a proper unit commitment scheduling of the production units while reducing the production cost during a definite period. The proposed strategy has been implemented on a the IEEE 9 bus test system containing 3 production unit and the results were promising compared to strategies based on meta-heuristic and deterministic methods.

1. Introduction

The main role of energy management is to ensure the production of active power in order to respond to the demand growth among a very little fuel cost [1,2]. Solving the Unit Commitment Problem (UCP) is very essential in electrical network planning. It can both optimize the daily operational planning of networks and reduce the total production cost among the improvement of the operating state of each unit leading to obtain the best unit commitment scheduling helping to respond to the power demand. Operations scheduling production units or Unit Commitment (UC) improve operational planning of the electrical grid while ensuring continuity of service [1-6]. The main purpose of solving the Unit Commitment problem is to schedule production units to respond to the consumed power with the minimization of the total production cost. The optimal planning [7-9] involves ensuring a better use of available generators subject to various constraints and guaranteeing the transfer of electrical energy from generating stations to the load. UC must satisfy the load demand, storage capability, minimum downtime startup and safety limits for each production unit.

The production scheduling comprises determining startup and each generation level for each unit in a given planning period [10-13]. Therefore, a study of literature [14,15] on methods which focus on unit commitment (UC) problem resolution shows that various optimization methods have examined this subject. Furthermore, Sasaki et al. demonstrated the possibility to use artificial neural network (ANN) to solve the UCP in which a large number of inequality constraints is processed. They have used the ANN to schedule generators and the dynamic programming to solve the load flow problem. The adopted strategy was compared to Lagrangian Relaxation (LR) and dynamic programming (DP) methods and the results offered a faster and cheaper solution compared to the LR and DP but it suffers from digital convergence because of the learning process. Certain works [16, 17] proposed a strategy based on tabu search method. They introduced new rules to generate an initial solution feasible to solve the Unit Commitment problem. This strategy consists on dividing the problem into two problems: the first combinatorial optimization problem is solved using tabu search algorithm and the second is a problem of nonlinear programming solved through the quadratic

*Corresponding Author: Sahbi Marrouchi, National Superior School of Engineers of Tunis (ENSIT), University of Tunis, Tunis, Tunisia
Email: sahbimarrouchi@yahoo.fr

programming routine. The structure resolution through Tabu search method is similar to that used by simulated annealing [18] even though TS is provided with a simplified configuration, so it is easy to pass from one optimization to the other. Indeed, the main advantage of the adopted strategy is to extend the search space provided for the best optimal solutions which are stored in the tabu list. This method has provided a lower production cost solution, but it's slower compared to the Lagrangian relaxation. However, T. Logenthiran et al. [19] have proposed a new approach based on particle swarm optimization (PSO) algorithm for solving the unit commitment problem. They presented three versions of particle swarm: binary particle swarm optimization (BPSO), improved binary particle swarm optimization (IBPSO) and combined use of particle swarm optimization and Lagrangian relaxation programming (LR-PSO) .The numerical results show that LR-PSO method has provided a lower production cost solution compared to LR, BPSO and IBPSO especially when the number of units important. Whereas, if the number of units is small, BPSO is taken as the best method since it has the lowest production cost compared to other algorithms. Other works, [20, 21] presented new approaches based on artificial intelligence to solve the UCP. The adopted approach combines two methods: tabu search and neural networks (ANN-TS) in order to get an optimal unit commitment scheduling allowing a minimal production cost in accordance to the constraints of the studied system. Artificial Neural networks provide a fast convergence to optimal solutions but it takes a lot of memory space because of the great number of constraints.

Cheng et al. [22] proposed a hybrid method based on the integration of the genetic algorithm in Lagrangian Relaxation Programming (LR-GA) to solve the problem of the planning of the operations of the production units. This integration consists in improving the Lagrange multipliers using the operators of the genetic algorithm to find a fast and an effective cost solution respecting all the constraints of the system. The implementation of this method requires two steps ; the first is to look for the minimum constraints of the Lagrange function under the multipliers constraint through dynamic programming. The second step consists on maximizing Lagrange's function while respecting the multiplier adjusted by the genetic algorithms. The experimental result of this method provides a faster and cheaper solution compared to the Lagrangian relaxation method (LR) and the tabu search method (TS). However, C. Christober et al. [23] presented a hybrid method combining the evolutionary programming the tabu search methods (EP-TS). The first has the advantage of a good convergence property, a significant acceleration based on the principle of traditional genetic algorithm and a high quality of solutions, but its major disadvantage is related to the dimensioning of the Unit Commitment problem. Tabu search method improves the status by avoiding imprisonment in the local minimum. The best solution is chosen by evolutionary strategy. Thus, the effort has been made to combine these two methods whose purpose is to meet the requirements of the generators commitment problem. The numerical results demonstrate that this method is efficient and accurate in terms of calculation time and minimization of the total production cost compared to the following methods: simulated annealing, taboo search, dynamic programming, evolutionary programming, Lagrange programming and LR- GA. By contrast, Kumar et al. [24] presented a hybrid method combining dynamic programming with Hopfield Neural Networks (DP-HNN). The proposed process consists on two steps: use of Hopfield neural networks from direct computation to generate the economic distribution (ED) and use of dynamic programming (DP) to plan generators. This approach provides a poor final solution in total of production cost over other methods, but is faster than LR, DP, GA, LR-GA and AS-GA. In addition, C. Asir et al. [25] presented a new approach based on artificial intelligence to solve the problem of allocation of units. This combines two methods: tabu research and Artificial Neural Networks (ANN-TS) and this in order to have an optimal solution that solves the problem of planning power with a minimum of total production cost with respecting all the constraints of the specified system. Neural networks provide a fast convergence solution but the programming of the algorithm takes up a lot of memory space because of the constraints of the problem. The tabu search is characterized by the flexibility of its memory and it is able to find good solutions. This algorithm gave a faster and cheaper result compared to tabu search, dynamic programming, Artificial neural networks, Lagrange programming, LR-GA, TS-GA and EP-TS methods. In addition, C. Asir [26] has developed another strategy which consists in integrating the tabu search with the genetic algorithm. The purpose of this technique is to find the planning of the production. This strategy depends on the exploitation of the total cost which can be minimized when it is subjected to a set of constraints. Tabu search can find good solutions using the tabu list while genetic algorithm is used to generate new solutions using its operators. The results demonstrate that this technique provides a low cost but slow solutions compared to ANN-TS. Alma et al. [27] proposed a hybrid approach combining fuzzy logic and the genetic algorithm (FL-AG) to solve the Unit Commitment problem. Genetic algorithm uses the actual coded chromosomes in contradiction with the most commonly used scheme which is binary coding. This method uses a strict priority order list in the genetic algorithm that generates different solutions. This list serves to reduce the size of the search space of the problem, while fuzzy optimization guides the entire search process in an uncertain environment (varying from load demand, renewable energy sources etc ...).The results of this technique offer a good final solution compared to dynamic programming and to the genetic algorithm.

According to our study, we thought to validate an approach to apprehend the whole unit commitment problem. To achieve this objective, our strategy for solving the Unit Commitment Problem is based on the combination of three stochastic optimization methods that are the Particle Swarm Optimization (PSO), the Tabu Search (TS) and Lagrangian Relaxation (LR) method in order to develop a proper unit commitment scheduling of the production units to minimize the production cost.

2. Notation

The notation used throughout the paper is stated below.

a_i, b_i, c_i : Coefficients of the production cost,

P_{ih} : Active power generated by the i^{th} unit h^{th} hour, $i = 1,2,3,....,N_g$ and $h = 1,2,3,....,H$

U_{ih} : On/Off status of the i^{th} production unit at the h^{th} hour, $U_{ih} = 0$ for the off state of one generating unit and $U_{ih} = 1$ for the operating status of one generating unit,

HSC_i : Hot start-up cost of the i^{th} unit,

CSC_i : Cold start-up cost of the i^{th} unit,

P_{rh} : System spinning reserve at the h^{th} hour,

P_{dh} : Amount of the consumed power at the h^{th} hour,

P_{Lh} : Total active losses at the h^{th} hour,

P_i^{min} : Minimum and maximum power produced by one generator,

P_i^{max} : Maximum power produced by one generator,

MUT_i : Continuously on-time of unit i .

MDT_i : Continuously down-time of unit i .

τ_i^{OFF} : Continuously off-time of unit i ,

SC_i : Cold start time of unit i .

N_g : Number of generating units,

H : Time horizon for UC (h).

3. Problem Formulation

Many works have been based on an analytical statement of the unit commitment problem [2,3,11, 18, 22]. We present in this paper a mathematical model of the unit commitment problem with limited security. This model is a mixed linear and constrained which has been adapted in several works [3, 9, 13, 15].

$$Min \left[F_T(P_{ih},U_{ih}) = \sum_{i=1}^{N_g} \sum_{h=1}^{H} [a_i P_{ih}^2 + b_i P_{ih} + c_i + ST_i(1-U_{i(h-1)})]U_{ih} \right] \quad (1)$$

Where;

ST_i: The starting cost of the i^{th} unit defined by:

$$ST_i = \begin{cases} HSC_i \ \ if \ \ MDT_i \le \tau_i^{OFF} \le MDT_i + SC_i \\ CSC_i \ \ if \ \ \tau_i^{OFF} \succ MDT_i + SC_i \end{cases} \quad (2)$$

The minimisation of the objective function is provided with the following constraints:

- System Constraints
  - Power balance constraints

$$\sum_{i=1}^{N_g} P_{ih} U_{ih} = P_{dh} \quad (3)$$

  - Spinning reserve constraints

$$P_{dh} + P_{rh} - \sum_{i=1}^{N_g} U_{ih} P_{ih} \le 0 \quad (4)$$

- Unit Constraints
  - Genration limits

$$P_i^{min}.U_i \le P_{ih}.U_i \le P_i^{max}.U_i \quad (5)$$

  - Minimum up-time constraint

$$U_{ih} = 1 \quad for \sum_{t=h-up_i}^{h-1} U_{ih} \le MUT_i \quad (6)$$

  - Minimum down-time constraint

$$U_{ih} = 0 \quad for \sum_{t=h-down_i}^{h-1} U_{ih} \le MDT_i \quad (7)$$

4. Methodology of resolution

In this paper, four optimization methods are available to solve the unit commitment problem; the first one uses the Particle Swarm Optimization (PSO). This strategy takes into account the advantage of PSO method for solving complex and nonlinear problems. The second method relies on the use of the Tabu Search approach (TS). The use of the TS approach is depicted to the flexibility of storage great memory of optimal solutions offered by this method. The third strategy shows the advantage of the Lagrangian relaxation providing the best convergence speed. Our strategy for solving the Unit Commitment Problem is based on the combination of three optimization methods that are the Particle Swarm Optimization (PSO), the Tabu Search (TS) and Lagrangian Relaxation (LR) method to find a good On / Off states scheduling of each production unit over a period of time leading to obtain a good production cost.

4.1. Particle Swarm Optimization

Particle swarm optimization provides a population based search procedure in which individuals called particles change their positions with time. This method is able to generate high quality of solutions within shorter calculation time and stable convergence characteristic than other stochastic optimization methods. The PSO model consists of a swarm of particles moving, figure 1, in a definite dimensional real-valued space of possible problem solutions [9,28,29].

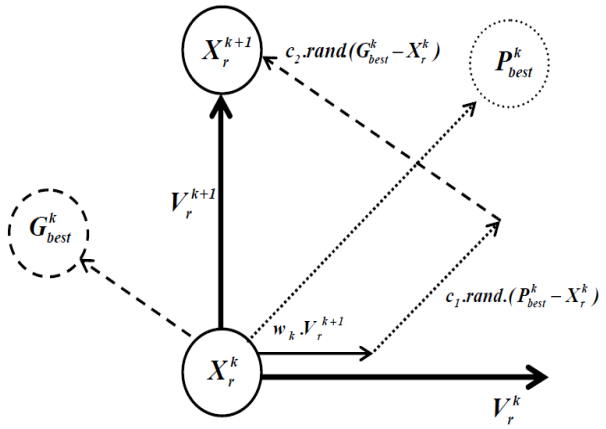


Figure 1. PSO research process

Every particle has a position $X_i = (x_i^1, x_i^2, ..., x_i^l)$ and a flight velocity $V_i = (v_i^1, v_i^2, ..., v_i^l)$. Indeed, each particle has its own best positions $P_{ibest} = (P_{ibest}^1, P_{ibest}^2, ..., P_{ibest}^l)$ and a global best position $G_{best} = (G_{best}^1, G_{best}^2, ..., G_{best}^l)$. Each time step is characterized by the update of the velocity and the particle is moved to a new position which is the sum of the previous position and the new velocity as shown in the following expression:

$$X_r^{k+1} = X_r^k + V_r^{k+1} \quad (8)$$

The update of the velocity from one particle to another is given by:

$$V_r^{k+1} = w_k.V_r^{k+1} + c_1.rand.(P_{best}^k - X_r^k) + c_2.rand.(G_{best}^k - X_r^k) \quad (9)$$

Where, c_1 et c_2 are acceleration constant, *rand* is a uniform random value between $[0,1]$, X_r^k and V_r^k are respectively the position and the velocity of one particle i at iteration k. w_k is the inertia weight factor defined by the following equation:

$$w_k = w_{max} - \frac{w_{max} - w_{min}}{k_{max}}.k \quad (10)$$

Where, w_{max} and w_{min} are the maximum and the minimum inertia weight factors respectively and k_{max} is the maximum number of iterations.

4.2. Tabu search

Tabu search uses a local or neighborhood search procedure to iteratively move from a solution X to a solution X' in the neighborhood of X, until some stopping criterion has been satisfied. To explore regions of the search space that would be left unexplored by the local search procedure, TS modifies the neighborhood structure of each solution as the search progresses [16,17]. The search for the optimal solution corresponding to minimal production cost consists on repeating an iterative process until reaching a stop criterion so as to find one solution neighbor to the optimal one as shown to the following equation:

$$P_{ibest} = P_{best} + \left[(b_i + 2.c_i.P_{best}).\left(\frac{P_i^{max} - P_i^{min}}{2} \right) \right] \quad (11)$$

The new neighborhood solutions $N^*(X)$ are determined through the use of memory structures. The search then progresses by iteratively moving from a solution X to a solution X' in $N^*(X)$. To determine the solutions admitted to $N^*(X)$, a tabu list (TL) memory is used, which is a short-term memory containing the solutions that have been visited in the recent past as less than the maximum number of iterations.

4.3. Lagrangian Relaxation

The Lagrangian relaxation solves the Unit commitment problem by relaxing or temporarily ignoring the constraints, power balance and spinning reserve requirements [13,15,30]. Therefore, to transform the complex nonlinear constrained problem into a linear unconstrained problem, we have considered the following Lagrangian function:

$$L(P_{ih}, U_i, \lambda_i) = \sum_{i=1}^{N_g}\sum_{h=1}^{H}[a_i P_{ih}^2 + b_i P_{ih} + c_i + ST_i(1 - U_{i(h-1)})]U_{ih} + \lambda_i.(P_d - \sum_{i=1}^{N_g} P_i U_{ih}) \quad (12)$$

Where, λ_i is the Lagrangian coefficient.

To establish our strategy, we have considered the partial derivatives of the Lagrangian function (12) with respect to each of the controllable variables equal to zero.

$$\frac{\partial L}{\partial P_{ih}} = \frac{\partial\left[[a_i P_{ih}^2 + b_i P_{ih} + c_i + ST_i(1 - U_{i(h-1)})]U_{ih}\right]}{\partial P_{ih}} - \lambda_i\left(\frac{\partial P_{dh}}{\partial P_{ih}} - U_{ih}\right) = 0 \quad (13)$$

$$\frac{\partial L}{\partial \lambda_i} = P_{dh} - \sum_{i=1}^{N_g} P_i U_{ih} = 0 \quad (14)$$

Equations (13) and (14) represent the optimality conditions necessary to solve equation systems (1) and (3) without using inequality constraints (equations (4) and (5)). Equation (13) can be written as follows:

$$\lambda_i = \frac{\frac{\partial\left[a_i P_{ih}^2 + b_i P_{ih} + c_i + ST_i(1 - U_{i(h-1)})]U_{ih}\right]}{\partial P_{ih}}}{\frac{\partial P_{dh}}{\partial P_{ih}} - U_{ih}} \quad ; i = 1,\ldots,N_G; h = 1,\ldots,H \quad (15)$$

4.4. Proposed Strategy

The process of the Unit Commitment problem resolution by the combined use of Tabu search, Particle swarm optimization and Lagrangian Relaxation (TS-PSO-LR) methods is carried out according to the flowchart in Figure 2.

The proposed strategy not only helps to reach the optimal solution as quickly as possible using the speed of the Lagrangian relaxation but also to proceed through PSO method to search effective solutions corresponding to a minimum production cost and this is obtained through a specific determination of the new velocity and then the next best position corresponding to the best amount of generated power produced by each unit when it's in the ON state.

In the proposed method, it's notable that P_{ibest} representing the best information of each particle and the history of each generated power P_{gi} of each production unit is preserved in the list $P_{best}List$. Herein, in spite of the possibility of the PSO method with the solution better than G_{best}^k around P_{best}^k, there is the possibility not to be searched enough that's why we have thought to use the history of P_{ibest}^l in $P_{best}List$.

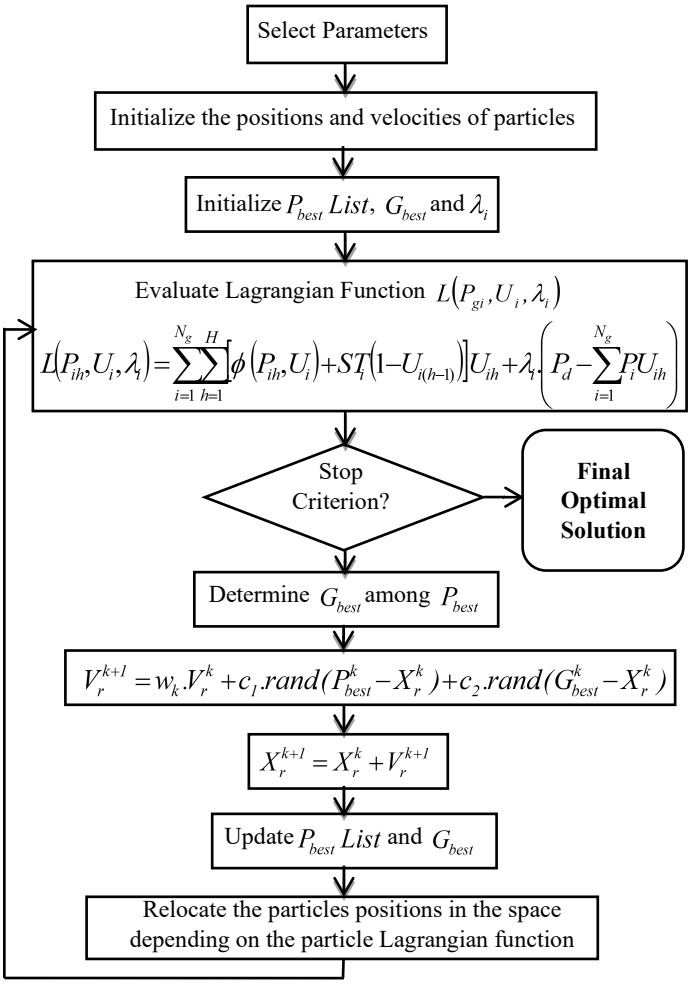


Figure 2. Flowchart of solving the unit commitment problem via Tabu Search, Particle Swarm Optimization and Lagrangian Relaxation

Whenever the particles lose the searching ability when the velocity V_r^k of one particle is very small, the TS-PSO-LR algorithm adapts the other $(P_{ibest}^2,...,P_{ibest}^l)$ instead of P_{ibest}^1 to update equation of velocity. This action increases the searching ability and helps to find more optimal solutions enabling a minimal production cost while considering a best unit commitment scheduling.

The proposed TS-PSO-LR strategy differs from other evolutionary computing techniques in providing an acceptable solution within a relatively short time and is likely to lead the search towards the most promising solution area. A step-by-step TS-PSO-LR for the optimization of the UC problem is outlined as follows:

Step 1: Initialization data for each unit production

Step 2: Create tabu lists $P_{best}List$

Step 3: Generate the position and the velocity (X_r^k, V_r^k) of each particle according to equations (8) and (9).

Step 4: Evaluation of each particle at an initial power value selected from the limit margins.

Step 5: Calculate the production cost function of each particle for each production unit,

Step 6: Calculate the Lagrangian coefficient λ according to the expression (15),

Step 7: Calculate the objective function of each particle according to equation (1),

Step 8: Save the best results in the tabu list $P_{best}List$

Step 9: If the stop criterion is satisfactory, the found values are those which corresponds to the desired optimal solution otherwise we return to step 3.

5. Simulation And Results

In order to test the performance of the optimization proposed method; the strategy has been applied to an IEEE electrical network 9 buses [13,15,31], having 3 generators, over a period of 48 hours. The characteristics of the different production units are given in Table 1.

Table 1: Characteristics of production units

U	P_{gimax} (MW)	P_{gimin} (MW)	a	b	c	MUT	MDT	HSC_i ($)	CSC_i ($)
1	582	110	0.0756	30.36	582	8	8	4500	9000
2	330	74	0.00031	17.26	970	8	8	5000	10000
3	115	25	0.00211	16.5	680	5	5	560	1120

In this paper, we have considered 48 successive periods in order to establish the temporal evolution of the power demand (Table 2).

Table 2: Amount of load required

H	Load (MW)	H	Load (MW)	H	Load (MW)	H	Load (MW)
1	353.2	**13**	993.2	**25**	833.2	**37**	682.2
2	378.5	**14**	913.2	**26**	813.2	**38**	715.2
3	463.2	**15**	853.2	**27**	763.2	**39**	773.2
4	573.2	**16**	725.2	**28**	713.2	**40**	843.2
5	628.2	**17**	613.2	**29**	626.2	**41**	883.2
6	693.2	**18**	580.2	**30**	547.2	**42**	911.2
7	713.2	**19**	673.2	**31**	503.2	**43**	945.2
8	753.2	**20**	730.2	**32**	473.2	**44**	960.2
9	843.2	**21**	835.2	**33**	433.2	**45**	1001.2
10	925.2	**22**	945.2	**34**	533.2	**46**	1003.2
11	963.2	**23**	1007.2	**35**	583.2	**47**	925.2
12	1013.2	**24**	893.2	**36**	627.2	**48**	823.2

Table 3: Comparative table of the different methods used to solve the UC problem

H	P_{dh} (MW)	Production Cost ($)				Unit Commitment scheduling			
		LR	PSO	TS	PSO-TS-LR	LR	PSO	TS	PSO-TS-LR
1	353.2	23000	14198	14012	**3424**	111	111	111	**111**
2	378.5	25000	12981	15396	**13504**	111	111	111	**110**
3	463.2	29000	19729	20582	**13261**	111	111	111	**110**
4	573.2	375000	20320	28638	**15693**	111	110	111	**110**
5	628.2	845000	25275	33228	**19913**	110	110	111	**110**
6	693.2	1278000	32907	39192	**28268**	110	110	111	**110**
7	713.2	1432000	33600	41136	**19099**	110	110	111	**111**
8	753.2	1770000	43686	45172	**25722**	110	110	111	**111**
9	843.2	2688000	44498	50251	**32251**	110	110	111	**111**
10	925.2	7267000	46943	51667	**41758**	111	111	111	**111**
11	963.2	8441000	52566	52325	**46546**	111	111	111	**111**
12	1013.2	10129000	53198	53197	**53172**	111	111	111	**111**
13	993.2	9434000	53198	53197	**50477**	111	111	111	**111**
14	913.2	6915000	51897	49806	**40681**	111	111	111	**111**
15	853.2	2804000	37910	47580	**33298**	110	111	110	**111**
16	725.2	1529000	25883	46852	**29262**	110	111	110	**110**
17	613.2	758000	21020	46852	**25554**	110	111	110	**110**
18	580.2	584000	16014	46852	**14909**	110	111	110	**110**
19	673.2	1134000	29433	47666	**23751**	110	111	110	**110**
20	730.2	1571000	32618	49493	**33494**	110	111	110	**110**
21	835.2	2597000	38861	53197	**42276**	110	111	110	**110**
22	945.2	7873000	52140	53197	**50791**	111	111	111	**111**
23	1007.2	9918000	53198	53197	**52967**	111	111	111	**111**
24	893.2	6349000	51158	53197	**37898**	111	111	111	**111**
25	833.2	4797000	35703	53197	**45178**	111	111	111	**111**
26	813.2	4327000	36113	48240	**28933**	111	111	111	**111**
27	763.2	1861000	28794	47371	**24434**	111	111	110	**111**
28	713.2	1432000	25401	46852	**29117**	111	111	110	**111**
29	626.2	833000	46373	43893	**38959**	110	110	110	**101**
30	547.2	432000	20724	43893	**31753**	110	110	100	**101**
31	503.2	260000	24171	43893	**24453**	110	110	100	**101**
32	473.2	30000	21804	43893	**21221**	110	110	100	**101**
33	433.2	28000	25661	43893	**17202**	110	110	100	**101**
34	533.2	33000	29609	46499	**30381**	111	110	100	**101**
35	583.2	485000	45736	46499	**33614**	111	110	101	**101**
36	627.2	1017000	31241	46499	**39093**	111	110	101	**101**
37	682.2	1198000	46649	53197	**22191**	111	110	101	**111**
38	715.2	1448000	36247	53197	**26763**	111	111	111	**111**
39	773.2	1955000	35234	53197	**29056**	110	111	111	**111**
40	843.2	2688000	39180	53197	**32182**	110	111	111	**111**
41	883.2	3169000	42243	53197	**36731**	110	111	111	**111**
42	911.2	6857000	51986	53197	**40051**	110	111	111	**111**
43	945.2	7873000	52083	53197	**52034**	110	111	111	**111**
44	960.2	8345000	50398	53197	**51313**	110	111	111	**111**
45	1001.2	9709000	53198	53197	**52755**	110	111	111	**111**
46	1003.2	9778000	53198	53197	**52831**	111	111	111	**111**
47	925.2	7267000	51916	53197	**41756**	111	111	111	**111**
48	823.2	2464000	34368	49284	**30006**	111	111	110	**111**

Simulation results shown in Table 3 have proved that the adopted optimization methods have helped to establish an appropriate On/Off scheduling operating states of the production units while respecting the time constraints.

Nevertheless, through these methods we have arrived to reach an optimal production cost. Based on Table IV, the production cost found by the hybrid method based on the combination between Tabu search, Particle Swarm Optimization and Lagrangian Relaxation methods (TS-PSO-LR) method among 48 hours is about 1.5800e+06 $ lower compared to that obtained through Tabu search (PC = 2.2350e+06 $) and through PSO (PC= 1.7813e+06 $) or through Lagrangian Relaxation (PC= 1.6405e+08$). This result shows the best performances of the adopted strategy in minimizing the production cost and proves that we can get promising results through hybridization.

Table 4: Production Cost and Time required to converge for each optimized method

	TS	PSO	LR	TS-PSO-LR
Production Cost ($)	2.2350e+06	1.7813e+06	1.6405e+08	1.5800e+06
Time (s)	1.536 s	2.432s	4539s	127.082 s

Furthermore, concerning the resolution time, TS and PSO methods has presented the best time of convergence to an optimal solution compared to our strategy which requires 127.082 s to reach the global optimum. Besides, through Lagrangian Relaxation method, the unit commitment problem requires a lot of time to converge and this is explained by the complexity of the problem.

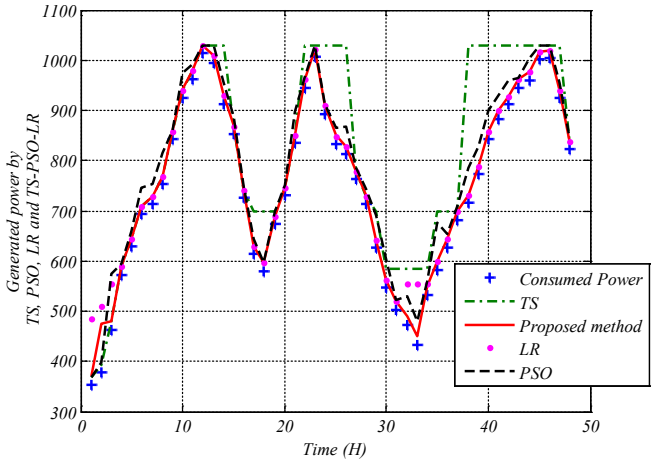


Figure 3. Generated power by TS, PSO, LR and TS-PSO-LR methods

It is interesting to note that the generated powers follow the optimum power quantities provided by the proposed optimization algorithm and the other optimization method. This demonstrates the high performance of the control algorithms adopted for the supervision of the system studied and proves the efficiency of the regulation loops for the different production units. In addition, the strategy adopts a permit to obtain sufficient and rapid planning in terms of convergence. Indeed, with the particular attention we offer you the considerable choice of input variables to particle swarm optimization method, we have managed via the proposed strategy to optimize the optimal solutions able to reduce the total cost of production.

Based on the results set out in Table 4, we find that our strategy has solved the Unit Commitment problem while addressing a planning of on-off states of production units. Planning that has complied with the constraints of each unit (minimum start-up MUT_i and shut-down times MDT_i). Moreover, we note that the power produced by the most powerful machine (615 MVA) remains unchanged throughout the 48 hours while the other production units vary to produce the amount of power demanded by the network.

Thanks to the simulation results, we can note that the Unit Commitment scheduling found by the hybrid approach has helped to obtain the minimum production cost compared to the other monotonous methods. This proves the main feature of the hybridization technique [32] that allows the combination of the advantages of various methods. Indeed, the tabu search, as table 4 shows, is found the most efficient regarding the convergence time but it has the highest total production cost. As for the Lagrangian relaxation, has allowed to have a very high production cost and requires a considerable time to converge. Whereas, the particle swarm optimization is conducted in a good convergence time but the cost is high compared to our TS-PSO-LR strategy. This comparison reflects the performance of the hybrid strategy, both in production cost and in convergence time.

Therefore, based on Figure 3, we can notice that the total amount of generated power by the production units is very similar to that consumed with a very limited amount of spinning reserve power compared to PSO, TS and LR methods where the generated powers are much higher than the amount requested. This proves the effectiveness of tracking of the consumed power per each hour and shows the performance of the algorithms enabling to get a minimal production cost. Besides, this minimal production cost has been established thanks to a good On/Off statements scheduling set for each production units (Figures 4 and 5). The organization is made through an estimation of the amount of load desired by the electric network, while taking into account of the allowable constraints.

We note that the optimal unit commitment scheduling found by our PSO-TS-LR strategy is characterized by the On status of the powerful unit (615 MVA) [3,6,8,9,13] throughout 48 hours and this is due to the minimum up and the minimum down-time constraints and to the power demand governed for each hour. We note that the unit commitment scheduling of the second production unit (370 MVA) is respecting the during minimum up/down time equal to 8 hours which proves the effectiveness of the control

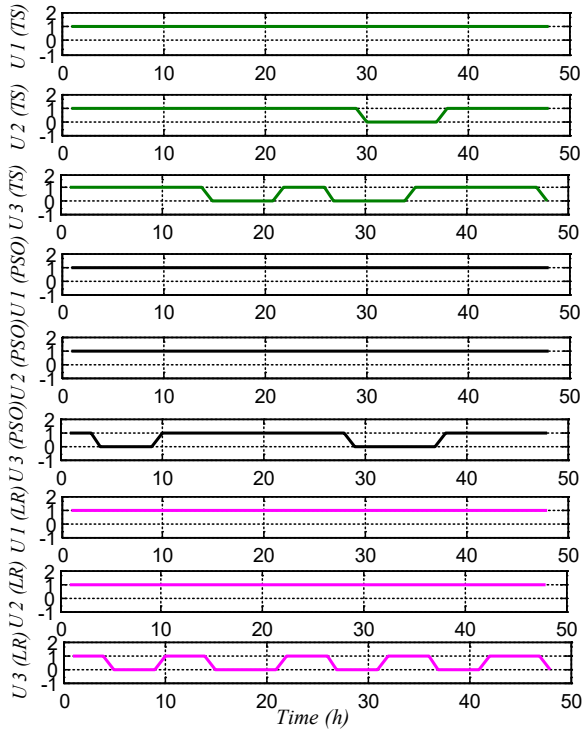


Figure 4. Optimal binary combination of units operation through TS method, PSO method and LR method

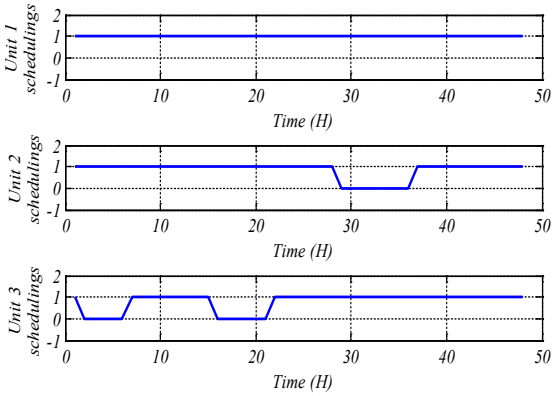


Figure 5. Optimal binary combination of units operation through hybrid TS-PSO-LR method

strategy. The third production unit undergoes Off status according to the production requirement and according to the power demand.

Furthermore, we confirm that our approach has allowed to select precisely the production units that should be available to respond to the load demand of the electrical network over a future period.

In addition, the adopted approach was promising both in terms of convergence to get the best optimal solutions to minimize the production cost and for an efficient unit commitment scheduling for the different production units, figure 6. Our strategy differs from other evolutionary computing techniques in providing an acceptable solution within a relatively short time and is likely to lead the search towards the most promising solution area.

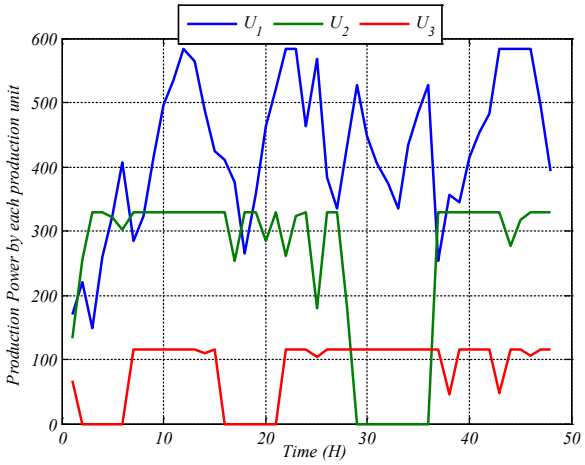


Figure 6. Generated Power by each production unit during 48 hours

6. Conclusion

This work shows the implementation of a new hybrid strategy which combines between Tabu Search, Particle Swarm Optimization and Lagrangian Relaxation. Herein, we have benefited from the rapidity of Lagrangian relaxation method, the storage solution with the memory of Tabu Search method and the flexibility to find optimal solution given by the Particle Swarm Optimization method. The proposed strategy TS-PSO-LR has presented high performances in optimizing the production cost and a capability of convergence to a global optimum as quick as possible compared to meta-heuristic (PSO, TS) and deterministic (LR) methods. In addition, the proposed strategy has ensured a proper unit commitment scheduling leading to get a minimal production cost. The right choice of the initial population suggests the possibility to obtain improvements in execution time. In addition, our strategy provides a fast enough time to converge to the optimal solution; which demonstrates the effectiveness of the adopted strategy compared to that obtained by Lagrangian Relaxation method and Particle swarm optimization methods.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Merlin, A., Sandrin, P. (1983) "A new method for unit commitment at Electricity De France", IEEE Trans on Power Apparatus and Systems; 102:1218-25.

[2] Lin, FT., Kao, C., Hsu, C. (1993) "Applying the genetic approach to simulated annealing in solving some NP-Hard problems". IEEE Trans on Systems, Man and Cybernetics; 23(6):1752-67.

[3] Sahbi Marrouchi, Moez ben hessine and Souad Chebbi, "New strategy based on Combined Use of Particle Swarm Optimization and Gradient methods to solve the Unit Commitment Problem", 15th IEEE International Conference on Environment and Electrical Engineering (EEEIC), Rome, Italy, 10-13 Juin 2015.

[4] Chunlin, D., Liu, Y., (2014) "Sample average approximation method for the chance-constrained stochastic programming in the transportation model of emergency management", Int. J. of Simulation and Process Modelling, Vol.9, No.4, pp.222 – 227.

[5] Richard, L. W., David, H. C. (2015) "A comprehensive method for solving finite-state semi Markov processes", Int. J. of Simulation and Process Modelling, Vol.10, No.1, pp.89 – 99.

[6] Rabeh, A., Sahbi, M., Moez, B.H., Houda, J., Souad, C., (2012) "Voltage Control Strategy of an Electrical Network by the Integration of the UPFC Compensator", International Review On Modelling and Simulation, vol. 5, no. 1, pp. 380–384.

[7] Galina, M., Vitaly, B., (2014) "Integrated planning and scheduling built on cluster analysis and simulation optimization", Int. J. of Simulation and Process Modelling, Vol.9, No.1/2, pp.81 – 91.

[8] Moez, B.H., Souad, B.S, (2014) "Accurate Fault Classifier and Locator for EHV Transmission Lines based on Artificial Neural Networks", Mathematical Problems in Engineering (MPE), ID 240565, vol 2014.

[9] Sahbi, M., Souad, C., (2014) "Combined Use of Particle Swarm Optimization and Genetic Algorithm Methods to Solve the Unit Commitment Problem", 16th International conference on Sciences and Techniques of Automatic control & computer engineering (STA), Monastir, Tunisia.

[10] Rahul, G., and Sharma, A.K. (2008) "Economic generation and scheduling of power by genetic algorithm", Journal of Theoretical and Applied Information Technology (JATIT).

[11] Aoki, K., Satoh, T., Itoh, M., Ichimori, T., and Masegi, K. (1987) "unit commitment in a large-scale power system including fuel constrained thermal and pumped-storage hydro", IEEE Transactions on Power Systems, Vol. PWRS-2, No. 4.

[12] Ouyang, Z., Shahidehpour, S. M. (1991) "An intelligent dynamic programming for unit commitment application", IEEE Trans on Power Systems; 6(3):1203-9.

[13] Sahbi, M., Souad, C., (2014) "A Comparative Study of Fuzzy Logic, Genetic Algorithm, and Gradient-Genetic Algorithm Optimization Methods for Solving the Unit Commitment Problem", Mathematical Problems in Engineering, vol. 2014, Article ID 708275, 14 pages, 2014. doi:10.1155/2014/708275

[14] Sasaki, H., Watanabe, M., Kubokawa, Yorino, J. N. and Yokoyama, R. (1992) "A solution method of unit commitment by artificial neural network", IEEE Transactions on Power Systems, Vol. 7, No. 3.

[15] Sahbi, M., Souad, C., (2014) "Unit Commitment Optimization Using Gradient-Genetic Algorithm and Fuzzy Logic Approaches", Complex system modeling and control through intelligent soft computations, Vol. 319, Springer book.

[16] Mantawy, A. H., Youssef, Y. L., Magid, L. A., and Shokri, S. Z., Selim, Z. (1998) "A unit commitment by Tabu search," Proc. Inst. Elect. Eng.Gen. Transm.Dist., vol. 145, no. 1, pp. 56–64.

[17] A. H. Mantawy. L. and Abdel MagidS. Z. Selim, (1998) "A unit commitment by Tabu search'', 1998 IEEE Proceedings online no. 1998.

[18] C. P. Cheng, C. W. Liu, C. C. Liu. (2002) "Unit commitment by annealing-genetic algorithm, Electrical Power and Energy Systems, vol. 24 pp 149-158.

[19] Logenthiran, T. and Dipti, S., (2010) ''Particle Swarm Optimization for Unit Commitment Problem'', International Conference on Probabilistic Methods Applied to Power Systems, pp 642 - 647.

[20] Charles, G., Christober A. R., (2009) ''Neural Based Tabu Search Method for Solving Unit Commitment Problem with Cooling-Banking Constraints'',Serbian Journal of electrical engineering Vol. 6, No. 1, May 2009, 57-74 UDK: 004.832.2, 2009.

[21] Senthil, K. and Palanisamy, V., (2006) ''A New Dynamic Programming Based Hopfield Neural Network to Unit Commitment and Economic Dispatch'' International Conference on Industrial Technology,pp 887 – 892, Mumbai.

[22] Chuan-Ping Cheng, Chih-Wen Liu, and Chun-Chang Liu, ''Unit Commitment by Lagrangian Relaxation and Genetic Algorithms'', IEEE Transactions on Power Systems, vol. 15, NO. 2, MAY 2000.

[23] C. Christober Asir Rajan and M. R. Mohan,"An Evolutionary Programming-Based Tabu Search Method For Solving The Unit Commitment Problem," IEEE Transactions on Power Systems, VOL. 19, NO. 1, FEBRUARY 2004.

[24] S. Senthil Kumar and V. Palanisamy, ''A New Dynamic Programming Based Hopfield Neural Network to Unit Commitment and Economic Dispatch'',IEEE Transactions on Power Systems, 2006.

[25] Charles Gnanakkan Christober Asir Rajan, ''Neural Based Tabu Search Method for Solving Unit Commitment Problem with Cooling-Banking Constraints'', Serbian Journal of electrical engineering Vol. 6, No. 1, May 2009, 57-74 UDK: 004.832.2, 2009.

[26] Christober C. Asir Rajan, ''Genetic algorithm based tabu search method for solving unit commitment problem with cooling — banking constraints'', Journal of electrical engineering, vol. 60, NO. 2, 2009.

[27] Alma Ademovic, Smajo Bisanovic and Mensur Hajro, '' A Genetic Algorithm Solution to the Unit Commitment Problem Based on Real-Coded Chromosomes and Fuzzy Optimization'', IEEE Transactions on Power Systems, 2010.

[28] Zhao, B., Guo, C. X., Bai, B. R., and Cao, Y.J. (2006) " An improved particle swarm optimization algorithm for unit commitment", International Journal of Electrical Power and Energy Systems, vol. 28, no. 7, pp. 482–490.

[29] Raglend, I.J., Raghuveer, C., Avinash, G.R., Padhy, N.P., and Kothari, D. P. (2010) "Solution to profit based unit commitment problem using particle swarm optimization", Applied Soft Computing Journal, vol. 10, no. 4, pp. 1247–1256.

[30] Chuan,-P.C., Chih, W.L., and Chun, C.L., (2000) ''Unit Commitment by Lagrangian Relaxation and Genetic Algorithms'',IEEE Transactions on Power Systems, vol. 15, NO. 2.

[31] Sahbi, M., Souad, C., (2014) "New strategy based on Fuzzy-Logic Approach to Solve the Unit-Commitment Problem", International Conference on Control, Engineering & Information Technology (CEIT), Monastir, Tunisia.

[32] Letizia, N.; Alessandro, C., Carlos, A., Rafael, D., (2014) "Hybrid approach for container terminals performances evaluation and analysis", Int. J. of Simulation and Process Modelling, Vol.9, No.1/2, pp.104–112.

.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 366-372 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Non-rigid Registration for 3D Active Shape Liver Modeling

Nesrine Trabelsi[*,1], Mohamed Ali Cherni[2], Dorra Ben Sellem[1,3,4]

[1] *Université de Tunis El Manar, Institute of Medical Technologies of Tunis , LR13 ES07 BTM, Tunis 1060 Tunisia*

[2] *Université de Tunis, LR13 ES03 SIME, ENSIT, Montfleury 1008 Tunisia*

[3] *Université de Tunis El Manar, Faculté de Médecine de Tunis, Tunis 1007 Tunisia*

[4] *Institut Salah AZAIEZ, Service de Médecine Nucléaire, Tunis 1006 Tunisia*

ARTICLEINFO

Article history:
Received: 10 October, 2017
Accepted: 18 January, 2018
Online: 10 February, 2018

Keywords:
3D Active Shape Model
B-Spline registration
CT-liver-scan
iso-surfaces
Marching Cubes
ROC-curves analysis

ABSTRACT

To avoid biopsies, doctors use non invasive medical techniques such as the computed tomography. Even that, the detection of the liver remains a big challenge because of the gray level and shape variations which depend on patients and acquisition modalites. In this work, we propose to create a 3D liver model in the training phase of 3D active shape model algorithm. This training model will be deformed according to any given 3D data for liver segmentation. The contribution of our work is the use of the Non-rigid registration with a B-spline registration on the training phase. We tested our method on an open access database ("3D-IRCADb") and on our database obtained from the radiology department of the National Oncology Institute of Tunis. Both data-sets showed the reliability of the method with an accuracy equal to 69.98% and 71.18% respectively for our database and "3D-IRCADb".

1 Introduction

Computed Tomography (CT) is a non invasive technique with a large field of view. It helps doctors to detect some hepatic disease. First, they identify the liver from the abdominal CT slice using two phases of enhancement: portal and arterial. In each time of enhancement, the gray level of the liver changes. In this work, we propose a 3D method to extract the liver from the CT slice according to its shape. We use different 3D CT exam with different phases of enhancement containing normal and pathological cases. The aim of our work is to use an advanced technique of segmentation to extract liver from images of any type of modality. For that reason, we propose to use the Active Shape Model (ASM). The ASM have been recognized as robust solution for a supervised technique of segmentation. Our contribution is to improve the performance of the 3D ASM algorithm with a pre-processing phase using a non rigid registration based on B-Spline transformation. The following paper is an extension of work originally presented in the 7th International Conference on Sciences of Electronics, Technologies of Information Telecommunication" [1]. In this work, we improve the evaluation of results using the ROC-Curve analysis and we ensure the reliability of our method using our database in addition to the "3D-IRCADb". Also, we compare the results not only with the Isosurfaces also with the Marching Cubes 3D reconstruction. The paper is organized as follow: Section 2 presents some literature related works. In section 3, we explain the 3D-ASM method and its different steps for 3D-model construction. the proposed method is applied on the open access database "3D-IRCADb" and our database. In section 4, we use the ROC-Curve analysis to evaluate the 3D liver modeling. First, we compare our 3D model with the Isosurfaces 3D reconstruction, then, with the Marching Cubes technique. Finally, we resume our proposed method and the obtained results on the conclusion.

2 Related works

The major point of interest of the Computed-aided diagnostic system is to solve difficulties of the medical field. The segmentation technique is the most important step in those system. It allows the physicians to extract the region of interest and to analyze it for more helpful medical information. In this con-

[*]Corresponding Author: Nesrine Trabelsi, nesrine.trabelsi@istmt.utm.tn

https://dx.doi.org/10.25046/aj030145

text, many techniques of segmentation have been developed such as morphological operations exposed on the work of Pham The Bao et al.[2]to extract the liver from the volumetric CT images. In [3], they improved morphological operator algorithm by using the watershed technique to segment the micro-tomographic trabecular bone. The watershed algorithm is widely used to identify the region of interest. In [4],P. Rodrigues et al. used it with an open software "Mevislab" to extract the liver from the CT exam with a 87% of accuracy. In [5], the author used the Fisher algorithm to improve the watershed algorithm for the radar images segmentation. Moreover, there are some intensity based techniques which detect the object according to the intensity repartition. In [6], the author constructed a model of the intensity distribution for the surface of the liver, cysts and lesions. Then, they calculated the probability of pixel belonging to each classification of different regions. In [7], Alom et al. used the "SGM growing slice method", an algorithm for 3D segmentation of the liver. Also, we note that algorithms of classification are used to segment the liver. For example, in the work [8], authors made a new texture feature extraction to improve liver classification using K-mean algorithm. While, in [9], the author used the "Support Vector Machine" algorithm and the surface distance maps for 3D-liver segmentation from CT-scans. Previously exposed works are based on the intensity of the liver and its gray level. In order to improve the segmentation technique, we propose a 3D-supervised-method for liver segmentation to overcome problems of gray level variation.

3 Methods and materials

In this section, we will explain the use of the shape context based on non-rigid surface registration with a B-Spline transformation in order to adapt the Active Shape Model for 3D CT liver segmentation. We start by a global description of the proposed method. Then, we detail each step. Figure 1 describes the different steps of our work to segment 3D CT liver. First, we note that the cases of the two databases are acquired with different parameters. So each exam has different number of slices, resolution and voxel size. To create the 3D model of the liver using the 3D ASM, we must uniform the volumetric data in the pre-processing phase. Several works combine different registration algorithm with a 3D segmentation method to extract a volumetric data such as to segment the left ventricle in [10] and the mandibular canal in [11]. In our work we propose to use a non-rigid registration with a B-Spline transformation. After the pre-processing phase, the 3D CT-scan has the same size of the matrix of vertices and the matrix of faces. Those two matrix will be used on the training phase of the 3D-ASM. Finally, we create the 3D liver model in the testing phase of the 3D-ASM.

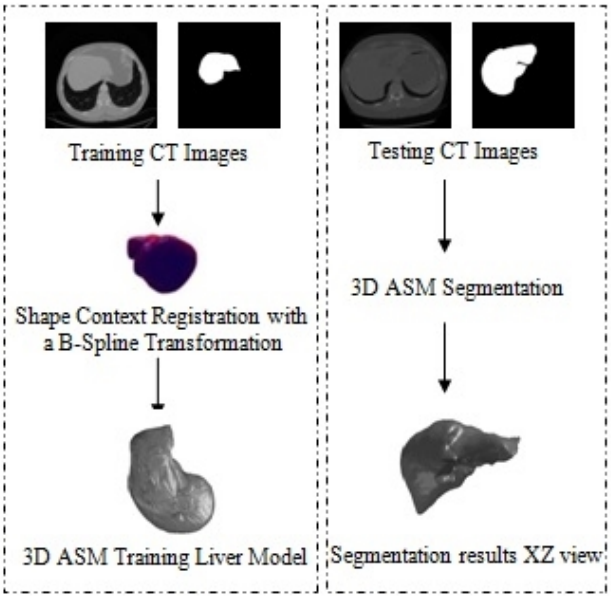


Figure 1: Diagram of the main work for 3D-ASM Liver segmentation

3.1 3D pre-processing CT-data

The CT-Scan measure the attenuation of the X-ray crossing the organ. These measurements are taken from different angles to produce a cross-sectional image called also slice separate with a thickness value. The total number of slices makes a volume information of the explored organ. However, doing the CT-Scan exam allows doctor to see details of the scanned organ in multi-planes (Coronal, axial and transverse). The obtained CT image is a matrix composed in pixels. Considering the the thickness between slices, each pixel represents a small volume element called also voxel. Its size depends on the matrix size, the thickness and the field of view. The CT images will be next stored as a numeric image. We note that the CT-images are coded according to thee Hounsfield units (UH) expressed by equation 1:

$$HU = \frac{\mu_{tissu} - \mu_{watter}}{\mu_{watter}} \quad (1)$$

In our work,we first change the UH to a gray level scaling using the DICOM header. We can define the gray value of a pixel as follow in equation (2):

$$Graylevel = \frac{UH \pm RescaleIntercept}{RescaleSlope} \quad (2)$$

where: The rescale intercept is equal to 0 for the IRCADb database and −1024 for our database; The rescale slope is equal to 1 for the both databases.
After changing the UH to the gray level scale, we uniform the slice number of the training data so all the training exams will have the same number of slices. We add a black slice at the end of the CT exam to reach the desired number. Then, we start the pre-processing step to achieve the same matrix size of

faces and vertices to all the training 3D data using the Shape Context Correspondence Point Model algorithm. It is based on a registration method. We can classify the registration into two classes: rigid and non-rigid transformation [12–14]. In our work, we employ a diffeomorphic B-Spline cubic transformation for no-rigid registration. We note : "C_1" a contour of an object 1 and "p_i" the number of points selected to define the contour; "C_2" a contour of an object 2 and "q_j" the number of points selected to define this contour. Then, we measure the cost function "C"between the two objects. It is defined in (3) as follow:

$$C(p_i, q_j) = \frac{1}{2} \sum_{k=1}^{k} \frac{[h_i(k) - h_j(k)]^2}{[h_i(k) + h_j(k)]} \quad (3)$$

With: h_i and h_j respectively histograms of the two shapes. After calculating the cost function, we minimize it to increase the degree of the similarity and the matching between the two shapes. To minimize the cost function, we choose a type of transformation in the registration phase. The type of transformation depends on the feature to be extracted from the object [15, 16], such as the match point algorithm[17], the 3D coordinate[18]and the intensity transformation using an iterative closet point algorithm [19, 20]. Figure 2 shows the different steps of the pre-processing phase. The first step of the registration phase using the Shape Context algorithm is to define the type of the transformation. In this work, we use the B-Spline cubic transformation. Then, we choose the mastery data which have the biggest number of slice and a floating data. This transformation is based on the surface matching with a spatial alignment of the two volumetric CT data. As an output of this phase, the CT exam will have the same size of vertices and faces matrix in order to be used on the next stage which is the 3D-ASM.

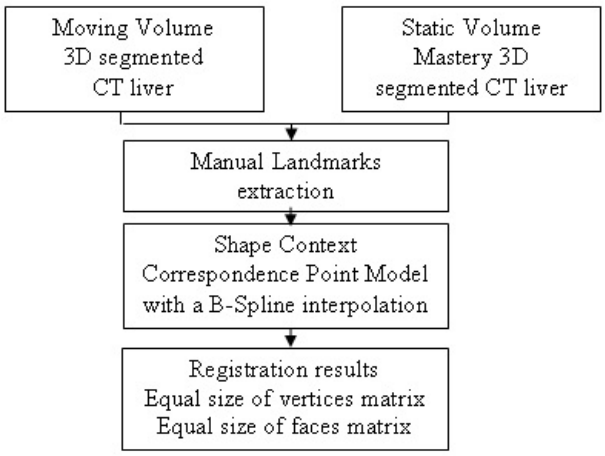


Figure 2: Pre-processing phase using Shape Correspondence Point Model

3.2 3D Active Shape Model

The 3D-ASM is an iterative supervised segmentation algorithm firstly proposed in [21] in 2002. It contains two phases: training and testing. Thanks to this algorithm based on the distribution point model (PDM) [22], we train our system to learn different liver shape variations. The input data in the training phase are 3D surface matrix containing the faces, the vertices of the 3D mesh and the binary volume of the extracted liver from the CT slice. At the beginning, we define for an object "O" a set of target point called landmark "n". The "n" points reform a vector of shape noted $X = (x, x_2, ..., x_n, y_1, y_2, ..., y_n, z_1, z_2, ..., z_n)^T$. After the identification of vector shape, we calculate from the 3D input data the covariance matrix. Then, we use the principal analyze components (PCA)[23]to identify the eigen-vectors"'x_i and their eigen-values. The PCA allows us to calculate the mean shape and its variations. Finally, we apply the PDM algorithm to create our training shape model noted $\tilde{X}$ in (4).

$$\tilde{X} = \bar{X} + \phi_s b_s \quad (4)$$

With: $\bar{X}$: aligned training shape; b_s:shape parameter vector of the shape. It is equal to $b_s = \pm m \sqrt{\lambda_i}$; $m \leq 3$; ϕ_s: eigenvectors corresponding to the eigenvalue of the covariance matrix. The next step in the 3D-ASM algorithm is to update the training model using the Mahalanobis distance[24]. Then, we apply the training deformed model to create our 3D liver model according to a given CT-data without doing a registration step.

4 Results and discussion

In this section, we start by a description of the databases used in this work and we present the obtained results. In order to evaluate the accuracy of our 3D liver modeling, we use the ROC-Curve analysis.

4.1 Description of databases

The proposed method was tested on two databases: our database (15 CT-exams) and the "3D-IRCADb" one (20 CT-exams). In [25], twenty radiologists expert describe the different parameters of the 3D IRCADb database. We found that the voxel size is various from one case to another from $0.57cm^3$ to $1.6cm^3$. Also in our database acquired at the radiology department of National oncology Institute "Salah AZAIEZ" of Tunis, the voxel size varies from 1.25cm*1.25cm*1.25cm to 1.4cm*1.4cm*1.4cm. For more efficiency, we use for the training phase, the 3D CT-scans with the same voxel size. In this work,for each database, we use the 80% of the CT-scan data having the same slice thickness for the training phase and 20% of 3D CT-scan for the testing phase.

4.2 Experiments

This study presents the use of the 3D-ASM for CT liver segmentation. For the beginning, we make a registration phase to recover some parameters such as rotation, scaling and translation in order to minimize the dissimilarity between two 3D CT-data. Figure 3 shows 3D mesh of the diffeomorphic B-Spline cubic registration. After the Shape Context registration, the output of this stage is used as input of the training phase of 3D-ASM algorithm. Different results of the 3D training shape are exposed in Figure 4.



Figure 3: 3D liver mesh of the B-Spline registration

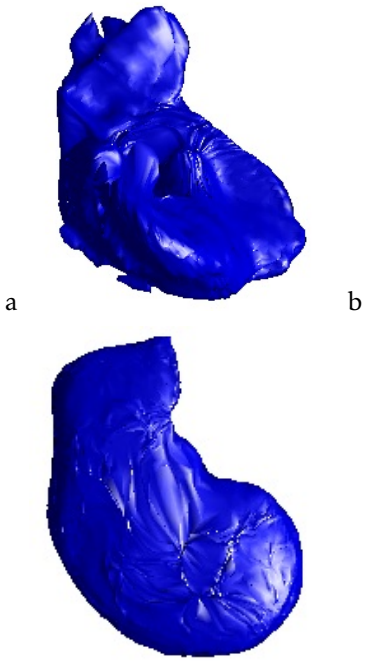


Figure 4: Training 3D model:a.Case from our database; b.Case from 3D-IRCADb database

To establish the proposed work, we use for the training phase of the 3D-ASM the cases which have the same thickness between slices and total number of slice. Then, we handle the corresponding Shape Context Points Model with a B-Spline transformation to create the same size of a 3D Liver grid. After the preprocessing step, we apply the 3D-ASM algorithm. To evaluate, in this work, the obtained results, we make a 3D reconstruction of the liver using the Isosurfaces and the Marching Cubes method.Then we compare them with the 3D-ASM of the liver. Figure 5 and figure 6 show different views of the 3D liver shape obtained from our database and the "3D-IRCADb" open access database.

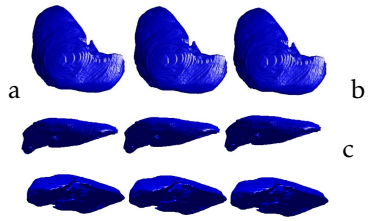


Figure 5: 3D liver segmentation of cases from "3D-IRCADb" database using respectively from the left to the right: the Isosurfaces, the Marching Cubes and the proposed method: a.XY view; b.XZ view; c.YZ view

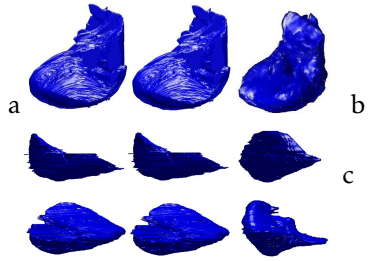


Figure 6: 3D liver segmentation of cases from our database using respectively from the left to the right: the isosurfaces, the Marching Cubes and the proposed method: a.XY view; b.XZ view; c.YZ view

After visual comparison, we use five parameters essentially used on ROC-curves analysis widely used in evaluation on the medical field [26]. These parameters are: sensitivity (SN), specificity (SP), accuracy (ACC), positive predicted value (PPV), negative predicted value (NPV) and area under the curve (AUC). We note the V_s is the segmented volume and V_{GT} the ground truth one. We define the True Positive(TP) as the degree of matching between the 3D mesh of V_s and the V_{GT} expressed in(5).

$$TP = V_s \cap V_{GT} \tag{5}$$

Its complement called False Negative (FN) is defined in (6).

$$FN = V_{GT} - TP \tag{6}$$

Moreover, we define the True Negative (TN) when we found a similarity on the false information. It is given in (7).

$$TN = 1 - (V_s \cup V_{GT}) \tag{7}$$

The False Positive (FP) is when our system detect a falsse mesh similiarity when there is no resemblance. It is given in (8).

$$FP = V_s - TP \tag{8}$$

According to those parameters, we calculate the sensitivity given in (9), the specificity given in (10) and the accuracy given in (11).

$$Sensitivity = \frac{TP}{TP+FN} \tag{9}$$

$$Specificity = \frac{TN}{TN+FP} \tag{10}$$

$$Accuraccy = \frac{TN+TP}{TN+TP+FN+FP} \tag{11}$$

We expose in tables 1 and 2 the different obtained values of the previously defined parameters. In this work, for each database, the liver was segmented by experts. So, we make a 3D reconstruction of these segmented volume using Isosurfaces and Marching Cubes techniques. Then, we apply the ROC-Curve independently. In the first hand, we apply it between our 3D results and 3D isosurfaces. On the second hand, we extract the different measure of the ROC-Curve between our 3D results and the Marching Cubes reconstruction.

	Our database	3D-IRCADb
Sensitivity (%)	68.38	63.8
Specificity (%)	71.57	59.36
PPV(%)	69.88	60.23
NPV (%)	70.85	59.52
Accuracy (%)	69.98	59.51
AUC	0.74	0.63

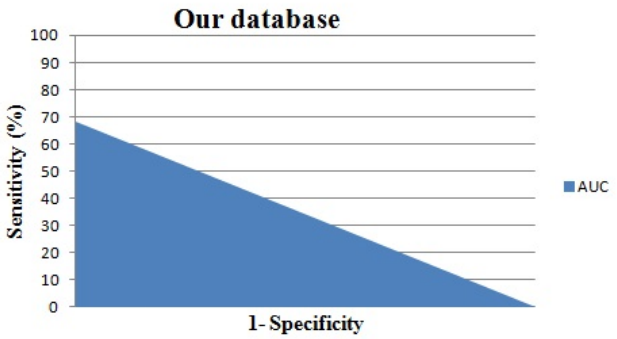
Table 1: Performances of the 3D-ASM with a B-spline registration versus the Isosurfaces reconstruction

	Our database	3D-IRCADb
Sensitivity (%)	75.16	50.06
Specificity (%)	65.93	49.94
PPV(%)	69.16	50.07
NPV (%)	72.33	50.76
Accuracy (%)	70.44	50.07
AUC	0.74	0.49

Table 2: Performances of the 3D-ASM with a B-spline registration versus the Marching Cubes reconstruction

Comparaing to the Isosurfaces and Marching Cubes methods, the 3D-ASM based on the Shape Context registration with a B-Spline registration achieve about 70% of accuracy and 71,57% of specificity. Those promising results, ensure the efficiency of the proposed method based on a non-rigid registration for the pre-processing phase. We found that we achieve better results with our database.

(a)



(b)

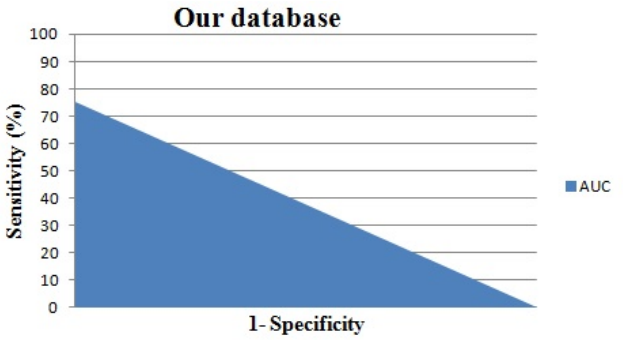
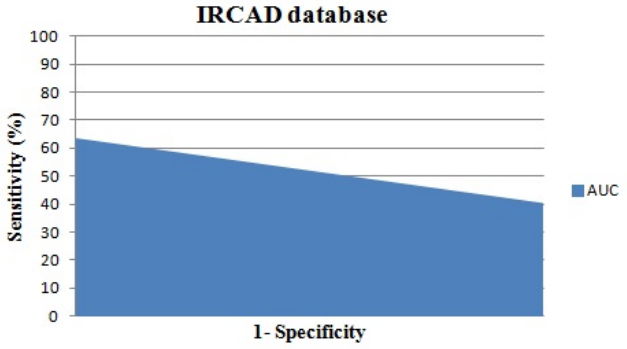


Figure 7: The area under the curve for cases from our database: a.versus Isosurfaces; b.versus Marching Cubes

(a)



(b)

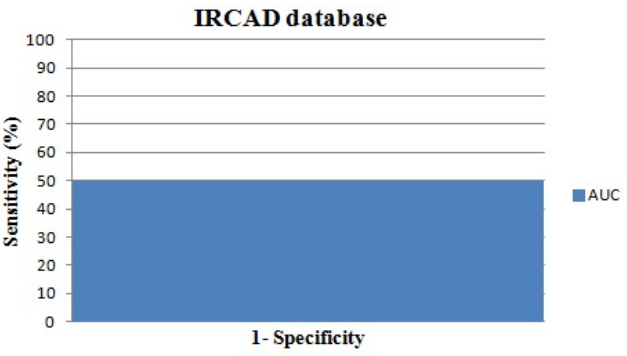


Figure 8: The area under the curve for cases from "3D-IRCADb"database: a.versus Isosurfaces; b.versus Marching Cubes

Our database contains pathological and normal cases. We used the normal cases for the training phase and we test the proposed method on pathological and normal cases. In fact, the 3D Active Shape Model is an advanced technique of 3D segmentation based on the shape. In this study, we prefer to use normals case with shape variation in order to create an efficient 3D learning model which will be deformed according to a given 3D CT-Scan exam. The inconvenient of the use of the pathological cases in the training phase is the no accurate liver shape since we can found one or several parts of the liver have been removed. The "3D-IRCADb" contains pathological cases without ablation which can be convincing for our study. To evaluate the efficiency of our proposed method using the "3D-IRCADb" database, we compare the 3D liver mesh obtained using our proposed method , the Isosurfaces and the Marching Cubes 3D reconstruction with the VTK files. The different results are exposed in the table 3. We found almost the same measure off the Area under the curve, presented in figure 9, but we achieve a slight improvement in terms of accuracy using the 3D Active Shape Model with a B-spline registration comparing to the Isosurfaces (Iso) and the Marching Cubes (MC).

	MC	Iso	Proposed Method
Sensitivity (%)	65.782	65.758	65.765
Specificity (%)	75.392	75.492	75.494
Accuracy (%)	70.629	70.648	71.182
PPV (%)	72.973	73.052	73.053
NPV (%)	68.775	68.744	68.736
AUC	0.757	0.756	0.751

Table 3: Comparison of the 3D liver mesh between the VTK files of "3D-IRCADb", the Isosurfaces, Marching Cubes and our proposed method using ASM with a B-spline registration.

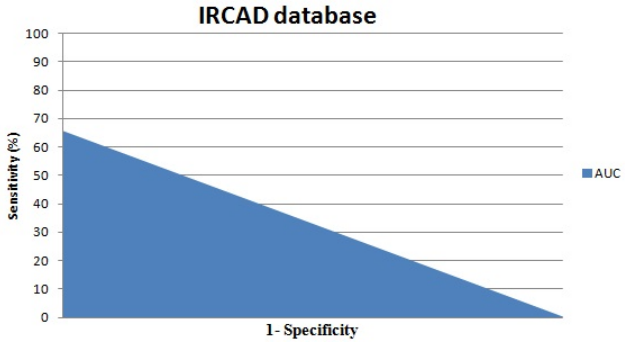


Figure 9: The Area under the curve: Comparison between our propose method and the VTK files of "3D-IRCADb" database

5 Conclusion

In this work, we used the Active shape Modeling for 3D CT-liver segmentation. In order to well create our model, we used in the pre-processing phase the Shape Context Corresponding Point Model with a B-Spline cubic interpolation. Evaluation of the proposed method has been performed by using 20 CT-exams from the "3D IRCADb"'database and 15 CT-exams from our database. We acheive a good results for 3D liver segmentation with a 70% of accuracy. Results obtained using our database are better than those obtained with the open access database "3D-IRCADb". This is due to the fact that the "3D-IRCADb" database is composed only by pathological cases contrariwise to our database which is composed by normal and pathological cases. As a conclusion, it would be better to use only normal cases or a much bigger number of normal cases than pathological cases to build a good training model. The proposed method brings to the doctors and the researchers, an approximate 3D model for normal liver cases. As a perspective, we attend to extract some feature from this 3D model for normal liver i order to detect liver pathologies according to the liver 3D mesh.

Conflict of Interest The authors declare no conflict of interest.

References

[1] N. Trabelsi, K. Aloui, and D. Ben Sellem. 3d active shape model for ct-scan liver segmentation. In *2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 161–165, Dec 2016. doi: 10.1109/SETIT.2016.7939859.

[2] Pham The Bao, Tran Hong Tai, Viet-Hang Duong, and Jia-Ching Wang. Liver segmentation from 3d abdominal ct images. In *Consumer Electronics-Taiwan (ICCE-TW), 2015 IEEE International Conference on*, pages 342–343. IEEE, 2015.

[3] W. A. Fourati and M. S. Bouhlel. Contour-based surface modeling and analysis of microtomographic trabecular bone images. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 423–427, March 2012. doi: 10.1109/SETIT.2012.6481951.

[4] P. Rodrigues, J. L. Vilaça, and J. Fonseca. An image processing application for liver tumour segmentation. In *1st Portuguese Biomedical Engineering Meeting*, pages 1–6, March 2011. doi: 10.1109/ENBENG.2011.6026097.

[5] B. Belkhaoui, A. Toumi, A. Khenchaf, A. Khalfallah, and M. S. Bouhlel. Segmentation of radar images using a combined watershed and fisher techniques. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 400–403, March 2012. doi: 10.1109/SETIT.2012.6481948.

[6] J. J. Chen, A. Kutluk, Y. T. Hu, and M. Hamit. Liver hydatid ct image segmentation based on localizing region active contours and modified parametric active contours. In *2014 7th International Conference on Biomedical Engineering and Informatics*, pages 217–221, Oct 2014. doi: 10.1109/BMEI.2014.7002773.

[7] M. Z. Alom, M. Mostakim, R. Biswas, and A. Chakrabarty. Automatic slice growing method based 3d reconstruction of liver with its vessels. In *Computer and Information Technology (ICCIT), 2013 16th International Conference on*, pages 338–344, March 2014. doi: 10.1109/ICCITechn.2014.6997361.

[8] J. Ma, F. Duan, and P. Guo. Improvement of texture image segmentation based on visual model. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 151–154, March 2012. doi: 10.1109/SETIT.2012.6481904.

[9] X. Zhang, J. Tian, D. Xiang, X. Li, and K. Deng. Interactive liver tumor segmentation from ct scans using support vector classification with watershed. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6005–6008, Aug 2011. doi: 10.1109/IEMBS.2011.6091484.

[10] C. Santiago, J. C. Nascimento, and J. S. Marques. Segmentation of the left ventricle in cardiac mri using a probabilistic data association active shape model. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7304–7307, Aug 2015. doi: 10.1109/EMBC.2015.7320078.

[11] F. Abdolali and R. A. Zoroofi. Mandibular canal segmentation using 3d active appearance models and shape context registration. In *2014 21th Iranian Conference on Biomedical Engineering (ICBME)*, pages 7–11, Nov 2014. doi: 10.1109/ICBME.2014.7043884.

[12] T. Huysmans, J. Sijbers, and V. Brigitte. Automatic construction of correspondences for tubular surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):636–651, April 2010. ISSN 0162-8828. doi: 10.1109/TPAMI.2009.93.

[13] M. Pereañez, K. Lekadir, I. Castro-Mateos, J. M. Pozo, Á. Lazáry, and A. F. Frangi. Accurate segmentation of vertebral bodies and processes using statistical shape decomposition and conditional models. *IEEE Transactions on Medical Imaging*, 34(8):1627–1639, Aug 2015. ISSN 0278-0062. doi: 10.1109/TMI.2015.2396774.

[14] Mohamed M Habib, Roshan A Welikala, Andreas Hoppe, Christopher G Owen, Alicja R Rudnicka, Adnan Tufail, Catherine Egan, and Sarah A Barman. Incorporating Spatial Information for Microaneurysm Detection in Retinal Images. *Advances in Science, Technology and Engineering Systems Journal*, 2(3):642–649, 2017. URL http://astesj.com/v02/i03/p82/.

[15] G. Pan, X. Zhang, Y. Wang, Z. Hu, X. Zheng, and Z. Wu. Establishing point correspondence of 3d faces via sparse facial deformable model. *IEEE Transactions on Image Processing*, 22(11):4170–4181, Nov 2013. ISSN 1057-7149. doi: 10.1109/TIP.2013.2271115.

[16] Olivier Commowick. *Design and Use of Anatomical Atlases for Radiotherapy*. Theses, Université Nice Sophia Antipolis, February 2007. URL https://tel.archives-ouvertes.fr/tel-00133432.

[17] M. Salehpour and A. Behrad. 3d face reconstruction by klt feature extraction and model consistency match refining and growing. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 297–302, March 2012. doi: 10.1109/SETIT.2012.6481932.

[18] A. Ghanbari, R. Abbasi-Asl, A. Ghaffari, and E. Fatemizadeh. Automatic b-spline image registration using histogram-based landmark extraction. In *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*, pages 1004–1008, Dec 2012. doi: 10.1109/IECBES.2012.6498119.

[19] Antonis D Savva, Theodore L Economopoulos, and George K Matsopoulos. Geometry-based vs. intensity-based medical image registration: a comparative study on 3d ct data. *Computers in biology and medicine*, 69:120–133, 2016.

[20] C. Xing and P. Qiu. Intensity-based image registration by nonparametric local smoothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2081–2092, Oct 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.26.

[21] B. van Ginneken, A. F. Frangi, J. J. Staal, B. M. ter Haar Romeny, and M. A. Viergever. Active shape model segmentation with optimal features. *IEEE Transactions on Medical Imaging*, 21(8):924–933, Aug 2002. ISSN 0278-0062. doi: 10.1109/TMI.2002.803121.

[22] Timothy F Cootes, Christopher J Taylor, et al. Statistical models of appearance for computer vision, 2004.

[23] T. Tateyama, A. H. Foruzan, and Y. W. Chen. Pca based statical shape model of the spleen. In *2009 Fifth International Conference on Natural Computation*, volume 6, pages 36–39, Aug 2009. doi: 10.1109/ICNC.2009.695.

[24] Fadlalla G. Elfadaly, Paul H. Garthwaite, and John R. Crawford. On point estimation of the abnormality of a mahalanobis index. *Computational Statistics & Data Analysis*, 99:115 – 130, 2016. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2016.01.014. URL http://www.sciencedirect.com/science/article/pii/S0167947316000232.

[25] *Database IRCAD', Http://www.ircad.fr/fr/recherche/3d-ircadb-01-fr, Accessed 01 May 2015.* Http://www.ircad.fr/fr/recherche/3d-ircadb-01-fr, 2015.

[26] Enoch Opanin Gyamfi and Yaw Marfo Missah. Pixel-Based Unsupervised Classification Approach for Information Detection on Optical Markup Recognition Sheet. *Advances in Science, Technology and Engineering Systems Journal*, 2(4):121–132, 2017. doi: 10.25046/aj020417. URL http://astesj.com/v02/i04/p17/.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 373-383 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Adaptive observer design for a class of nonlinear systems with time delays

Ahlem Sassi[*1,2], Michel Zasadzinksi[1], Harouna Souley Ali[1], Kamel Abderrahim[2]

[1]*Centre de Recherche en Automatique de Nancy (CRAN, CNRS UMR 7039), Université de Lorraine, IUT de Longwy, 54400 Cosnes et Romain, France*

[2] *Commande Numérique des PRrocédés Industriels (CONPRI), National School of Engeneering in Gabes, Street Omar Ibn Elkhattab 6029 ZRIG Gabes, Tunisia*

ARTICLEINFO

Article history:
Received: 21 November, 2017
Accepted: 24 January, 2018
Online: 10 February, 2018

Keywords:
Time delays
Nonlinear system
Adaptive observer
LMI
Lipschitz function
Bilinear system

ABSTRACT

This paper deals with the design of adaptive observer for a class of nonlinear systems with time delays. Within this work, we develop an adaptive observer for a bilinear time delay system, then we extend those results in the presence of Lipschitz nonlinear functions in the system's dynamics. In the stability analysis of the estimation errors, we combine a Linear Parameter Varying (LPV) approach in the presence of time delays with a polytopic approach. The obtained stability conditions are given in terms of the solvability of Linear Matrix Inequalities (LMIs) on the vertices of a convex polytope. Numerical examples are finally given to show the effectiveness and feasibility of our results.

1 Introduction

This paper is an extension of work originally presented at the 6th International Conference on Systems and Control and entitled *"Full order adaptive observer design for time delay bilinear system"* [1]

During the last decades, several theoretical results with interesting applications were focused on the design of observers for nonlinear systems [2], [3], [4], [5], [6], [7] [8], [9] (and references there in). Due to the difficulty of setting the nonlinear behaviour in a system's dynamics and the non availability of all the state vector component, the design of observer is a challenging and open problem.

Indeed, since the presence of time delay is often encountered in the industrial processes, it is necessary to integrate it in the system model for a best modelization of these systems. The presence of time delay should not be neglected, because it can affect the systems performances and may lead, in certain cases, to its instability [10]. For these reasons, this class of systems has been intensively studied in the literature [11], [12], [13], [14] and will be considered also in the present work.

Apart the presence of time delays in a nonlinear model, another difficulty may appear: the presence of some unknown parameters, which should be taken into account in order to make the model more accurate. Hence, we propose in this work to design an adaptive observer, which does not only estimate the state vector of a nonlinear time delay system, but also the unknown parameters which affect its dynamics. Thanks to its ability to handle some challenging applications such as in robust and fault tolerant control, the adaptive observer allows to cope with the lack of knowledge on the system's unknown parameters. Classically, an adaptive observer may provide a suitable estimation of the states and the unknown parameters under some appropriate excitation conditions [3]. There are two major approaches to design adaptive observer. The first approach is based on the elaboration of an adaptation law derived from the stability analysis of a state observer. The convergence of the unknown parameters can be ensured under some persistent excitation [15], [16], [17]. The second approach consists in designing a state observer for an augmented system, where the state dynamics model is augmented with the dynamics of its unknown pa-

[*]Corresponding Author: Ahlem Sassi, CRAN, IUT de Longwy, 186 Rue de Lorraine 54400 Cosnes et Romain, France. Email:ahlemsessi@yahoo.com

https://dx.doi.org/10.25046/aj030146

rameters. [18], [19], [20], [21].

In this paper, we consider an adaptive observer design for a class of nonlinear time delay systems where the dynamics are nonlinear in the states and in the unknown parameters and contains some bilinear terms. The nonlinearities considered satisfied a Lipschitz condition. Here, the convergence of the adaptive observers is treated in a more tractable way and does not require to satisfy the persistent excitation condition as in [22], [23], [24] and contrary to many previous contributions. In a first time, an adaptive observer will be proposed in the absence of some nonlinearities. The considered system represents a time delay bilinear system affected by unknown parameters. The bilinear systems do appropriately model some physical processes than linear or nonlinear ones. Nevertheless, the fact of considering a bilinear model does not suffice, due to the presence of nonlinear behaviour which ought to be integrated. For that reason, we consider in a second time the presence of some additional Lipchitz nonlinear functions. It is necessary to study those results separately because the design of adaptive observer in the first case does not derive from the second one due to the presence of some unmeasured state variables in the nonlinear functions, which make the problem more conservative.

This paper is structured so that the statement of the problem and some useful formulas are presented in the section 2. Section 3 is devoted to the design of adaptive observer for a bilinear time delay system. This result is extended in section 4, by the addition of nonlinear Lipschitz functions in the delayed bilinear system dynamics. In section 5, a discussion of the obtained results is made by comparison with some previous ones. Finally, in section 6, the obtained results will be applied to numerical examples to show their effectiveness.

Notations. Throughout this paper, $\mathbb{R}^n$ denotes an n-dimentional Euclidean space, and $\|\cdot\|$ the associated Euclidean norm [25], where

$$\|x\| = \sqrt{x^T x}, \quad \forall x \in \mathbb{R}^n \tag{1}$$

and using expression (1), the following induced matrix norm given by

$$\|A\| = \max_{\|x\|\neq 0} \frac{\|Ax\|}{\|x\|} \tag{2}$$

is used in this paper where $A \in \mathbb{R}^{n\times m}$.
$(*)$ will denotes the transpose of the off-diagonal parts of a matrix.

2 Problem statement

In this work, a class of nonlinear time delay systems is investigated where the state space model is given by the following equation

$$\dot{x}(t) = A_0 x(t) + \sum_{i=1}^{m} A_i u_i x(t) + \ell(x,u) + A_{d_0} x(t-\tau_0) + \sum_{i=1}^{m} A_{d_i} u_i(t) x(t-\tau_i) + Bu(t) + Gg(x,u)\theta \tag{3a}$$

$$\dot{\theta} = 0 \tag{3b}$$

$$y = Cx(t) \tag{3c}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $\theta \in \mathbb{R}^q$, $y \in \mathbb{R}^p$ are the states vector, the input control, the unknown parameters vector and the output vector, respectively. τ_i for $i = 0,\dots,m$ are known constants delays. The matrices: $A_i \in \mathbb{R}^{n\times n}$, $A_{d_i} \in \mathbb{R}^{n\times n}$, for $i = 0,\dots,m$, $B \in \mathbb{R}^{n\times m}$, $G \in \mathbb{R}^{n\times r}$ and $C \in \mathbb{R}^{p\times n}$ are known with constant values.
The following assumptions are given and will be used in the sequel

Assumption 1. *The input $u(t)$ is bounded such that $u(t) \in \mathcal{U} \subset \mathbb{R}^m$, where*

$$\mathcal{U} = \{u : t \to \mathbb{R}^m / \forall t \in \mathbb{R}^+, u_{i,min} \leq u_i(t) \leq u_{i,max}, \mu_{i,min} \leq \dot{u}_i(t) \leq \mu_{i,max}\} \tag{4}$$

Assumption 2. *The function $\ell(x,u) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is Lipschitz in x, $\forall u \in \mathcal{U}$, i.e there exists a positive scalar b_1 such that*

$$\|\ell(x,u) - \ell(x^*,u)\| \leq b_1 \|x - x^*\| \tag{5}$$

The function $g(x,u) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{r\times q}$ is bounded $\forall x \in \mathbb{R}^n$ and $\forall u \in \mathcal{U}$, that is there exists a scalar b_g such that

$$\|g(x,u)\| \leq b_g$$

and is Lipschitz in x, $\forall u \in \mathcal{U}$, i.e there exists a positive scalar b_2 where

$$\|g(x,u) - g(x^*,u)\| \leq b_2 \|x - x^*\| \tag{6}$$

Assumption 3. *The unknown parameters are supposed to be bounded such that there exists a scalar $b_3 > 0$ verifying*

$$\|\theta\| \leq b_3$$

Since assumption 1 holds, we have put the inputs $u_i(t)$, for $i = 1,\dots,m$, and their derivatives in the same vector

$$\delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_m \\ \delta_{m+1} \\ \vdots \\ \delta_{2m}^T \end{bmatrix} = \begin{bmatrix} u_1 \\ \vdots \\ u_m \\ \dot{u}_1 \\ \vdots \\ \dot{u}_m \end{bmatrix} \tag{7}$$

one can see that the vector δ belongs to a convex polytope, described by

$$\mathcal{P} = [u_{1,\min}, u_{1,\max}] \times \dots \times [u_{m,\min}, u_{m,\max}] \times [\mu_{1,\min}, \mu_{1,\max}] \times \dots \times [\mu_{1,\min}, \mu_{1,\max}] \tag{8}$$

Let us note Φ the set of vertices of the convex polytope $\mathcal{P}$, where its cardinality is equal to 2^{2m}, and described by

$$\Phi = \{\sigma = [\phi_1,\dots,\phi_{2m}]^T \in \mathbb{R}^{2m} / \forall i \in [0,m], \phi_i \in \{u_{i,\min}, u_{i,\max}\} \text{ and } \forall i \in [m+1,2m], \phi_i \in \{\mu_{i,\min}, \mu_{i,\max}\}\} \tag{9}$$

In what follows, we point out the problem of the design of adaptive observer, which allows a simultaneous estimation of the states and the unknown parameters. As these observers are usually designed to be applied in robust and fault tolerant control, the add of the term $Gg(x,u)\theta$ in the system dynamics allows to model uncertainties affecting the system in the case of robust control, or may be used to model faults in the case of fault detection and isolation [18], [26], this term is more general than only the term $G\theta$ as considered in [1] and in the section 3 of the present work.

In a first time, we propose an adaptive observer for system (3) without considering the presence of the nonlinear Lipschitz functions, i.e $\ell(x,u) = 0$ and $g(x,u) = 1$. In a second time, a more general adaptive observer is proposed for the considered class of nonlinear time delay systems (3). The obtained results do not lead to the results of the section before (see the discussion in section 5), which justify the structure of this work.

Before starting the observer design, let us give some useful relations used in this paper. For $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, the following well known inequalities hold

$$2\|x\|\,\|y\| \leq \beta x x^T + \frac{1}{\beta} y y^T, \quad \forall \beta > 0 \tag{10}$$

$$xz^T + zx^T \leq cxx^T + \frac{1}{c} zz^T, \quad \forall c > 0 \tag{11}$$

3 Observer Design without Lipschitz nonlinearities

In this section, we consider system (3) without Lipschitz nonlinear functions, i.e. $\ell(x,u) = 0$ and $g(x,u) = 1$. Thus, an adaptive observer is proposed under the following form

$$\dot{\hat{x}}(t) = A_0\hat{x}(t) + \sum_{i=1}^{m} A_i u_i \hat{x}(t) + Bu(t) + A_{d_0}\hat{x}(t-\tau_0) + \sum_{i=1}^{m} A_{d_i} u_i(t)\hat{x}(t-\tau_i) + G\hat{\theta}(t) + L_x(y(t) - C\hat{x}(t)) \tag{12a}$$

$$\dot{\hat{\theta}}(t) = L_\theta(y(t) - C\hat{x}(t)) \tag{12b}$$

where $\hat{x}(t) \in \mathbb{R}^n$ and $\hat{\theta} \in \mathbb{R}^q$ are the estimated states vector and the estimated unknown parameters vector, respectively. $L_x \in \mathbb{R}^{n\times p}$ and $L_\theta \in \mathbb{R}^{q\times p}$ are the observer's gains to be determined.

As a full order observer, we propose this structure of the observer, since it is the most commonly used in the literature, and had shown performance results. Within this observer, only two observer gains have to be determined, contrary to the full order observer structure as proposed in [23], which may be cumbersome to compute, due to the number of matrices to be computed.

The estimation error has the following dynamics

$$\dot{e}(t) = H(u)e(t) + \sum_{i=0}^{m} H_{d_i} e(t-\tau_i) \tag{13}$$

where

$$e(t) = \begin{bmatrix} e_x(t) \\ e_\theta(t) \end{bmatrix} = \begin{bmatrix} x(t) - \hat{x}(t) \\ \theta(t) - \hat{\theta}(t) \end{bmatrix}$$

$$H(u) = \begin{bmatrix} \sum_{i=1}^{m} A_i u_i - L_x C & G \\ -L_\theta C & 0 \end{bmatrix} \tag{14}$$

$$H_{d_i}(u) = \begin{bmatrix} A_{d_i} u_i & 0 \\ 0 & 0 \end{bmatrix} \tag{15}$$

Notice that we considered $u_0(t) = 1$ for reasons of simplification. Then, system (12) is an adaptive observer for the delayed considered system described by (3) with $\ell(x,u) = 0$ and $g(x,u) = 1$, if and only if the estimation error system described by (13) is asymptotically stable. The stability of the estimation error $e(t)$ and the computation of the observer's gains L_x and L_θ are ensured via the following theorem.

Theorem 1. *Assume that assumption 1 holds. System* (12) *represents an adaptive observer to system* (3) *(with* $\ell(x,u) = 0$ *and* $g(x,u) = 1$*), and the estimation errors system* (13) *is quadratically stable for* $\sigma^j \in \Phi$, $j = 1,\ldots,2^{2m}$, *if there exist matrices*

- $P(\sigma^j) \in \mathbb{R}^{(n+q)\times(n+q)}$ *where*

$$P(\sigma^j) = P^T(\sigma^j) = \sum_{i=0}^{m} \sigma_i^j \begin{bmatrix} P_{i1} & P_{i2} \\ P_{i2}^T & P_{i3} \end{bmatrix} > 0 \tag{16}$$

  $P_{i1} \in \mathbb{R}^{n\times n}$, $P_{i2} \in \mathbb{R}^{n\times q}$ *and* $P_{i2} \in \mathbb{R}^{q\times q}$, *for* $i = 0,\ldots,m$,

- $M \in \mathbb{R}^{(n+p)\times(n+p)}$, *given by*

$$M = \begin{bmatrix} M_{11} & S_2 M_{22} \\ S_1 M_{11} & M_{22} \end{bmatrix} \tag{17}$$

  where $M_{11} \in \mathbb{R}^{n\times n}$ *and* $M_{22} \in \mathbb{R}^{q\times q}$ *are nonsingular matrices.* $S_1 \in \mathbb{R}^{q\times n}$ *and* $S_2 \in \mathbb{R}^{n\times q}$ *are some tuning matrices.*

- $Y_1 \in \mathbb{R}^{n\times p}$ *and* $Y_2 \in \mathbb{R}^{q\times p}$

and a positive scalar γ, *such that the following LMIs holds*

$$\begin{bmatrix} \alpha^j_{(1,1)} & \alpha^j_{(1,2)} & \alpha^j_{(1,3)} & \alpha^j_{(1,4)} & \alpha^j_{(1,5)} \\ * & \alpha^j_{(2,2)} & \alpha^j_{(2,3)} & \alpha^j_{(2,4)} & \alpha^j_{(2,5)} \\ * & * & \alpha_{(3,3)} & \alpha_{(3,4)} & \alpha^j_{(1,5)} \\ * & * & * & \alpha_{(4,4)} & \alpha^j_{(2,5)} \\ * & * & * & * & \frac{-1}{\gamma} I_k \end{bmatrix} < 0 \tag{18}$$

where

$$\alpha^j_{(1,1)} = \sum_{i=1}^{m} \sigma^j_{m+i} P_{i1} + \sum_{i=0}^{m} M_{11} A_i \sigma_i^j + \sum_{i=0}^{m} A_i^T M_{11}^T \sigma_i^j - Y_1 C - C^T Y_1^T - S_2 Y_2 C - C^T Y_2^T S_2^T + \sum_{i=0}^{m} M_{11} A_{d_i} \sigma_i^j + \sum_{i=0}^{m} A_{d_i}^T M_{11}^T \sigma_i^j \tag{19a}$$

$$\alpha^j_{(1,2)} = \sum_{i=1}^{m} \sigma^j_{m+i} P_{i2} + M_{11}G + \sum_{i=0}^{m} A_i^T \sigma_i^j M_{11}^T S_1^T + \sum_{i=0}^{m} A_{d_i}^T M_{11}^T S_1^T \sigma_i^j - C^T Y_2^T - C^T Y_1^T S_1^T \quad (19b)$$

$$\alpha^j_{(2,2)} = \sum_{i=1}^{m} \sigma^j_{m+i} P_{i3} + S_1 M_{11} G + G^T M_{11}^T S_1^T \quad (19c)$$

$$\alpha^j_{(1,3)} = \sum_{i=0}^{m} \sigma_i^j P_{i1} + \sum_{i=0}^{m} A_i^T M_{11}^T \sigma_i^j + \sum_{i=0}^{m} A_{d_i}^T M_{11}^T \sigma_i^j - M_{11} - C^T Y_1^T - C^T Y_2^T S_2^T \quad (19d)$$

$$\alpha^j_{(1,4)} = \sum_{i=0}^{m} \sigma_i^j P_{i2} - M_{12} + \sum_{i=0}^{m} A_i^T M_{11}^T S_1^T \sigma_i^j - C^T Y_1^T S_1^T - C^T Y_2^T + \sum_{i=1}^{m} A_{d_i}^T M_{11}^T S_1^T \sigma_i^j \quad (19e)$$

$$\alpha^j_{(2,3)} = \sum_{i=0}^{m} \sigma_i^j P_{i2}^T - S_1 M_{11} + G^T M_{11}^T \quad (19f)$$

$$\alpha^j_{(2,4)} = \sum_{i=0}^{m} \sigma_i^j P_{i3} - M_{22} + G^T M_{11}^T S_1^T \quad (19g)$$

$$\alpha_{(3,3)} = -M_{11} - M_{11}^T \quad (19h)$$

$$\alpha_{(4,4)} = -M_{22} - M_{22}^T \quad (19i)$$

$$\alpha_{(3,4)} = -S_2 M_{22} - M_{11}^T S_1^T \quad (19j)$$

$$\alpha^j_{(1,5)} = [M_{11}A_{d_0}\sigma_0^j, 0, M_{11}A_{d_1}\sigma_1^j, 0, \ldots, M_{11}A_{d_m}\sigma_m^j, 0] \quad (19k)$$

$$\alpha^j_{(2,5)} = [S_1 M_{11} A_{d_0}\sigma_0^j, 0, S_1 M_{11} A_{d_1}\sigma_1^j, 0, \ldots, S_1 M_{11} A_{d_m}\sigma_m^j, 0] \quad (19l)$$

The observer gains are expressed by

$$L_x = M_{11}^{-1} Y_1 \quad (20a)$$

$$L_\theta = M_{22}^{-1} Y_2 \quad (20b)$$

Proof. The proof of the theorem 1 will be developed into two steps:

- In a first step, we give the stability conditions using a Lyapunov Krasovskii approach for LPV time delay systems, which leads to the resolution of an inequality.
- In a second step, we compute the observers matrices and we transform the obtained inequality into LMIs, using a Polytopic approach.

First Step. Let us consider a Lyapunov Krasovskii function candidate with this form

$$V(e) = \begin{bmatrix} e \\ \dot{e} \end{bmatrix}^T F(u) E \begin{bmatrix} e \\ \dot{e} \end{bmatrix} + \frac{1}{\gamma} \sum_{i=0}^{m} \int_{-\tau_i}^{0} \int_{t}^{t+\beta} \dot{e}^T(s)\dot{e}(s) ds d\beta \quad (21)$$

where $F(u) = \begin{bmatrix} P(u) & M \\ 0 & M \end{bmatrix}$, with $P(u) \in \mathbb{R}^{(n+p)\times(n+p)}$, $P(u) = P(u)^T > 0$ and M is a matrix with a structure described by (17). $E = \begin{bmatrix} I_{(n+q)} & 0 \\ 0 & 0 \end{bmatrix}$ and γ is a positive scalar.

Let us note $\varepsilon(t) = \dot{e}(t)$, and we rewrite system (13) under a descriptor form as follows

$$\begin{bmatrix} I_{(n+q)} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{e}(t) \\ \dot{\varepsilon}(t) \end{bmatrix} = \begin{bmatrix} 0 & I \\ H(u) + \sum_{i=0}^{m} H_{d_i} & -I \end{bmatrix} \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix} + \sum_{i=0}^{m} \begin{bmatrix} 0 \\ H_{d_i} \end{bmatrix} \int_{t}^{t-\tau_i} \varepsilon(\kappa) d\kappa$$

Then, using the latter equation, we differentiate the Lyapunov function in t, which leads to

$$\dot{V}(t) = \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix}^T \mathbb{H} \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix} + \sum_{i=0}^{m} \beta_i(t) - \frac{1}{\gamma} \sum_{i=0}^{m} \int_{t}^{t-\tau_i} \dot{e}^T(s)\dot{e}(s) ds \quad (22)$$

where

$$\mathbb{H} = \begin{bmatrix} h_{(1,1)} & h_{(1,2)} \\ * & h_{(2,2)} \end{bmatrix} \quad (23)$$

with

$$h_{(1,1)} = \dot{P}(u) + MH(u) + H^T(u)M^T + \sum_{i=0}^{m} MH_{d_i} + \sum_{i=0}^{m} H_{d_i}^T M^T$$

$$h_{(1,2)} = P(u) - M + H^T(u)M^T + H_{d_i}^T M^T$$

$$h_{(2,2)} = -M - M^T$$

and

$$\beta_i(t) = \begin{bmatrix} e^T(t) & \dot{e}^T(t) \end{bmatrix} F(u) \begin{bmatrix} 0 \\ H_{d_i} \end{bmatrix} \int_{t}^{t-\tau_i} \dot{e}(s) ds + \int_{t}^{t-\tau_i} \dot{e}^T(s) ds \begin{bmatrix} 0 & H_{d_i}^T \end{bmatrix} F^T(u) \begin{bmatrix} e(t) \\ \dot{e}(t) \end{bmatrix} \quad (24)$$

Using inequality (11), we majorate β_i as follows

$$\beta_i(t) \le \gamma \begin{bmatrix} e^T(t) & \dot{e}^T(t) \end{bmatrix} F(u) \begin{bmatrix} 0 \\ H_{d_i}(u) \end{bmatrix} \begin{bmatrix} 0 \\ H_{d_i}(u) \end{bmatrix}^T F^T(u) \times \begin{bmatrix} e(t) \\ \dot{e}(t) \end{bmatrix} + \frac{1}{\gamma} \int_{t}^{t-\tau_i} \dot{e}^T(s)\dot{e}(s) ds \quad (25)$$

Replacing inequality (25) in the expression of the derivative of the Lyapunov function $V(e)$ described by (22), implies

$$\dot{V}(e) \le \begin{bmatrix} e(t) \\ \dot{e}(t) \end{bmatrix}^T \begin{bmatrix} \overline{h}_{(1,1)} & \overline{h}_{(1,2)} \\ * & \overline{h}_{(2,2)} \end{bmatrix} \begin{bmatrix} e(t) \\ \dot{e}(t) \end{bmatrix}$$

where

$$\overline{h}_{(1,1)} = \dot{P}(u) + MH(u) + H^T(u)M^T + \sum_{i=0}^{m} MH_{d_i} + \sum_{i=0}^{m} H_{d_i}^T M^T + \gamma \sum_{i=0}^{m} MH_{d_i} H_{d_i}^T M^T$$

$$\overline{h}_{(1,2)} = P(u) - M + H^T(u)M^T + H_{d_i}^T M^T + \gamma \sum_{i=0}^{m} MH_{d_i}H_{d_i}^T M^T$$

$$\overline{h}_{(2,2)} = -M - M^T + \gamma \sum_{i=0}^{m} MH_{d_i}H_{d_i}^T M^T$$

The negativity of the derivative of the Lyapunov-Krasovskii function is equivalent to the following inequality:

$$\begin{bmatrix} \overline{h}_{(1,1)} & \overline{h}_{(1,2)} \\ * & \overline{h}_{(2,2)} \end{bmatrix} < 0$$

Applying the Schur complement on the latter inequality allows to get the following inequality,

$$\begin{bmatrix} a_{(1,1)} & a_{(1,2)} & a_{(1,4)} \\ * & a_{(1,3)} & a_{(1,4)} \\ * & * & \frac{-1}{\gamma} I_k \end{bmatrix} < 0 \tag{26}$$

where

$$a_{(1,1)} = \dot{P}(u) + MH(u) + H^T(u)M^T + \sum_{i=0}^{m} MH_{d_i}(u) + \sum_{i=0}^{m} H_{d_i}^T(u)M^T$$

$$a_{(1,2)} = P(u) - M + H^T(u)M^T + \sum_{i=0}^{m} H_{d_i}^T M^T$$

$$a_{(1,3)} = -M - M^T$$

$$a_{(1,4)} = [MH_{d_0}(u), \ldots, MH_{d_m}(u)]$$

with $k = (m+1)(n+q)$.

Second step. Now, using the polytopic approach, we consider that assumption 1 holds and that the inputs and their derivatives belong to the convex polytope $\mathcal{P}$ defined by (8). Then, using the vector δ given by (7), we rewrite the system's matrices, the Lyapunov matrix and its derivative as follows

$$P(u) = P(\delta) = P_0 + \sum_{i=1}^{m} \delta_i P_i, \tag{27a}$$

$$\dot{P}(u) = \overline{P}(\delta) = \sum_{i=1}^{m} \delta_{m+i} P_i \tag{27b}$$

$$H(\delta) = \begin{bmatrix} \sum_{i=0}^{m} A_i \delta_i - L_x C & G \\ -L_\theta C & 0 \end{bmatrix} \tag{27c}$$

$$H_{d_i}(\delta) = \begin{bmatrix} A_{d_i}\delta_i & 0 \\ 0 & 0 \end{bmatrix} \tag{27d}$$

Then, we compute inequality (26) in the whole set of the vertices of the polytope Φ. By taking the matrices $P^j(\sigma_i)$ and M with the form (16) and (17) respectively, we obtain the results given in the proof, where the observer's gains are given by (20). □

Now, as discussed in the introductory section, we consider in the next section a more general class of systems with Lipschitz nonlinearities.

4 Observer Design with Lipschitz nonlinearities

In this section, the objective is to design an adaptive observer for system (3) with the following general structure in order to estimate simultaneously the states vector x and the unknown parameters θ. For that reason, the following adaptive observer is considered

$$\dot{\hat{x}}(t) = \left(A_0 + \sum_{i=1}^{m} A_i u_i\right)\hat{x}(t) + Bu(t) + \ell(\hat{x},u) + A_{d_0}\hat{x}(t-\tau_0) + \sum_{i=1}^{m} A_{d_i}u_i(t)\hat{x}(t-\tau_i) + Gg(\hat{x},u)\hat{\theta}(t) + L_x(y - C\hat{x})$$

$$\dot{\hat{\theta}} = L_\theta(y - C\hat{x}) \tag{28}$$

where $\hat{x} \in \mathbb{R}^n$ and $\hat{\theta} \in \mathbb{R}^q$ are the estimates of the states x and the unknown parameters θ respectively. $L_x \in \mathbb{R}^{n\times p}$ and $L_\theta \in \mathbb{R}^{q\times p}$ are the observer's gains.

The convergence analysis made in this section is different from the section above, due to the presence of the Lipschitz nonlinear functions $\ell(x,u)$ and $g(x,u)$.

In order to ensure the convergence of the proposed adaptive observer, we had to choose the gains L_x and L_θ which guarantee that the errors $e_x(t) = x(t) - \hat{x}(t)$ and $e_\theta = \theta - \hat{\theta}$, converge to zero, in other words, the estimated vectors $\hat{x}$ and $\hat{\theta}$ converge to their actual values. Thus, this section is devoted to obtain the stability conditions in term of LMI.

Let us give the dynamics of the estimation errors $e_x(t)$ and $e_\theta(t)$ as follows

$$\dot{e}_x(t) = \left(A_0 + \sum_{i=1}^{m} A_i u_i - L_x C\right)e_x(t) + (\ell(x,u) - \ell(\hat{x},u)) + A_{d_0}e_x(t-\tau_0) + \sum_{i=1}^{m} A_{d_i}u_i e_x(t-\tau_i) + G\left(g(x,u)\theta - g(\hat{x},u)\hat{\theta}\right) \tag{29a}$$

$$\dot{e}_\theta(t) = -L_\theta C e_x(t) \tag{29b}$$

After adding and subtracting the term $\rho G e_\theta(t)$ to equation (29a), where ρ is a positive scalar, we obtain the following augmented error system

$$\dot{e}(t) = H(u)e(t) + \sum_{i=0}^{m} H_{d_i}(u)e(t-\tau_i) + H_\ell + \overline{G}\,\overline{H}_g \tag{30}$$

where H_{d_i} is defined by (15) and

$$H(u) = \begin{bmatrix} \sum_{i=1}^{m} A_i u_i - L_x C & \rho G \\ -L_\theta C & 0 \end{bmatrix}$$

$$H_\ell = \begin{bmatrix} \ell(x,u) - \ell(\hat{x},u) \\ 0 \end{bmatrix}$$

$$\overline{H}_g = H_g - \begin{bmatrix} \rho e_\theta(t) \\ 0 \end{bmatrix}$$

$$H_g = \begin{bmatrix} g(x,u)\theta - g(\hat{x},u)\hat{\theta} \\ 0 \end{bmatrix}$$

$$\overline{G} = \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix}$$

with $u_0(t) = 1$.

The quadratic stability of the estimation error system is ensured via the following theorem

Theorem 2. *Assuming that assumptions 1, 2 and 3 hold. System* (28) *is an adaptive observer for system* (3)*, and the estimation error is quadratically stable for $\sigma^j \in \Phi$, if there exist matrices*

- $P(\sigma^j) \in \mathbb{R}^{(n+q)\times(n+q)}$ *where*

$$P(\sigma^j) = \sum_{i=0}^{m} \sigma_i^j \begin{bmatrix} P_{i1} & P_{i2} \\ P_{i2}^T & P_{i3} \end{bmatrix} > 0 \tag{31}$$

  $P_{i1} \in \mathbb{R}^{n\times n}$, $P_{i2} \in \mathbb{R}^{n\times q}$ *and* $P_{i2} \in \mathbb{R}^{q\times q}$, *for* $i = 0,\dots,m$,

- $M \in \mathbb{R}^{(n+p)\times(n+p)}$, *such that*

$$M = \begin{bmatrix} M_{11} & S_2 M_{22} \\ S_1 M_{11} & M_{22} \end{bmatrix} \tag{32}$$

  where $M_{11} \in \mathbb{R}^{n\times n}$ *and* $M_{22} \in \mathbb{R}^{q\times q}$. $S_1 \in \mathbb{R}^{q\times n}$ *and* $S_2 \in \mathbb{R}^{n\times q}$ *are some tuning matrices.*

- $Y_1 \in \mathbb{R}^{n\times p}$ *and* $Y_2 \in \mathbb{R}^{q\times p}$

and scalars $\rho > 0$, $c_1 > 0$, $c_2 > 0$ *and* $\gamma > 0$, *such that the LMIs* (33) *hold (see next page), for* $j = 1,\dots,2^{2m}$, *where the blocks* $\alpha^j_{(1,3)}$, $\alpha^j_{(1,4)}$, $\alpha^j_{(1,5)}$, $\alpha^j_{(2,5)}$, $\alpha_{(3,3)}$, $\alpha_{(3,4)}$ *and* $\alpha_{(4,4)}$ *are given by* (19d), (19e), (19k), (19l), (19h), (19j) *and* (19i), *respectively. The blocks* $\omega^j_{(1,1)}$, $\omega^j_{(1,2)}$, $\omega^j_{(2,2)}$, $\omega^j_{(2,3)}$, $\omega^j_{(2,4)}$, $\omega^j_{(1,5)}$, $\omega^j_{(1,6)}$, $\omega^j_{(2,5)}$ *and* $\omega^j_{(2,6)}$ *are as follows*

$$\begin{aligned}
\omega^j_{(1,1)} &= \sum_{i=1}^{m} \sigma_{m+i} P_{i1} + \sum_{i=0}^{m} M_{11} A_i \sigma_i + \sum_{i=0}^{m} A_i^T M_{11}^T \sigma_i \\
&\quad - Y_1 C - C^T Y_1^T - S_2 Y_2 C - C^T Y_2^T S_2^T \\
&\quad + \sum_{i=0}^{m} M_{11} A_{d_i} \sigma_i + \sum_{i=0}^{m} A_{d_i}^T M_{11}^T \sigma_i \\
&\quad + \frac{b_1^2}{c_1} I_n + 2\frac{b_2^2 b_3^2 + b_2 b_3 b_g}{c_2} I_n \\
\omega^j_{(1,2)} &= \sum_{i=1}^{m} \sigma^j_{m+i} P_{i2} + \rho M_{11} G + \sum_{i=0}^{m} A_i^T \sigma_i^j M_{11}^T S_1^T \\
&\quad + \sum_{i=0}^{m} A_{d_i}^T M_{11}^T S_1^T \sigma_i^j - C^T Y_2^T - C^T Y_1^T S_1^T \\
\omega^j_{(2,2)} &= \sum_{i=1}^{m} \sigma_{m+i} P_{i3} + \rho S_1 M_{11} G + G^T M_{11}^T S_1^T \\
&\quad + 2\frac{b_g^2 + b_2 b_3 b_g + \rho^2}{c_2} I_q \\
\omega^j_{(2,3)} &= \sum_{i=0}^{m} \sigma_i^j P_{i2}^T - S_1 M_{11} + \rho G^T M_{11}^T \\
\omega^j_{(2,4)} &= \sum_{i=0}^{m} \sigma_i^j P_{i3} - M_{22} + \rho G^T M_{11}^T S_1^T
\end{aligned}$$

$$\begin{bmatrix} \omega^j_{(1,5)} & \omega^j_{(1,6)} \\ \omega^j_{(2,5)} & \omega^j_{(2,6)} \end{bmatrix} = P(\sigma^j)$$

The observer gains are expressed by

$$L_x = M_{11}^{-1} Y_1 \tag{34a}$$

$$L_\theta = M_{22}^{-1} Y_2 \tag{34b}$$

Proof. We rewrite the error system described by (30) in a descriptor form as follows

$$\begin{aligned}
\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{e}(t) \\ \dot{\varepsilon}(t) \end{bmatrix} &= \begin{bmatrix} 0 & I \\ H(u) + \sum_{i=0}^{m} H_{d_i} & -I \end{bmatrix} \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix} \\
&\quad + \sum_{i=0}^{m} \begin{bmatrix} 0 \\ H_{d_i} \end{bmatrix} \int_t^{t-\tau_i} \varepsilon(k)dk + \begin{bmatrix} 0 \\ H_\ell \end{bmatrix} + \begin{bmatrix} 0 \\ \overline{G}\overline{H}_g \end{bmatrix}
\end{aligned} \tag{35}$$

in order to put it in the Lyapunov function candidate $V(e)$, which will be chosen as in the first part, under the form (21). A computation of the derivative of the Lyapunov function $V(e)$ yields to

$$\begin{aligned}
\dot{V}(e) &= \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix}^T \begin{bmatrix} \overline{W}_{(1,1)} & W_{(1,2)} \\ * & -M - M^T \end{bmatrix} \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix} + \sum_{i=0}^{m} \beta_i(t) \\
&\quad + \alpha_1(e) + \alpha_2(e) - \frac{1}{\gamma} \sum_{i=0}^{m} \int_t^{t-\tau_i} \dot{e}^T(s)\dot{e}(s)ds
\end{aligned}$$

where $\beta_i(t)$ is defined by (24) and

$$\begin{aligned}
\overline{W}_{(1,1)} &= \dot{P}(u) + MH(u) + \sum_{i=0}^{m} MH_{d_i}(u) + H^T(u)M^T \\
&\quad + \sum_{i=0}^{m} H_{d_i}^T(u) M^T
\end{aligned} \tag{36a}$$

$$W_{(1,2)} = P(u) - M + H^T(u)M^T + \sum_{i=0}^{m} H_{d_i}^T M^T \tag{36b}$$

$$\alpha_1(e) = \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u) \begin{bmatrix} 0 \\ H_\ell \end{bmatrix} + \begin{bmatrix} 0 \\ H_\ell \end{bmatrix}^T F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \tag{36c}$$

$$\alpha_2(e) = \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u) \begin{bmatrix} 0 \\ \overline{G}\overline{H}_g \end{bmatrix} + \begin{bmatrix} 0 \\ (\overline{G}\overline{H}_g) \end{bmatrix}^T F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \tag{36d}$$

However, to give an upper bound to the derivative of the Lyapounov function, we had to give an upper bound to some terms. For the term $\beta_i(t)$, an upper bound was given by the inequality (25). So, we proceed, in the sequel, by giving an upper bound to the terms $\alpha_1(e)$ and $\alpha_2(e)$.

Using inequality (11), $\alpha_1(e)$ can be upper bounded as follows

$$\begin{aligned}
\alpha_1(e) &\le \frac{1}{c_1} \begin{bmatrix} 0 \\ H_\ell \end{bmatrix}^T \begin{bmatrix} 0 \\ H_\ell \end{bmatrix} + c_1 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \\
&= \frac{1}{c_1} H_\ell^T H_\ell + c_1 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix}
\end{aligned}$$

However the product $H_\ell^T H_\ell$ can be majorated by the following expression using (1) and (5)

$$\begin{aligned}
H_\ell^T H_\ell &= [\ell(x,u) - \ell(\hat{x},u)]^T [\ell(x,u) - \ell(\hat{x},u)] \\
&\le b_1^2 e_x^T(t) e_x(t)
\end{aligned}$$

$$\begin{bmatrix}
\omega^j_{(1,1)} & \omega^j_{(1,2)} & \alpha^j_{(1,3)} & \alpha^j_{(1,4)} & \omega^j_{(1,5)} & \omega^j_{(1,6)} & M_{11} & S_2M_{22} & 0 & M_{11}G & 0 & \alpha^j_{(1,5)} \\
* & \omega^j_{(2,2)} & \omega^j_{(2,3)} & \omega^j_{(2,4)} & \omega^j_{(2,5)} & \omega^j_{(2,6)} & S_1M_{11} & M_{22} & 0 & S_1M_{11}G & 0 & \alpha^j_{(2,5)} \\
* & * & \alpha_{(3,3)} & \alpha_{(3,4)} & 0 & 0 & M_{11} & S_2M_{22} & 0 & M_{11}G & 0 & \alpha^j_{(1,5)} \\
* & * & * & \alpha_{(4,4)} & 0 & 0 & S_1M_{11} & M_{22} & 0 & S_1M_{11}G & 0 & \alpha^j_{(2,5)} \\
* & * & * & * & \frac{-1}{c_1}I_n & 0 & \dots & & \dots & & \dots & 0 \\
* & * & * & * & * & \frac{-1}{c_1}I_q & 0 & \dots & & & \dots & 0 \\
* & * & * & * & * & * & \frac{-1}{c_1}I_n & 0 & \dots & & \dots & 0 \\
* & * & * & * & * & * & * & \frac{-1}{c_1}I_q & 0 & \dots & \dots & 0 \\
* & * & * & * & * & * & * & * & \frac{-1}{c_2}I_{n+q} & 0 & \dots & 0 \\
* & * & * & * & * & * & * & * & * & \frac{-1}{c_2}I_n & 0 & 0 \\
* & * & * & * & * & * & * & * & * & * & \frac{-1}{c_2}I_q & 0 \\
* & * & * & * & * & * & * & * & * & * & * & \frac{-1}{\gamma}I_k
\end{bmatrix} < 0 \quad (33)$$

then,

$$\alpha_1(e) \le e^T(t)\mathbb{N}_1 e(t) + c_1 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \quad (37)$$

where

$$\mathbb{N}_1 = \begin{bmatrix} \frac{b_1^2}{c_1}I_n & 0 \\ 0 & 0_{q\times q} \end{bmatrix} \quad (38)$$

We note $\overline{\overline{G}} = \begin{bmatrix} 0 & 0 \\ 0 & \overline{G} \end{bmatrix}$, and we give an upper bound to α_2 as follows:

$$\begin{aligned}
\alpha_2(e) &\le \frac{1}{c_2}\begin{bmatrix} 0 \\ \overline{H}_g \end{bmatrix}^T \begin{bmatrix} 0 \\ \overline{H}_g \end{bmatrix} + c_2 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)\overline{\overline{G}}\,\overline{\overline{G}}^T F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \\
&= \frac{1}{c_2}\overline{H}_g^T\overline{H}_g + c_2 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)\overline{\overline{G}}\,\overline{\overline{G}}^T F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix}
\end{aligned}$$

where

$$\overline{H}_g^T\overline{H}_g = \|\overline{H}_g\|^2 \le 2\left(\|H_g\|^2 + \rho^2 e_\theta^T(t)e_\theta(t)\right)$$

with

$$\begin{aligned}
H_g^T H_g &= \left(g(x,u)\theta - g(\hat{x},u)\hat{\theta}\right)^T \left(g(x,u)\theta - g(\hat{x},u)\hat{\theta}\right) \\
&\le \left((g(x,u)\theta - g(\hat{x},u)\theta) + \left(g(\hat{x},u)\theta - g(\hat{x},u)\hat{\theta}\right)\right)^T \\
&\times \left((g(x,u)\theta - g(\hat{x},u)\theta) + \left(g(\hat{x},u)\theta - g(\hat{x},u)\hat{\theta}\right)\right) \\
&\le b_2^2 b_3^2 e_x^T(t)e_x(t) + b_g^2 e_\theta^T(t)e_\theta \\
&+ b_2 b_3 b_g\left(\|e_x^T(t)\|\,\|e_\theta(t)\| + \|e_\theta^T(t)\|\,\|e_x(t)\|\right) \\
&\le b_2^2 b_3^2 e_x^T(t)e_x(t) + b_g^2 e_\theta^T(t)e_\theta \\
&+ b_2 b_3 b_g(e_x^T(t)e_x(t) + e_\theta^T(t)e_\theta)
\end{aligned}$$

The latter inequalities lead to

$$\alpha_2(e) \le 2e^T(t)\mathbb{N}_2 e(t) + c_2 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)\overline{\overline{G}}\,\overline{\overline{G}}^T F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \quad (39)$$

where

$$\mathbb{N}_2 = \begin{bmatrix} \frac{b_2^2 b_3^2 + b_2 b_3 b_g}{c_2}I_n & 0 \\ 0 & \frac{b_g^2 + b_2 b_3 b_g + \rho^2}{c_2}I_q \end{bmatrix} \quad (40)$$

Hence,

$$\begin{aligned}
\dot{V}(e) \le \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T \Bigg(&\begin{bmatrix} \overline{W}_{(1,1)} + \mathbb{N}_1 + 2\mathbb{N}_2 & W_{(1,2)} \\ * & -M - M^T \end{bmatrix} \\
&+ c_1 F(u)F^T(u) + c_2 F(u)\overline{\overline{G}}\,\overline{\overline{G}}^T F^T(u) \\
&+ \gamma \sum_{i=0}^{m} F(u) \begin{bmatrix} 0 \\ H_{d_i}(u) \end{bmatrix} \begin{bmatrix} 0 \\ H_{d_i}(u) \end{bmatrix}^T F^T(u) \Bigg) \begin{bmatrix} e \\ \varepsilon \end{bmatrix}
\end{aligned}$$

Applying the Schur complement on the latter inequality, leads to inequality (41) (see next page), where the block $W_{(1,1)}$ is described by

$$\begin{aligned}
W_{(1,1)} = \dot{P}(u) + MH(u) + H^T(u)M^T + \sum_{i=0}^{m} MH_{d_i}(u) \\
+ \sum_{i=0}^{m} H_{d_i}^T(u)M^T + \mathbb{N}_1 + 2\mathbb{N}_2
\end{aligned}$$

$W_{(1,2)}$, $\mathbb{N}_1$ and $\mathbb{N}_2$ are given by (36b), (38) and (40), respectively.

Finally, using notations (27) and the information on the inputs and their derivatives, we compute inequality (41), to extract the observer gains via LMIs, which completes the proof. □

5 Discussion

1. In the literature, two major approaches have been developed to tackle the design of adaptive observer. These approaches are essentially based on:

   (a) the elaboration of a parameter adaptation law. Here, the unknown parameter vector is deduced from the stability analysis of a state observer and the convergence property of the parameter error is obtained by a persistence of excitation type constraint. Many contributions deal with this approach as in [5], [15], [17], etc.

$$\begin{bmatrix} W_{(1,1)} & W_{(1,2)} & P(u) & M & 0 & M\overline{G} & MH_{d_0} & MH_{d_1} & \dots & MH_{d_m} \\ * & -M-M^T & 0 & M & 0 & M\overline{G} & MH_{d_0} & MH_{d_1} & \dots & MH_{d_m} \\ * & * & -c_1^{-1}I_{n+q} & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ * & * & * & -c_1^{-1}I_{n+q} & 0 & 0 & 0 & 0 & \dots & 0 \\ * & * & * & * & -(c_2)^{-1}I_{n+q} & 0 & 0 & 0 & \dots & 0 \\ * & * & * & * & * & -(c_2)^{-1}I_{n+q} & 0 & 0 & \dots & 0 \\ * & * & * & * & * & * & -\gamma^{-1}I_{n+q} & 0 & \dots & 0 \\ * & * & * & * & * & * & * & -\gamma^{-1}I_{n+q} & \ddots & \vdots \\ * & * & * & * & * & * & * & * & \ddots & 0 \\ * & * & * & * & * & * & * & * & * & -\gamma^{-1}I_{n+q} \end{bmatrix} < 0 \tag{41}$$

(b) an augmented system for which the adaptive observer design is elaborated. In this case, the system dynamics are augmented with the dynamics of its unknown parameters as in [20], [26], [27], [21].

In our work, the design of our adaptive observers are based on the approach described in item (b). However, those results are established by assuming a Lyapunov function where the derivative depends of both the state and parameters errors. So, the assumption of persistent excitation is not required in our work since the matrix appearing in the derivative of the Lyapunov function is not block diagonal, unlike in [15] where the boundedness of this derivative depends only of the state error terms.

2. The use of a descriptor approach and augmenting the estimation error system as done in the present work give more additional degrees of freedom to the problem resolution and allow to overcome the problem of the product between the Lyapunov matrix $P(u)$ and the system's dynamic matrix $H(u)$ (see [28]).

3. First, due to the form of the block diagonal terms $\alpha_{(3,3)}$ and $\alpha_{(4,4)}$ appearing in both LMIs (18) and LMIs (33), the matrices M_{11} and M_{22} should be nonsingular matrices to satisfy the LMIs constraints. Notice that adding and subtracting the term $\rho G e_\theta(t)$ from equation (29a) allow to have the matrix M_{11} in the term $\omega_{(2,2)}$ in the LMI (33) given in theorem 2.

   Second, the matrix M is chosen with the form given by (17) for the following reasons:

   (a) If the matrix M is chosen under the following form

   $$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

   where $M_{11} \in \mathbb{R}^{n\times n}$ and $M_{22} \in \mathbb{R}^{q\times q}$, $M_{12} \in \mathbb{R}^{n\times q}$ and $M_{21} \in \mathbb{R}^{q\times n}$, we will come across the following problem

   $$\begin{bmatrix} Y_{11} \\ Y_{21} \end{bmatrix} = \begin{bmatrix} M_{11} \\ M_{21} \end{bmatrix} L_x$$

   $$\begin{bmatrix} Y_{12} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} M_{12} \\ M_{22} \end{bmatrix} L_\theta$$

   Then, the existence of the observer's gains depends on the following rank conditions

   $$\operatorname{rank}\begin{bmatrix} M_{11} \\ M_{21} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} M_{11} & Y_{11} \\ M_{21} & Y_{21} \end{bmatrix}$$
   $$\operatorname{rank}\begin{bmatrix} M_{12} \\ M_{22} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} M_{12} & Y_{12} \\ M_{22} & Y_{22} \end{bmatrix}$$

   which add some non-convex constraints to satisfy in theorem 1 and 2.

   (b) Putting the matrix M with a diagonal form, i.e. $M_{12} = 0$ and $M_{21} = 0$, implies that the blocks $\alpha_{(2,2)}$ and $\omega_{(2,2)}$ in LMIs (18) and (33) respectively, will be written as follow

   $$\alpha^j_{(2,2)} = \sum_{i=1}^{m} \sigma^j_{m+i} P_{i3}$$
   $$\omega^j_{(2,2)} = \sum_{i=1}^{m} \sigma_{m+i} P_{i3} + \frac{b_g^2 + b_2 b_3 b_g + \rho^2}{c_2} I_q$$

   and one can see that the LMIs (18) and (33) can not be satisfied.

   So, to avoid the above rank constraints, we set $M_{21} = S_1 M_{11}$ and $M_{12} = S_2 M_{22}$ where matrices S_1 and S_2 are a priori chosen tuning parameters. This leads to $Y_{21} = S_1 Y_{11}$ and $Y_{12} = S_2 Y_{22}$.

   However, one can see that, unlike matrix S_1, matrix S_2 does not appear in the diagonal blocks $\alpha_{(2,2)}$ and $\omega_{(2,2)}$ of LMIs (18) and (33), respectively. By the way, we can set $S_2 = 0$. Whereas, we need the condition $S_1 \neq 0$.

4. The obtained results in theorem 2 may appear as an extension of theorem 1. However, it is not the case. For the observer design in section 3, the nonlinear Lipschitz functions are taken as

   $$\ell(x,u) = 0 \ \text{ and } \ g(x,u) = 1$$

   which imply that assumption 2 will be as follows

$$b_1 = 0, \quad b_2 = 0 \text{ and } b_g = 1$$

Applying this assumption on the results of theorem 2, do not lead to the results obtained in theorem 1, due to the fact that this assumption does not cancel some terms in the block $\omega^{j}_{(2,2)}$ of the LMIs (33). In addition, the bound b_3 of the unknown parameters θ is not required in the design of adaptive observer in section 3.

6 Numerical examples

To illustrate the efficiency and the feasibility of our results, we give in the sequel some numerical examples. Let us consider a bilinear time delay system, with

$$A_0 = \begin{bmatrix} -4 & 2 \\ -1 & -1.52 \end{bmatrix} \quad A_1 = \begin{bmatrix} -0.4 & 0.8 \\ 0.27 & -0.4 \end{bmatrix}$$
$$A_{d_0} = \begin{bmatrix} -0.1 & 0.01 \\ -0.11 & -0.68 \end{bmatrix} \quad A_{d_1} = \begin{bmatrix} -0.8 & 0.08 \\ -0.2 & -0.04 \end{bmatrix}$$

We assume that the system is controlled by one bounded input control $u(t) = u_1(t)$

$$-0.2 < u(t) = u_1(t) = 0.2\sin(0.1t) < 0.2$$

with

$$B = \begin{bmatrix} 0.8 \\ 0.01 \end{bmatrix}$$

One can see, that the derivative is also bounded, such that

$$-0.02 < \dot{u}(t) < 0.02$$

Then, we assume that the system is affected by one constant unknown parameters θ, where

$$G = \begin{bmatrix} -0.09 \\ 0.9 \end{bmatrix}$$

The time delays are constant and known such that

$$\tau_0 = 0.9, \text{ and } \tau_1 = 0.4$$

The available measurement vector is given by

$$y(t) = x_1(t)$$

The Lipschitz nonlinear function $g(x,u)$ is bounded and chosen as follows

$$-0.2 \leq g(x,u) = 0.2\sin(u(t) - x_1(t)) \leq 0.2$$

and the function $\ell(x,u)$ is given by

$$\ell(x,u) = \begin{bmatrix} \sin(x_1(t)e^{-0.2u(t)}) \\ \sin(x_2(t)e^{-0.2u(t)}) \end{bmatrix}$$

The unknown parameters θ will be assumed to be bounded by $b_3 = 0.5$ only for the simulation in section 6.2.

6.1 Numerical results related to the first observer design without Lipschitz nonlinearities

By computing the LMIs (18), given in theorem 1, in the set of the vertices of the convex polytope $\mathcal{P}$, using the toolbox "Lmilab", we obtain the following matrices

$$P_0 = \begin{bmatrix} 8.1168 & -1.8823 & -1.7727 \\ -1.8823 & 7.9346 & -3.7539 \\ -1.7727 & -3.7539 & 8.8302 \end{bmatrix}$$
$$P_1 = \begin{bmatrix} 1.1724 & -1.0393 & -0.37411 \\ -1.0393 & 1.1054 & 0.36291 \\ -0.37411 & 0.36291 & 0.081736 \end{bmatrix}$$
$$M_{11} = \begin{bmatrix} 2.3048 & 0.15176 \\ 0.15176 & 2.4665 \end{bmatrix}$$
$$M_{22} = 7.1197$$
$$S_1 = \begin{bmatrix} -0.9 & -1.12 \end{bmatrix}$$
$$S_2 = \begin{bmatrix} -0.003 \\ 0.02 \end{bmatrix}$$
$$Y_1 = \begin{bmatrix} -2.8248 \\ -3.067 \end{bmatrix}$$
$$Y_2 = 5.6283$$
$$\gamma = 0.1$$

which yield to the following observer's gains

$$L_x = M_{11}^{-1} Y_1 = \begin{bmatrix} -1.1484 \\ -1.1728 \end{bmatrix}$$
$$L_\theta = M_{22}^{-1} Y_2 = 0.79053$$

Figures 1, 2 and 3 show that the observer gives a suitable estimation of the states x_1 and x_2 and the unknown parameters θ. They prove also that even if the values of the unknown parameters change, but still constant, the estimation of the states and the unknown parameter converge to their real values.

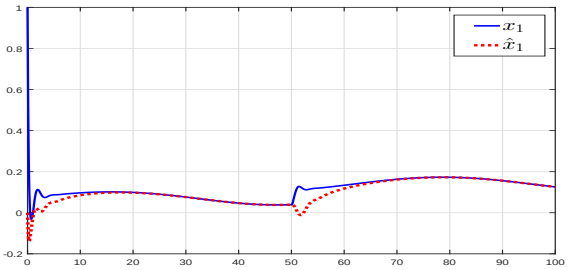


Figure 1: Variation of the state x_1 (—) and its estimation $\hat{x}_1$ ($\cdots$).

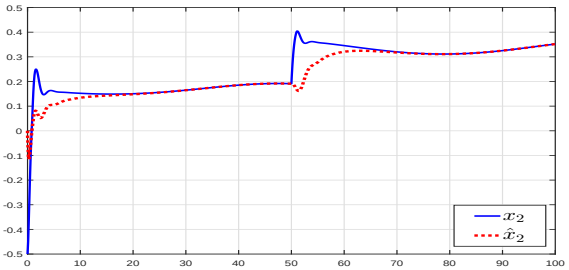


Figure 2: Variation of the state x_2 (—) and its estimation $\hat{x}_2$ ($\cdots$).

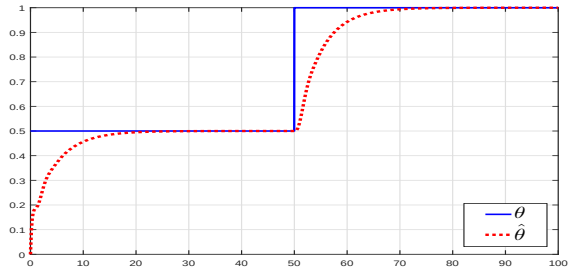


Figure 3: Variation of the unknown parameter θ (—) and its estimation $\hat{\theta}$ ($\cdots$).

6.2 Numerical results related to the second observer design with Lipschitz nonlinearities

By setting the tuning matrices S_1 and S_2 as follows

$$S_1 = \begin{bmatrix} -0.21963 & -0.36278 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 0.11754 \\ 0.049037 \end{bmatrix}$$

and fixing $b_1 = 0.2$, $b_2 = 0.2$ and $\rho = 0.1$, the resolution of the LMIs (33) given in theorem 2 using the toolbox "sdpt3" of Matlab, leads to the following observers gains

$$L_x = M_{11}^{-1} Y_1 = \begin{bmatrix} 0.41697 \\ -1.0876 \end{bmatrix}$$

$$L_\theta = M_{22}^{-1} Y_2 = 1.0292$$

with

$$P_0 = \begin{bmatrix} 1.1844 & -0.28959 & -0.032828 \\ -0.28959 & 0.61048 & 0.015679 \\ -0.032828 & 0.015679 & 0.084515 \end{bmatrix}$$

$$P_1 = \begin{bmatrix} 0.22818 & -0.07406 & -0.024134 \\ -0.07406 & 0.10067 & -0.0056328 \\ -0.024134 & -0.0056328 & -0.00030159 \end{bmatrix}$$

$$M_{11} = \begin{bmatrix} 0.31247 & 0.030714 \\ 0.030714 & 0.4203 \end{bmatrix}$$

$$M_{22} = 0.19653$$

$$Y_1 = \begin{bmatrix} 0.096884 \\ -0.44431 \end{bmatrix}$$

$$Y_2 = 0.20227$$

$$c_1 = 1.12, \quad c_2 = 0.58, \quad \gamma = 0.054$$

Using the obtained gains into the observer yields to a suitable estimation of the states $x(t)$ and the unknown parameter θ. The following figures show the effectiveness of our design.

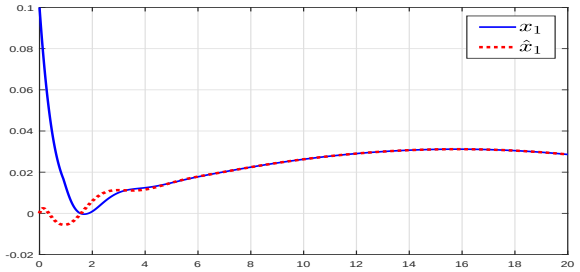


Figure 4: Variation of the state x_1 (—) and its estimation $\hat{x}_1$ ($\cdots$).

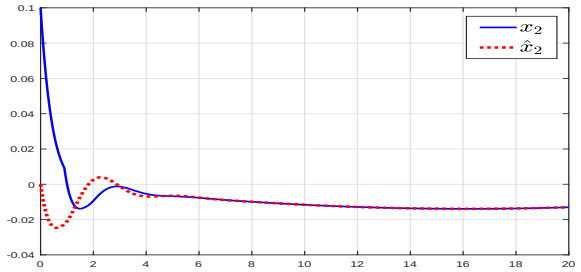


Figure 5: Variation of the state x_2 (—) and its estimation $\hat{x}_2$ ($\cdots$).

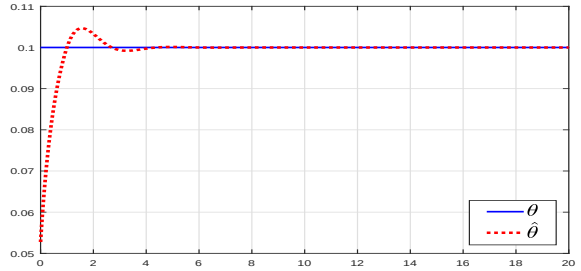


Figure 6: Variation of the unknown parameter θ (—) and its estimation $\hat{\theta}$ ($\cdots$).

7 Conclusion

The proposed adaptive observer for a class of nonlinear time delay system developed in this paper allows a suitable estimation of the state and the unknown parameter vectors, simultaneously. In a first time, a bilinear system with time delays is considered, which was extended by the addition of some nonlinear Lipshitz functions. The observer gains are obtained by solving LMIs in the vertices of a convex polytope. The simulation results show the performances and feasibility of the proposed approach. An extension of our adaptive observer design was presented in a second time, when a nonlinear Lipschitz functions interfere in the dynamics of a bilinear time delay system.

References

[1] A. Sassi, H. Souley Ali, M. Zasadzinski, and K. Abderrahim, "Full order adaptive observer design for time delay bilinear system." *In the Proceedings of the 6th International Conference on Systems and Control*, pp. 267–272, June 2017.

[2] G. Bastin and M. Gevers, "Stable adaptive observers for nonlinear time-varying systems." *IEEE Transactions on Automatic Control*, vol. 33, no. 7, pp. 650–658, 1988.

[3] G. Besançon, "Remarks on nonlinear adaptive observer design." *Systems and Control Letters*, vol. 41, pp. 271–280, 2000.

[4] Y. Choi, H. Shim, Y. Son, and J. Seo, "Design of an adaptive observer for a class of nonlinear systems," *International Journal of Control, Automation and Systems*, vol. 1, no. 1, pp. 28–34, 2003.

[5] M. Ekramian, F. Sheikholeslam, S. Hosseinnia, and M. Yazdanpanah, "Adaptive state observer for Lipschitz nonlinear systems," *Systems and Control Letters*, vol. 62, pp. 319–323, 2013.

[6] L. Xie and P. Khargonekar, "Lyapunov-based adaptive state estimation for a class of nonlinear stochastic systems," *Automatica*, vol. 48, pp. 1423–1431, 2012.

[7] L. Zhao, X. Li, , and P. Li, "Adaptive observer design for a class of mimo nonlinear systems," *Proceedings of the 10th Wold Congress on Intelligent Control and Automation*, pp. 2198–2203, 2012.

[8] S. Ibrir, W. Xie, and M. C. Su, "Observer-based control of discrete-time Lipschitizian nonlinear systems: Application to one-link flexible joint robot." *International Journal of Control*, vol. 78, no. 6, pp. 385–395, 2005.

[9] A. Zemouche and M. Boutayeb, "On LMI conditions to design observers for Lipschitz nonlinear systems." *Automatica*, vol. 49, pp. 585–591, 2013.

[10] J. Richard, "Time-delay systems : An overview of some recent advances and open problems," *Automatica*, vol. 39, no. 10, pp. 1667–1694, 2003.

[11] M. Kadhraoui, M. Ezzine, H. Messaoud, and M. Darouach, "Design of unknown input functional observers for delayed singular systems with state variable time delay," *Transactions on Systems and Control*, vol. 10, pp. 503–509, 2015.

[12] C. Briat, O. Sename, and J. Lafay, "Design of LPV observers for LPV time-delay systems: an algebraic approach." *International Journal of Control*, vol. 84, no. 9, pp. 1533–1542, 2011.

[13] E. Fridman, "Tutorial on Lyapunov-based methods for time-delay systems," *European Journal of Control*, vol. 20, pp. 271–283, 2014.

[14] K. Gu, V. Kharitonov, and J. Chen, *Stability of Time-Delay Systems*. Birkhäuser Boston, 2003.

[15] Y. Cho and R. Rajamani, "A systematic approach to adaptive observer synthesis for nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 42, no. 4, pp. 534–537, 1997.

[16] H. Wang, Y. Dong, and W. Qin, "Adaptive observer design for a class of Lipschitz nonlinear systems," *In the proceedings of the 30th Chinese Control Conference*, pp. 665–669, July 2011.

[17] M. Pourgholi and V. Majd, "Robust adaptive observers design for lipschitz class of nonlinear systems." *International Journal of Electrical and Computer Engineering*, vol. 6, no. 3, pp. 275–279, 2012.

[18] Q. Zhang and B. Deylon, "A new approach to adaptive observer design for mimo systems," *American Control Conference*, vol. 2, pp. 1545 – 1550, 2001.

[19] G. Besançon, J. D. León-Morales, and O. Huerta-Guevara, "On adaptive observers for state affine systems." *International Journal of Control*, vol. 79, no. 6, pp. 581 – 591, 2010.

[20] M. Alma and M. Darouach, "Adaptive observer design for a class of linear descriptor systems." *Automatica*, vol. 50, pp. 578–583, 2014.

[21] M. Farza, I. Bouraoui, T. Ménard, and R. Ben Abdennour, "Adaptive observer for a class of uniformly observable systems with nonlinear parametrization and sampled outputs." *Automatica*, vol. 50, pp. 2951–2960, 2014.

[22] M. Alma, H. S. Ali, M. Darouach, and N. Gao, "An H_∞ adaptive observer design for linear descriptor systems." *2015 American Control Conference*, pp. 4838–4843, May 2015.

[23] A. Sassi, H. Souley Ali, S. Bedoui, K. Abderrahim, and M. Zasadzinski, "Design of a functional adaptive observer for bilinear delayed systems." *13th IFAC workshop on Time Delay Systems TDS*, vol. 49, no. 10, pp. 25–30, 2016.

[24] A. Sassi, M. Zasadzinski, H. Souley Ali, and K. Abderrahim, "Design of an H-infinity adaptive observer for bilinear delayed systems." *The 20th World Congress of the International Federation of Automatic Control (IFAC WC)*, vol. 50, no. 1, pp. 2923–2928, July 2017.

[25] H. Khalil, *Nonlinear Systems, Third Edition*. USA: Prentice Hall, 2002.

[26] Q. Zhang, "Revisiting different adaptive observers through a unified formulation." *Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference.*, pp. 3067–3072, December 2005.

[27] T. Ahmed Ali, R. Postoyan, and F. Lamnahbi Lagarrigue, "Continuous-discrete adaptive observers for state affine systems." *Automatica*, vol. 45, pp. 2986–2990, 2009.

[28] B. Gérard, H. Souley Ali, M. Zasadzinski, and M. Darouach, "H-infinity filter for bilinear systems using LPV approach," *IEEE Transactions on Automatic Control*, vol. 55, no. 7, pp. 1668–1674, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 384-393 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

A Joint Safety and Security Analysis of message protection for CAN bus protocol

Luca Dariz[*,1], Gianpiero Costantino[2], Massimiliano Ruggeri[1], Fabio Martinelli[2]

[1]*CNR-IMAMOTER, Via Canal Bianco 28, Ferrara, Italy*

[2]*CNR-IIT, Via G. Moruzzi, 1, Pisa, Italy*

ARTICLEINFO

Article history:
Received: 15 November, 2017
Accepted: 10 January, 2018
Online: 10 February, 2018

Keywords:
Automotive
CAN bus
Security by design
Security-properties
Safety

ABSTRACT

One of the prominent challenges of the automotive-transportation system is represented by the integration of security and safety properties within protocols, applications and connectivity mechanisms. A joint safety/security design can sometimes expose to trade-offs, since their requirements may not match perfectly or even be incompatible. This paper analyses an example of security and safety design, by combining integrity with encryption considering the constraints of a typical CAN protocol and real-time traffic. The analysis is presented modelling differently attackers, packet fragmentation issues and the residual probability of error of the combined scheme.

1 Introduction

The common thought about cars is that they are mechanical devices employed by passenger to move from a place to another. This is not true any more, or at least is not completely true since cars, or in general vehicles, in the last lustrum evolved to offer several services and connections that turn them into Cyber-Physical Systems (CPS) by the combination of sensing/actuation, processing, storing, and networking capabilities, as described by Fortino et al. [1]. Applications, sensors, park and driving assistants are integrated into commercial vehicles and they are considered standard features present in entry-level model of cars. Features, such as Internet connectivity, enlarge the attack surface of vehicles and, in particular, traditional communication protocols developed to work on isolated environments could not maintain the same level of robustness when new variables are taken into account. This is especially true if one considers that connectivity itself can be provided in multiple heterogeneous ways, which are not always predictable by the manufacturer anymore. For example, in [2], the author proposed a connectivity solution based on smartphones. On the other hand, practically all Electronic Control Units (ECUs) on a car are connected to one or more CAN busses, which is the traditional interface for intra-vehicle communications. The CAN bus use messages whose payload is at maximum 64bits length, and depending on the payload set, it can enable specific functionality of the vehicle, for instance enabling the accelerator of a car. The CAN bus protocol was not designed to embed security properties such as: *Authentication*, *Integrity* and *Confidentiality*, and the security aspects are left to higher layers of the protocols-stack, for instance the application level. An example of attack on the CAN bus protocol is that performed on a Jeep Cherokee by Valasek and Miller in 2015 [3], where the authors showed how to hack and remotely control a Jeep Cherokee. This attack exploit a security flaw in the *In-vehicle Infotainment* (IVI) system of the car to access the CAN bus network of the car. To fix this flaw, the Fiat Chrysler was forced to push a software update [4].

Security issues are a major challenge for connected vehicles, as reported in [5], and this is recognised also to have a relevant impact on vehicle's safety [6]; it is clear that *security* and *safety* are fundamental in the design of an intelligent transportation system, especially regarding the communication protocols and message protection schemes. It is important to note that the combination of safety and security issues is not exclusive of vehicular systems, but it affects the whole IoT world, especially systems which functionality is considered critical (e.g. medical devices).

In this work, we propose a solution to turn the CAN bus protocol as a *Security by Design* protocol by integrating authentication, integrity and confidential-

[*]Corresponding Author: Luca Dariz, CNR-IMAMOTER, Via Canal Bianco 28, Ferrara, Italy

https://dx.doi.org/10.25046/aj030147

ity properties. Our solution proposes the adoption of a Message Authentication Code (MAC), targeted for CAN bus messages, that is then encrypted with an additional key. The message created guarantees authentication and integrity through the MAC, and confidentiality with the additional encryption. Our defence strategy is studied to be applied against a model of attacker that runs both a Honest-But-Curious (HBC) or Fully Malicious attack strategy. Furthermore, our solution is evaluated from a safety point of view, in particular regarding the residual probability of error (P_{re}). This is necessary since the outcome of the security MAC i.e. accept or reject a particular message is a form of error detection which could reveal also transmission errors (e.g. caused by noise), and the message containing the MAC could bring safety-critical information. For example, this scheme could be applied to SAE J1939 Torque/Speed Control 1 message, which embeds a 4-bit checksum within the 8-byte CAN payload. The residual probability of error is first evaluated using an ideal block cipher model, then simulation results are presented for a specific implementation choice. We show that averaging over the secret keys and over the possible messages, the value of P_{re} depends only on the length of the integrity tag used to decide whether the message is valid or not, and this is *independent* from how the integrity tag is generated. On the other hand, we show how the worst case combination of key and message reduces massively the ability of this scheme to detect transmission errors, assuming a simplified channel model. The worst-case P_{re} is then simulated, with considerations on the difficulty of finding the worst-case combination of key and message.

The main contributions of this paper is to analyse a message protection protocol from both a security and safety point of view, and highlight the trade-offs that result from the analysis. This paper is an extension of work originally presented in *IEEE MT-ITS* [7] and our contributions are summarized as follows:

- We improved the description of the analytical model for residual error probability by providing more details on the error model, new explanation on the computation of P_{re} and analytical results;
- We expanded the security analysis and we better modelled the attacker;
- We improved the safety analysis giving more details on the fragmentation and retransmission issues;
- We added more results and better described the analytical model depending on the statistical distribution of errors in the simple case of binary symmetric channel;
- We redraw the plots with the simulation results, using colours, to be easier to understand.

This paper is structured as follows: section 2 reviews the state of the art with regard to MAC algorithms and best practices, as well as error detection mechanisms; section 3 presents the message protection scheme discussed in this paper, discussing some design choices. In section 4, we introduce the attacker model and we make a security consideration on the MAC size. Section 5 exposes two aspects usually relevant for the safety of the system, that are packet fragmentation and the probability of residual error P_{re}. Section 6 discusses other message protection schemes and their security and safety properties, compared to the scheme proposed here. Finally, section 7 concludes the paper with some motivation for future research directions.

2 Related Works

The following works refers to lightweight Message Authentication Code solutions for devices that have limited computational resources, like processors and memory. Chowdhury and Dasbit in [8] introduce *LMAC*, a Lightweight Message Authentication Code (MAC) of 64bits dimension for Wireless Sensor Network that uses hash based symmetric key MAC. The authors show that LMAC is secure against passive and active attacks and it has a low overhead compared with other similar solutions. Another 64bits MAC, called *Chaskey* is presented by the authors of [9]. Chaskey uses 128bits key to generate a MAC which length is of 64bits or more. The authors say that Chaskey generates MAC that are suitable for 32bit Microcontroller and that it does not suffer of MAC truncation [10]. In [11] the authors show two versions of lightweight MAC of 64 and 128bits called *TuLP-64* and *TuLP-128* that are resources efficient and are though for body sensor networks. Then, in [12] the authors presented a lightweight MAC suitable for Smart Grid communications in which two devices reach mutual authentication by sharing a session key exchanged using Diffie-Hellman and a hash-based authentication code technique.

Regarding safety properties, such as the residual probability of error, Schiller and Mattes have analysed different ways of using nested CRC codes, for example in [13]. However, when it comes to cryptographic algorithms the kind of errors treated are related to security properties and possible vulnerabilities, see for example the survey in [14]. To the best of our knowledge, an explicit model for the residual probability of error of a system using cryptographic algorithms has not been developed, although the statistical properties of symmetric ciphers are sometimes studied in-depth, but always in the context of security, see for example [15].

3 Message protection scheme

The main scheme analysed in this paper is based on encryption, and is represented in Figure 1. In general, a message μ with length μ_{size} bits is combined with an integrity tag $\tau = H(k_2, \mu)$ of length τ_{size} , where k_2

is the *authentication key*. The combined $\mu\|\tau$ is then encrypted to obtain the ciphertext $c = ENC(k_1, \mu\|\tau)$, where k_1 is the *encryption key*; the ciphertext is then transmitted on the CAN bus. The receiver receives c', decrypts it and checks if $\tau' = H(k_2, \mu')$, with $\mu'\|\tau' = DEC(k_1, c')$, to decide if the message is valid.

There can be some variations in this scheme; for example the integrity code can be appended to μ to form the plaintext (also known as MAC-then-encrypt approach) with $c = ENC(k_1, \mu\|\tau)$, or it can be excluded from encryption (encrypt-then-MAC approach) with $c = ENC(k_1, \mu)\|\tau$. In both cases the MAC has to be computed using the original message μ. Sometimes the MAC-then-encrypt approach is considered less secure, for example see [16], but in this paper the scheme has no padding and a fixed length of the message, so these considerations do not apply. Considering the CAN bus, the first approach is more practical since there exist encryption algorithms with 64-bit block size, equal to the maximum payload of a CAN message; in this case there is no need for additional data to perform the encryption, and the ciphertext is computed as $c = ENC(k_1, \mu\|\tau)$. On the other hand, if the plaintext is different from the block size (like in the encrypt-then-MAC approach), the cipher must be used in counter mode, or a stream cipher can be used. Either way, there needs to be additional information shared between the sender and the receiver, e.g. a nonce, to perform the encryption; in this case the ciphertext is computed as $c = ENC(k_1, I, \mu)\|\tau$ with I being the shared information.

Another variant is to avoid the use of two different shared keys and define the integrity code as $\tau = H(\mu)$, where $H(\cdot)$ is a hash function like SHA1 or an error-detection code of the CRC family.

In this paper we consider only different possibilities for the integrity tag $\tau = H(\mu)$, which can be a proper Message Authentication Code, a hash function or a CRC; we do not consider then the encrypt-then-MAC approach, so we can define the plaintext $m = \mu\|\tau$.

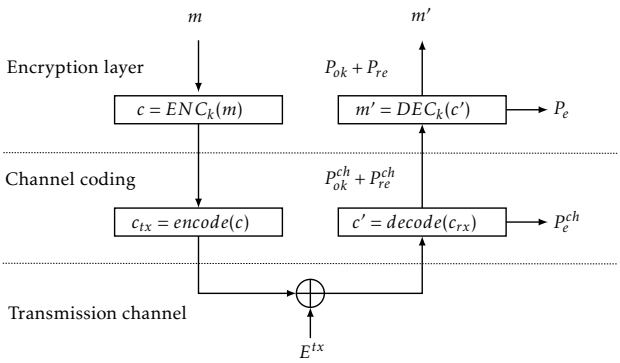


Figure 1: Message protection scheme.

[1] Explanation cited from https://www.krackattacks.com

4 Security considerations

When cars were not connected to the Internet, the attack surface was limited to local entry points that could make cars vulnerable to security attacks. Since connectivity has been embedded as feature inside cars, the attack surface is increased and it is not limited to local attacks, but also remote ones make a vehicle more vulnerable. This provides a higher impact to the security and safety of cars' passengers, and in addition, gives to attackers more and more kinds of attacks that they are able to exploit.

When, we talk about to local attacks, we refer to physical inputs to the vehicle that can be represented, for instance, by the auxiliary jack of the CD-Rom, USB ports, and the ODB2 connector. All these physical connectors require that an attacker has the opportunity to access the car and manipulate or alter the input to perform its attack. This, however, is possible but the impact is limited and less transparent to passenger. On the other side, remote attacks made on the wireless inputs enlarge the attack surface and may make attacks more relevant. Wireless inputs can be: the Bluetooth, Wi-Fi and cellular connections. In particular, exploiting the cellular connection, it would be possible to remotely access the In-vehicle infotainment (IVI) system of a car, and from that reading and writing into the CAN bus [3]. In fact, the IVI system are consoles that often run an operating system such as Windows CE, Linux, and may even run Android. So, attackers may exploit known vulnerabilities of unpatched version of those operating systems to perform their attacks. On this scenario, the Key Reinstallation Attacks (KRACKs) [17] has demonstrated that a vulnerability in the WPA2 protocol could allow *"attackers to use this novel attack technique to read information that was previously assumed to be safely encrypted. This can be abused to steal sensitive information such as credit card numbers, passwords, chat messages, emails, photos, and so on*[1]*."*. Thus, using KRACKs and exploiting an unpatched version of the Wi-FI provided by the IVI, an attacker could jump in the car's network getting sensitive information, and using that access as gateway to force and get in the CAN bus network of the vehicle.

4.1 Attacker Model

In this paper, we analyse the CAN protocol from the security and safety point of views. In particular, this section aims at modelling attackers, defining which are the attacks that they may exploit.

In the modelling phase, we consider that attackers may have local or remote access to the vehicle to compromise the CAN bus by forging or altering messages that may be considered valid by recipients. For instance, attackers may be able to exploit a weakness of the authentication module to remotely access the CAN bus network using a classic IP connection, and

once inside the vehicle, they forge valid message or even altering their contents.

To minimize the power of attackers, our defence strategy foresees that more attackers are not able to forge valid messages, keeping enabled confidentiality of proper messages generated. Also, we aim at identifying messages that were altered, i.e., *losing of integrity*. Thus, our defence strategy is based on three security properties that are:

Authentication: A recipient should be able to verify whether a message is sent by a legitimate sender;

Integrity: A recipient should be able to verify whether a message has been altered during its transmission;

Confidentiality: it guarantees that the content of a message is not revealed to an illegitimate entity, as it can happen with the Man-in-the-Middle (MITM) attack;

We apply our defence strategy against attackers who can play the following roles:

Honest-but-Curious (HBC): Also known as *Passive Attack*; an attacker may exploit the information legitimately gleaned by capturing messages exchanged over the CAN bus network, but he/she will not perform any malicious activity to harvest it.

Fully Malicious (FM): Also known as *Active Attack*; an attacker is able to forge or alter messages that are considered valid, after a verification step, by the recipient. So, the attacker strategy is to succeed in at least one of the following attacks:

- *Impersonation* attack: the attacker is able to assume the identity of one of the legitimate parties;
- *Guessing* attack: the attacker is able to forge a valid MAC after a number of trials.
- *Replay* attack: the attacker is able to re-use valid messages with a malicious or fraudulent aim;
- *Sniffing* attack: the attacker is able to read the content of any messages exchanged through the CAN bus network:

In Table 1, we combine the defence strategy to each attack presented above aiming at blocking or mitigating the corresponding attack.

4.2 Security considerations on encrypted MAC and MAC size

We introduced the Message Authentication Code in the CAN-message payloads to provide *by-design authentication* and *integrity* properties. At the same time, by encrypting $\mu||\tau$, we added the *confidentiality* property in messages to avoid that an attacker sniffs content messages exchanged among ECUs. However, even not considering encryption on messages, but even only the first two properties, an attacker is capable to sniff messages, but she will not easily forge valid messages due to the MAC. With encryption, the attacker's knowledge is still minor, and the probability to forge a valid message is linked to the *guessing attack* plus the encryption of $\mu||\tau$.

Dworkin in [18] points out the importance of choosing a robust MAC to be resistant against the guessing attack. In particular, Dworkin explains that a sound MAC is provided with a size greater than 64bits, i.e., $\tau_{size} \geq$ 64-bits. However, due to the standard CAN bus payload restriction, i.e., 64bit in total, it is very hard to keep that inequality true and we need a workaround to guarantee the security properties with a specific level of risk. Thus, the workaround can be implemented through two different strategies: i) concatenating a MAC which size is at least 64-bit, or ii) limiting the number of repeated trials of an attacker before considering invalid the key that generates the MAC. The first solution may cause the fragmentation issue that we detail better in §5.1. Instead, the second solution can be the best candidate for our workaround on τ_{size}. To this purpose, in [18], Dworkin illustrates how to calculate the right τ_{size} depending on the following two bounds:

MaxInvalids: as the limit on the number of trials that an attacker can perform before the key is retired;

Risk: the highest acceptable probability for an inauthentic message to be accepted as valid;

Then, due the above parameters, the τ_{size} should satisfy the following inequality:

$$\tau_{size} \geq lg(\frac{MaxInvalids}{Risk}). \quad (1)$$

Our goal is to satisfy the inequality 1 with a value of τ_{size} that is greater than or equal to 16. From inequality 1, the system can tolerate up to 30 (2^5) messages before considering the key invalid, and the system can accept 2^{-11}, i.e., $Risk = 2048$, chance of inauthentic messages. So, considering that the payload-size of a CAN-message is generically 48bits (μ_{size}) and the maximum bandwidth of the communication channel is 64bit ($bandwidth_{max}$), we obtain that $\mu_{size} + \tau_{size} \leq bandwidth_{max}$. The inequality 1 defines that the lowest condition to have a $\mu||\tau$ message for the CAN bus protocol is to have $MaxInvalids = 2^5$ and $Risk = 2048$.

5 Safety considerations

5.1 Fragmentation

It is well known that a standard CAN message is too short for the proper implementation of many security properties, since the maximum allowed payload length is 8 bytes. However there are use cases where using a second CAN message (e.g. for authentication

Attack	*Description*	**Defence**
Impersonation	Generating messages being identified as legitimate party	Authentication
Guessing	Forging or altering messages that are valid by the recipient	Confidentiality, Authentication, Integrity
Replay	Re-use messages that are considered valid by the recipient	Authentication
Sniffing	Read content of messages	Confidentiality

Table 1: Summary of attack and defence strategies.

purposes) is not acceptable since it introduce unnecessary latency in the complete reception of a message; this is mainly due to the low speed of CAN bus, which is typically 250 kB/s on vehicular networks, and can reach the maximum of 1 MB/s. Another reason is the increase of residual error rate in the communication (see for example the appendix D.5.2 of ISO 15998 [19]); however, this increase could be tolerated more easily since, to a first approximation, the P_{re} of a two-message scheme is roughly the double of the P_{re} of a single message. With these limitations, common application requirements allow for a rather limited security level achievable without changing the network communication protocol stack, and is fundamentally due, for a point-to-point communications, to the limited payload length of a single CAN 2.0 bus message. The use of CAN-FD could relax these limitations, but even in this case the maximum payload length is 64 bytes.

5.2 Retransmissions

In order to tolerate packet loss, communication protocols usually employ some form of re-transmission handling. Basically this means that if an error is detected on the packet, or if a timeout for packet reception expires, the transmitter sends again the same packet. Different variants of ARQ schemes (Automatic Repeat reQuest) exist, but the essential principle is that a data packet is sent more than one time without modifications. While this feature is usually handled transparently at link or transport layer, it is a potential security vulnerability, since an attacker could either inject errors or transmit duplicated packets to perform a replay attack (§4.1). Similar attacks are not hypothetical but have been demonstrated in reality, one of the most recent examples being the key reinstallation attack against WPA2 [17]. It is then desirable to handle retransmissions at the application level, if any, so sensitive security information like nonces, which must be used exactly one time, are handled properly. It is also worth considering that fragmentation makes retransmissions more complicated to handle efficiently, since here a small portion of a packet can be lost.

5.3 Probability of Residual Error

The analysis of the probability of residual error P_{re} in a communication protocol is usually a difficult task. In order to simplify the model, it is assumed that the channel coding is independent from the encryption scheme and it is possible to obtain a measure of P_{re} for channel coding, which correspond to P_{re}^{ch} with reference to Figure 1. In this way, it is possible to focus only on the performance of the message protection scheme discussed in this paper. Here P_{re} is obtained assuming Shannon's ideal cipher model, which has been used in other works like [20], and the results are compared with simulations where real algorithms are used for encryption and integrity.

5.3.1 Error model

According to Shannon's model, an ideal cipher is a random family of permutations, chosen independently for each possible key. More precisely, suppose $\mathcal{K}$ is the set of all keys and $\mathcal{M}$ is the set of all messages. An ideal block cipher is a map $ENC : \mathcal{K} \times \mathcal{M} \to \mathcal{M}$ where, for each key $k \in \mathcal{K}$, the encryption function $ENC_k(\cdot) = ENC(k, \cdot)$ is a random permutation on the message set $\mathcal{M}$ (independent of any other permutation). The same considerations apply to the decryption function $DEC_k(\cdot) = DEC(k, \cdot)$, which being the inverse of a random permutation is itself a random permutation.

In this context, it is more useful to fix a specific pair (k, m) of key and message, and consider the cipher text message set as:

$$\mathcal{C}' = \{c' : c' = c + e', e' \in \mathcal{E}'\} \tag{2}$$

while the plain text message set is:

$$\mathcal{M}' = \{m' : m' = m + e, e \in \mathcal{E}\} \tag{3}$$

where the set $\mathcal{E}'$ is the set of all possible undetected error vectors after channel decoding and before decryption, and $\mathcal{E}$ is the set of all error vectors after decryption. The addition here correspond to a bitwise XOR. While the elements of $\mathcal{E}'$ correspond basically to undetected transmission errors, the elements of $\mathcal{E}$ can be seen as undetected transmission errors transformed by the decryption function. For this reason, the distribution of the values in $\mathcal{E}$ depends on the distribution of $\mathcal{E}'$, in a different way for each different pair (k, m), since the function $E(k, \cdot)$ will correspond to a different and independent permutation. The relation between $\mathcal{E}'$ and $\mathcal{E}$ is then of the form $ENC : \mathcal{K} \times \mathcal{M} \times \mathcal{E}' \to \mathcal{E}$ which can be rewritten as the map $ENC_{(k,m)} : \mathcal{E}' \to \mathcal{E}$ identified by a specific pair of values (k, m). The explicit form can be derived from Figure 1 as

$$\begin{aligned} e = ENC_{(k,m)}(e') &= DEC(k, c') && (4)\\ &= DEC(k, c + e') && (5)\\ &= DEC(k, ENC(k, m) + e') + m && (6) \end{aligned}$$

Being $ENC(k,\cdot)$ and $DEC(k,\cdot)$ random permutations, equation 6 represent a random map from $\mathcal{E}'$ to $\mathcal{E}$. Each pair (k,m) correspond then to a different map $ENC_{(k,m)}$, independent from other maps. This means that each error vector e' on the transmission channel correspond to a random error vector e after decryption. Assuming that the pair (k,m) is chosen uniformly from the set $\mathcal{K}\times\mathcal{M}$, the random error vector e is itself uniformly distributed. In practice, the relation between e and e' is highly non-linear, since it strongly depends on the ENC and DEC functions, which in real block ciphers are usually highly non-linear functions themselves. This means that the approximation of uniform distribution of e, for any given e' and varying (k,m), is accurate as long as these conditions apply:

1. the real block cipher approximates a random permutation;
2. the messages in $\mathcal{M}$ are uniformly distributed;
3. the keys in $\mathcal{K}$ are uniformly distributed.

Condition 1 is itself an important security property, as shown for example in [21]. Condition 2 can be false depending on the context; for example, in a closed industrial control network the message set is usually limited; furthermore, messages in $\mathcal{M}$ can include integrity checks (as in the scheme discussed here) which alter the distribution of the plaintext. Condition 3 is another important security property, although some block ciphers have *weak keys* (e.g. DES and Blowfish). These conditions can be relaxed to having either the messages or the key uniformly distributed, provided that condition 1 still apply.

5.3.2 Computation of P_{re}

In this section, we assume these conditions are satisfied. Given that, as shown in section 3, the original message is $m = \mu\|\tau$, and the received message is $m' = \mu'\|\tau'$, the probability of residual error can be defined as the joint probability

$$P_{re} = P(\mu \neq \mu', \tau' = H(\mu')) \tag{7}$$

which correspond to the probability of having a correct integrity tag ($\tau' = H(\mu')$) but the decoded message is different from the original ($\mu \neq \mu'$) due to transmission errors. Considering that $ENC_{(k,m)}$ is a random map, each element of $\mathcal{C}'$ can correspond to all the possible values of $\mathcal{M}'$. If the elements of $\mathcal{C}'$ are uniformly distributed (such as when considering a random message attack) the probability that $\tau' = H(\mu')$, averaged over all the possible (k,m), can be computed by counting as

$$P^{avg}(\tau' = H(\mu')) = \frac{2^{\mu_{size}}}{2^{\mu_{size}+\tau_{size}}} = \frac{1}{2^{\tau_{size}}} \tag{8}$$

which is the probability of guessing a valid message. On the other hand, when transmission errors are considered, only one of the elements of $\mathcal{C}'$ correspond to the correct message, while all other elements have cumulative probability P_{re}^{ch}. The residual probability of error in this case can be computed as

$$P_{re}^{avg} = P_{re}^{ch}\frac{2^{\mu_{size}}-1}{2^{\mu_{size}+\tau_{size}}-1} \approx \frac{P_{re}^{ch}}{2^{\tau_{size}}}. \tag{9}$$

Both equations 8 and 9 are valid without making any assumption on the actual algorithm H used for computing τ.

This measure of P_{re} is, however, a measure for the average case, while from the safety point of view it is necessary to consider the worst case, that is:

$$P_{re}^{wc} = \max_{\substack{k\in\mathcal{K}\\ m\in\mathcal{M}}} P_{re} \tag{10}$$

This again can be computed by counting, but this time the actual distribution of $\mathcal{E}$, given by transmission errors, must be considered, specifically the one which maximise equation 10, which correspond to the worst-case pair (k^{wc}, m^{wc}). Using the ideal cipher model, the distribution in $\mathcal{E}$ can be obtained from a permutation of the distribution in $\mathcal{E}'$, so a simpler way to compute P_{re}^{wc} is to take the $2^{\mu_{size}}-1$ error vectors of $\mathcal{E}'$ with higher probability $P_{e'}$ and sum their probability. Clearly, the upper bound for the various $\mathcal{E}$ is

$$P_{re}^{wc} \leq P_{re}^{ch} \tag{11}$$

where equality would mean that the encryption procedure, even with an integrity check, is not effective at all in detecting transmission errors when (k^{wc}, m^{wc}) are used, because the relevant transmission errors e cause the plaintext to have an non-correctable error. More specifically, this correspond to the case where the $2^{\mu_{size}}-1$ most probable error vectors in $\mathcal{E}$ cover practically the whole amount of P_{re}^{ch}. The fact that this is independent from the particular algorithm H is because of the random map; in other words it is impossible to design an efficient integrity algorithm for this scheme as the distribution of e can not be known a priori.

This consideration exposes the trade-off between safety (as related to protection against random errors) and security (as related to protection against a malicious adversary). In the first case the statistical distribution of transmission errors is usually concentrated on a restricted set of values, and error detection codes are designed to detect the most probable errors, achieving a very high detection rate. On the other hand, if a malicious adversary is considered, all error patterns must be distributed ideally uniformly, otherwise an attacker may exploit its statistical characteristics.

5.3.3 Analytical results

To illustrate this problem, we consider a simplified model with a binomial distribution $\mathcal{E}' \sim B(n,k,p)$ (which would correspond in Figure 1 to the case with no channel coding and a Binary Symmetric Channel with probability of bit error p). The worst case P_{re}^{wc} can then be obtained by first listing the probability P_l

of each error vector e'_l with l bit set (note that must be $l > 0$); this list is then sorted incrementally and then the first values are taken, one error vector at a time, until exactly $2^{\mu_{size}} - 1$ error vectors are chosen. As shown in figure 2, where the worst-case residual probability of error is computed with the algorithm exposed above for different values of p in a binomial distribution, the value of P_{re} rapidly increases as the distribution of errors is more concentrated. Here τ_{size} is fixed to 64 bytes; similar results are obtained for different values. In figure 3 instead, μ_{size} is fixed to 256 bits and the worst-case P_{re} is plotted, and here the effect of the concentration of error distribution is also visible with effects similar to figure 2. This algorithm can be extended to the case where channel coding allows approximating the residual error distribution, for example if a CRC with a known minimum hamming distance H_{CRC} is used (some examples are available from the work of Koopman, see [22], for different payload lengths). In this case, the error vectors e_l with $l < H_{CRC}$ are not considered since they are detected by channel coding. However if channel coding include other information in the CRC (such as the length of the data payload) this approximation must be further refined.

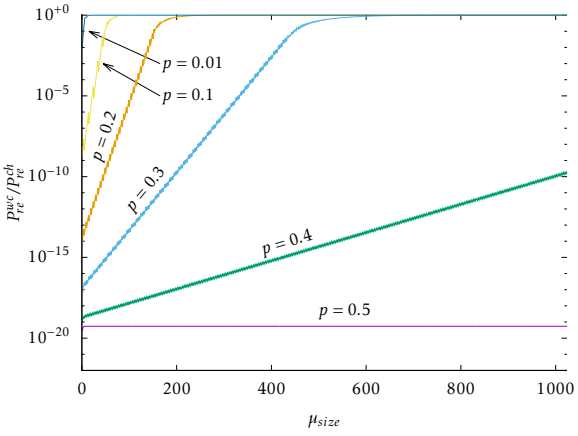


Figure 2: Normalised worst-case P_{re} obtained through computation, for different values of p. τ_{size} is fixed to 64 bytes.

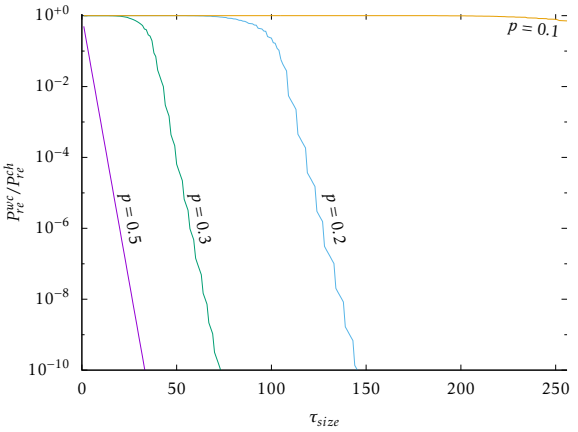


Figure 3: Normalised worst-case P_{re} obtained through computation, for different values of p. μ_{size} is fixed to 256 bytes.

5.3.4 Simulation results

The simulations have been performed applying the encryption scheme to randomly selected (k, m) and using different error vectors to obtain an approximation of P_{re}. For the ENC cipher, DES and AES have been used, while for H we used CRC, a truncated SHA1 hash function and a truncated HMAC scheme based on SHA1. The simulation have been implemented as a C++ program using the Nettle v2.7.1 cryptographic library. The simulations were run on a Intel i7 8-core laptop with 8 GB RAM; for each pair (k, m) a number of error patterns e' are generated, with the respective probability, and the normalised output P_{re} is computed as the ratio between the erroneous messages that resulted in $\tau = H(\mu)$ and the total number of erroneous messages. The plotted results represent a normalised $\frac{P_{re}}{P_{re}^{ch}}$, where the value of 1 represent equality in equation 11. The value of τ_{size} varies from 0 to 16; greater values, which in theory correspond to a lower P_{re}, have not been simulated since they would have taken too much time to yield a result with reasonable precision; however the results are still meaningful with respect to the theoretical model. The complexity of the simulation has three factors: k, m and e'. For example, using 1000 different keys, 1000 different messages and 10000 error patterns the total number of iterations is $1000 \cdot 1000 \cdot 10000 = 10^{10}$. However each value of P_{re} is evaluated using 10000 samples, so lower values close to 10^{-4} will have a lower accuracy. This explains the convenience to simulate with low τ_{size}.

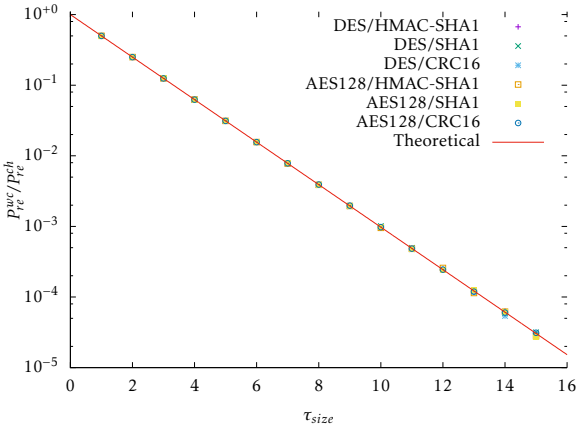


Figure 4: Normalised average P_{re} obtained through simulation and from equation 9, for different values of τ_{size}.

In Figure 4 the normalised P_{re}^{avg} is plotted, both resulting from equation 9 and from simulations. The correspondence between the theoretical model and the simulation results is very good, and the results are independent either from the H algorithm used to compute τ and the encryption algorithm ENC. Only a small glitch is visible for $\tau_{size} = 14$, presumably due to the relatively small number of iterations. The simulations results are accurate because all the P_{re} are av-

eraged over the pairs (k,m).

In Figure 5 and 6 the normalised P_{re}^{wc} is plotted with varying τ_{size}. The numerical values, displayed with a continuous line, are evaluated with the algorithm described in section 5.3.2, while the simulation results are taken as the highest value of P_{re} among all tested (k,m). In this case the simulations do not match the theoretical model; the reason is that while the theoretical model assumes that (k^{wc},m^{wc}) is known, in practice this is not true, although for some ciphers it could be feasible to calculate it. In this case a great number of (k,m) combinations are randomly chosen and the worst case is considered. However, being unable to scan all the (k,m) space, it is unlikely to find the worst case but only a "bad" pair (k,m) is found, for which P_{re} differ significantly from the average case.

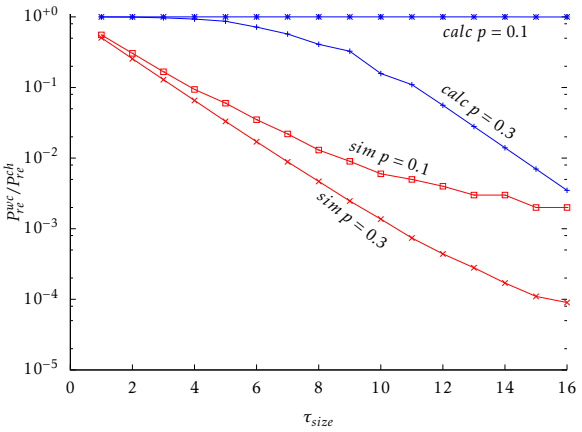


Figure 5: Normalised worst-case P_{re} obtained with $\mathcal{E}' \sim B(n,k,p)$ through calculation and simulation using a DES/SHA1 scheme, with $\mu_{size} + \tau_{size} = 64$, for different values of τ_{size}.

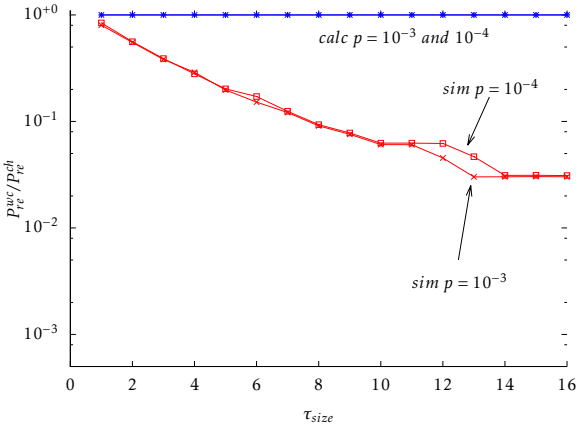


Figure 6: Normalised worst-case P_{re} obtained with $\mathcal{E}' \sim B(n,k,p)$ through calculation and simulation using a DES/SHA1 scheme, with $\mu_{size} + \tau_{size} = 64$, for different values of τ_{size}.

In Figure 7 and 8 the normalised P_{re}^{wc} is plotted with varying p. For low values of p, the value of P_{re}^{wc} does not change a lot, since the most probable error patterns are always the ones with 1 bit error. On the other hand, with higher p the value of P_{re}^{wc} approaches P_{re}^{avg}, which is reached with $p = 0.5$, corresponding to a uniform distribution. The issue of finding the worst case (k,m) is then different depending on the bit error probability of $B(n,k,p)$. For low p, approximately under 0.1, it is easier to find a pair (k,m) with high P_{re} since the most probable error patterns are the ones with only 1 bit error and are exactly $\mu_{size} + \tau_{size}$. On the other hand, for higher p, the most probable error patterns are a much great number, because it is easier to find more than one bit error. This explains the difficulty of finding the pair (k^{wc},m^{wc}) to simulate P_{re}^{wc} accurately.

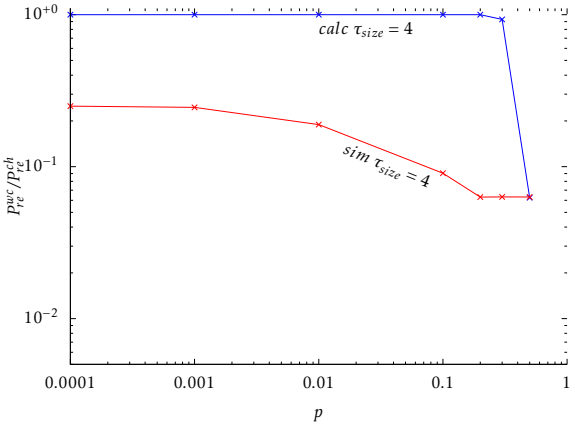


Figure 7: Normalised worst-case P_{re} obtained with $\mathcal{E}' \sim B(n,k,p)$ through calculation and simulation using a DES/SHA1 scheme, with $\mu_{size} + \tau_{size} = 64$, for different values of p.

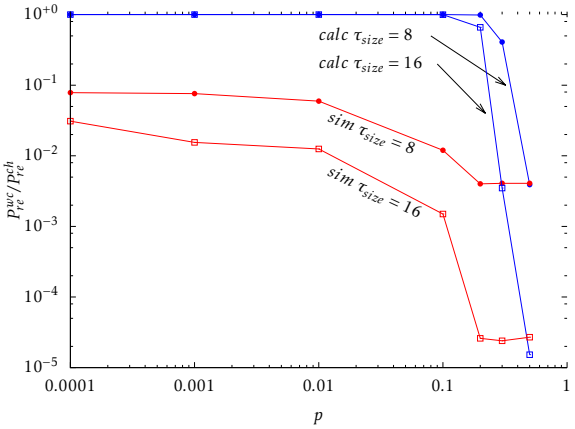


Figure 8: Normalised worst-case P_{re} obtained with $\mathcal{E}' \sim B(n,k,p)$ through calculation and simulation using a DES/SHA1 scheme, with $\mu_{size} + \tau_{size} = 64$, for different values of p.

6 Discussion

In Table 2 different message protection schemes are compared and ordered with decreasing security properties. The message protection schemes addressed

Scheme	Resists to	Security properties	Leaking out	Safety properties
ENC + MAC	Fully Malicious with chosen plaintext	Confidentiality Authentication Integrity	-	Strongly depends on (k, m)
ENC + CRC	Fully Malicious with chosen plaintext	Confidentiality Integrity	-	Strongly depends on (k, m)
plain + MAC	Fully Malicious	Authentication Integrity	Plaintext to FM and HBC attackers	Depends on $H()$
plain + CRC	Honest-But-Curious	-	Plaintext to FM and HBC attackers	Good under common channel assumption [13]
plain	-	-	Plaintext to FM and HBC attackers	-

Table 2: Summary of message protection schemes.

in this paper correspond to the ENC+MAC and ENC+CRC schemes, depending on the choice of $H()$, to resist a fully malicious attacker with chosen plaintext. The main alternative scheme, which does not consider the Confidentiality property, is evaluated for reference, based on literature work. Here the trade-off appears clear comparing the ENC+CRC and plain+CRC scheme; while the first has better security properties, the latter has better safety properties under common channel models, because CRC codes are designed specifically for correcting transmission errors.

7 Conclusion

In this paper, we have presented an analysis showing that a security property, like encryption, directly influences the probability of residual error, which is a safety property. On the other hand, the restricted size of the CAN bus payload forces the length of a MAC code to respect the fragmentation constraints which can be imposed by real-time requirements. With respect to similar schemes without encryption, by using a second CRC in addition to that one at the physical layer, the combination *ENC+CRC* performs worse; this is due to the intrinsic properties of block ciphers, which transform the distribution of errors to uniform on average. Other message protection schemes could have a less drastic impact on the worst-case error detection capability, or even this error detection capability could be embedded into the encryption algorithm itself, but then the risk is to offer the possibility for a side-channel attack.

Future works include the design and study of different message protection schemes, to offer a better trade-off between safety and security, for example improving the performance of the integrity tag τ with respect to transmission errors. Additionally, an experimental testbed on a real CAN bus would be useful to assess the performance of the protocol.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgement This work has been partially supported by the GAUSS (MIUR, PRIN 2015) and by the H2020 EU funded NeCS (GA 675320).

References

[1] G. Fortino, A. Rovella, W. Russo, C. Savaglio, "On the Classification of Cyberphysical Smart Objects in the Internet of Things" in *Proceedings of the 5th International Workshop on Networks of Cooperating Objects for Smart Cities (UBICITEC 2014)*, Berlin, Germany, Apr 14, 2014, pp. 86-94.

[2] G. Aloi, G. Caliciuri, G. Fortino, R. Gravina, P. Pace, W. Russo, C. Savaglio, "Enabling IoT interoperability through opportunistic smartphone-based mobile gateways" in *Journal of Network and Computer Applications*, Volume 81, 2017, Pages 74-84. https://doi.org/10.1016/j.jnca.2016.10.013

[3] C. Valasek and C. Miller, "Remote Exploitation of an Unaltered Passenger Vehicle" in *DEFCON 23*, 2015.

[4] A. Greenberg, "After jeep hack, chrysler recalls 1.4m vehicles for bug fix," Online, Jul 2015. [Online]. Available: https://www.wired.com/2015/07/jeep-hack-chrysler-recalls-1-4m-vehicles-bug-fix/

[5] B. McCluskey, "Connected cars: the security challenge for autonomous vehicles," *IET Engineering and Technology Magazine*, Feb 2017. [Online]. Available: https://eandt.theiet.org/content/articles/2017/02/connected-cars-the-security-challenge-for-autonomous-vehicles/

[6] C. Kim, "Safety challenges for connected cars," *IEEE Transportation Electrification Newsletter*, Jun 2016. [Online]. Available: http://tec.ieee.org/newsletter/june-2016/safety-challenges-for-connected-cars

[7] L. Dariz, M. Selvatici, M. Ruggeri, G. Costantino, and F. Martinelli, "Trade-off analysis of safety and security in CAN bus communication," in *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017, Naples, Italy, June 26-28, 2017*, pp. 226–231. [Online]. Available: https://doi.org/10.1109/MTITS.2017.8005670

[8] A. R. Chowdhury and S. DasBit, "Lmac: A lightweight message authentication code for wireless sensor network," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6. [Online]. Available: https://doi.org/10.1109/GLOCOM.2015.7417118

[9] N. Mouha, B. Mennink, A. Van Herrewege, D. Watanabe, B. Preneel, and I. Verbauwhede, *Chaskey: An Efficient MAC Algorithm for 32-bit Microcontrollers*. Cham: Springer International Publishing, 2014, pp. 306–323. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-13051-4_19

[10] N. Ferguson, "Authentication weaknesses in gcm," Comments submitted to NIST Modes of Operation Process, May 2005.

[11] Z. Gong, P. Hartel, S. Nikova, S.-H. Tang, and B. Zhu, "Tulp: A family of lightweight message authentication codes for body sensor networks," *Journal of Computer Science and Technology*, vol. 29, no. 1, pp. 53–68, 2014. [Online]. Available: http://dx.doi.org/10.1007/s11390-013-1411-8

[12] M. M. Fouda, Z. M. Fadlullah, N. Kato, R. Lu, and X. S. Shen, "A lightweight message authentication scheme for smart grid communications," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 675–685, Dec 2011. [Online]. Available: http://dx.doi.org/10.1109/TSG.2011.2160661

[13] F. Schiller and T. Mattes, "Analysis of nested crc with additional net data by means of stochastic automata for safety-critical communication," in *2008 IEEE International Workshop on Factory Communication Systems*, May 2008, pp. 295–304. [Online]. Available: http://dx.doi.org/10.1109/WFCS.2008.4638714

[14] A. Barenghi, L. Breveglieri, I. Koren, and D. Naccache, "Fault injection attacks on cryptographic devices: Theory, practice, and countermeasures," *Proceedings of the IEEE*, vol. 100, no. 11, pp. 3056–3076, Nov 2012. [Online]. Available: http://dx.doi.org/10.1109/JPROC.2012.2188769

[15] A. Rimoldi, "On algebraic and statistical properties of aes-like ciphers," Ph.D. dissertation, University of Trento, 2010. [Online]. Available: http://eprints-phd.biblio.unitn.it/151/

[16] S. Vaudenay, "Security flaws induced by cbc padding - applications to ssl, ipsec, wtls ..." in *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques: Advances in Cryptology*, ser. EUROCRYPT '02. London, UK, UK: Springer-Verlag, 2002, pp. 534–546. [Online]. Available: http://dl.acm.org/citation.cfm?id=647087.715705

[17] M. Vanhoef and F. Piessens, "Key reinstallation attacks: Forcing nonce reuse in WPA2," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM Press, Nov. 2017, pp. 1313–1328. [Online]. Available: http://dx.doi.org/10.1145/3133956.3134027

[18] M. Dworkin, "Recommendation for block cipher modes of operation: The cmac mode for authentication," NIST Special Publication 800-38B, May 2005.

[19] ISO, "Earth-moving machinery - machine-control systems (mcs) using electronic components - performance criteria and tests for functional safety," The International Organization for Standardization, Genève, Switzerland, Tech. Rep. ISO 15998, 2008.

[20] C. Petit, F.-X. Standaert, O. Pereira, T. G. Malkin, and M. Yung, "A block cipher based pseudo random number generator secure against side-channel key recovery," in *Proceedings of the 2008 ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS '08. New York, NY, USA: ACM, 2008, pp. 56–65. [Online]. Available: http://doi.acm.org/10.1145/1368310.1368322

[21] L. Grassi, "New approaches for distinguishers and attacks on round-reduced aes," Cryptology ePrint Archive, Report 2017/832, 2017, http://eprint.iacr.org/2017/832.

[22] P. Koopman, "Best crc polynomials," 2017, https://users.ece.cmu.edu/~koopman/crc/.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 394-403 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Decentralized Control Approaches of Large-Scale Interconnected Systems

Rabeb Ben Amor*, Salwa Elloumi

Advanced Systems Laboratory in Polytechnic School of Tunisia, BP. 743, 2078, Marsa, Tunisia

ARTICLEINFO

Article history:
Received: 29 November, 2017
Accepted: 23 January, 2018
Online: 10 February, 2018

Keywords:
Interconnected systems
Decentralized stabilization
LMI

ABSTRACT

In this paper, we investigate the decentralized control problem for large-scale interconnected systems. The synthesis of the decentralized controller consists in determining gains which ensure the stability of the global system. To calculate these gains, three approaches are presented. Our main contribution is to develop a new decentralized stabilization approach which the decentralized local gains are calculated and formulated via the resolution of linear matrix inequalities (LMIs) problem. A numerical simulation comparison of the three methods is performed on an interconnected double-parallel inverted pendulum.

1 Introduction

This paper is an extension of the work originally we presented in the International Conference on Advanced Systems and Electric Technologies, 2017 [1]. This work treats three approaches dealing with the decentralized control of interconnected systems.

In fact, large-scale interconnected systems have received considerable attention in recent years due to its presence in several fields such as power electronics, robotics, communication, aerospace, transportation networks, manufacturing processes, biochemical applications and others. Designing a centralized control for these systems may not be efficient due to the modular nature of the system that can prevent the sharing of information between the various subsystems. Thus it is important to decompose the large-scale system into several subsystems. This decomposition which can be physical or mathematical, can make structures easier to control. This includes the implementation of decentralized control law.

In this way, it is necessary to decompose the global system into a number of interconnected subsystems for which, instead of a single centralized controller, a set of independent decentralized controllers is built. Thanks to its structure, the decentralized control has several advantages, mainly: the minimization of the information rate processed by the control units, the simplicity of the developed control laws compared to the centralized case and the improvement of the reliability of data transfer using only local information.

Many works in literature have been devoted to the decentralized control problems for interconnected systems. The decentralized adaptive control has been studied in [2–6]. The robust decentralized control is presented in [7–9]. The decentralized control using sliding mode approach is developed in [10–13].

Decentralized stabilization problem is the subject of our work. This problem is extensively studied in the literature and different design approaches were proposed accordingly [14–18]. To ensure the stability of the interconnected system formed by n subsystems, it is necessary to verify the local stability at each subsystem as well as the overall stability taking into account their interconnection.

The main contribution of this paper consists in developing some conditions allowing the synthesis of decentralized control laws that will ensure the stability of the overall interconnected system. In this way, we propose in this work a new decentralized stabilizing control approach for the interconnected systems. Indeed, the outcomes of this development are formulated in terms of linear matrix inequalities (LMIs).

The presented methods in our paper are applied to the physical system of two inverted pendulums interconnected by a spring. For the design of the decentralized control scheme, each pendulum should be seen as a subsystem. Many works used the typical system easily isolated into two subsystems to approve the validity of their proposed decentralized control approaches [19–22].

The rest of the manuscript is structured according to the following outline : The second section is reserved to formulate the problem and present the

*Corresponding Author : Rabeb Ben Amor, Email: rabebbenamor1@gmail.com, Salwa Elloumi, Email: salwa.elloumi@laposte.net

https://dx.doi.org/10.25046/aj030148

studied interconnected system formed by two parallel inverted pendulums coupled by a spring. In section 3, decentralized control approaches for the interconnected systems are presented, which are the decentralized quadratic optimal control and the decentralized pole-placement control. The last part of this section focuses on the development of a new decentralized stabilization control approach by using the Linear Matrix Inequalities Formulation. Section 4 is devoted to the implementation of the decentralized control approaches presented and developed in the previous section on the studied system. A comparative study between the three control approaches is presented to prove the validity of the new proposed approach. Finally, conclusions and some perspectives are given in the fifth section.

2 Problem Formulation and Description of the Studied Dynamic System

2.1 Problem Formulation

Large-scale interconnected systems are represented as follows:

$$\dot{x}_i = A_i x_i + B_i u_i + \sum_{\substack{j=1 \\ j\neq i}}^{n} H_{ij} x_j, i = 1, 2, .., n \qquad (1)$$

where $x_i \in \mathbb{R}^{n_i}$ and $u_i \in \mathbb{R}^{n_i}$ denote the state vector and the control vector of i^{th} subsystem, respectively.
$A_i \in \mathbb{R}^{n_i \times n_i}$ is the state matrix and $B_i \in \mathbb{R}^{n_i \times m_i}$ is the control matrix of each subsystem.
H_{ij} represents the term of interconnection between the i^{th} subsystem and the other subsystems.

The global interconnected system composed of N subsystems can be rewritten in a compact form as follows:

$$\dot{x} = Ax + Bu + Hx \qquad (2)$$

where:

- $x^T = [x_1^T, x_2^T, .., x_n^T]$ is the state vector of the global system ;
- $u^T = [u_1^T, u_2^T, .., u_n^T]$ is the control vector of the global system ;
- $A = diag[A_i], B = diag[B_i]$;
- H is the matrix formed by the terms of interconnection having the following form

$$H = \begin{bmatrix} 0 & H_{12} & \cdots & H_{1n} \\ H_{21} & 0 & & H_{2n} \\ \vdots & \vdots & \ddots & \\ H_{n1} & \cdots & \cdots & 0 \end{bmatrix}$$

2.2 Description of the Studied System : Double Inverted Pendulums Coupled by a Spring

We present in this section the description of the studied system formed by two interconnected inverted pendulum and its dynamic modeling.
In this system, two identical inverted pendulums of mass m directly mounted on the motor shafts in parallel where τ_1 and τ_2 are the input torques of each motor. These pendulums are connected to each other by an elastic spring of constant k which is mounted at the height a.
θ_1 and θ_2 are the angular displacements and of the pendulums from vertical.
New particular movements appear compared to the single movement of the individual pendulum. The interconnected inverted pendulums system is shown in figure 1 [23].

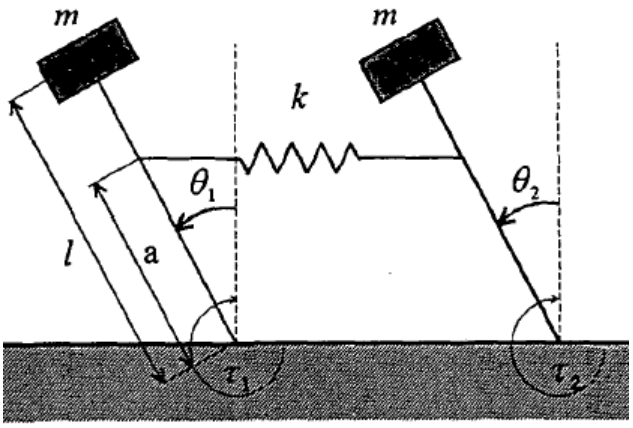


Figure 1: Modeling of the parallel inverted pendulum

The Lagrangian is defined as the difference between the kinetic energies and the potential energies of the system.

The kinetic energy for each pendulum is described by the following form:

$$T_i = \frac{1}{2} J_i \dot{\theta}_i^2 \qquad (3)$$

where J_i is the moment of inertia of the i^{th} pendulum and $\dot{\theta}_i$ is the angular velocity of i^{th} pendulum.

The total kinetic energy of the global system is then:

$$T = \frac{1}{2} J_1 \dot{\theta}_1^2 - \frac{1}{2} J_2 \dot{\theta}_2^2 = -\frac{1}{2} ml^2 \dot{\theta}_1^2 - \frac{1}{2} ml^2 \dot{\theta}_2^2 \qquad (4)$$

The potential energy for each mass is represented as follows:

$$V_i = mgl(1 - cos\theta_i) \qquad (5)$$

The potential energy of the spring is calculated using Hooke's law:

$$V_{spring} = \frac{1}{2}kx^2 = \frac{1}{2}k(-a\,sin\theta_1 + a\,sin\theta_2)^2 \quad (6)$$

The total potential energy of the system is given by:

$$\begin{aligned} V &= mgl(1-cos\theta_1) + mgl(1-cos\theta_2) \\ &+\tfrac{1}{2}k(-a\,sin\theta_1 + a\,sin\theta_2)^2 \end{aligned} \quad (7)$$

The Lagrangian of the interconnected studied system is written as follow:

$$\begin{aligned} L &= T - V \\ &= -\tfrac{1}{2}ml^2\dot{\theta}_1^2 - \tfrac{1}{2}ml^2\dot{\theta}_2^2 - mgl(1-cos\theta_1) \\ &-mgl(1-cos\theta_2) + \tfrac{1}{2}k(a\,sin\theta_1 - a\,sin\theta_2)^2 \end{aligned} \quad (8)$$

The Euler-Lagrange equations are given by:

$$\begin{cases} \frac{d}{dt}[\frac{\partial L}{\partial \dot{\theta}_1}] - \frac{\partial L}{\partial \theta_1} = \tau_1 \\ \\ \frac{d}{dt}[\frac{\partial L}{\partial \dot{\theta}_2}] - \frac{\partial L}{\partial \theta_2} = \tau_2 \end{cases} \quad (9)$$

Using Lagrange equations (9), we can easily show that the nonlinear equations of motion of the parallel inverted pendulum system are:

$$\begin{cases} -ml^2\ddot{\theta}_1 + mgl\,sin\theta_1 - ka^2[cos\theta_1(sin\theta_1 - sin\theta_2] = \tau_1 \\ -ml^2\ddot{\theta}_2 + mgl\,sin\theta_2 - ka^2[cos\theta_1(sin\theta_2 - sin\theta_1] = \tau_2 \end{cases} \quad (10)$$

Assuming a small angular displacement, the nonlinear equations of motion (10) can be replaced by the following linear model around the equilibrium point $\theta_1 = \theta_2 = 0$:

$$\begin{cases} -ml^2\ddot{\theta}_1 + mgl\,\theta_1 - ka^2(\theta_1 - \theta_2) = \tau_1 \\ -ml^2\ddot{\theta}_2 + mgl\,\theta_2 - ka^2(\theta_2 - \theta_1) = \tau_2 \end{cases} \quad (11)$$

So, the dynamics of the studied system composed of the two interconnected inverted pendulums are described by the following equations:

$$\begin{cases} -ml^2\ddot{\theta}_1 = mgl\theta_1 - ka^2(\theta_1 - \theta_2) - \tau_1 \\ -ml^2\ddot{\theta}_2 = mgl\theta_2 - ka^2(\theta_2 - \theta_1) - \tau_2 \end{cases} \quad (12)$$

For the design of the decentralized control scheme, each pendulum should be seen as a subsystem. Equations (12) can be writen into state equations with a standard choice of state variable for the i^{th} pendulum:

$$x_i(t) = \begin{bmatrix} \theta_i(t) \\ \dot{\theta}_i(t) \end{bmatrix}$$

The system consisted of two interconnected inverted pendulums is then described by the following state equations:

$$\begin{cases} \dot{x}_1 = A_1x_1 + B_1u_1 + H_1x_2 \\ \dot{x}_2 = A_2x_2 + B_2u_2 + H_2x_1 \end{cases} \quad (13)$$

with

- x_1, x_2 the state vectors of the subsystems ;
- u_1, u_2 the control vectors of the subsystem such as the input torque of each motor ;

The matrices and interconnection terms of the i^{th} subsystem are given by:

$$A_i = \begin{bmatrix} 0 & 1 \\ \frac{g}{l} - \frac{ka^2}{ml^2} & 0 \end{bmatrix}, B_i = \begin{bmatrix} 0 \\ \frac{-1}{ml^2} \end{bmatrix}$$
$$H_i = \begin{bmatrix} 0 & 0 \\ \frac{ka^2}{ml^2} & 0 \end{bmatrix}, i = 1, 2$$

The global system formed by two identical inverted pendulums coupled by a spring can be expressed by the following global state representation:

$$\dot{x} = Ax + Bu + Hx \quad (14)$$

where

- $x^T = [x_1^T, x_2^T]$ is the state vector,
- $u^T = [u_1^T, u_2^T]$ is the control vector.
- $A = diag(A_1, A_2)$ is the characteristic matrix:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{g}{l} - \frac{ka^2}{ml^2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{g}{l} - \frac{ka^2}{ml^2} & 0 \end{bmatrix}$$

- $B = diag(B_1, B_2)$ is the control matrix:

$$B = \begin{bmatrix} 0 & 0 \\ \frac{-1}{ml^2} & 0 \\ 0 & 0 \\ 0 & \frac{-1}{ml^2} \end{bmatrix}$$

- H is the term of interconnection:

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{ka^2}{ml^2} & 0 \\ 0 & 0 & 0 & 0 \\ \frac{ka^2}{ml^2} & 0 & 0 & 0 \end{bmatrix}$$

with

m The mass of each pendulum, in Kg
l The rod length, in m
a The connecting position of the spring, in m
k The stiffness of spring, in N/m
g The acceleration of gravity, in $m.s^{-2}$
θ_i The angular displacement of the i^{th} pendulum, in Rad
τ_i The input torque of i^{th} motor, in $N.m$

3 Decentralized Control Approaches of Interconnected Systems

Possible control strategies for large-scale interconnected systems are generally based on a decentralized solution. A decentralized control structure applied to a process of n interconnected subsystems is shown in Fig 2.
The decentralized control partitions the measurement information and elaborates a local and independent control law for each subsystem.
It is necessary to check the stability of the interconnected system by examining two main aspects:
- Local stability: at each subsystem.
- Overall stability: taking into account the interconnections.

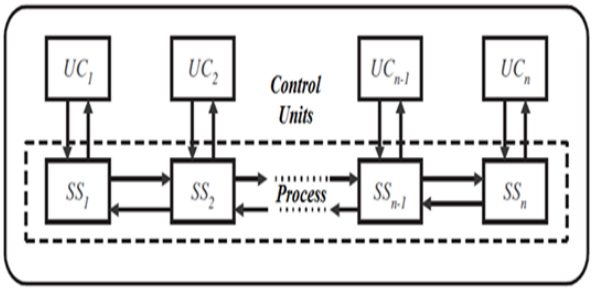


Figure 2: Decentralized control structure

The synthesis of the decentralized controller consists in determining the local gains K_i which ensure the stability of the overall closed-loop system.
To respect the decentralized information structure constraint, each subsystem is controlled by the local control law:

$$u_i(x) = -K_i x_i \quad i = 1,..,n \tag{15}$$

which leads to the following global control law of the overall system (2):

$$u(x) = -Kx \tag{16}$$

where $K = diag(K_1, K_2, .., K_n)$ is the block diagonal control gain matrix. Using global state-feedbacks, we get the closed loop system dynamics as following:

$$\dot{x} = A_f x \tag{17}$$

where:

$$A_f = \begin{bmatrix} A_1 - B_1K_1 & H_{12} & \cdots & H_{1n} \\ H_{21} & A_2 - B_2K_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1} & \cdots & \cdots & A_n - B_nK_n \end{bmatrix}$$

To calculate the gains K_i, different approaches can be considered.

3.1 Decentralized Quadratic Optimal Control

The decentralized control synthesis consists in considering the decoupled subsystems defined by the following state equations:

$$\dot{x}_i = A_i x_i + B_i u_i \tag{18}$$

and minimizing the modified quadratic criteria [24]:

$$J_i = \frac{1}{2}\int_0^\infty e^{2\alpha t}(x_i^T Q_i x_i + u_i^T R_i u_i)dt \tag{19}$$

Let $Q_i(n_i \times n_i), i = 1,..,n$ semi positive definite matrices, $R_i(m_i \times m_i), i = 1,..,n$ positive definite matrices and α a positive real.
The decentralized optimal control laws for each isolated subsystem can be expressed as a linear state feedback:

$$u_i = -K_i x_i, \quad i = 1, 2, .., n \tag{20}$$

where

$$K_i = R_i^{-1} B_i^T P_i \tag{21}$$

and P_i is the symmetric positive definite matrix solution of the following algebraic Riccati equation:

$$A_i^T P_i + P_i A_i - P_i(B_i R_i^{-1} B_i^T)P_i + 2\alpha P_i + Q_i = 0 \tag{22}$$

These decentralized state feedbacks applied to the interconnected system lead to the following global state representation:

$$\dot{x} = (A - BR^{-1}B^T P)x + Hx \tag{23}$$

where $R^{-1} = diag[R_i^{-1}]$ and $P = diag[P_i]$.
A sufficient condition to guarantee the stability of the overall system taking into account the interconnections, is given by the following theorem which proof is detailed in Appendix A.

Theorem 1 [24]:
The decentralized control law (16) is globally and asymptotically stabilizable for system (17) if the matrix F, given by(24), is positive definite.

$$\begin{aligned} F &= 2\alpha P + W - (PH + H^T P) \\ W &= Q + PBR^{-1}B^T P, \quad Q = diag[Q_i] \end{aligned} \tag{24}$$

3.2 Decentralized Pole-Placement Control

Pole-placement technique is a controller design method in which we determine the places of the closed loop system poles on the complex plane by setting a controller gain.
In this work we will apply this method for interconnected systems composed of n different subsystems that can be easily isolated. Firstly, it is necessary to verify the local stability.

For each subsystem, Ackermann's formula is used to find the control gain matrices.

Theorem 2: Ackermann's formula [25]
The controllability matrix C can be formed from:

$$C = [B\ AB...A^{n-1}B]$$

The feedback matrix K can be found as:

$$K = [0\ 0...1]C^{-1}P_d(A)$$

where P_d is the desired characteristic polynomial.

Using the local gain matrices obtained by Akermann's formula for each subsystem, the matrix in closed loop A_f of the global system taking into account the interconnection is given by:

$$A_f = \begin{bmatrix} A_1 - B_1K_1 & H_{12} & \cdots & H_{1n} \\ H_{21} & A_2 - B_2K_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1} & \cdots & \cdots & A_n - B_nK_n \end{bmatrix}$$

Stability condition:
In order to be stable, the eigenvalues of the system $\dot{x} = A_f x$ must all lie strictly in the left half of the complex s-plane. That means, the eigenvalues must all have strictly negative real parts.

3.3 Synthesis of a Decentralized Stabilization Control

This section deals with the global asymptotic stabilization of linear interconnected systems within the framework of Linear Matrix Inequalities (LMIs). We present the development of a new decentralized control approach.
To compute the gain matrix K, so that the closed loop system (17) is asymptotically stable, let consider the quadratic Lyapunov function represented by the following form:

$$V(x) = x^T Px \tag{25}$$

where P is a positive definite symmetric matrix of the following form:

$$A = \begin{bmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & P_n \end{bmatrix}$$

The time derivative of $V(x)$ is developped as :

$$\begin{aligned} \dot{V}(x) &= \dot{x}^T Px + x^T P\dot{x} \\ &= x^T A_f^T Px + x^T PA_f x \\ &= x^T (A_f^T P + PA_f)x \end{aligned} \tag{26}$$

The global asymptotic stability of system (17)provided with the decentralized control law (16) is ensured when the time derivative $\dot{V}(x)$ is negative definite which is equivalent to:

$$A_f^T P + PA_f < 0 \tag{27}$$

We note this expression by $\breve{A} = A_f^T P + PA_f$ with

$$A_f = \begin{bmatrix} A_1 - B_1K_1 & H_{12} & \cdots & H_{1n} \\ H_{21} & A_2 - B_2K_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1} & \cdots & \cdots & A_n - B_nK_n \end{bmatrix}$$

$\breve{A}$ can be written as:

$$\breve{A} = \begin{bmatrix} (A_1 - B_1K_1)^T & H_{12}^T & \cdots & H_{1n}^T \\ H_{21}^T & (A_2 - B_2K_2)^T & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1}^T & \cdots & \cdots & (A_n - B_nK_n)^T \end{bmatrix}$$

$$\begin{bmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & P_n \end{bmatrix} + \begin{bmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & P_n \end{bmatrix}$$

$$\begin{bmatrix} A_1 - B_1K_1 & H_{12} & \cdots & H_{1n} \\ H_{21} & A_2 - B_2K_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1} & \cdots & \cdots & A_n - B_nK_n \end{bmatrix}$$

So,

$$\breve{A} = \begin{bmatrix} A_1^T P_1 + P_1 A_1 - K_1^T B_1^T P_1 - P_1 B_1 K_1 & P_1 H_{12} + H_{21}^T P_2 \\ P_2 H_{21} + H_{12}^T P_1 & \ddots \\ \vdots & \vdots \\ P_n H_{n1} + H_{1n}^T P_1 & P_n H_{n2} + H_{2n}^T P_2 \end{bmatrix}$$

$$\begin{bmatrix} \cdots & P_1 H_{1n} + H_{n1}^T P_n \\ \ddots & \vdots \\ \ddots & \vdots \\ \cdots & A_n^T P_n + P_n A_n - K_n^T B_n^T P_n - P_n B_n K_n \end{bmatrix}$$

$$< 0 \tag{28}$$

Multiplying (28) on the right and then on the left by P^{-1} where P^{-1} is also symmetric positive definite matrix, inequality (28) becomes:

$$\begin{bmatrix} P_1^{-1}A_1^T + A_1P_1^{-1} - P_1^{-1}K_1^T B_1^T - B_1K_1P_1^{-1} & H_{12}P_2^{-1} + P_1^{-1}H_{21}^T \\ H_{21}P_1^{-1} + P_2^{-1}H_{12}^T & \ddots \\ \vdots & \vdots \\ H_{n1}P_1^{-1} + P_n^{-1}H_{1n}^T & \cdots \end{bmatrix}$$

$$\begin{bmatrix} \cdots & H_{1n}P_n^{-1} + P_n^{-1}H_{n1}^T \\ \ddots & \vdots \\ \ddots & \vdots \\ \cdots & P_n^{-1}A_n^T + A_nP_n^{-1} - P_n^{-1}K_n^T B_n^T - B_nK_nP_n^{-1} \end{bmatrix}$$

$$< 0 \tag{29}$$

It should be noted that the inequality matrix (29) has nonlinearities that are difficult to solve. We then use the changes of variables (30) and (31).

$$S_i = P_i^{-1} \tag{30}$$

$$L_i = K_iP_i^{-1} \tag{31}$$

Thus, the problem (29) can be rewritten to the form of linear matrix inequalities :

$$\begin{bmatrix} \breve{a}_{11} & \breve{a}_{12} & \cdots & \breve{a}_{1n} \\ \breve{a}_{21} & \breve{a}_{22} & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \breve{a}_{n1} & \breve{a}_{n2} & \cdots & \breve{a}_{nn} \end{bmatrix} < 0 \tag{32}$$

where:
$\breve{a}_{11} = S_1 A_1^T + A_1 S_1 - B_1 L_1 - L_1^T B_1^T$
$\breve{a}_{12} = H_{12} S_2 + S_1 H_{21}^T$
$\breve{a}_{1n} = H_{1n} S_n + S_1 H_{n1}^T$
$\breve{a}_{21} = H_{21} S_1 + S_2 H_{12}^T$
$\breve{a}_{22} = S_2 A_2^T + A_2 S_2 - B_2 L_2 - L_2^T B_2^T$
$\breve{a}_{n1} = H_{n1} S_1 + S_n H_{1n}^T$
$\breve{a}_{n2} = H_{n2} S_2 + S_n H_{2n}^T$
$\breve{a}_{nn} = S_n A_n^T + A_n S_n - B_n L_n - L_n^T B_n^T$

In order to find the gain matrices K of the decentralized control law, we have to solve the following LMI problem:

$$\begin{cases} S_i > 0 \quad i = 1, .., n \\ (32) \end{cases} \tag{33}$$

The following result is proved::
The interconnected system (17) provided with the decentralized control law (16) is asymptotically stable if LMI problem (33) is feasible.

4 Simulation results

This section is devoted to the implementation of the three decentralized control approaches exposed and developed in the previous section.
It consists in studying the stability by decentralized quadratic optimal control, decentralized pole-placement control and decentralized stabilization control based on LMI applied to the interconnected inverted pendulums system (Figure1), presented in Section2. The parameters of the studied system are summarized in Table 1.
In last party of this section, we carry out a comparative study between these three studied decentralized approaches to confirm the validity and the efficiency of the proposed approach.

Parameter	Value	Unit
m	0.4489	Kg
l	0.325	m
a	0.21	m
k	340.22	N/m

Table 1: The studied system parameters

Using the numerical parameters, model (14) of interconnected system composed of two parallel inverted pendulums is given by:

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -286.2486 & 0 & 316.4332 & 0 \\ 0 & 0 & 0 & 1 \\ 316.4332 & 0 & -286.2486 & 0 \end{bmatrix} x + \begin{bmatrix} 0 & 0 \\ -21.0903 & 0 \\ 0 & 0 \\ 0 & -21.0903 \end{bmatrix} u \tag{34}$$

To improve the performance of the studied system, we will apply the different studied approaches to guarantee an adequate stabilization.

4.1 Application of the decentralized optimal control approach

For this decentralized control, we focus on minimizing the modified quadratic criteria (19) for each separate pendulum.
The weighting factors are selected as follows:

$$\alpha = 0.2,$$
$$R_1 = R_2 = 0.0043$$
$$Q_1 = Q_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

The positive definite solution P_i of the Ricatti equation for each inverted pendulum is obtained by solving the equations (22):

$$P_1 = \begin{bmatrix} 0.0724 & 0.0014 \\ 0.0014 & 0.0002 \end{bmatrix}$$
$$P_2 = \begin{bmatrix} 0.0725 & 0.0014 \\ 0.0014 & 0.0002 \end{bmatrix} \tag{35}$$

Using (20) and (21) we obtain the decentralized control gain matrices:

$$K1 = [-7.0068 \ -0.8247]$$
$$K2 = [-7.0068 \ -0.8247]$$

To guarante the stability of the overall interconnected double-inverted pendulum, we should verify the theorem(24) when calculating the matrix F:

$$F = \begin{bmatrix} 0.3360 & -0.0278 & 0 & 0 \\ -0.0278 & 0.0030 & 0 & 0 \\ 0 & 0 & 0.3360 & -0.0278 \\ 0 & 0 & -0.0278 & 0.0030 \end{bmatrix}$$

The eigenvalues of the matrix F are given by:

$$\begin{bmatrix} 0.3383 \\ 0.0007 \\ 0.3388 \\ 0.0007 \end{bmatrix}$$

We can easily verify that matrix F is positive definite, so the decentralized control law stabilizes asymptotically the overall interconnected system (17).

The performances of the controlled system are shown in Figure3. The curves present the evolution of

the system state variables with decentralized optimal control, when some perturbations occur on θ_1 and θ_2. From the simulation results shown in these curves, it can be seen that the decentralized optimal control is able to enhance stability of the studied system in approximately 0.6 seconds.

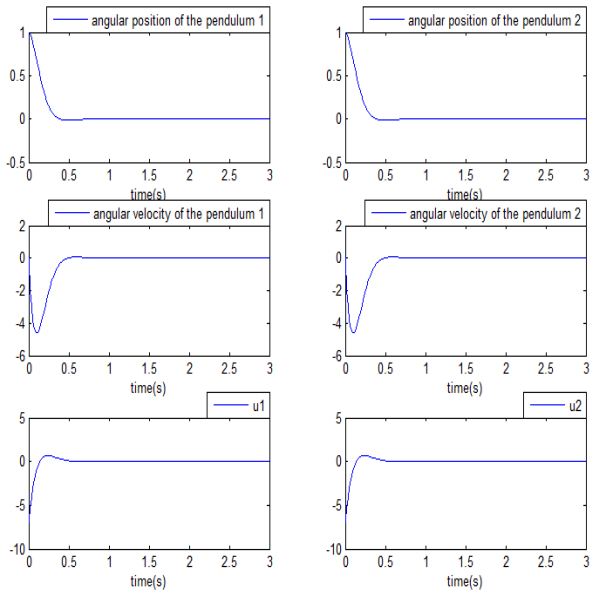


Figure 3: Evolution of the system state variables and corresponding decentralized optimal control signals

4.2 Application of the decentralized pole-placement control approach

In order to apply the decentralized pole-placement for the studied interconnected system, we shall firstly decompose the system into two decoupled inverted pendulums.
Thus, the dynamical model of the isolated subsystems is given by:

$$\dot{x}_1 = \begin{bmatrix} 0 & 1 \\ -286.24 & 0 \end{bmatrix} x_1 + \begin{bmatrix} 0 \\ -21.09 \end{bmatrix} u_1 \qquad (36)$$

$$\dot{x}_2 = \begin{bmatrix} 0 & 1 \\ -286.24 & 0 \end{bmatrix} x_2 + \begin{bmatrix} 0 \\ -21.09 \end{bmatrix} u_2 \qquad (37)$$

Since both (A_1, B_1) and (A_2, B_2) are controllable, we can move their poles to any desired locations, we choose the following desired eigenvalues:

$$\lambda_1^1 = -24; \lambda_2^1 = -18; \lambda_1^2 = -24; \lambda_2^2 = -12$$

In this case we calculate the local gains using the Ackermann's formula [25] :

$$K1 = [-6.9108 \ \ -1.9914]$$

$$K2 = [-0.0830 \ \ -1.7096]$$

Figures 4 and 5 present the evolution of the state variables and their corresponding pole-placement control signals for each isolated pendulum. From the simulation results shown in these curves, we can verify the local stability at each decoupled pendulum.

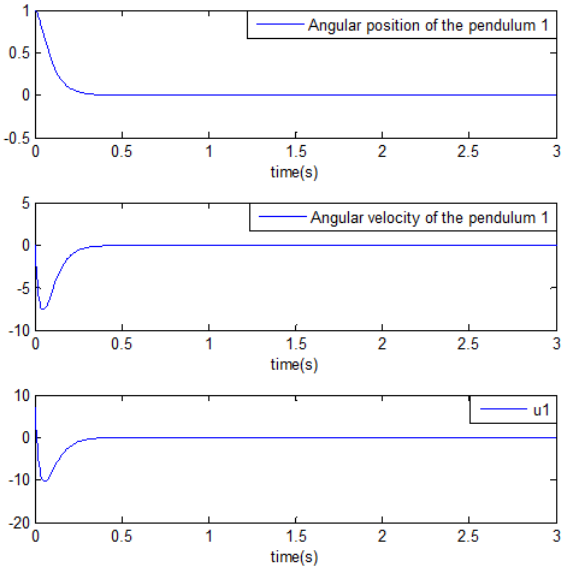


Figure 4: Evolution of the state variables and corresponding pole-placement control signals for the first decoupled pendulum

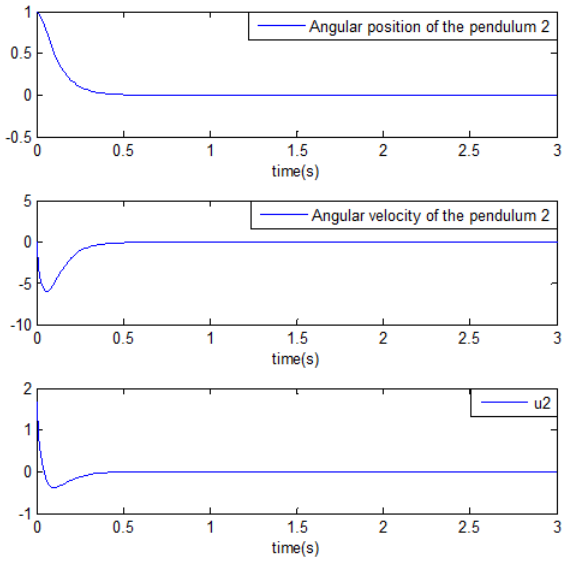


Figure 5: Evolution of the state variables and corresponding pole-placement control signals for the second isolated pendulum

After having applied the formula of Ackermann for each isolated decoupled pendulum, we obtain the closed loop matrix A_f of the overall interconnected system.

$$\dot{x} = A_f x = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -432 & -42 & 316.4332 & 0 \\ 0 & 0 & 0 & 1 \\ 316.4332 & 0 & -288 & -36 \end{bmatrix} x \qquad (38)$$

The eigenvalues of the system (38) have strictly negative real parts:

$$\lambda_1 = -0.9487$$
$$\lambda_{2,3} = -19.5639 \pm 17.0966i$$
$$\lambda_4 = -37.9236$$

Thus, the overall interconnected system provided with such a decentralized control law is asymptotically stable.
The curves in Figure6 illustrate the evolution of the system state variables and the corresponding decentralized pole placement control signals of the double inverted pendulum coupled by a spring, subjected to the same perturbations on the variable θ_1 and θ_2.
From the simulation results shown in these curves, it can be seen that the decentralized control is able to enhance stability of studied system in approximately 4 seconds.

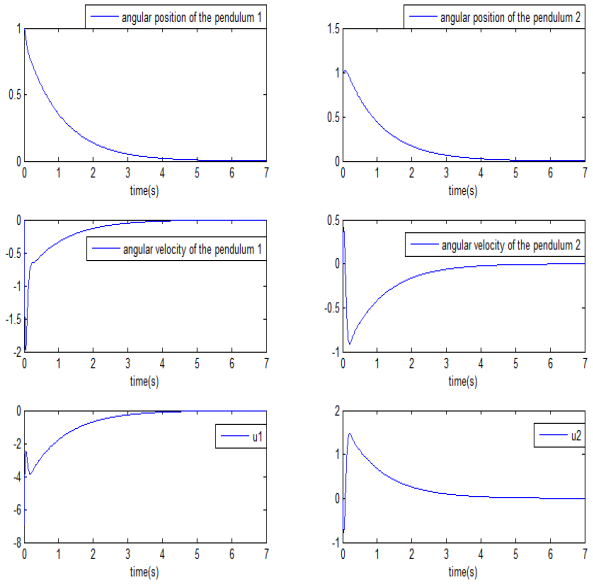


Figure 6: Evolution of the system state variables and corresponding decentralized pole-placement control signals

4.3 Application of the decentralized stabilization control approach

We consider the application of the proposed decentralized stabilizing control developed in section 3.3. on the studied system formed by two inverted pendulums coupled by a spring.
In this part, we solve the proposed LMI formulation in order to find the decentralized gains of the double inverted pendulum.
So we obtain:

$$\begin{cases} S_1 > 0 \\ S_2 > 0 \\ \begin{bmatrix} \breve{a}_{11} & \breve{a}_{12} \\ \breve{a}_{21} & \breve{a}_{22} \end{bmatrix} < 0 \end{cases} \quad (39)$$

where :
$\breve{a}_{11} = S_1 A_1^T + A_1 S_1 - B_1 L_1 - L_1^T B_1^T$
$\breve{a}_{12} = H_{12} S_2 + S_1 H_{21}^T$
$\breve{a}_{21} = H_{21} S_1 + S_2 H_{12}^T$
$\breve{a}_{22} = S_2 A_2^T + A_2 S_2 - B_2 L_2 - L_2^T B_2^T$

By solving problem LMI (39) we obtain the decentralized control gain matrices:

$$K1 = [-8.4386 \ \ -1.0841]$$

$$K2 = [-31.6377 \ \ -2.4752]$$

The evolution of the state variables of the dynamic system composed of two interconnected inverted pendulums with decentralized control by LMI is depicted in Figure 7.
It is clearly seen, from these curves, that the proposed decentralized stabilization control approach is efficient, it allows the best stabilization of the studied system despite the strong disturbances affecting the interconnection between its subsystems.

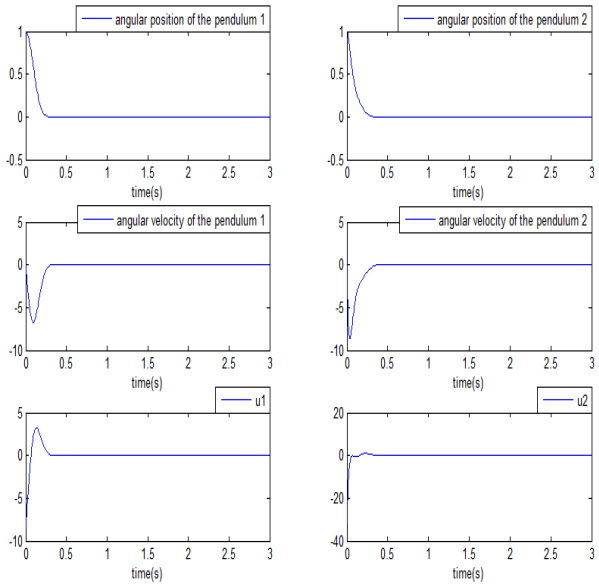


Figure 7: Evolution of the system state variables and corresponding decentralized stabilizing control signals

4.4 Comparative study of the three approaches

We present in this section a comparative study between the three decentralized control approaches studied in section 3.
The visualization of the curves in figure 3, 6 and 7 presenting the evolution of the system state variables and the corresponding control signals, submitted to the same perturbations, shows that the three studied decentralized control approaches can improve the stability of the interconnected system with double inverted pendulums coupled by a spring.

However, we find some disadvantages when applying the decentralized quadratic optimal and decentralized pole-placement on interconnected system. We were obliged to decompose the overall system into

a number of isolated subsystems and then to determine gain matrices that verify local stability for each subsystem. Then we present some sufficient conditions on the obtained gains to guarantee the global stability of the overall system taking into account the interconncetion terms. Indeed, we note an advantage for the proposed new stabilization decentralized approach using LMI problem which calculation of the local gains takes account of the interconnections terms.
On the other hand, when we compare the stabilization times of the three presented approaches, we note that our proposed approach is able to stabilize the system more quickly than other approaches.

5 Conclusion

This extended paper is devoted to the decentralized control techniques of large-scale interconnected systems. In this context, we have presented and studied some decentralized control approaches which objective is to synthesize the gains matrices in order to guarantee the stability of the global system. Our contribution focuses on the development of a new decentralized stabilization control approach based on linear matrix inequalities LMI.

The different approaches studied and formulated in this paper have been applied and validated on a double-parallel inverted pendulum coupled by a spring.
The simulation results have shown that it is possible to ensure the stability and improve the performance of the studied system controlled by each of the decentralized control laws relating to the proposed methods when some sufficient conditions are verified.
Comparative study presented in the fourth section has confirmed the validity and the efficiency of the proposed approach based on LMI which succeeded to ensure quickly the stability of the system and calculated the local gains taking account of the interconnections terms.

Many interesting directions for future research remain. One of the possible perspectives is to develop decentralized control nonlinear approaches for multi-robot cooperative system manipulating a common object.

Appendix A

The proof of the ***Theorem 1*** is based on Lyapunov direct method. Let V be the Lyapunov function defined by the following quadratic form:

$$V(x) = x^T P x \tag{40}$$

Using (23), The time derivative of V (x) is developped as :

$$\begin{aligned} \dot{V} &= x^T(A^T P - PBR^{-1}B^T P)x + x^T H^T P x \\ &\quad + x^T(PA - PBR^{-1}B^T)x + x^T PHx \end{aligned} \tag{41}$$

So (41) becomes

$$\begin{aligned} \dot{V} &= x^T(A^T P + PA - 2PBR^{-1}B^T P)x \\ &\quad + x^T H^T P x + x^T PHx \end{aligned} \tag{42}$$

Then, using the expression (22) in (42), we obtain:

$$\dot{V} = -x^T[2\alpha P + W - (PH + H^T P)]x \tag{43}$$

To ensure the asymptotic stability of system (23) , $\dot{V}$ should be negative definite, then which is equivalent to the matrix F:

$$\begin{aligned} &F = 2\alpha P + W - (PH + H^T P) \\ &W = Q + PBR^{-1}B^T P, \;\; Q = diag[Q_i] \end{aligned} \tag{44}$$

should be positive definite.

References

[1] R. Ben Amor, S. Elloumi, "On decentralized control techniques of interconnected systems-application to a double-parallel inverted pendulum" in International Conference on Advanced Systems and Electric Technologies(IC-ASET), Tunisia, 2017. https://doi.org/10.1109/ASET.2017.7983671

[2] C. Wang, C. Wen, Y. Lin, W. Wang, "Decentralized adaptive tracking control for a class of interconnected nonlinear systems with input quantization" Automatica 81 (2017) 359?368, 2017. https://doi.org/10.1016/j.automatica.2017.03.010

[3] R. Huang, J. Zhang, Z. Linc, "Decentralized adaptive controller design for large-scale power systems" Automatica 79 (2017) 93?100, 2017. https://doi.org/10.1016/j.automatica.2017.01.022

[4] H. Wu, "Decentralised adaptive robust control schemes of uncertain large-scale time-delay systems with multiple unknown dead-zone inputs" IET Control Theory Appl., Vol. 11 Iss. 9, pp. 1360-1370, 2017. https://doi.org/10.1049/iet-cta.2016.1277

[5] F. Sahami, M. U. Salamci, "Decentralized model reference adaptive control design for nonlinear systems; application to cancer treatment" in 17th International Carpathian Control Conference(ICCC), Slovakia, 2016. https://doi.org/10.1109/CarpathianCC.2016.7501173

[6] A. Hernndez-Mndez, J. Linares-Flores, H. Sira-Ramrez, "Decentralized adaptive control for interconnected boost converters based on backstepping approach" in Energy Conversion Congress and Exposition (ECCE), USA, 2016. https://doi.org/10.1109/ECCE.2016.7854996

[7] H. Wu, "Decentralised robust stabilisation of uncertain large-scale interconnected time-delay systems with unknown upper bounds of uncertainties" International Journal of Systems Science, 2015. https://doi.org/10.1080/00207721.2015.1029569

[8] H. Rtibi, S. Elloumi, "Robust decentralized nonlinear control for multimachine power systems" in International Conference on Advanced Systems and Electric Technologies(IC-ASET), Tunisia, 2017. https://doi.org/10.1109/ASET.2017.7983739

[9] H. Rtibi, S. Elloumi, N. Benhadj Braiek "A robust stabilization based on decentralized guaranteed cost control approach for a class of uncertain interconnected systems" in 4th International Conference on Systems and Control(ICSC), Tunisia, 2015. https://doi.org/10.1109/ICoSC.2015.7152782

[10] W. J. Liu, "Decentralized Sliding Mode Control for Multi-Input Complex Interconnected Systems Subject to non-smooth Nonlinearities" Asian Journal of Control, Vol. 20, No. 5, 2017. https://doi.org/10.1002/asjc.1627

[11] M. Fathallah , F. Abdelhedi, N. Derbel, "A synchronizing second order sliding mode control applied to decentralized time delayed multiagent robotic systems: Stability Proof" Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 160-170, 2017. https://doi.org/10.25046/aj020321

[12] F. Abdelhedi, N. Derbel, "Adelay-dependent distributed SMC for stabilization of a networked robotic system exposed to external disturbances" Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 513-519, 2017. https://doi.org/10.25046/aj020366

[13] M. Cucuzzella, G. P. Incremona, A. Ferrara, "Decentralized Sliding Mode Control of Islanded AC Microgrids with Arbitrary Topology" IEEE Transactions on Industrial Electronics Vol. 64, Iss. 8, 2017. https://doi.org/10.1109/TIE.2017.2694346

[14] Y. Li, J. Li "Decentralized stabilization of fractional order T-S fuzzy interconnected systems with multiple time delays" Journal of Intelligent and Fuzzy Systems 30 (2016) 319?331, 2016. https://doi.org/10.3233/IFS-151758

[15] J. Lin, "Decentralized Stabilization of Uncertain Large-scale Fractional Order Interconnected Systems via Output Feedback" in American Control Conference (ACC), USA, 2016. https://doi.org/10.1109/ACC.2016.7525336

[16] J. Liu, V. Gupta, "On Stabilization of Decentralized Systems Across Analog Erasure Links" IEEE Transactions on Automatic Control Vol. 62 Iss. 3, 2017. https://doi.org/10.1109/TAC.2016.2575833

[17] D. Liu , D. Wang, H. Li, "Decentralized Stabilization for a Class of Continuous-Time Nonlinear Interconnected Systems Using Online Learning Optimal Control Approach" IEEE Transactions on Neural Networks and Learning Systems, 2014. https://doi.org/10.1109/TNNLS.2013.2280013

[18] L. Bakule, M. de la Sen, M. Papik and B. Rehak, "Decentralized stabilization of symmetric systems with delayed observer-based feedback" in American Control Conference, USA, 2013. https://doi.org/10.1109/ACC.2013.6580888

[19] H. Ferdowsi, S. Jagannathan, "Decentralized Fault Tolerant Control of a Class of Nonlinear Interconnected Systems"International Journal of Control, Automation and Systems 15(1) (2017) 1-10, 2017. https://doi.org/10.1007/s12555-015-0384-5

[20] B. Zhao, D. Wang, G. Shi, D. Liu, Y. Li, "Decentralized control for large-scale nonlinear systems with unknown mismatched interconnections via policy iteration" IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017. https://doi.org/10.1109/TSMC.2017.2690665

[21] B. Zhao, D. Liu, Y.Li, Q. Wei, R. Song, "Adaptive Dynamic Programming Based Decentralized Tracking Control for Unknown Large-scale Systems" in Proceedings of the 36th Chinese Control Conference, China, 2017. https://doi.org/10.23919/ChiCC.2017.8027913

[22] C. Wei, J. Luo, H. Dai, Z. Yin, J. Yuan, 'Low-complexity differentiator-based decentralized fault-tolerant control of uncertain large-scale nonlinear systems with unknown dead zone"Nonlinear Dynamics , Vol. 89, Iss. 4, 2017. https://doi.org/10.1007/s11071-017-3605-z

[23] U.S. Park, J.H. Kim, S.J. Kim, J.W. Choi, J.S. Kim, "Development of a Parallel Inverted Pendulum System and Its Control" in 38th SICE Annual Conference, Japan, 1999. https://doi.org/10.1109/SICE.1999.788676

[24] M.S. Mahmoud, M.F. Hassan, M.G. Darwish, "Large scale control systems, theories and techniques" Edition Marcel Dekker, New York, 1985.

[25] J. Ackermann, "Sampled-Data Control Systems" Springer-Verlag, Germany, 1985.

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 404-411 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Security Analysis and the Contribution of UPFC for Improving Voltage Stability

Asma Meddeb*, Hajer Jmii, Souad Chebbi

LATICE laboratory of National High School of Engineers of Tunis, Tunis 1008, Tunisia, Tunis University.

ARTICLE INFO

Article history:
Received: 30 October, 2017
Accepted: 02 January, 2018
Online: 10 February, 2018

Keywords:
Power system security assessment
Contingency analysis
Newton Raphson power flow method
UPFC

ABSTRACT

The occurrence of many failures in the power system can lead to power instability and affects the system parameters to go beyond its operating limits. It may lead to obstructing the secure operations and reliability of power systems. Ensuring power system security needs proper actions to be taken for the undesirable contingency. Thus, security analysis is important tasks in modern energy management systems. This paper proposes an approach based on the Newton Raphson power flow method for power system security analysis. Firstly, the contingencies will be specified to assess their impact on the transient stability. Secondly, the selected contingencies will be classified in the order of severity. In addition, the integration of the Unified Power Flow Controller (UPFC) to enhance the transient stability of the power system is considered. The proposed method is implemented on the IEEE-14 bus system. We performed this case study using the well-known software EUROSTAG.

1. Introduction

Security analysis of a power system plays a significant role in the growth and development of modern societies. Given that the power system is a large-dimension complex system, it might be the site of various disturbances, making its chaotic behavior and difficult to control [1]. Thus, ensuring the security of the electric network becomes an essential task for operators and researchers.

Following a violent disturbance enormous damage may occur in the power system and impacts directly its normal operation. We can note that the concept of network security is broader than the stability of networks [2]. It is defined as the ability of power system to maintain its stable operation after each disturbance or unfavorable change of the system [3]. However, the level of security is deemed satisfactory if, firstly, it is able to ensure the production-consumption balance within certain acceptable limits related to line's power flows, voltage in each node and operating points of generation units and if, on the other hand, it is able to persist in a more or less violent disturbance.

Therefore, the system operator must ensure that the power system operates at a secure operating point where all equipment is within their safe limits. If the failure is not controlled suitably, this could lead to a disastrous event such as major blackouts and impact then the economic and environmental requirements [4].

*Corresponding Author: Asma Meddeb, National High School of Engineers of Tunis (ENSIT), Tunis 1008, Tunisia, +21697467529, asma.meddeb@ensit.rnu.tn

The advent of FACTS devices has been a boon to the power industry. These devices provide fast and effective control of the various parameters of the power system to improve their stability and preventing other related cascade accidents [5-9]. The UPFC is one of the most versatile FACTS devices, providing independent control of the voltage, real and reactive power of the transmission under its supervision. This device enhances the power system security.

N. Hashim *et al*. [10] have presented an analysis of transient stability of IEEE 14 bus test system. They studied the effect of the fault location on system stability; and they analyzed the characteristics of the machine states including speed, rotor angle, output electrical power and the terminal voltage of the machine after a three-phase short-circuit which occurs at different points of the system. The IEEE-14 bus system has been also studied in [11], wherein both authors have studied the effect of fault location and critical clearing time on the system stability. In order to achieve this, they analyzed the behavior of the synchronous machine, in particular, the angular position of the rotor. They proposed to rapidly isolate the faulted part to increase stability margin.

A. Zerigui [12] provided a solution to the problem of transient stability. To test a constrained optimal power flow, and estimate critical clearing time she developed a new analytical function. She applied her strategy to test large networks up to 145 buses. While in [13] the authors just used nine-bus IEEE system to confirm their proposed method. They presented a global approach to transient stability constraint optimal power flow based on SIME model.

www.astesj.com
https://dx.doi.org/10.25046/aj030149

Likewise, authors in [14] analyzed the transient stability for IEEE-9 bus test system. Then, they proposed a simple control method based on the equal area criterion for calculating critical clearing time. To improve the stability margin, the temporal evolution of the frequency and voltage, with and without load shedding, is performed.

For three decades, J-C Chow et al [15] have formulated the contingency classification problem into a pattern recognition problem and then design a Hopfield model to detect a prescribed set of patterns. This optimization method, which uses the linear programming technique, considers only the steady-state security assessment problem.

In order to explain the contribution in deepening understanding of the power system security problem, the authors in [16], proposed an efficient steady-state contingency classification methodology based on Rough Set Theory. Through this analysis, they could classify the system operating in four different states.

Abdulrazzaq, A. [17] used the Newton Raphson load flow method for the power system contingency ranking only for the line outage.

In most cases, the researchers deal with the security analysis in the steady case, considering the execution time of computational simulations performed. The main purposes of security analysis are the fast identification of critical contingencies and their evaluation related to the severity level [18-20].

This paper is an extension of work originally published in the Proceedings of the International Conference on Sciences of Electronics, Technologies of Information and Telecommunications [21].

This paper focuses on the contingency analysis that comprises three steps. Firstly, we create the contingencies list containing all sets of possible contingencies that may occur in a power system. Secondly, the classification of severe contingencies from the list that may lead to the supply and security constraints violations is achieved. Lastly, the evaluation of contingencies is necessary. In addition, a curative action using UPFC devices to avoid the effects of most severe contingencies is proposed. To verify our proposed approach, we perform simulations on an IEEE 14-bus power system.

2. Analysis of Electrical Networks security

The security analysis of an electrical network is related to its dynamic state or steady state following the disturbance. In dynamic security analysis, the transition from the existing operating condition for the new operating conditions and the fact that during the transient state should not be cascading outages are considerations of interest. The dynamic security analysis is normally done in a deterministic approach using analytical tools such as load flow calculation, dynamic simulation, etc. Indeed, these tools use a detailed model of the electrical system to determine the system dynamic response with respect to each of the analyzed contingencies. Analysis methods of the dynamic security of electrical networks are classified into four groups: methods based on numerical simulation, pattern recognition methods, direct methods and hybrid methods [22-25]. In this paper, we propose heuristic method because it is easier to set up and it gives quick and good solutions to difficult problems.

2.1. Operating States of a Power System

In security analysis, power system may operate in different states, namely normal, emergency, alert, extreme state and restorative [26, 27]. A system is said to be in normal operating state when it satisfies equality and inequality sets of constraints. This is equivalent to the fact that there is no overloaded equipment and all the variables of the system are within the normal operating limits [28, 29]. Following every single contingency the system must be able to operate in a secure way while respecting all the constraints. However, if the contingency is highly dangerous or the system generation drops below the required amount, the system operation may be insecure indicating an alert state. In this case, the equality and inequality constraints are still satisfied nevertheless; preventive action is needed to restore the normal state. In case of preventive control failure, the system security level may be under the permissible limit. Therefore, the inequality constraints are no longer maintained and the system enters in an emergency state. It is to be mentioned that in case of a sufficiently serious contingency, the system transits directly to the extremis state.

To overcome the emergency state, control measures should be initiated immediately. If these actions are not efficient or are not taken at the adequate time, the state of the system is said in extremis. This state is characterized by violated equality and inequality constraints which may cause the loss of some parts of the system or a total system blackout. Otherwise, the rapid and efficient application of control actions brings the system to the restorative state. In that instance, the implementation of other control measures may lead the system either to the normal state or the alert state.

2.2. UPFC device

UPFC (Unified Power Flow Controller) is the most powerful and versatile device. It can control three parameters either individually or in appropriate combinations at its series connected output while maintaining reactive power support at its shunt connected input device. Therefore, the aim of UPFC is to enhance the usable transmission capacity of lines and control the power flow [30-35].

The model of UPFC implemented in a transmission line is shown in Figure 1. A series converter connects with the line through an insertion transformer. A shunt converter connects with the line through a second transformer. The DC terminals of the two converters are linked together, and their common DC voltage is supported by a capacitor bank. A mathematical model of the UPFC is well-detailed in Reference [36].

3. Mathematical Model of power system

In order to analyze, simulate, design and control the electric power system operation, steady-state and dynamic system performances properly are explained.

3.1. Description of system

In general, a power system can be modeled by the following set of nonlinear differential-algebraic equations:

$$\dot{x} = F(x, y, \lambda)$$
$$0 = G(x, y, \lambda) \quad (1)$$

Where x is a vector of state variables, y is a vector of network variables, and λ represent the control and parameter variable which may be used to control or tune power system performance.

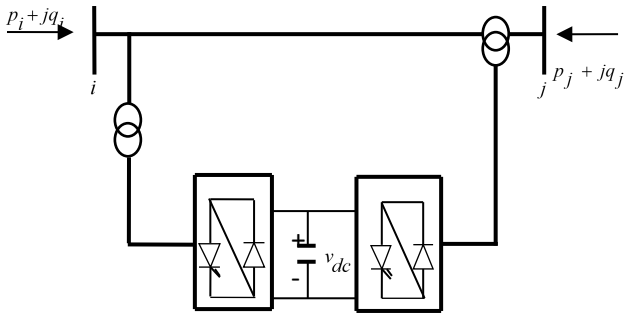


Figure 1. UPFC model.

The differential equation F represents governing dynamics of power systems, which is associated with generators, excitation systems, and speed governor. The differential variables x consist of the states of dynamic components such as rotor angles and rotor speed.

$$x = [\delta, \omega]^T \quad (2)$$

The algebraic equation g represents the network power balance of power systems. The algebraic states y includes bus voltages and bus phase angles. It is also known as control variables.

$$y = [V, \alpha]^T \quad (3)$$

The active and reactive powers injected at any bus are given by Equation (4). Therefore, it is a function of voltage magnitudes and angles for all the buses, as well as the admittance, magnitude, and angle, for the transmission network.

$$\begin{cases} P_i = V_i . \sum_{i=1}^{n} Y_{ij} . V_j . \cos(\alpha_i - \alpha_j - \theta_{ij}) \\ Q_i = V_i . \sum_{i=1}^{n} Y_{ij} . V_j . \sin(\alpha_i - \alpha_j - \theta_{ij}) \end{cases} \quad (4)$$

Where Y_{ij},θ_{ij} are, respectively, the admittance magnitude and angle for the transmission line i-j, V_i, V_j are the voltages at bus i and j, α_i, α_j are the voltages angles, and n is the number of studied network buses.

The active and reactive power functions are evaluated by this load flow equations:

$$\begin{cases} P_{Gi} - P_{ci} - P_i = 0 \\ Q_{Gi} - Q_{ci} - Q_i = 0 \end{cases} \quad (5)$$

P_{ci}, Q_{ci}, P_{Gi} and Q_{Gi} are respectively active and reactive power consumed and produced for bus i.

Equations (6) give the output active and reactive power. These equations show that P and Q are functions of the bus voltages and rotor angle. X_T and E_q represent respectively transfer reactance between buses and internal voltage of the generator.

$$\begin{cases} P = \frac{E_q . V_i}{X_T} . \sin\delta \\ Q = \frac{V_i^2}{X_T} - \frac{E_q . V_i}{X_T} . \cos\delta \end{cases} \quad (6)$$

3.2. Model of Machine

The generator model is represented by the four-order model as follows:

$$\begin{cases} \frac{d\delta}{dt} = \omega - 1 \\ M . \frac{d^2\delta}{d^2 t} + \frac{d\delta}{dt} . D = P_m - P_e \\ \frac{dE'_q}{dt} = \frac{1}{T'_d} . (E_{fd} - E'_q + (X_d - X'_d) . i_d) \\ \frac{dE'_d}{dt} = \frac{1}{T'_q} . (-E'_d + (X_q - X'_q) . i_q) \end{cases} \quad (7)$$

δ is the rotor angle of the machine, ω is the rotor speed, M is an inertia coefficient of machine, D represents the damping coefficient of machine, Pm, Pe are the mechanical and electrical power of the machine, E'_d, E'_q are the voltage behind the direct and quadrature axis transient reactance X_d', X_q' respectively, T'_{d0} is the d-axis open circuit transient time constant, T'_{q0} is the q-axis open circuit transient time constant, and E_{fd} represents a field voltage.

3.3. Security Contraints

The security constraints are based on the operating limits which have to be satisfied for normal operation of the power system. In order to operate the system within an acceptable security domain, the basic security constraints are as follows:

$$V_{min} \le V \le V_{max} \quad (8)$$

$$P_{gmin} \le P_g \le P_{gmax} \quad (9)$$

$$Q_{gmin} \le Q_g \le Q_{gmax} \quad (10)$$

$$\delta_{min} \le \delta \le \delta_{max} \quad (11)$$

4. Proposed Method

Considering the fact that the electrical network operates under two main types of constraints; constrained supply (C1) of energy (C2) which imposes that all consumers must be supplied and operations constraints require that the system variables should be within specified operating limits.

During a normal operating state, all constraints are satisfied. It is characterized by a sufficient level of stability margins so that the system can withstand a single contingency. When the system satisfies all the supply and security constraints without reserve power generation, it passes in an alert state. This means that in the event of a contingency, at least one inequality constraint will be violated. Therefore, preventive actions have to be taken to bring the system to a secure state. The power system enters an emergency state from an alert state when a contingency occurs. In this case, all the equality constraints are satisfied and at least one inequality constraint is violated. Consequently, the system requires an immediate implementation of corrective actions to prevent the damage of equipment. However, once the severity of the disturbance is very high, both equality and inequality constraints are violated. Thus, the system passes in an extreme state.

In this paper, we consider the time taken by the system to transit from a secure state into another one where is not secure as an index for ranking serious contingencies. Therefore, the contingency is deemed severe when both constraints of the network are not satisfied and, it is potentially harmful if only the constraints of system operation are not satisfied. In the case of both constraints are respected, the contingency is recognized as being harmless.

To achieve this aim, a fast strategy is proposed to identify power system dynamic behavior using the time domain simulation method. A corrective control to restore power system operating equilibrium after disturbances has been also proposed. This strategy is based on a deterministic approach using analytical tools such as load flow calculation and dynamic simulation. Newton-Raphson method is used to solve load flow problem. In this method, optimal solution can be obtained using iterative method. It usually converges faster than other methods. Then, the identification of properties by the best solutions and introduce them as boundaries of the problem are mainly fulfilled. Firstly, the assessment of different operating state according to a set of selected contingencies is required. Then, security assessment is based primarily on steady-state load flow analysis and transient analysis of the power system. The based approach gives priority to the most severe cases for thorough analysis, and the stable cases are previously eliminated from the list.

Figure 2 depicts the flowchart of the proposed method.

5. Simulation

5.1 Case Study

Figure 3 shows the single line diagram of IEEE 14-bus system. It consists of five synchronous machines. There are eleven loads in the system. Three step-up transformers, one of which is three winding transformer. The generators are modeled as an ideal voltage source behind the synchronous reactance of the machines. The model of transmission lines considers the resistance and the reactance. The loads are modeled as constant impedance. The all data for simulation were selected from [37]. We performed all simulations using software package EUROSTAG [38]. This software is a powerful tool dedicated to dynamic simulations. Its major advantage is the high rapidity of its algorithm.

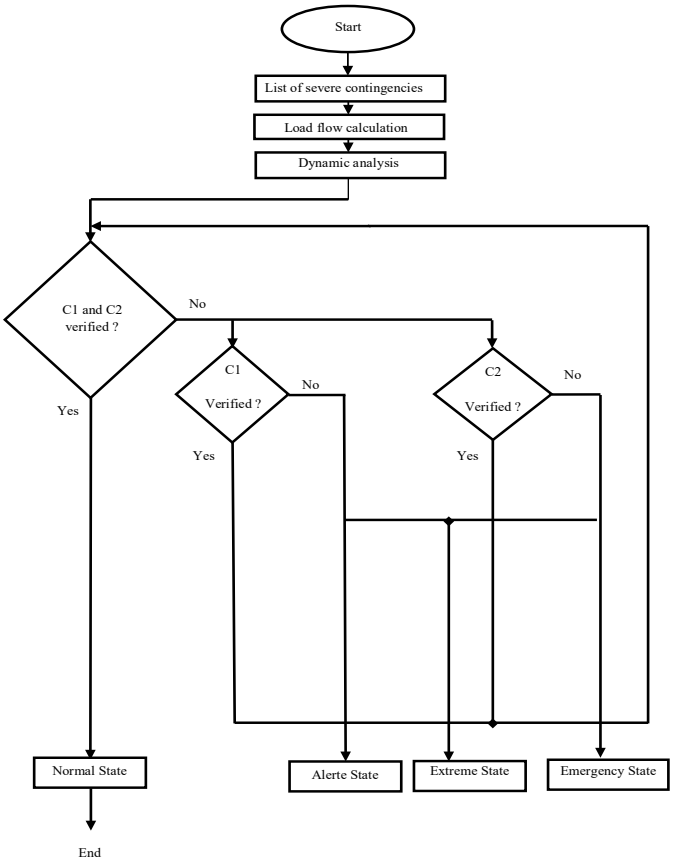


Figure 2. Flow chart of proposed method.

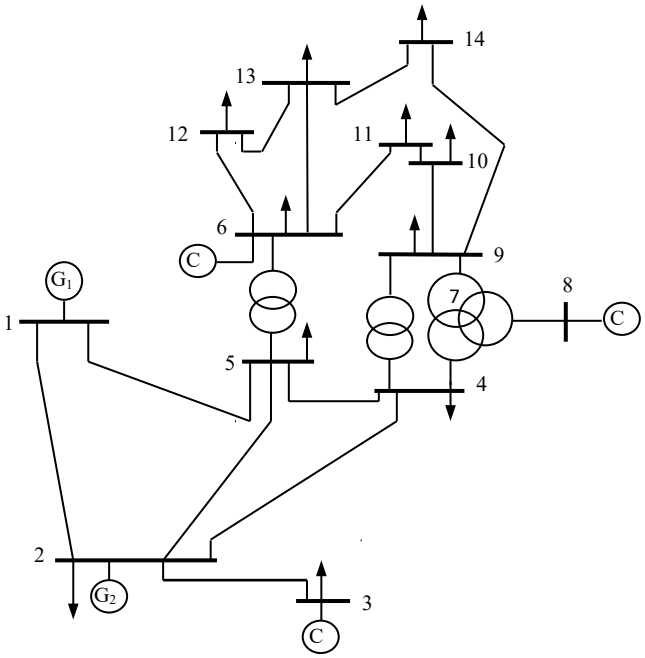


Figure 3. Single line diagram of IEEE 14-bus system.

5.2 Contingency Classification Method

Contingency analysis should be performed for the unexpected and severe events that may occur in power system. Table 1 presents the contingencies list containing the contingencies that may occur in a power system. For the contingencies classification, we considered two criteria; maintaining the network characteristics in their acceptable ranges, and ensuring a balance between production and consumption. A contingency is deemed very dangerous when both constraints of the network are not satisfied, which corresponds to an extreme state. It is potentially dangerous if only the constraints of system operation are not satisfied; the network is in an emergency state. Whenever the system goes to alert state under severe contingency; is that the constraint of energy supply is not fully respected. If both constraints are respected the contingency is recognized as being harmless; as a result, the network is stable. Thus, this contingency is ineffective on the loss of system and it can be rejected.

Table 1. Contingency Index.

Contingency	Index
Bolted fault	I
Impedant fault	II
Increase of load by 10%	III
loss of generator G1	IV
loss of generator G2	V
Opening of line 8-7	VI
Opening of line 2-3	VII

5.3 Simulations Results

At time 250s, we simulated different types of fault and we followed the temporal transition of the system. Table 2 defines the emergency state, alert and the extreme state of network study. We can notice that the range of stability is different for each type of fault. Loss of generator G1 remains the scariest of all faults on network stability. In order to have reasonable accuracy in the classification of contingencies, the detailed assessment of the contingencies is necessary. The aim is to detect the worst case for the transient stability of the test system. Thus, we tested the impact of line opening, loss of power generator, and the severity of three-phase short-circuits.

5.3.1 Line opening scenario

We should first look at how the line opening affects the transient stability of the system. To achieve that, we propose the tripping of lines one by one. All system parameters (frequency, voltages at different buses, rotor angle of the two machines) must be within the allowable range to predict the security of the system.

Therefore, all security practice is followed. For the opening of line 1-5, the behavior of the two machines is different; G1 accelerates while the machine G2 decelerates. Both machines accelerate, in case VII and oscillations are much less than line 1-5. In the case of opening the line 9-14, a voltage drop at bus 14 passes from 0.971pu to 0.928pu, oscillations amplitude of the rotor angle is important. With the opening of line 8-7, we note a separation of the network in two sub-networks; the first sub-network includes the machines G1, G2, G3, and G6. The second involves only the machine G8. Following this fault, a drop in voltage appears at the terminals of all the consumer nodes and an increase of 2.5% of the voltage at bus 8 with a slight oscillation of voltages.

Table 2. Operating modes of the test system

State	Alert State	Emergency state	Extreme State
I	[250.1 250.4]	] 250.4 250.413]	]250.413 250.414]
II	[251.5 252]	] 252 253]	] 253 255]
III	-	-	-
IV	] 250.07 250.114]	] 250.114 250.13]	] 250.13 250.15]
V	] 250.9 251.5]	] 251.5 257]	] 257 260]
VI	-	-	-
VII	-	-	-

We note that the variation in the speed and power of the two generators have a damped oscillatory regime. Then we can say that the system is stable under these conditions because generator speed and hence rotor angle regain their stability after few highly damped oscillations. In conclusion, the opening of a single line does not disrupt the stability of the network; this contingency may be classified as non-dangerous.

5.3.2 Loss of machine scenario

In order to study which of two machines is critical, we simulated separately, the case of the loss of machine G1 and G2 and followed the transient behavior of the system. As shown in Table 3, in the first case, the two machines are unstable, and the system exhibits a total voltage collapse. While in the second case, only machine G2 becomes unstable, and we note the oscillations of the rotor angle and acceptable electrical power. There are thus among the two machines disturbed during the fault, a machine that is more stable than the other after the elimination of fault. This is confirmed by the necessary critical clearing time that leads to instability. The critical clearing time must be equal to 0.114s for that the machine G2 becomes unstable, while the machine G1 maintains its stability during the definitive loss of G2. The machine G1 does undergo smaller oscillations than the machine G2 in the post-fault configuration. Due to its high nominal power, and that it is chosen as the reference generator. We confirm that machine G1 is critical.

5.3.3 Impact of location of the short-circuit

We will now proceed to study the impact of location of the fault on security margin of the network. A bolted three-phase fault was applied respectively to the nodes 2, 6, 5 and 14 at time t = 250s. The time duration of this fault is 100ms. The choice of nodes is relevant, bus 2 is producer node, a compensator is connected to bus 6, and bus 5 is a consumer node that is linked to two producer nodes and bus 14 is a weak node. Table 4 summarizes all the results of simulations. We noticed that during a fault at bus 2, the system

puts only 0.413s to pass in an extreme state, bus 6 take 5.3s, the bus 5 puts 1.05s and bus 14 passes to an extreme state after 1s.

Let us note that transient system behavior is associated with large oscillations of both machines outputs. Therefore, a short circuit applied at bus 2 and bus 14 impacts greatly the stability of the network. These fault types are therefore classified as a dangerous contingency.

Table 3. State of both machine G1 and G2

Parameters	Loss of G1	State of G2	Loss of G2	State of G1
f(Hz)	40	unstable	49.84	Stable
δ1 (deg)	diverge		15.94	
δ2 (deg)	50		diverge	

Table 4. Variation of Fault Location

Bus	Alert state	Emergency State	Extreme State
2	] 250.1 250.4]	] 250.4 250.413]	] 250.413 250.414]
6	] 250.1 253.7]	] 253.7 255. 3]	] 255.3 260]
5	] 250.1 250.8]	] 250.8 251.05]	] 251.05 251.1]
14	] 250.1 250.8]	] 250.8 251]	] 251 251.3]

5.3.4 *Load increase*

Similarly, we have performed simulations of load variation from 5% to 100%. Table 5 illustrates the stability intervals for three states of the system. We notice that the system satisfies both constraints energy supply and operating for 40% load increase on the system. On the other hand, from 70% the system loses the balance between production and consumption. For a 100% increase in load, the system is subject to voltage collapse. This is explained by the fact that the machine G1, during normal operation, generates only 233MVA (47%) while its nominal capacity is 615MVA. Thus, we note that the machine G1 has reserves to provide the required active power, but obviously, can only satisfy a large load variation.

It is noteworthy that in a multi-machine network, stability for the frequency range is ± 0.4Hz, the critical load of the system is thus 40%. Then, we can indicate that the system is highly vulnerable to voltage collapse in the event of heavy load increase. Furthermore, if we bear in mind that when transition time is small in an emergency or extreme state, this contingency is more severe and requires the treatment.

Numbered contingencies III, V, and VII pose no risk because the network remains stable. As against the others should be analyzed as they can cause loss of synchronism of system. Potentially dangerous contingencies are maintained in a waiting list because some of them can be harmful. Whereas the dangerous contingencies, I, II and IV must be treated. Figure 4 shows the impact of an increase of load on frequency value and rotor angle of the two machines.

Table 5. Stability Margin Depending on Load Increasing

State	Alert State	Emergency state	Extreme State
Stability margin	] 5 40]	] 40 70]	] 70 100]

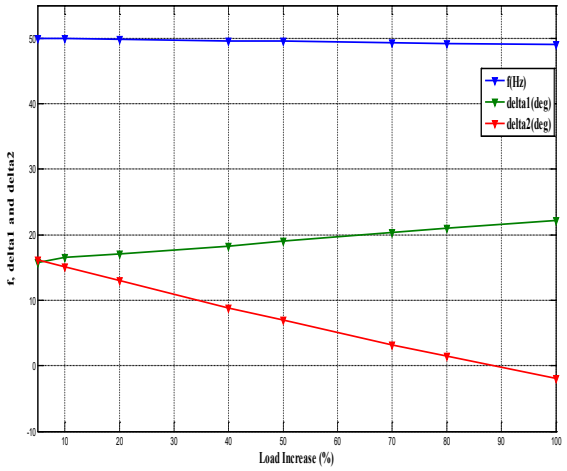


Figure 4. Impact of an increase of load on frequency value and rotor angle of the two machines.

5.4 *Contribution of UPFC on the voltage stability*

The enhancement of voltage stability of the test system using the UPFC is studied. We connect UPFC to the system at different locations. Power dimension of UPFC is ± 200 MVAR. Two scenarios are performed. At first, we consider a total load increase of 40%. Then, a three-phase fault is applied at bus 14. The fault duration is 100ms.

The UPFC was integrated in the middle of line 2-3. Then, it is connected in the middle of line 1-2. Lastly, the UPFC is integrated in the middle of line 9-14.

Table 6 shows the voltage level of all bus under load increase of 40%. Regardless of its location, the UPFC improve almost the voltage level of all bus.

Figure 5 shows the effect of the UPFC on the behavior of voltage at bus 9 and bus 2 with and without the UPFC. The addition of the UPFC has remarkably improved the behavior of voltage.

Here, we consider a bolted three-phase fault at bus 14. The fault duration is 100ms.

Figure 6 and Figure 7 respectively illustrate the temporal evolution of voltages at bus 2 and bus 9 with and without the UPFC.

It should be noted that a short circuit at bus 14 causes a voltage drop which reaches its minimum beyond a certain threshold value and regains its initial value after highly damped oscillations in the presence of the UPFC. Therefore, the UPFC provides the best control of voltage and a good damping of oscillations.

Table 6. Load Increase of 40%

	Sans FACTS	UPFC in line 2-3	UPFC in line 1-2	UPFC in line 9-14
BUS1	1.060	1.060	1.060	1.060
BUS2	1.040	1.040	1.040	1.044
BUS3	1.003	0.985	1.003	1.004
BUS4	1.014	0.992	1.013	1.017
BUS5	1.022	1.003	1.022	1.026
BUS6	0.996	0.982	0.996	1.014
BUS7	1.011	0.992	1.011	1.011
BUS8	1.082	1.067	1.082	1.018
BUS9	0.978	0.959	0.978	1.085
BUS10	0.972	0.954	0.972	0.972
BUS11	0.980	0.963	0.980	0.980
BUS12	0.977	0.962	0.977	0.977
BUS13	0.971	0.956	0.971	0.971
BUS14	0.952	0.935	0.952	0.952

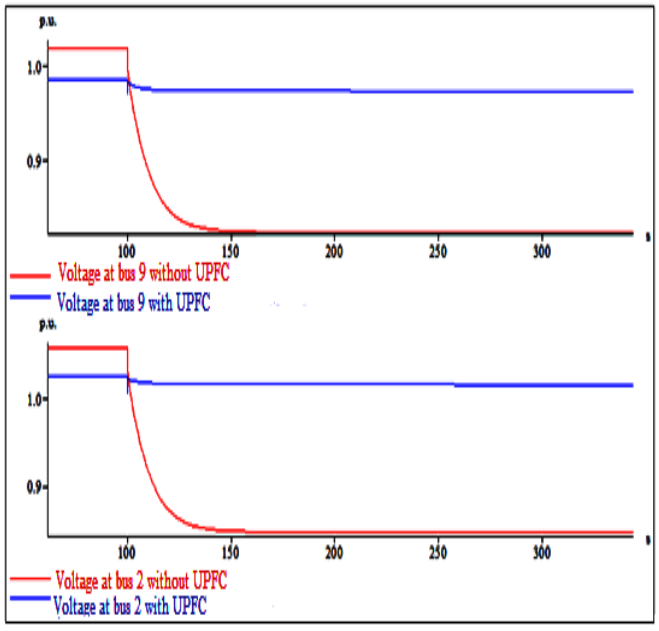


Figure 5 Voltage magnitudes at bus 9 and bus 2 with and without the UPFC.

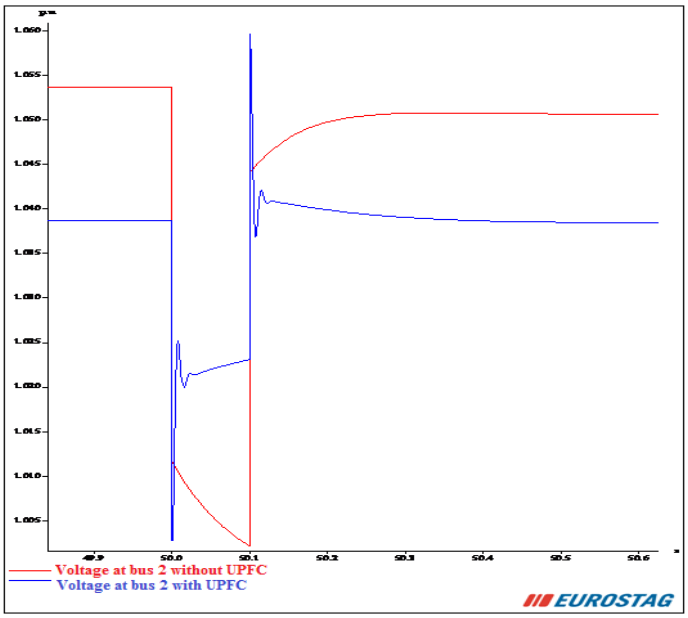


Figure 6 Voltage magnitudes at bus 2 with and without the UPFC.

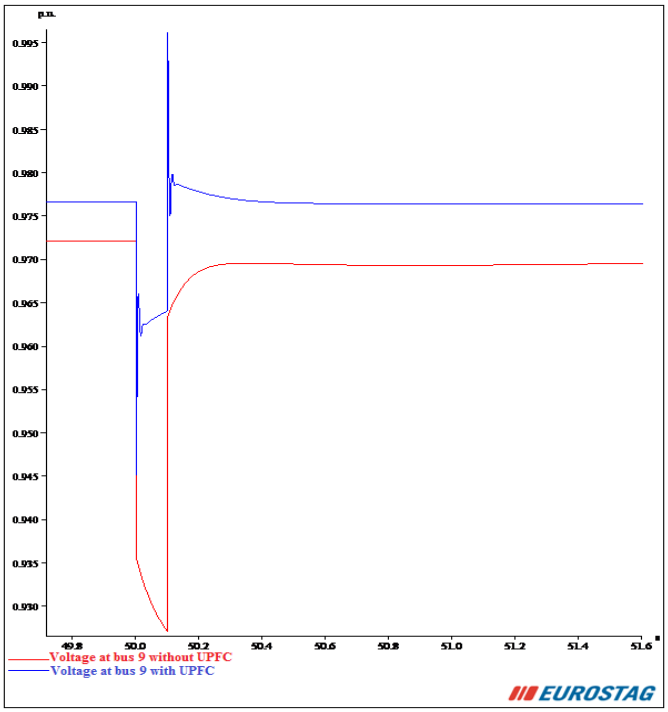


Figure 7 Voltage magnitudes at bus 9 with and without the UPFC.

6. Conclusions

This paper presented a classification of contingencies as well as curative action against faults affecting the voltage stability of the system. Extensive studies of the behavior of IEEE 14-bus system for different contingency cases revealed that is subject to instabilities under well-defined fault conditions. We have demonstrated through this paper that power system suffered from severe contingencies leading to their instability. Thus, we investigated effective methods to improve the stability of power network. A comprehensive analysis of the network security testing has allowed us to classify contingencies in an increasing order of

severity. Some of these contingencies may cause serious instability on the network, and others may not be critical. As a result, it is necessary to focus attention on a limited number of contingencies and seek remedial action through UPFC device, which has a role to reduce the possibility of voltage collapse and limit damage of fault, as overloads and short -circuit. Therefore, the UPFC was most effective in improving voltage stability and reducing the harmful effects of dangerous contingencies.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] K. Alcheikh-Hamoud, "Modeling of large interconnected power systems: application to the safety analysis in a competitive environment," Phd. Engineering Sciences. Institut National Polytechnique de Grenoble - INPG, 2010.

[2] S.r. Gongada ,T.S. Rao, P.Mallikarjuna rao, S. Salima, "Power System Contingency Ranking using Fast Decoupled Load Flow Method," International Conference on Electrical, Electronics, and Optimization Techniques, 2016.

[3] E.Ciapessoni, D.Cirio, S.Massucco, A.Morini, A.Pitto and F.Silvestro, "Risk-Based Dynamic Security Assessment for Power System Operation and Operational planning,", MDPI journal , Energies, 2017.

[4] S. Burada, D. Joshi, and K.D. Mistry, "Contingency Analysis of Power System by using Voltage and Active Power Performance Index," 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, 2016.

[5] Abido, M. A., "Power system stability enhancement using FACTS controllers: A review," The Arabian journal for science and engineering, 34(1B),153-172,2009.

[6] A. Meddeb, H. Jmii, S. Chebbi, "UPFC and SVC Devices for Transient Stability Enhancement,"4th International Conference on Automation, Control Engineering and Computer Science (ACECS) Proceedings of Engineering and Technology – PET, pp.82-87, Vol.19, 2017.

[7] M. Benaissa, S. Hadjeri, S.A. Zidi, "Impact of PSS and SVC on the Power System Transient Stability," Advances in Science, Technology and Engineering Systems Journal, pp. 562-568, Vol. 2, No. 3, 2017.

[8] K.Y. Lee, M. Farsangi, and H.Nezamabadi-pour, "Hybrid of Analytical and Heuristic Techniques for FACTS Devices in Transmission Systems," Power Engineering Society General Meeting, IEEE, 2007.

[9] R.Dubey, S.Dixit, and G.Agnihotri, "Optimal Placement of Shunt FACTS Devices Using Heuristic Optimization Techniques: An Overview," International Conference on Communication Systems and Network Technologies, 2014.

[10] N. Hashim N. Hamzah, M.F. Abdul Latip and A.A. Sallehhudin, "Transient stability analysis of the IEEE 14-Bus test system using dynamic computation for power systems (DCPS)," Third International Conference on Intelligent Systems Modelling and Simulation 2012.

[11] P.K. Iyambo, and R. Tzoneva, "Transient Stability Analysis of the IEEE 14-Bus Electric Power System," Africon conference, IEEE, 2007.

[12] A. Zerigui, "Optimal distribution of electricity production with transient stability constraint established by a statistical approach," Phd. thesis,Univ. Québec, Feb 2015.

[13] X. Tu, L.A. Dessaint, and I. Kamwa, "A global approach to transient stability constrained optimal power flow using a machine detailed model," Can. J. Elect. Comput. Eng., Vol. 36, No. 1, Winter 2013.

[14] R.Kamdar, M.Kumar and G.Agnihortri, "Transient stability analysis and enhancement of IEEE-9bus system," Electrical & Computer Engineering: An International Journal (ECIJ) Volume 3, Number 2, June 2014.

[15] J.C .Chow, R. Fischl, M. Kam, H.H. Yan and S. Ricciardi, "An Improved Hopfield Model for Power System Contingency Classification," Department of Electrical and Computer Engineering Drexel University, Philadelphia, PA 19104, USA.

[16] C. I. Faustino Agreira, C.M. Machado Ferreira, J. A. Dias Pinto, and F. P. Maciel Barbosa, "Application of the Rough Set Theory to the Steady – State Contingency Classification," Power Tech, June 2005, IEEE Russia.

[17] Abdulwahhab A. A. 'Contingency ranking of power systems using a performance index', International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 02 Issue: 02, 2015.

[18] Y. Chen, F.Luo, Y. Xu and Jing Qiu4, "Self-adaptive differential approach for transient stability constrained optimal power flow," IET Gener. Transm. Distrib, pp. 1–10, 2016.

[19] De. Gan, R. J. Thomas, and R. D. Zimmerman, "Stability-Constrained Optimal Power Flow," IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 15, NO. 2, MAY 2000.

[20] F. Mohseni-Kolagar and H. Miar-Naimi , " Transient Analysis of Bang-Bang Phase locked Loops without Cycle Slipping for Frequency Step Inputs," 2012, 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT).

[21] A.Meddeb, H.Jmii, and S.Chebbi, "Heuristic analysis and contingencies classification of case Study IEEE 14-bus," 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp.128–133, 2016.

[22] L. H. Fink and K. Carslen, "Operating Under Stress and Strain," IEEE Spectrum Magazine , 1978

[23] P. M. Anderson and A. A. Fouad, "Power system control and stability," Book, IEEE Press, 1994.

[24] P.Kundur, "Power system stability and control, book" McGraw-Hill,Inc 1993.

[25] M.Pavella, D.Ernest, and D. Ruizvega, "Transient Stability of Power Systems A Unified Approach to Assessment and Control," Book, Kluwer Academic Publishers Boston/Dordrecht/London, 2000th Edition.

[26] Y. Zeng, and Y.Yu,"A Practical Direct Method for Determining Dynamic Security Regions of Electrical Power Systems,", International conf on power system Tech proceeding, IEEE 2002.

[27] P. A. Kaplunovich, K. S. Turitsyn, "Statistical properties and classification of N – 2 contingencies in large scale power grids," 2014 47th Hawaii International Conference on System Science, 2014.

[28] K. Cheong Sou, H. Sandberg, and K. Henrik Johansson, "Electric Power Network Security Analysis via Minimum Cut Relaxation," 2011 50th IEEE Conference on Decision and Control and European Control Conference, Orlando, FL, USA, December 2011.

[29] A. Singh, A. Uddin Ahmad, "Control Reactive Power Flow with UPFC Connected Using Different Distance Transmission Line," Copyright to IJAREEIE DOI: 10.15662/IJAREEIE, 0409034, 2015.

[30] Visakha, K., Thukaram, D., & Jenkins, L, "Application of UPFC for system security improvement under normal and network contingencies," Electric Power Systems Research, 70(1), 46-55, 2004.

[31] Thukaram, D., Jenkins, L., & Visakha, K. , "Improvement of system security with unified-power-flow controller at suitable locations under network contingencies of interconnected systems," IEE Proceedings-Generation, Transmission and Distribution, 152(5), 682-690, 2005.

[32] Gupta, S., Tripathi, R. K., & Shukla, R. D, "Voltage stability improvement in power systems using facts controllers: State-of-the-art review," Power, Control and Embedded Systems (ICPCES), pp. 1-8, 2010.

[33] Kamarposhti, M. A., Alinezhad, M., Lesani, H., & Talebi, N, "Comparison of SVC, STATCOM, TCSC, and UPFC controllers for static voltage stability evaluated by continuation power flow method," In Electric Power Conference, EPEC, IEEE Canada pp. 1-8, 2008.

[34] S.V.R.Kumar, and S. S.Nagaraju, "Transient Stability Improvement using UPFC and SVC," ARPN Journal of Engineering and Applied Sciences, Vol. 2, No. 3, June, 2007.

[35] B.Bhattacharyya, V.K.Gupta, and S.Kumar, "UPFC with series and shunt FACTS controllers for the economic operation of a power system," Ain Shams Engineering Journal, 2014.

[36] X.Ping Zhang, C. Rehtanz, and B. Pal, "Flexible AC Transmission Systems: Modelling and Control," DOI 10.1007/978-3-642-28241-6, Springer-Verlag Berlin Heidelberg, 2012.

[37] S.K. Kodsi and C. A. Canizares, "IEEE 14 Bus system with facts controllers," IEEE Trans, 2003.

[38] Eurostag, Eurostag Software Release Notes, Tractebel-EDF, Release 5.1, Dec 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 412-417 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

The method of correlation investigation of acoustic signals with priority placement of microphones

Bohdan Trembach*,[1], Roman Kochan[1], Rostyslav Trembach[2]

[1]*L'viv Polytechnic National University, Department of Specialized Computer Systems, 79000, L'viv, Ukraine*

[2]*Ternopil Ivan Puluj National Technical University, Department of Automation of Technological Processes and Productions, 46001, Ternopil, Ukraine*

ARTICLE INFO

Article history:
Received: 13 November, 2017
Accepted: 24 January, 2018
Online: 10 February, 2018

Keywords:
Acoustic signals
Correlations
Special processors
Hamming space

ABSTRACT

Examples of analytical calculations of the system characteristics of the hardware and time complexity of the correlation system based on a certain number of microphones and the corresponding number of interrelations are presented. The structural solutions of the hardware special processor implementation of such class of multichannel devices for recognition and identification of types and the spatial location of sources of acoustic signals are developed. The structural model of the spatial identification of sources of acoustic signals in Cartesian coordinates of a two-dimensional Hamming space with the priority placement of microphones as receivers of acoustic signals is proposed.

1. Introduction

This paper is an extension of work originally reported in 14th International Conference the Experience of Designing and Application of CAD Systems in Microelectronics (CADSM) [1].

Development of theoretical foundations of information technology and software - hardware correlation signal processing is actual scientific - applied problem is to be solved in many industries. Identification of sources of acoustic signals (SAS) relative spatial placement of microphones - receivers of acoustic signals (RAS) is also included. The given problem is a primary-industry-technical task of the special technique [2–4].

Analysis of the known research results. In [5–8] the authors created successful but far from optimal solution of such problem is working out localization accumulated information systems. The determination of spatial parameters (θ) azimuth and distance to the SAS (φ) the use of a certain number of (q) correlates for a given number of those (m) microphones is taken as the base of the system.

The example of the structure of Acoustic Localization by Accumulated Correlation (ALAC) system [5] is shown in Figure 1.

It should be noted that the number of required correlators for a given number of chaotic space placed microphones RAS is determined by the condition of symmetry correlation matrix (1).

*Corresponding Author: Bohdan Trembach, L'viv Polytechnic National University, Department of Specialized Computer Systems, 28 a S. Bandera str., 79000, Lviv, Ukraine. Email: trembach.bogdan@gmail.com

$$\begin{array}{ccccc} MK & 1 & 2 & 3 & 4 \\ 1 & 1 & R_{12} & R_{13} & R_{14} \\ 2 & R_{21} & 1 & R_{23} & R_{24} \\ 3 & R_{31} & R_{32} & 1 & R_{34} \\ 4 & R_{41} & R_{42} & R_{43} & 1 \end{array}, \quad (1)$$

where $R_{ii} = 1$, are placed along the diagonal;

$R_{ij} = R_{ji}$ – are identical, according to the symmetry of the matrix (1).

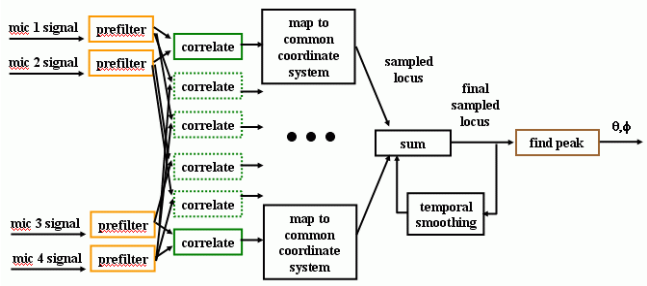


Figure 1. Structure of the ALAC

So, according to the information technology of the correlation processing of the SAS signals in ALAC system, the required amount of (q) correlates for a given number of the (m) microphones is determined by the expression:

$$q = (m^2 - m)/2. \quad (2)$$

The graph of dependency (2) with different numbers of microphones is shown in Figure 2.

For implementation of each correlator in ALAC system the integrated assessment multiplicative correlation function is used by the expression:

$$L(q) = G(\int_{\tau(i,q)-\frac{w}{2}}^{\tau(i,q)+\frac{w}{2}} x_i(t) \times x_j(t-\tau_{i,q}+\tau_{j,q})dt) + \quad (3)$$

$$+ \alpha V_E,$$

where: q – the identifier of the SAS; G – the integrated cross-correlation function; $x_i(t)$ and $x_j(t-\tau_{i,q})$ – current and delayed on the time interval $\pm\tau_{i,q}$ acoustic signals (AS) accordingly; αV_E – damping energy coefficient of the cross-correlation function on the interval $\tau_{i,q}$.

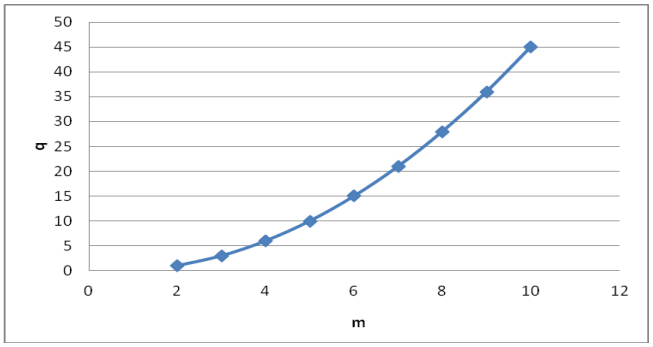


Figure 2. The dependence of the required number of correlators on the number of microphones in the system

The analysis of the analytic expression (3) allows to determine that the implementation of R_{xx} correlates in ALAC system is based on a time delay, $x_i(t)$ multiplying and $x_j(t-\tau)$ integrating analog signals, essentially limiting its functionality, simplifying and increasing the speed and accuracy also prevents it implementation based on digital microelectronics crystals and programmable integrated-circuit logic (FPGA).

The purpose of the work is to develop and explore systemic and structural characteristics of digital special processor computing centered multiplicative correlation function.

2. Formulation of the problem

The purpose of the work is to develop a digital correlator to determine the location of the acoustic signal source.

In order to study the principles of improving and optimizing system features of digital correlators as the basic components of the discovering system, analysis of system features of special processor and computing digital multiplicative correlation estimates by the expression is carried out [9]:

$$R_{xx}(j) = \frac{1}{n}\sum_{i=}^{n} \overset{\circ}{x} \times \overset{\circ}{x}_{i-j}; \quad j \in \overline{0,m}, \quad (4)$$

where: $R_{xx}(j)$ – is centered autocorrelation function; $\overset{\circ}{x}_i$ and $\overset{\circ}{x}_{i-j}$ – are centered digital value and analog signals $x_i(t)$ and $x_j(t-\tau)$; n – is the volume of the sample data set $\left\{\overset{\circ}{x}_i\right\}$; m – is the number of points of correlation function $R_{xx}(j)$; j – is discrete digital delay unit point x_{i-j} in time.

The example of the interaction in time of the digital signals $\overset{\circ}{x}_i$ and $\overset{\circ}{x}_{i-j}$, where C – constant threshold and corresponding asymptotic of the correlation function $R_{xx}(j)$ is shown in Figure 3.

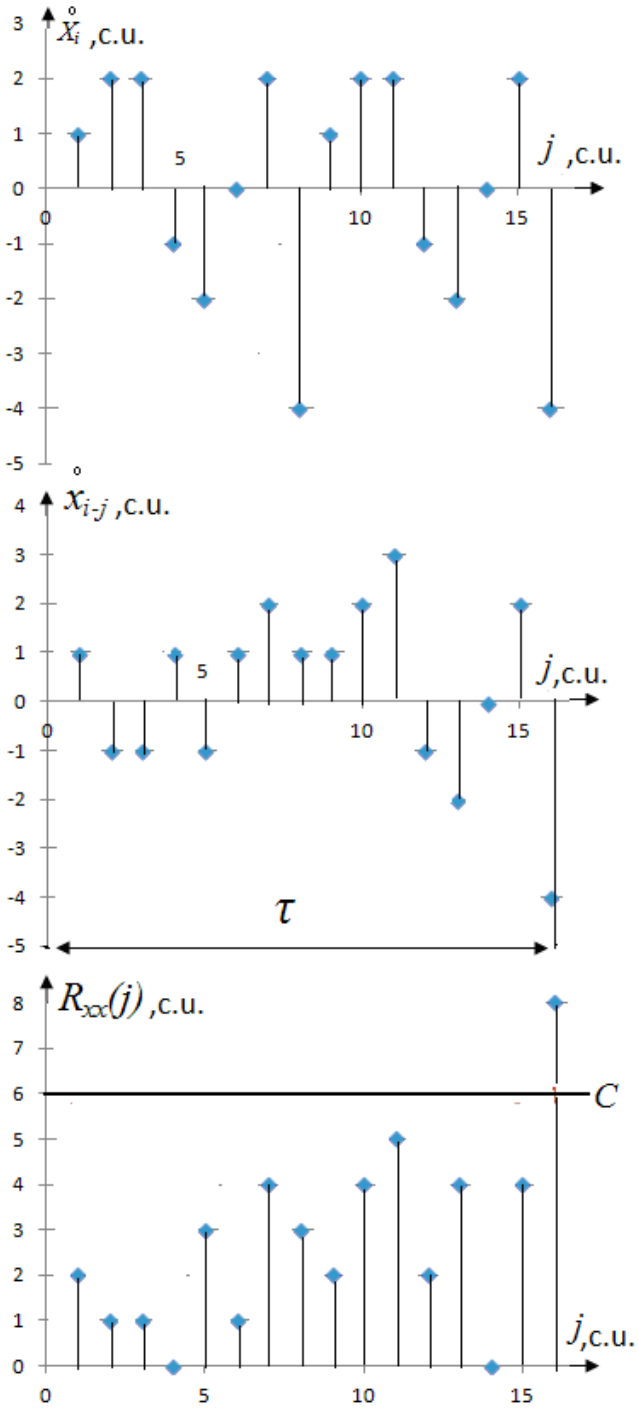


Figure 3. Example of temporal interaction sampled in time and amplitude quantized digital signals and corresponding digital asymptotic estimates multiplicative correlation function in conventional units (c.u.)

Figure 3 shows that at the time of coincidence the current τ digital streams until the moment $\overset{\circ}{x}_i$ of the reception of acoustic signal $x_i(t)$ by a remote microphone SAS level of correlation between the previously obtained a SAS close to the microphone signal $x_j(t-\tau)$ to digital value j – s correlator outputs do not exceed the threshold level constant C and at the time of singing falling at the time fixed main lobe function $R_{xx}(j)$ and accordingly j a numeric value that corresponds to the duration Δt used for direction finding SAS. Obviously, depending on the structure of the stream of digital samples $\overset{\circ}{x}_i$ that reflect the analog signal generated $x(t)$ by a remote SAS in the vicinity of the main lobe $R_{xx}(j)$ functions will emerge sufficiently large lateral lobes, which require separate examination for specific researched objects that have SAS.

In [10] the authors showed that in general for calculation the correlation function $R_{xx}(j)$ according to expression (4) based on a processing arrays of digital data x_i which representing the converted analog-to-digital converter (ADC) input analog signals $x(t)$ you should do the following:

1) define the digital evaluation expectation:

$$M_x = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad (5)$$

2) calculate the array centered values:

$$\overset{\circ}{x}_i = x_i - M_x;\ i \in 1,n; \quad (6)$$

3) calculate the variance:

$$D_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - M_x)^2 = \frac{1}{n}\sum_{i=1}^{n}\overset{\circ}{x}_i^2; \quad (7)$$

4) calculate the correlation function $R_{xx}(j)$ centered on the expression (4);

5) calculate the normalized correlation function:

$$\rho_{xx}(j) = \frac{R_{xx}(j)}{D_x};\ i \in \overline{1,m}; \quad (8)$$

6) perform the comparison of the digital value $\rho_{xx}(j)$ which changes in boundaries $-1 \le \rho_{xx}(j) \le +1$, with boundary constant $0 < C \le 1$:

$$\rho_{xx}(j) > C_0;\ \rho_{xx}(\neq j) < C_0; \quad (9)$$

7) register a digital value j which corresponds to the time delay Δt of the acoustic signal $x(t)$ between two microphones placed at different distance from SAS.

This correlation algorithm of digital processing acoustic signals on a base of multiplicative function (4) is greatly simplified if before the analog-to-digital converter (ADC) pre-differentiation analog $x(t)$ signal is performed and it is passed through a device of automatic gain control.

Such pre-processing of the analog signals $x_i(t)$ and $x_j(t-\tau)$, which are formed on the outputs of the microphones allows to remove from algorithm of calculating operations (5–8) and immediately calculate correlation digital integrated assessment (4) and compare it values at all points j of the constant threshold $R_{xx}(j) > C$.

As it will be shown later, the numerical value of this constant is selected due to the given parameters of digital correlator, which implements the calculation of the multiplicative function $R_{xx}(j)$. Such parameters are:

1) the number of quantization levels of analog signals

$$A = 2^k;\ -\frac{A}{2} \le \overset{\circ}{x}_i \le +\frac{A}{2};$$

2) binary output bit ADC k

$$k = 4,8,10,12;$$

3) the sample size of the database $\left\{\overset{\circ}{x}_i\right\}$

$$n = 2^r;\ r = 4,5,6,7,8,...;$$

4) the number of points correlation function

$$m = 2^l;\ l = 8,10,12....$$

The basic structure of the investigated correlation processing of acoustic signals special processor designed for their processing is shown in Figure 4.

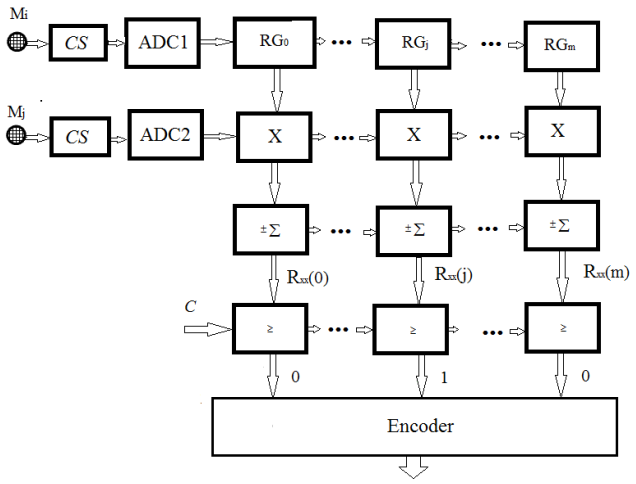


Figure 4. The basic structure of the digital special processor for correlation processing of acoustic signals

The following notions are used in Figure 4: M_i; M_j – RAS microphones; *CS* – conditioners signals; *ADC* – analog-to-digital converter; $RG_0 ... RG_m$ – multi bit shift register source $\pm\overset{\circ}{x}_i$; X – digital binary multiplier; $\pm\sum$ – reversing saving up adder; $\ge$ – module comparing numerical values of the correlation function $R_{xx}(j)$ threshold constant C; encoder – encoder codes Haar – Rademacher.

The work of the digital correlator is as follows. Acoustic signals which are taken by microphone M_i and with a certain time M_j lag are converted to electrical signals which passing

through conditioners signals (CS) are filtered, normalized on power and enter the inputs of the first ADC1 and second (ADC2) alternating analog - digital converters parallel type.

Output k – bit binaries ADC1 come to k – bit input of the multibit shift register (MSR) where RG_j are stored in memory registers, the outputs are fed to the first $j-x$ digital multipliers inputs (X). Binary k – bit codes that are generated at the output ADC2 with a certain time delay corresponding to the time delay of acoustic signals M_j, which are simultaneously fed to all second input ($j-x$) k – bit binary multipliers (X), the outputs of which received $2k$ – bit binaries enter the relevant inputs $(2k+\log_2 n)$ – bit reversible accumulating adder ($\pm\sum$), where codes are received digital values point correlation function $R_{xx}(j)$.

Simplification of operations of division by sample size data set (n) in formula (4) is achieved throughout its multiplicity of 2 degrees, allowing to get the average value of digital codes point correlation function by discarding four junior level in the original binary code accumulating $\pm\sum$. Obtained codes $R_{xx}(j)$ with the bit $2k$, or with less precision $2k–r$, where r = 2,4,6, ... are compared to the corresponding ($j-x$) module compare ($\geq$), the output of which is formed m – bit position code Haar type (00 ... 10 ... 0) , the position "1" of which corresponds to the numerical values of the time delay of acoustic signals received Δt by spatially located microphones M_i and M_j.

For example, when the number of points of the correlation function m = 4096 binary code Δt has 12 bits, that corresponds to Δt = 0.0025 measurement accuracy and uncertainty, and at m = 1024, corresponds to 10-bit binary code and the accuracy does not exceed Δt = 0.001.

The structural scheme shown in Figure 4 can be used for measuring the distance to the SAS by gradient method in case when the source and signal receivers are located anyone line. The example of gradient method is demonstrated in Figure 5.

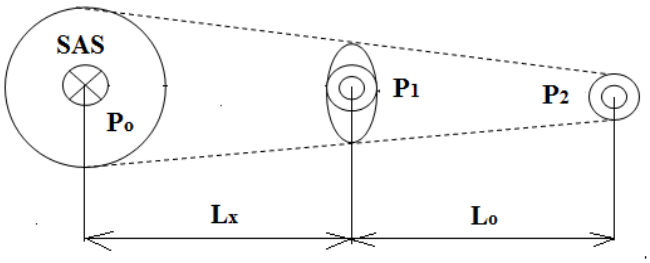


Figure 5. Spatial diagram of SAS capacity changes

The distance to the source is defined by formula

$$L_x = L_0 + \frac{\ln \Delta P}{\alpha},$$

where: α – is the coefficient of acoustic sound energy attenuation in the atmosphere; $\Delta P = P_1 - P_2$ - capacity difference of the signals received from the first and second microphones.

3. Structure of the proposed correlator

The analysis of the expression (4) and the structure of the corresponding digital special processor shows that the presence of the centering operation and the multiplication of alternating digital codes $\pm\overset{\circ}{x}_i \times \pm\overset{\circ}{x}_{i-j}$ complicates the implementation of such a special processor and significantly reduces its performance compared to its functionally equivalent implementation based on the Hamming distance estimation according to the expression

$$G_{xx}(j) = \frac{1}{n}\sum_{i=1}^{n} | x_i - x_{i+j} |, \qquad (10)$$

where centering and multiplication operations are not applied. The application of the Hamming distance estimation based on the calculation of the modular correlation function allows to realize the operation of determining the modular difference between the two digital values $|x_i - x_{i-j}|$ in the microelectronic performance proposed in [9] according.

The purpose of improving the ALAC correlation system is:

1) implementation of the digital representation and correlation acoustic signals processing;
2) reduction of the number of digital correlations of the system;
3) reduction of algorithmic and hardware complexity and increase of speed of digital correlators;
4) implementation of the tabular method for the identification of the spatial location of sources of acoustic signals based on time delay of acoustic signals between microphones;
5) adaptation of the digital correlator to the form of an acoustic signal generated by different sources of sound.

To meet these objective, a method of correlation processing of the acoustic signal with priority placement of microphones [11,12] is proposed on the basis of the multiplicative function (4) converted to the normalized form (8).

Despite successful problem solution for determining time delay of acoustic signals between microphones Δt, such structure of a special processor does not allow to determine independently the spatial coordinates of the source of acoustic signal since it requires more than one base. That is, the number of hardware equipment of such a system increases almost twice. In addition, the use of multiplicative correlation functions $R_{xx}(j)$ and $\rho_{xx}(j)$ requires the implementation of modules for the multiplication and accumulation of reversible interchangeable digital data which considerably complicates the algorithm for processing digital data and the hardware implementation of such a system.

In order to optimize the characteristics of the investigated correlation system the authors propose to implement it on the basis of three microphones, with the priority of spatial placement of one of the microphones at the testbed, and the application of a modular correlation function Gij which allows to identify time delays between acoustic signals Δt_1, Δt_2 and Δt_3 based on the estimation of the Euclidean distance in the Hamming space.

Implementation of the principle of digital processing of acoustic signals allows to reduce significantly the hardware complexity of a single two-channel correlator. The structure of the system is shown on Figure 6.

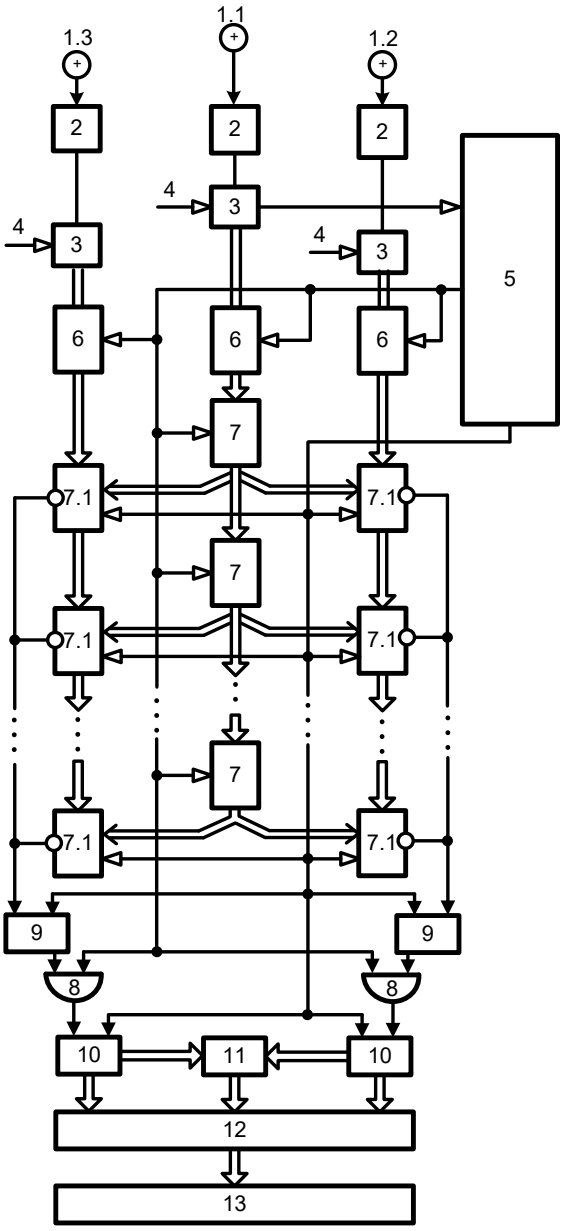


Figure 6. Flowchart of multichannel device for the calculation of modular correlation function

Multi-channel device for calculation of modular correlation function consists of: 1.1, 1.2, 1.3 - respectively: the first priority, second and third receivers of acoustic signals; 2 - amplification automatic adjustment; 3. matched filter of acoustic signals; 4 - reference acoustic signal input; 5 - synchronizer; 6 - parallel-type ADC with output source codes in the binary numeral system of the Rademacher theoretic-numerical basis; 7 - multichannel shift register; ; 7.1 - reverse-accumulation adder; 8 - logical elements AND; 9 - threshold storage of modular differences; 10- RS - triggers; 11 - binary counters; 12 - modular-difference adder; 13 coordinate system based on read-only memory (ROM).

At the beginning of the operating cycle of the device signal S_o of the first output synchronizer 5 generates initial pulse, which resets the memory registers of all cumulative adders of modular differences 9, trigger inputs 10, and binary counters 11 to the zero state.

Input analogue acoustic signals x (t), y (t), z (t), which are generated by a remote source of acoustic signals, enter the input of the acoustic signal receiver 1.1 which is spatially closer to the source of acoustic signals and with some delay in time Δt_1 and Δt_2. Correspondingly, they enter the inputs of related acoustic signal receivers 1.2 and 1.3. Being formed on the outputs of the acoustic signal receivers 1.1, 1.2 and 1.3, electrical signals come to the inputs of the corresponding devices of automatic amplification adjustment 2. Electric signals are produced at their output terminals and are normalized by amplitude and positive indicative potential. Generated, output signals of the automatic amplification adjustment devices 2 flow to the first inputs of matching filters 3. The second filter inputs are connected to the input terminals of reference acoustic signal input 4, and the outputs are connected with the first inputs of corresponding ADC 6.

During the operation cycle of the device the clock signals of the second output of the synchronizer 5 Sx synchronize the formation of source codes x_i, y_i and z_i on the outputs of the corresponding ADCs 6, the corresponding shifts of the digital codes x_{i-j} in the multichannel shift register 7, and pulses coming from the outputs of the corresponding logic elements AND 8 on the inputs of the corresponding counters 10. At the same time, corresponding threshold amounts $\sum_{i=1}^{n}\left|x_{i-j}-y_i\right| \bmod P;\ \sum_{i=1}^{n}\left|x_{i-j}-z_k\right| \bmod P$, s are formed in the accumulated aggregates of the modular differences 9 of the first and second groups, the excess of which causes the formation of a zero potential which converts the corresponding trigger 10 into a single state of the S-input in one of the channels of each group on the inverted outputs of adder accumulator of modular differences 9.

The accumulated amount of impulses in the first counter 11 Δt_1 and in the second counter 11 Δt_2 enters the first and second inputs of the coordinate system 12, and the resulting modular difference in the modular increment adder 12, enters the third input of the coordinate system 13, the output in the form of a code Δt_3. This is the output of the device. Figure 7 shows a timeline diagram of the formation and processing of acoustic signals with the registration of time delays between three microphones Δt_1, Δt_2 and Δt_3 and corresponding values of the source sound coordinates selected from ROM.

For example, at $\Delta L = 1$ м; $i \in \overline{1,256}$, $j \in \overline{1,256}$, which corresponds to the binary codes i and j with 8 bit capacity, the number of sound identified sources with coordinates C_{ij} equals 128^2=8192, and with $i \in \overline{1,64}$ and $j \in \overline{1,128}$ respectively $32 \times 64 = 2048$, and with $i \in \overline{1,16}$ and $j \in \overline{1,32}$ equals respectively $8 \times 16 = 128$. Hence, it corresponds to the spatial dimensions of the testbed targets 256×256; 64×128 and 8×16 m.

4. Research results

The application of multiple autonomous systems of parallel direction finding of acoustic signal sources with respectively

small number of nodes in the Hamming space allows to reduce significantly the accuracy requirements of the digital representation of values Δt_1, Δt_2 and Δt_3.

The required memory for tactical representation of coordinates of the acoustic signal source at $i \in \overline{1,256}$ and $j \in \overline{1,256}$ equals 256 Kbytes.

Generalized chart of acoustic signal processor operation in the Hamming distance based on the modular correlation function is shown on Figure 8.

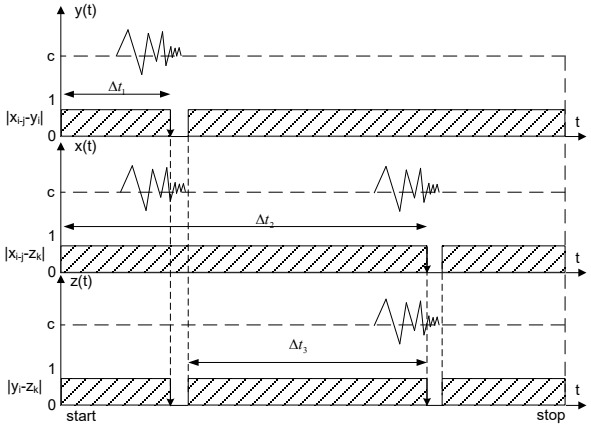


Figure 7. Timeline diagram of code correlation-modular formation of acoustic signal time delay.

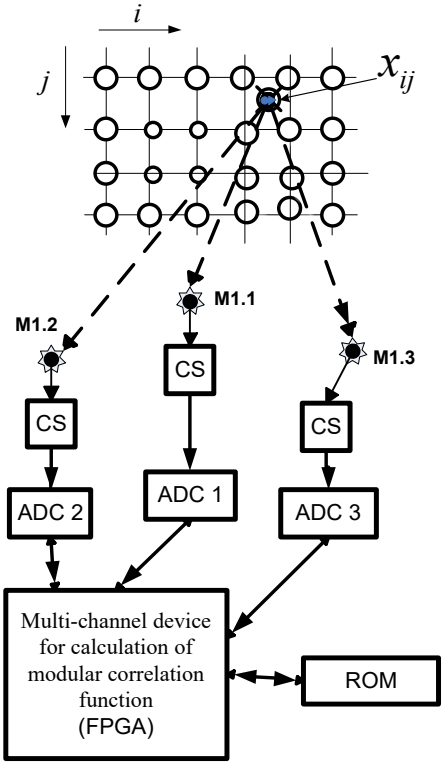


Figure 8. Basic structure of digital special processor for correlation processing of acoustic signals in the Hamming distance

The accuracy of the proposed digital correlator depends on:

- the resolution of a multichannel shift register (Figure 6, element 7);
- the size of square of a Hamming space x_{ij} (Figure 8), which is stored in ROM.

5. Conclusions

The purpose of the work was to develop a structure of a digital correlator to determine the location of the acoustic signal source.

Concept design of an automatic system to determine the location of the acoustic signal source has been proposed. It is based on open architecture and provides connection of multiple autonomous acoustic receivers to the system of correlation special processor using wireless communication channels. This allows automatic collection, processing and data transfer regarding operational conditions in the coverage area of acoustic receivers.

The proposed method of optimizing the structure of multi-channel digital correlates with priority spatial placement of a microphone and application module correlation function to process acoustic signals can significantly simplify the algorithm of calculations, reduce the hardware complexity correlates. This allows enhancing its performance, justifying feasibility and effectiveness of these solutions in the established system of monitoring sources of acoustic signals. The proposed chart allows reducing the number of correlators by 3 times.

Conflict of Interest

No conflict of interest in this paper has been identified.

References

[1] Bohdan Trembach, Roman Kochan, Rostyslav Trembach, "The method of correlation investigation of acoustic signals with priority placement of microphones", in 14th IEEE International Conference on The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Polyana, Ukraine, 2017.

[2] A.M. Krivosheev, V.N. Petrenko, A.I. Prikhodko, Fundamentals of artillery reconnaissance, Sumy State University, 2014.

[3] R. Kochan, B. Trembach, "The concept of distributed automated system of sound artillery intelligence-based provider" Modern information technologies in the sphere of security and defence, 1, 59–63, 2016.

[4] P.E. Trofymenko, Y.G. Filipenko, Sound intelligence station – 100 years, Sumy State University, 3, 198–202, 2009.

[5] S. T. Birchfield and D. K. Gillmor. Acoustic Localization by Accumulated Correlation. [Online]: http://www.ces.clemson.edu/~stb/research/acousticloc/

[6] S. T. Birchfield. A Unifying, "Framework for Acoustic Localization" in 12th European Signal Processing Conference (EUSIPCO). Vienna, Austria, 2004.

[7] S. T. Birchfield and D. K. Gillmor, "Fast Bayesian Acoustic Localization", in the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, Florida, 2002.

[8] S. T. Birchfield and D. K. Gillmor, "Acoustic Source Direction by Hemisphere Sampling", in the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, USA, 2001.

[9] I. Albanskiy, V. Pikh, T. Zavedyuk, G. Korniychuk, "Theory and Special Processors of Spectral Cosine Fourier Transformation Based on Various Correlation Functions in Hamming Space" in 11th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), Slavske, Ukraine, 2014.

[10] Trembach R.B., Trembach B.R., Sydor A.V., Device of adding multi-bit binary numbers, Ukrainian patent №.117789, 2017.

[11] Bohdan Trembach, Roman Kochan, Rostyslav Trembach, "Multiplex digital correlator with high priority deployment of one of the acoustic signal receivers", Scientific Journal of TNTU, 4(84), 99-104, 2016.

[12] Bohdan Trembach, Roman Kochan, Rostyslav Trembach "Methods of structural design optimization of software hardware problem identification of the spatial parameters of acoustic signals sources", Scientific Journal of KNU, 1(245), 136-139, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 418-425 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Systematic Tool Support of Engineering Education Performance Management

Aneta George*, Liam Peyton, Voicu Groza

School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa ON K1N6N5, Canada

ARTICLE INFO

Article history:
Received: 16 November, 2017
Accepted: 24 January, 2018
Online: 10 February, 2018

Keywords:
Data analysis
Accreditation
Graduate attributes
Performance management

ABSTRACT

Engineering schools must adopt or develop their own systems and processes for graduate attribute assessment. In this paper, we take a systems engineering approach to graduate attribute assessment and propose a system architecture and tool-supported continuous improvement process with key algorithms and mathematical analysis to process the data and provide performance management reporting. Over several iterations, we have introduced and evaluated improved systems support in a collection of tools called the Graduate Attribute Information Analysis system (GAIA). GAIA integrates course indicators as well as external indicators from a variety of sources. It provides a tool-supported continuous improvement process with templates and notifications for all deliverables. There are sound algorithms and tool support and built-in mathematical analysis for data collection and reporting that includes quantitative and qualitative data; weighted grading; historic trend analysis; improved visualization of results; and standardized reports at both the course level and the program level that can be used either for accreditation or to inform program improvement.

1. Introduction

The Canadian Engineering Accreditation Board (CEAB) requires engineering programs to collect data and assess 12 graduate attributes (GA) as part of a continuous improvement process to ensure the programs are adequately preparing students to be licensed as professional engineers in Canada [1]. The CEAB does not specify how graduate attributes should be measured or how the data should be managed. Engineering schools must adopt or develop their own systems and processes. Radloff, de la Harpe, Dalton, Thomas and Lawson report that for over a decade, academic faculty finds GA assessment challenging [2]. They see the need for faculty to develop a shared understanding of how to integrate GA assessment within the teaching of their courses.

This paper is continuation of the work originally presented in 2017 14th International Conference on Engineering of Modern Electric Systems (EMES) [3]. In this paper, we take a systems engineering approach to graduate attribute assessment and propose a system architecture and tool-supported continuous improvement process with key algorithms and mathematical analysis to process the data and provide performance management reporting. Our research methodology to validate our approach is an iterative combination of action and design science research methodology. We work with the faculty of software, computer and electrical engineering programs in the School of Electrical Engineering and Computer Science at the University of Ottawa to perform graduate attribute assessment, while at the same time identifying gaps, and prototyping improved tool support in our lab. Over several iterations, we have introduced and evaluated improved systems support in a collection of tools called the Graduate Attribute Information Analysis system (GAIA).

In our initial assessment of systems support for graduate assessment at uOttawa, we did not encounter the "academic resistance" discussed by Chanock in his article on developing GA assessment criteria [4]. However, we did identify that existing systems, tools and processes already in place at University of Ottawa School of Engineering tended to complicate rather than support the task. The key issues identified were cognitive overload, lack of clarity, and lack of defined systems and processes.

The first version of GAIA was introduced in 2015 and has progressed through three iterations or versions. The architecture integrates course indicators (such as tests, assignments, quizzes, exams or selected questions) from any learning management system (LMS) as well as external indicators from a variety of sources (such as student surveys, employer evaluation or different types of feedback forms). There is a systematic tool-supported

*Aneta George, SEECS, University of Ottawa, Ottawa ON K1N6N5, Canada

continuous improvement process with templates and notifications for all deliverables. There are sound algorithms and tool support and built-in mathematical analysis for data collection and reporting that includes quantitative and qualitative data; weighted grading; historic trend analysis; improved visualization of results; and standardized reports at both the course level and the program level that can be used either for accreditation or to inform program improvement.

2. Literature Review

We researched and compared solutions from two sources - engineering institutions from countries members of Washington accord; and engineering schools across Canada.

There have been several attempts by different universities to create their own tool that will inform student learning, serve accreditation, and inform program development. In 2004, Maxim describes an assessment plan for student performance in three undergraduate engineering programs - computer science, information systems, and software engineering at the University of Michigan at Dearborn,[5]. It lists course learning outcomes (LO), the measurement instruments used to assess them, student achievement on each particular outcome, and the average score [5]. It gained popularity because of its ability to serve simultaneously as a grade book and LO evaluation tool.

The University of West Georgia, US, created a custom-designed software tool to collect, analyze and report assessment data for program requirements and for accreditation purposes. The tool called COMPASS supported an existing open-source classroom management system by adding the ability to map course LO. It allows for review and analysis of collected assessment data, but lacks a direct reporting feature. The data needs to be retrieved and formatted in order to produce a course assessment report. This complicated the process of data analysis and its implementation for informing program improvement [6].

The Curtin University of Technology in Perth, Australia followed a similar approach in developing the Outcomes Database web-based tool in 2005. It mapped course LO, unit LO, generic graduate attributes and assessment rubrics. The reports it generated helped provide an outcomes focused assessment [7]. Although the paper does not specify any method of tool evaluation, it does show that the Outcomes Database was successfully implemented across courses that share common units in Information Technology, Computer Science and Software engineering.

In his thesis, Essa proposes a custom-developed ABET Course Assessment Tool (ACAT) at the University of Nevada Reno that further improves the Michigan approach [8]. The goal of ACAT is to streamline the course assessment process and standardize reporting. To validate the design and user interface of the tool, Computer Science and Engineering faculty members perform a usability study. Usability of the tool was tested, based on International Organization for Standards criteria [9] that measure effectiveness, efficiency and satisfaction in a specified context of use by specific users. The results showed that the tool created is an improvement over the existing manual process used to assess GA. This study is one of the earliest we have identified that attempts to theoretically compare technology adoption issues between three different tools – COMPASS, Outcomes Database and ACAT. A comparison of the features in two off-the-shelf systems, Compass and Outcomes Database, and the ABET Course Assessment Tool (ACAT), shows 100% compatibility for ACAT, 33.3% for Compass and 66.7% for Outcomes Database.

In 2015, the University of Notre Dame, Australia introduced an outcomes-based curriculum mapping system, Prudentia©. It allows for constructive alignment between different learning outcomes and informs assessment and instructional methodology [10]. The week point of the tool as described by the author is its dependence on the quality of the curriculum framework itself.

Deferent statistical ways to analyze and interpret assessment data was presented by the Office of Academic Planning and Assessment (OAPA) at the University of Massachusetts [11].

Integrating data by cohort using analytical methods and correlations, a method similar to the one used in GAIA, is the subject of collaborative research between Northern Arizona University, Christopher Newport University and James Madison University [12].

Weber addresses the major concern in data analysis – reliability of the results [13]. He explores the use of T-test, ANOVA and ANCOVA to compare different treatments of assessment data for reliability.

Approaches by Canadian universities involve adopting suitable vendor products, adapting tools and processes in place and evolving them into their own learning management system that suits the institutional needs.

A comparison of different GA assessment compatible software tools available on the market was provided in a series of papers by Kaupp, Frank and Watts [14, 15, 16]. Considering the ability of each tool to handle and process data, they classify several outcomes-based assessment support vendor products into five categories – Learning Management Systems (LMS), Learning Content Management Systems (L/CMS), Assessment Platforms (AP), Analytics Systems (AS) and Curriculum Mapping Tools (CMT). Their study concludes that none of the tools is able to manage the GA assessment data independently and they can only address a specific aspect of the GA process [17]. Being distinct from the actual assessment process was identified as a common general weakness for all vendor tools. Issues like duplication of grading, need for uploading/re-entering data by instructors, agreement with other tools or systems or data record forms and most of all incompatibility with diverse nature of student assessment turns them into an additional tool rather than major carrier of assessment analysis for program improvement. Furthermore, the authors compiled evaluation criteria for tool adoption identifying that popularity of the tool should not be used as measure for its functionality. The choice should rather be made based on the compatibility of the tool with institutional needs and systems.

Identifying the process of measuring GA performance as the most difficult step in meeting accreditation requirements. Saunders and Mydlarski from McGill University discuss adopting current institutional resources and evolving them into a software information system [18].

Queen's University, University of Calgary, University of Toronto, Concordia University, University of British Columbia, University of Manitoba and Dalhousie University outlined and

compared their institutional approaches to accreditation requirements in a joint publication [19]. In the research, Concordia University is identified as one of the first Canadian engineering schools to develop their own Learning Management System when enabled them to collect data and allow for sharing between users.

Developing our algorithms, we explored the different ways for administering assessment data presented by Carleton University [20], University of Alberta [21] and University of Calgary [22]. Carleton University and University of Alberta merged the twelve CEAB graduate attributes with respective indicators, measures and rubrics. University of Alberta adds an additional step to the process by involving sub-categories associated with learning objectives.

3. GAIA

Engineering programs in Canada need to demonstrate that their graduates possess twelve specific attributes [23]: engineering knowledge; problem analysis; investigation; design; use of engineering tools; individual and teamwork; communication skills; professionalism; impact of engineering on society and the environment; ethics and equity; economics and project management; life-long learning. An Accreditation Board (AB) Report indicates that accreditations completed in 2015 were the first ones, which included Graduate Attribute Assessment and Continual Improvement compliance [24].

In developing the architecture for GAIA we had the following objectives:

- find a way to use (when possible) and/or modify (when needed) assessment tools and rubrics already in place;
- integrate GAs, key performance indicators (KPI), assessment tools, measurement criteria, course information sheets, data collection and analysis into one information system;
- measure GA performance and allow data to inform a continuous improvement process for each program
- generate reports and perform mathematical analysis to inform program improvement;
- be user-friendly and time-efficient;
- minimize cognitive overload for any data collection or analysis task.

3.1. System Architecture

GAIA provides three types of performance management support – collecting data, processing data and generating reports. Figure 1 below shows the GAIA architecture. It includes an academic platform (faculty administrator, program coordinator, program professors, course professors and students) focused on in-class evaluation and a non-academic platform (employers, co-op office, alumni and students) focused on evaluation mechanisms external to the class room. Data collected through the Academic Platform (AP) is associated with program-related courses and traditional courses typically supported by a Learning Management System (LMS). The Non-Academic Platform (NAP) deals with all other sources of data using Registration Management System (RMS), university-run surveys and outside survey sources. GAIA's performance is managed by a system administrator to assure regular assessment data flow and support reporting of results in a timely manner.

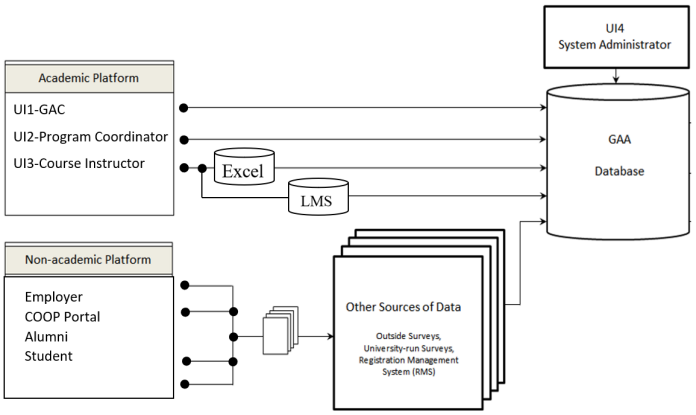


Figure 1: GAIA Architecture

Tool support needs to provide the GAA database (DB) with four types of user interfaces – Graduate Attribute Committee (GAC, UI1), program coordinator (UI2), course instructor (UI3) and system administrator (UI4). Furthermore, it needs to accommodate data fed by different machine interfaces for learning management systems (LMS) and tools already in place for the process of data recording and reporting, as well as qualitative data reported through the COOP portal.

Figure 2 shows the data processing performed by GAIA. It uses different algorithms to process graduate attribute assessment data and generate reports for accreditation agencies in addition to supporting a continuous improvement process for each engineering program. It applies different algorithms to process qualitative and quantitative data input by faculty directly, or imported from a variety of sources. We have identified four MI sources for GA assessment data – LMSs, COOP portal, employer evaluation reports and student surveys. GAIA's architecture pulls external data from these sources in three different ways:

- Using CSV files
- SQL server
- ODBC compliant sources

GAIA's architecture also supports Excel's External Data feature. Once data is initially placed at the location selected by the user, the Refresh button from the Excel Data tab allows for automated updates reflecting data changes at the source. This process is usually set up and performed by the system administrator. GAIA also allows the use of Open Database Connectivity (ODBC) as a programming interface to pull data from different database management systems (DBMS) such as Blackboard or D2L, both currently used at the University of Ottawa.

3.2. Tool Supported Continuous Improvement Process

GAIA supports a systematic continuous improvement process for each program. Creating and implementing improvement steps are initiated and overseen by the members of respective Program Curriculum committees. Success is reported and gaps identified based on consecutive cycles of data analysis performed supported by GAIA as shown in Figure 3 below.

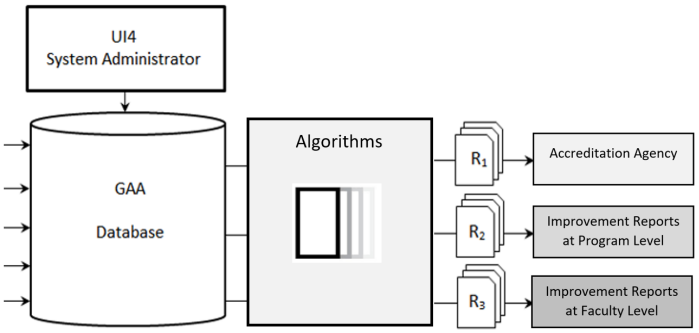


Figure 2: Data Processing for Reporting

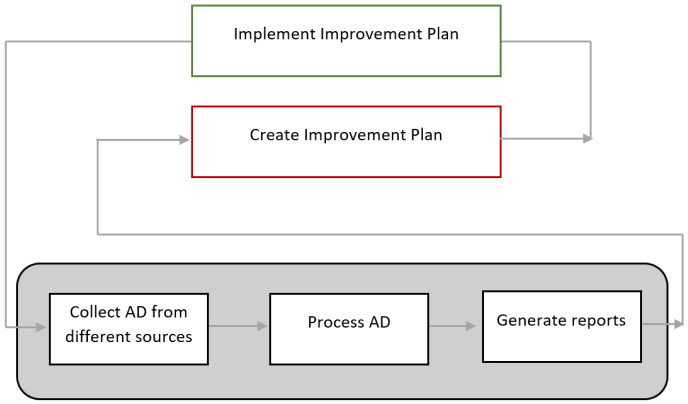


Figure 3: GAIA-supported Continuous Improvement Cycle

The top-down approach, illustrated in Figure 4, mandated by a common continuous improvement process (CIP) leads to a common set of indicators that can be shared across programs. There is a standardized program report that charts a common set of indicators across the 12 graduate attributes. The data for those indicators is obtained from measuring student achievement on the learning outcomes compiled into the common set of high-level indicators specified in the Program Reports. Initially, each program had its own set of performance indicators, which was problematic when trying to achieve consistent reporting at the faculty level that could support cross-program comparisons. Each program aggregated data is clustered into three levels:

- Level I (Course Level): Learning Outcomes presented at the bottom row in Figure 4;

- Level II (Program Level): Performance Indicators, indicated in Figure 4 in black ovals;

- Level III (Meeting Accreditation Requirements): Graduate Attributes analysis for accreditation and program improvement.

At level III we have the 12 graduate attributes specified by CEAB for graduate attribute assessment. Each program has its own set of measurement criteria used to report on achievement for a particular performance indicator (KPI). Selected KPIs are integrated to report on respective graduate attributes (GA). Those indicators are determined by the measurement of learning outcomes for particular courses. The compilation into indicators is mandated in a standardized fashion by the faculty to ensure that achievement is reported in a consistent fashion within a standardized process of continuous improvement.

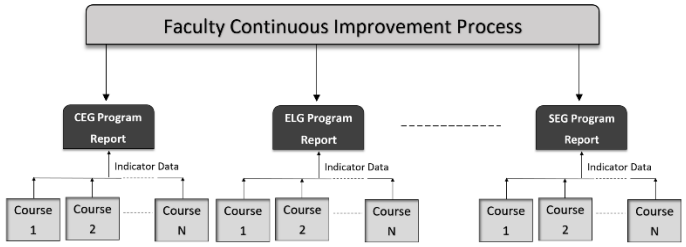


Figure 4. Faculty Continuous Improvement Process.

3.3. Tool-Supported Data Collection and Reports

GAIA provides data collection and reports at two different levels using Course Data Entry Form (CDEF) for individual courses and the Program Report Form (PRF) for an entire engineering program. Data for CDEF is provided by UI3 or input through CVS, SQL server or ODBC compliant sources. Data to Data for PRF is automatically processed and integrated from all the CDEFs for a program. CDEF provides reports and analysis for individual courses. PRF provides reports and analyses across the entire program.

As discussed in section 3.1, GAIA generates three major types of reports – reports for accreditation, improvement reports at program level and improvement reports at faculty level. They are presented in table and graph forms to improve visibility and usefulness. Different tool components, algorithms and data sources are involved in each.

A compiled data report is generated at course level and reflects GA achievement for courses which typically take two semesters to complete. Example for such course is a capstone project. Figure 5 and Figure 6 show a sample report generated for the purpose of this paper in table and graph form respectively.

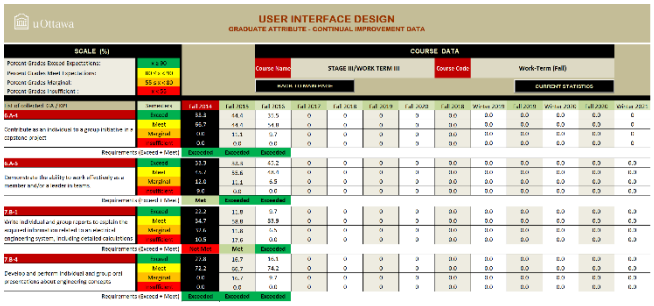


Figure 5. CDEF Table Report Form.

As illustrated in Figure 5 above the report uses color coding to better indicate the level of achievement for the course toward accreditation requirements. Using the CEAB meeting requirements scale with 60% - 80% boundaries. All courses that show achievement per GA below the lower bound did not meet expectations. Their status is shown in red. Above the upper bound of 80% indicates exceeding expectations (shown in dark green shading). The accepted level for meeting accreditation requirements is indicated in light green.

The current statistics graph illustrated in Figure 7, provides a graphical presentation of a historic trend of data collected for selected course. It allows for easy comparison of achievement for

an entire accreditation cycle and gives a better visualization of the overall graduate attribute assessment against meeting accreditation requirements. The Report button allows users to see a historic trend of data in table form.

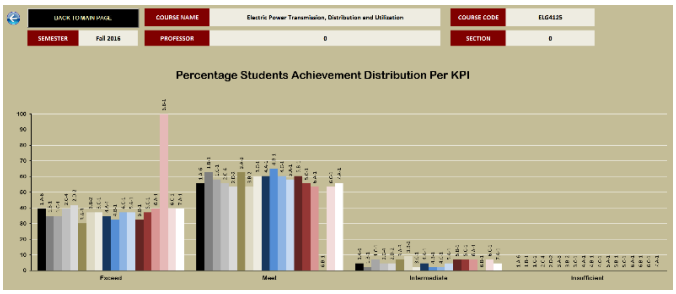


Figure 6. CDEF Graph Report Form.

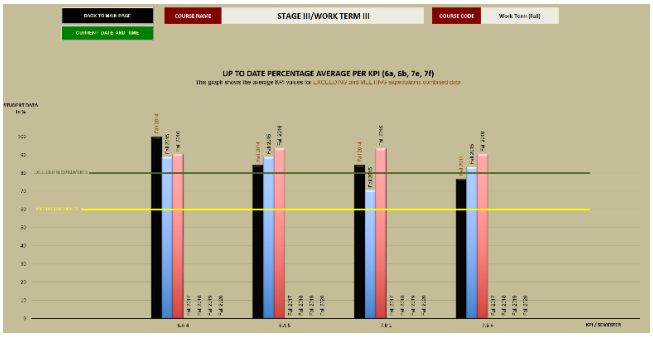


Figure 7. CDEF Current Statistics Report Form.

The Program Report Form (PRF) provides similar statistics which reflect achievement at program level. These reports are used by the curriculum committee members to inform program improvement. The PRF is a read-only workbook. It collates the assessment data provided by CEDF for all courses on a semester basis. This component of the tool measures and reports the cumulative impact individual courses have on overall program performance. It is generated in table and graph form per graduate attribute. A sample report is shown on Figure 8.



(a) (b)

Figure 8. PRF generated report (a) table form, (b) graph form.

A special report form is used to track COOP work-placement data. Data is fed into GAIA from the University of Ottawa COOP Portal using CVS, SQL server or ODBC compliant sources. The report is used to inform on students' ability to secure their first COOP position. A sample of such report form is simulated for the purpose of this paper and is shown on Figure 9.

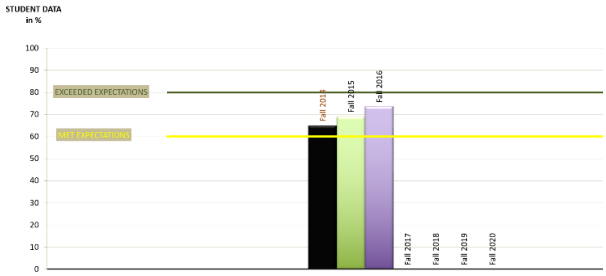


Figure 9. Work-Term Placement Data Report Form.

GAIA allows for comparison between cohorts by generating two reports: the Graduate Attribute Report per Cohort (GAR/C) and the Course Progression Report per Cohort (CPR/C). The former arranges averaged GA data per attribute, while the latter tracks students' achievement as student progress in their program. Data from both cohort reports are used for comparison to provide a historic data trend for further curriculum development. Furthermore, the COOP Progress Report per cohort (COOPR/C) adds reliability in analyzing students' employability and professional skills assessment provided by employers. Table 1, Table 2 and Table 3 below illustrate the datum of cohort reports. It is based on SEG course sequence selected by the program GA Committee to report on GA achievement. Color codes indicate the relative level of the course within the program as follows:

- blue shading is used for Year 5 courses
- green shading is used for Year 4 courses
- peach shading is used for Year 3 courses
- magenta shading is used for Year 2 courses
- no shading is used for Year 1 courses

Table 1. Graduate Attribute Report per Cohort (GAR/C)

	Graduate Attribute											
Courses	GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10	GA11	GA12
	SEG 3101	SEG 3101	SEG 3103	SEG 491 1	SEG 2105	SEG 2105	SEG 2911	SEG 491 1	SEG 2911	SEG 2911	SEG 4105	SEG 4911
		SEG 491 1			SEG 3101	SEG 4105	SEG 491 1	SEG 1911				SEG 1911
					SEG 3102	SEG 491 1						

Table 2. Course Progression Report per Cohort (CPR/C)

	Academic Year				
Course	Year 1	Year 2	Year 3	Year 4	Year 5
	SEG1911	SEG2105	SEG2106	SEG3101	SEG4911
		SEG3103	SEG2911	SEG3102	SEG4105
		SEG3125		SEG4145	

Table 3. COOP Progress Report per Cohort (COOPR/C)

Academic Year				
Year 1	Year 2	Year 3	Year 4	Year 5
SEG1911	COOP Placement I	SEG2901	SEG3901	SEG3902

To illustrate the mechanism of data sorting behind the three cohort reports we generated random data that is being used for demonstration purposes in all tables and graphs. To follow the course progression within a cohort GAIA combines a historic trend data per respective courses according to the course sequence of the specific engineering program (i.e. software engineering in this sample). To generate GAR/C GAIA for cohort N, GAIA will use

Year (N) data for SEG4911 and SEG4105, Year (N-1) data for SEG3101, SEG3102 and SEG4145; Year (N-2) data for SEG2106 and SEG2911; Year (N-3) data for SEG2105, SEG3103, SEG3125; and Year (N-4) data for SEG1911. This is illustrated in Table 4 below.

Table 4. Use of historic trend data for generating cohort reports (GAR/C and CPR/C)

Cohort (Year)	Course	Graduate Attribute
GA Data (N)	SEG4911	2, 4, 6, 7, 8, 12
	SEG4105	6, 11
GA Data (N – 1)	SEG3101	1, 2, 5
	SEG3102	5
	SEG4145	11
GA Data (N – 2)	SEG2106	Xxx
	SEG2911	7, 9, 10
GA Data (N – 3)	SEG2105	5, 6
	SEG3103	3
	SEG3125	Xxx
GA Data (N – 4)	SEG1911	8, 12

The data-supported framework for cohort reports is illustrated in Figure 10. It represents a combination of methodologies described in Figure 1, Figure 2, Figure 4 and Table 4.

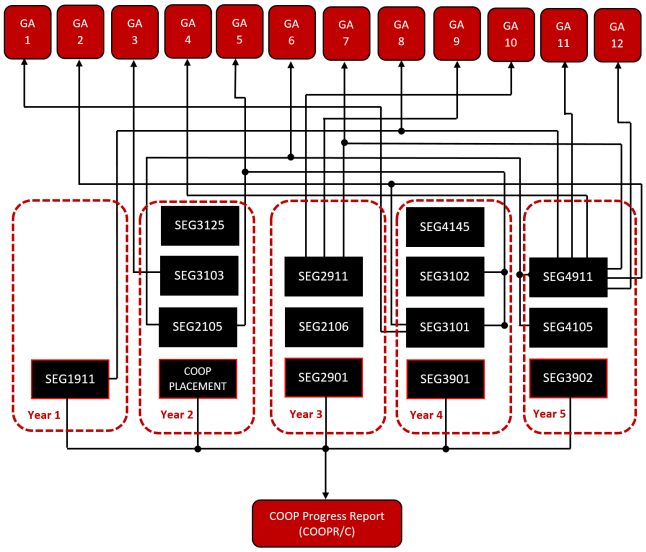


Figure 10. Data-Support Framework for Cohort Reports.

4. Results and discussion

GAIA's improvement and further development is based on ongoing research on latest developments in the area of data management and analytics and the input provided by its users. Mapping the goals stated in section 3 has enabled us to establish an ongoing user-centered tool evaluation. It addresses the different users (U1, U2, U3 and U4 as listed in Figure 1) targeting their specific context of use. Thus, GAIA needs to reflect specific user requirements and specific program requirements for each

university actor. The GAA DB needs to accommodate different types of input data, analyze qualitative and quantitative data, and produce reports has to meet the latest accreditation requirements and inform a continuous program improvement process. Input on the report efficiency and efficiency of the improvement process are being provided by the GA Committee members per program. The ongoing evaluation process as described above is illustrated in Figure 11.

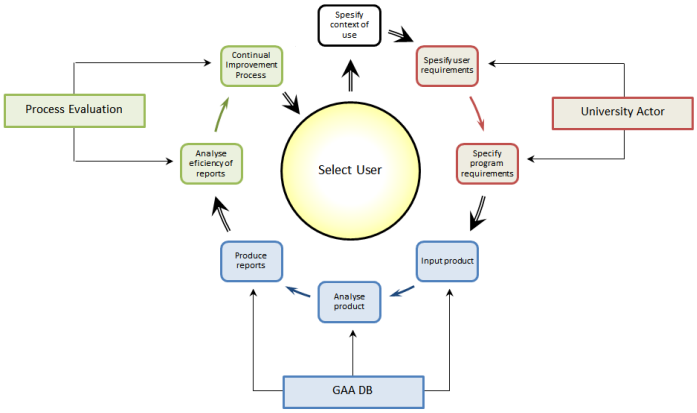


Figure 11. User-Centered Tool Evaluation Process

The following list of evaluation criteria are used for tool evaluation:

4.1. Perceived Ease of Use (PEOU)

This is a core variable, which measures the degree of belief that mastering the tool will not require any extra effort. In the TAM 2 model this construct is defined as a direct determinant of perceived usefulness [25]. Example for its use in academia is the study performed by Park in 2009 [26].

4.2. Perceived Usefulness of the Tool (PU)

Perceived Usefulness is the degree to which a user believes the tool will help successfully complete the task and excel at their job performance. This is a core variable, used to measure adoption in academia assessing learning performance, academic productivity and supporting learning process [26].

4.3. Attitude Toward Using the Tool

The attitude toward use measures the user's feeling about performing the task using the tool. It shows the user's desire to actually use the tool, their positive or negative evaluation of performing the behavior. It measures the ability to perform the task faster, improve user performance when using the tool, using the tool is related to increase of productivity and effectiveness. As a core variable, it is a part of almost every TAM construct set. Samples for its use in academia can be found in the studies by Kim, Park and Tsai [26, 27, 28].

4.4. Behavior Intention to Use the Tool (BI)

It measures the strength of the user's intention to use the tool or the degree of one's willingness to use the tool. It is one of the best indicators of the real usage of the tool. In other words, the actual use depends on the user's intentions to apply effort. It is a combined measure of the wish to finish the task and planning its use in the future.

4.5. Perceived Usage of the Tool

This is the amount of time interacting with the tool and the frequency of its use. Measuring this variable has highest

importance for Faculty Administrator (UI1) as manager interested in evaluating the impact of the tool as a whole.

4.6. Experience Using the Tool

Prior experience was identified as a determinant of behavior in 1980 by Ajzen and Fishbein long before Davis proposed the technology acceptance model [29]. Measuring the variable is mentioned in several studies [30, 31]. According to the studies, experienced users show strong correlation to perceived usefulness of the tool and the behavioral intention to use the tool. Being related to the number of years using the tool, this construct provides valuable information in terms of university users being ready to deal with measuring graduate attributes an ongoing process in a long term.

4.7. Complexity of the Tool

This variable measures the extent to which the user expects to use the tool without any additional effort. It has to do with the tool being difficult to understand and use. Complexity is measured in terms of time taken for the task and integration of tool usage results into existing tool. Complexity is inversely proportional to perceived usefulness (PU) and perceived usage as aspects of adoption. It is also measured by the extent to which the user realizes the possibility of computer crash or data loss.

4.8. Mandatory versus Voluntary Use of Tool

This variable is measured by the extent to which adoption is perceived as a mandatory or non-mandatory task. It is positively related to the behavioral intention to use the tool. In our particular case, the use of the tool is mandatory (requirements enforced by CAEB), so definitely imposed on Faculty Administrator (UI1) and Program Coordinator (UI2). Therefor this variable is not-applicable for them and should be noted as such in their respective evaluation criteria. Participation is mandatory for UI1 and UI2. It becomes mandatory for UI3 if their course is included. At the same time, UI3 cannot easily be coerced for tenured professors so UI3 adoption is critical. Similarly, indirect UI for students, employers etc. cannot be coerced so adoption into process is also critical. The point of this is how to weight the importance of various aspects of the tool. The reporting to CEAB is mandatory and critical so that UI adoption is both constrained and important, but adoption by administrator and coordinator will be coerced so not so critical.

4.9. Interoperability

This measure the tool compatibility with other systems or tools used simultaneously.

4.10. Handling Assessment Data

This measure reflects on tool's ability to serve as database and allow for historic trends of analysis and comparison. The ability to use quantitative and qualitative data is to be evaluated as well.

4.11. Reporting Ability

Tests the tool against its ability to produce reports on course and program levels. Quality of reports and their use to inform program improvement is part of this criteria.

4.12. Alignment with CEAB accreditation requirements

This construct provides direct information on the user's belief that the work performed will serve the need it was intended for. Analysis of the results will have a wide range of application – from indication for improving the tool, to explanation about the user's attitude and perceived intentions

5. Conclusions and Future Work

Creation of GAIA reflected the need for our engineering programs to respond to CEAB accreditation requirements. Ever since it follows the changes implemented to the criteria to provide best support for reporting graduate attributes. Flexibility of the tool and the fact that it is onside made allows for immediate modifications to take place. It is constantly improved following the requests and recommendations from instructors, program coordinators or faculty administrators as well.

In future work, a structural equation model showing the relation between major and external variables in terms of hypotheses will be added. Such a graphical representation of the research model will clearly show the relationships between constructs and specify hypotheses with respect to those relationships. The structural equation modeling (SEM) technique can be employed using the LISREL program. Then direct/indirect effect and t-values will have to be calculated to identify the state of each hypothesis as "Supported" or "Not-supported". Secondary future research will also include further improvement of the model and and more case studies to validate. The goal will be to achieve technology adoption results consistent with expectations. Such results will allow us to measure the tool satisfaction level and predict its usability in specifically constrained contexts.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to acknowledge funding from NSERC and the School of Electrical Engineering and Computer Science at the University of Ottawa which supported this research. We would especially like to thank Claude D'Amours, Emad Gad, Voicu Groza, Stéphane Somé and Marc-André Daoust for their partnership and support.

References

[1] Engineers Canada, "Canadian Engineering Accreditation Board Accreditation Criteria and Procedures," 2014, pp. 1-118. Engineers Canada, Ottawa, ON, K2P2K3, ISSN 1708-8054

[2] A. Radloff, Barbara de la Harpe, Helen Dalton, Jan Thomas, Anne Lawson, "Assessing Graduate Attributes: Engaging Academic Staff and their students," in ATN Assessment Conference 2008: Engaging Students in Assessment, 2008, p. 90.

[3] A. George, L. Peyton, V. Groza, "Graduate Attribute Information Analysis System (GAIA) – From Assessment Analytics to Continuous Program Improvement - Use of Student Assessment Data in Curriculum Development and Program Improvement" in 14th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, Romania 2017. https://doi.org/10.1109/EMES.2017.7980398

[4] K. Chanock, "Developing criteria to assess graduate attributes in students' work for their disciplines," J. Learn. Dev. High. Educ., no. 6, p. 16, 2013.

[5] Bruce R. Maxim, "Closing the loop: assessment and accreditation", Journal of Computing Sciences in Colleges, vol. 20, issue 1, October 2004, pp. 7-18.

[6] Abunawass A., Lloyd W., Rudolph E., "COMPASS: a CS program assessment project", ACM SIGCSE Bulletin, vol. 36, issue 3, September 2004. Pp. 127-131, NY, USA. {DOI: 10.1145/1026487.1008031}.

[7] Konsky B., Loh A., Robey M., Gribble S., Ivins J., Cooper D., "The Benefit of Information Technology in Managing Outcomes Focused Curriculum Development Across Related Degree Programs", Eighth Australasian Computing Education Conference (ACE2006), Hobart, Tasmania, Australia, January 2006.

[8] Eugene O. Essa, "ACAT: ABET Course Assessment Tool," University of Nevada, Reno, 2010, 73 pp., 2010. Available as of Apr. 19, 2017 from http://csi.fau.edu/wp-content/uploads/2013/03/Essa_ACAT_MS-Thesis_U-of-Nevada_Reno_2010.pdf

[9] ISO 13407, "Human-centered design processes for interactive systems", International Standards, Reference number ISO 13407:1999(E), www.isostandards.com.au. Summary online. Accessed July 10, 2016 from http://www.ash-consulting.com/ISO13407.pdf

[10] Steketee C., "Prudentia©: An outcomes curriculum mapping system", Teaching and Learning Forum 2015 - Teaching and Learning Uncapped, Category Professional Practice, pp. 9, 2015.

[11] M. Stassen, K. Doherty and M. Poe, "Program-based review and assessment. Tools and technoques for program improvement", Umass Amherst, pp. 1-62, Fall 2001

[12] S. L. Pieper, K. H. Fulcher, D. L. Sundre and T. D. Erwin, "What do I do with the data now? Analyzing assessment information for accountability and improvement", Research and Practice in Assessment, vol. 2, Issue 1, pp. 1-8, Jan. 2008

[13] E. Weber, "Quantifying student learning: How to analyze assessment data", Ecology 101, pp. 501-511, October 2009

[14] J. Kaupp, B. Frank, and C. Watts, "Evaluation of software tools supporting outcomes-based continuous program improvement processes," in 2013 Canadian Engineering Education Association (CEEA13) Conf., 2013. https://doi.org/10.24908/pceea.v0i0.4861

[15] Jake Kaupp and Brian Frank, "Evaluation of Software Tools Supporting Outcomes-Based Continuous Program Improvement Process. Part II," Canadian Engineering Education Association (CEEA14) Conf. 2015. https://doi.org/10.24908/pceea.v0i0.5856

[16] J. Kaupp and B. Frank, "Evaluation of software tools supporting outcomes-based continuous program improvement processes. Part III," in 2015 Canadian Engineering Education Association (CEEA15) Conf., 2015. https://doi.org/10.24908/pceea.v0i0.5764

[17] K. Chanock, "Developing criteria to assess graduate attributes in students' work for their disciplines," J. Learn. Dev. High. Educ., no. 6, p. 16, 2013.

[18] A. L. Saunders and L. B. Mydlarski, "Evolution of graduate attribute assessment and continuous program improvement in the Faculty of Engineering at McGill University," in 2015 Canadian Engineering Education Association (CEEA15) Conf. 2015. https://doi.org/10.24908/pceea.v0i0.5763

[19] J. Kaupp, B. Frank, R. Brennan, S. Mccahan, L. Narayanan, P. Ostafichuck, N. Sepehri, and K. C. Watts, "A Comparison of Institutional Approaches to CEAB Graduate Attribute Requirements" in Canadian Engineering Education Association (CEEA12) Conf., 2012. https://doi.org/10.24908/pceea.v0i0.4642

[20] J. Harris, D. Russell, and A. Steele, "Progress on Defining the CEAB Graduate Attributes at Carleton University, in Canadian Engineering Education Association Conference, 2011. http://dx.doi.org/10.24908/pceea.v0i0.3628, pp. 1-5, 2011

[21] E. Csorba, D. Chelen, N. Yousefi, N. Andrews, and C. More, "Graduate Attributes at the University of Alberta," a report of the Committee on the Learning Environment (CLE) Subcommittee on Attributes and Competencies, http://www.provost.ualberta.ca/en/~/media/provost/Documents/Information/GraduateAttributes.pdf, pp. 1-22, June 5, 2013

[22] "UC Graduate Attributes Framework", Faculty of Science, U of Calgary, http://www.ucalgary.ca/science/teachinglearning/graduate_attributes_framework, 2017

[23] CEAB, "2014 Canadian Engineering Accreditation Board Accreditation Criteria and Procedures", pp. 118, ISSN 1708-8054, Engineers Canada, 2014. Online "https://engineerscanada.ca/sites/default/files/2014_accreditation_criteria_and_procedures_v06.pdf. Accessed February, 2018.

[24] Iachiver, G., "Accreditation Board Report, EC Board Meeting – September 30, 2015, https://engineerscanada.ca/sites/default/files/AB-Report-September-2015.pdf

[25] Venkatesh V. and Davis D., "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies", Management Science © 2000 INFORMS, vol. 46, no. 2, Feb., 2000, pp.186-204.

[26] Park S., "An Analysis of the Technology Acceptance Model in Understanding University Students' Behavioral Intention to Use e-Learning", Educational Technology & Society, 2009, vol. 12, issue 3, pp. 150–162.

[27] Kim Y., Chun J., Song J., "Investigating the role of attitude in technology acceptance from an attitude strength perspective", International Journal of Information Management, Volume 29, Issue 1, February 2009, Pages 67–77.

[28] Tsai Wen-Chia, "A study of consumer behavioral intention to use e-books: The Technology Acceptance Model perspective", Innovative Marketing, Volume 8, Issue 4, 2012.

[29] Ajzen I., Fishbein M., "Understanding attitudes and predicting social jj..abehavior", Prentice Hall, 1980, pp. 278. {ISBN: 0139364439, 9780139364433}.

[30] Yang, H.-d., and Yoo, Y. "It's all about attitude: revisiting the technology acceptance model," Decision Support Systems, vo.38, issue 1, Oct 2004, pp. 19-31.

[31] Alharbi S., Drew S., "Using the Technology Acceptance Model in Understanding Academics Behavioural Intention to Use Learning Management Systems ", (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 5, No. 1, pp. 13, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 426-433 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Co-designed accelerator for homomorphic encryption applications

Asma Mkhinini[1,2,3*], Paolo Maistri[1], Régis Leveugle[1], Rached Tourki[3]

[1]*Univ. Grenoble Alpes, CNRS, Grenoble INP[**], TIMA, F-38000 Grenoble, France*

[2]*Univ. of Monastir, EμE, 5019 Monastir, Tunisia*

[3]*Univ. of Sousse, Eniso, BP 264 Erriadh 4023, Tunisia*

ARTICLE INFO

Article history:
Received: 17 November, 2017
Accepted: 16 January, 2018
Online: 10 February, 2018

Keywords:
Homomorphic encryption
Hardware accelerator
Modular polynomial multiplication
High Level Synthesis

ABSTRACT

Fully Homomorphic Encryption (FHE) is considered as a key cryptographic tool in building a secure cloud computing environment since it allows computing arbitrary functions directly on encrypted data. However, existing FHE implementations remain impractical due to very high time and resource costs. These costs are essentially due to the computationally intensive modular polynomial multiplication. In this paper, we present a software/hardware co-designed modular polynomial multiplier in order to accelerate homomorphic schemes. The hardware part is implemented through a High-Level Synthesis (HLS) flow. Experimental results show competitive latencies when compared with hand-made designs, while maintaining large advantages on resources. Moreover, we show that our high-level description can be easily configured with different parameters and very large sizes in negligible time, generating new designs for numerous applications.

1. Introduction

This paper is an extension of the work originally presented in 2017 IEEE International Symposium on Design and Diagnostics of Electronic Circuits and Systems [1].

Homomorphic encryption is one of the most significant advances in cryptography in the last decade. It allows arbitrary computations on ciphertexts without compromising the content of the corresponding plaintexts. Thus, data can remain confidential while it is processed, enabling useful tasks to be accomplished with data being stored in untrusted environments. Considering the recent growth in the adoption of the cloud computing and the large deployment of the internet of things, homomorphic cryptography will have a major impact on preserving security and privacy in the coming years. Enterprise customers in the medical and financial sectors, for example, can potentially save money and streamline business processes by outsourcing not only the storage but also the computation of their data to public clouds.

Since the introduction of the first fully homomorphic encryption (FHE) scheme by Gentry [2] in 2009, we have noticed substantial research in the area, for the purpose of designing new homomorphic encryption algorithms, improving the schemes, their implementations, and their applications. Among them, the schemes that are based on Ring Learning With Errors (RLWE) [3] [4] [5] are among the most efficient homomorphic schemes because of their simpler structure, strong hardness assumptions, reduced key size, and reduced ciphertexts expansion with respect to previous schemes. Many of these RLWE-based schemes have been implemented in software [6] [7] [8] [9]. Results report very large latencies and resources consumption. So, in order to improve the performance of homomorphic encryption schemes, there has been research into the hardware acceleration of various homomorphic schemes and their building blocks. To date, there have been few hardware implementations for cryptosystems based on the RLWE problem. The corresponding architectures were mainly designed for fixed and small length operands and optimized for a restricted set of parameters [10], [11], [12], making them limited in terms of target applications and security requirements.

This paper presents a flexible and configurable accelerator implementing modular polynomial multiplication; the main performance bottleneck in RLWE-based homomorphic schemes. The work describes a software/hardware (SW/HW) co-designed architecture based on a High-Level Synthesis (HLS) approach. By combining HLS and SW/HW partitioning, we are able to easily configure our modular multiplier with large parameters (larger than those seen in the literature) suited for high security requirements. In addition, our modular polynomial reducer can be defined as any generic (cyclotomic) polynomial, allowing optimizations in the homomorphic context. We demonstrate the efficiency of the approach on many designs of full modular

*Asma Mkhinini, Univ. Grenoble Alpes, CNRS, Grenoble INP[**], TIMA, F-38000 Grenoble, France, asma.mkhinini@univ-grenoble-alpes.fr Institute of Engineering Univ. Grenoble Alpes

polynomial multipliers satisfying different area/latency trade-offs. So, our results can guide a designer in his choice of the appropriated configuration with respect to the targeted application.

The paper is organized as follows. In section 2, the background information is introduced. Section 3 presents the related works. Section 4 describes our proposed design. Implementation details are reported and discussed in section 5.

2. Theoretical background

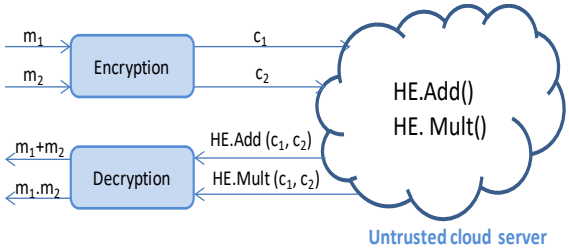
2.1. Homomorphic encryption

The purpose of homomorphic encryption is allowing computations on encrypted data. This means that if a user has a function called f and wants to get $f(m_1, m_2, \dots, m_t)$ for some plaintext messages $(m_1, m_2, \dots, m_t)$, it is possible to instead compute on the corresponding ciphertexts$(c_1, c_2, \dots, c_t)$obtaining a result which decrypts to $f(m_1, m_2, \dots, m_t)$.

Formally, if $\text{Encrypt}(m_i) = c_i (i = 1..t)$ then when evaluating a function f homomorphically on$(c_1, c_2, \dots, c_t)$, we get:

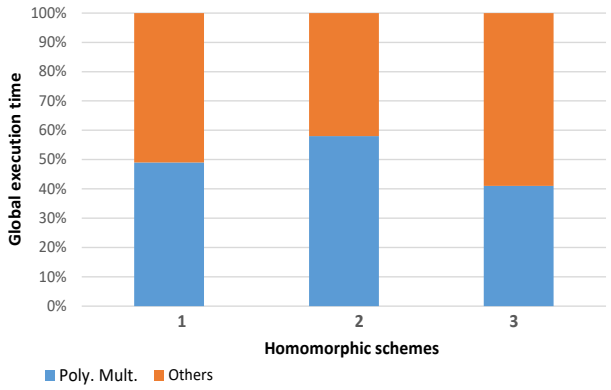
$$\text{Decrypt}\,[f(c_1, \dots, c_t)] = f(m_1, \dots, m_t)$$

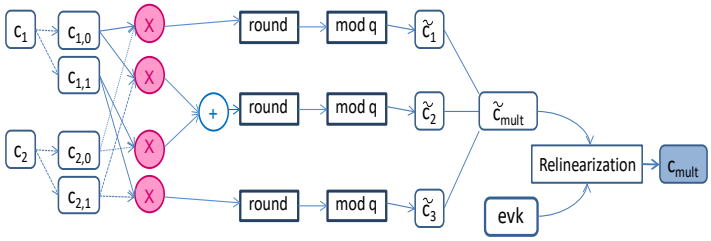
Since every function f can be expressed as a series of additions and multiplications over some algebraic structure, a homomorphic encryption scheme can be defined as an augmented encryption scheme with two additional functions HE.Add() and HE.Mult() to add or multiply on ciphertexts, that result in a ciphertext encrypting respectively the sum or the product of the underlying plaintexts. Figure 1 shows an example of application of FHE in the context of cloud computing. When the number of successive additions HE.Add() and multiplications HE.Mult() can be unlimited during the evaluation step (i.e., computation of the function f), the scheme is known as a fully homomorphic encryption scheme. This generally requires to periodically refresh the ciphertext, otherwise it will be impossible to decrypt. This operation is performed in the encrypted domain and is called "bootstrapping".



2.2. Analysis of software implementations of homomorphic schemes

In order to define the most frequent and time consuming functions during the execution of homomorphic encryption schemes, we profiled existing software implementations [6] [8] of three RLWE-based cryptosystems. Figure 2 shows the profiling results. The analysis reports that polynomial multiplication consumes 41% to 58% of the total execution time. The polynomial multiplication is needed in the encryption, decryption, and evaluation (homomorphic multiplication) steps, as well. For instance, we resume in Figure 3 the operation flow of homomorphic multiplication of two ciphertexts (c_1, c_2) in case of scheme 2 (see Figure 2). In practice, we need 4 polynomial multiplications in order to get c_{mult}. For this reason, a hardware acceleration of this function is of great interest. Hence, an optimized implementation of the modular polynomial multiplication is the target of this paper.





2.3. Modular polynomial multiplication in RLWE-based schemes

In RLWE-based cryptosystems, the primitives are defined over a modular polynomial ring of the form $R_q = \mathbb{Z}_q[x]/f(x)$ where f(x) is a specific irreducible polynomial (cyclotomic polynomial) of degree n and q is an integer modulus ($q > 0$). Parameters n and q define respectively the degree and the coefficient size of polynomials in R_q. Operating in R_q requires reductions modulo q and modulo f(x).

Let $A(x) = a_{n-1}x^{n-1} + \dots + a_0 \equiv (a_{n-1}, \dots, a_0)$ and $B(x) = b_{n-1}x^{n-1} + \dots + b_0 \equiv (b_{n-1}, \dots, b_0)$ be two polynomials of R_q. Computing $C(x) = A(x) \times B(x)$ in R_q needs to first compute the polynomial multiplication of A(x) and B(x) and then reduce the result modulo $(q, f(x))$.

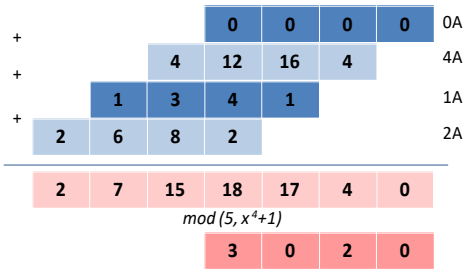
Here is a simple example where $n = 4$, $q = 5$ and $f(x) = x^4 + 1$. We choose:
$A(x) = x^3 + 3x^2 + 4x^1 + 1 \equiv (1,3,4,1)$ and
$B(x) = 2x^3 + 1x^2 + 4x^1 + 0 \equiv (2,1,4,0)$
As shown in figure 4, the result $C(x) = A(x) \times B(x)$ mod $(q, f(x))$ is equal to C_{inR} (3,0,2,0).

2.4. Parameters Derivation

The schemes based on the RLWE problem are governed by a number of inter-related parameters. The modulus q and the degree n are chosen in order to satisfy a given security level λ and a given multiplicative depth L (defined as the maximal number of multiplications that the scheme can handle before it becomes

necessary to apply the bootstrapping procedure). The derivation of these parameters is getting increasing attention lately, in order to provide easy-to-use guidelines for real world applications. In a nutshell, the methodology for parameters extraction aims at sizing these parameters in order to respond to the desired trade-off between security, efficiency, and correctness. Real use-cases of homomorphic cryptosystems define requirements for the multiplicative depth L and the security level λ, then one needs to choose the corresponding values of n and q. Figure 5 illustrates the wide space of practical parameters for RLWE-based schemes with different constraints on L. These configurations are extracted from [6] where the authors explain how to choose these parameters in order to guarantee correctness and security against lattice attacks. They use a lattice basis reduction algorithm based on the van de Pol and Smart approach. This algorithm determines an upper bound on the modulus in a given dimension and for targeted numbers of multiplicative depth L, to ensure a given security level.

For example, with L set to 1, polynomials with a degree around 1024 and coefficients on less than 100 bits can be sufficient. But another scheme requires at least a degree $n = 101853$ and a coefficients size $\log_2(q) = 278$ bits to achieve a security level $\lambda = 80$ bits and a multiplicative depth of 20. Consequently, if we want to target a large set of real applications, our design must be flexible and accept such variations of the parameters.

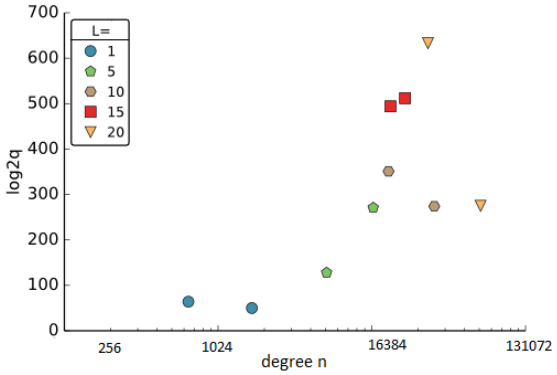


3. Related works

Two principal ways are employed in order to accelerate homomorphic cryptosystems: hardware implementation and GPU (Graphic Processing Unit) acceleration. Hardware accelerators focus mainly on accelerating the most complex functions of homomorphic encryption schemes. There has been some research already conducted into hardware implementations of RLWE-based schemes and their related building blocks. Almost all of them focus on the polynomial multiplication.

Many of these implementations have used the Number Theoretic Transform (NTT) or the Negative Wrapped Convolution (NWC) to perform polynomial multiplication in R_q efficiently. NTT and NWC are two special forms of the Fast Fourier Transform (FFT), known as the asymptotically fastest algorithm for computing polynomial multiplication.

In [10], Doröz et al. propose an implementation of the modular polynomial multiplication computed with the NTT algorithm and a Barrett reducer. A pre-computation based on the Chinese Remainder Theorem (CRT) is performed on input polynomials to reduce the size of coefficients. The overall architecture is based on an array of units, which gives some flexibility to process several residue polynomials in parallel. They evaluate their architecture on polynomials of fixed degree $n = 2^{15}$, and fixed coefficients size $log_2(q) = 32$ bits. Their accelerator was dedicated for a specific homomorphic scheme. In [11], Chen et al. present an optimized design of the modular polynomial multiplication.



All computations are carried out in the FFT domain on polynomials with degree n ∈ [256, 2048] and coefficients size ∈[20, 29]. They provide a selection method for the parameter set supporting efficient modular reduction, meeting at the same time the security requirements for RLWE and Somewhat Homomorphic Encryption (SHE) schemes. SHE means that the maximum number of successive operations in the encrypted domain is limited. Though efficient, this selection leads to many restrictions on the polynomials supported by the design: polynomial modular reduction is in fact computed with respect to the common choice $f(x) = x^n + 1$. When FFT multiplication using NWC is employed, they show that the modular reduction $(x^n + 1)$ is eliminated; this elimination comes at the expense of pre- and post-computation steps. A hardware architecture for the modular polynomial multiplication is described in [13]. They provide a fast unit for polynomial operations using CRT and NTT for multiplication combined with an optimized memory access scheme and a Barrett reduction method. The implemented unit can be used to instantiate a specific encryption scheme. Results are provided for $n = 32768$ and $log_2(q) = 1228$ bits. The authors of [14] use the Karatsuba algorithm to implement the modular polynomial multiplication in hardware. They demonstrate that for various degrees and coefficient sizes, Karatsuba can be a good alternative to FFT. Lastly, Jayet-Griffon et al. [12] consider the polynomial multiplication of 512-degree polynomials with 32-bit coefficients. They analyze and compare three algorithms (Karatsuba, FFT, and Schoolbook): the Schoolbook method is shown as the most efficient for a hardware implementation, due to its simple and regular structure. Modular reductions were not covered in their work.

All the aforementioned related works reported significant speed up factors when compared with software implementations. This speed up illustrates that further research into hardware implementations could greatly improve the performance of FHE schemes. However, they are almost all designed for fixed length operands and optimized for one specific type of multiplication algorithm (mainly FFT-based algorithms) which puts restrictions on the parameters selection. This makes them limited to some specific schemes and application domains. Besides, the operand

sizes are not very large, which has a direct impact on the security level of the cryptosystem. When modular polynomial reductions are performed, the simplest (and more limiting) choice is often selected.

Consequently, existing designs can be considered as proof-of-concept implementations only suitable for homomorphic encryption with small multiplicative depth circuits and low computation complexity. In the homomorphic context and because of its rapid growth evolution, new designs with a wide range of parameters are needed. These parameters are very large, consume large amounts of memory, and require many resources in order to perform efficient computations. Thus, memory storage and available resources in a target device should be taken into consideration, especially when manufacturing a specific circuit is not affordable and FPGA (Field Programmable Gate Array) platforms are therefore intended.

4. Software/Hardware design description

4.1. General presentation

We propose a hybrid and flexible SW/HW design based on a generic polynomial multiplier. This design is basically constructed from two parts: a dedicated hardware accelerator, and the software running on a general-purpose processor. Our solution aims at improving the overall performance and supporting much larger parameter sets than previous designs while optimizing resources for a given computation performance level. As in most previously published approaches, we will demonstrate our solution on FPGA-based implementations. However, the same approach may also be used with pre-characterized libraries to generate an application-specific integrated circuit (ASIC). The general-purpose processor can be implemented on-chip, or the hardware part of the accelerator can be connected to a computer. The interface between hardware and software can thus be performed, for example, through a high performance AXI (Advanced eXtensible Interface) bus, when software is running on an embedded processor, or a PCI (Peripheral Component Interconnect) express bus in the case of a computer processor. The choice should take into account the performance targeted for the global design, the communication and post-processing overheads, as well as the implementation constraints.

4.2. Design configurability

Our accelerator has been designed to support polynomials of any degree and any coefficients size. This goal has been met thanks to four hierarchy levels. As shown in Figure 6, level 1 and level 3 compute the product of two polynomials with large degree, while levels 2 and 4 deal with the product of coefficients with large sizes. This approach allows us to design efficient implementations for the lower (smaller) blocks, and configurable algorithms for the upper (larger) ones.

The designer starts by defining the input parameters (degree, coefficients size and irreducible polynomial used for the modular reduction).

In order to compute the multiplication of large input polynomials, we first represent the inputs as sets of polynomials of smaller fixed degree K. Then, we compute the pairwise products of each pair of sub-polynomials, using a hardware block for the multiplication of polynomials of fixed degree K based on the Schoolbook algorithm. This algorithm is not the fastest from a pure algorithmic point of view (asymptotic complexity), but it has three main advantages: it is more efficiently implemented on FPGA targets [12], it does not require pre- or post-processing and it does not impose any specific constraints thus allowing the possibility of optimizations such as batching (see section 5.3). An additional reason for choosing this algorithm is that the proposed decomposition limits the degree of the polynomials at level 3, so the degrees required to take full advantage of the asymptotic complexity of other algorithms is not reached.

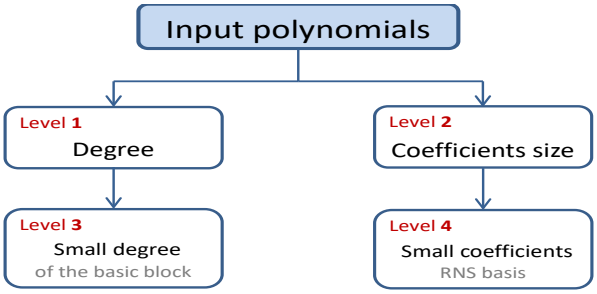


Figure 6. Multi-level design

The product of large coefficients is calculated through a RNS (Residue Number System) approach [15]. The advantage of RNS is that computations can be performed in parallel, that can result in a significant speed-up. We convert each large coefficient into a set of several values of smaller size by applying the RNS transformation. This technique can be easily made (almost) independent of the coefficient size: in order to support larger coefficients, it is sufficient to add a new element to the existing RNS basis without need to change the underlying architecture. A more detailed description of our approach is discussed in [16].

Let us mention that for small degrees and/or small coefficients sizes, level 1 and/or level 2 are optional (see Figure 6), as the algorithm can operate directly on full size operands.

4.3. SW/HW Partitioning

Our hybrid design implements part of the computations in software and part of them in hardware. The adopted partitioning is summarized in Figure 7. On the software-level, we consider simple and cheap computations such as input representations in subset polynomials, RNS basis generation, and polynomial additions; additionally, polynomial modular reduction is also computed on the main processor, in order to take advantage of the flexibility of a software implementation.

The hardware part implements the product of large coefficients in the RNS domain, and the Schoolbook multiplication algorithm at the polynomial level.

In order to reach our goals in terms of flexibility (size of parameters, but also implementation target), the hardware blocks are generated with a High Level Synthesis (HLS) tool.

4.4. High Level Synthesis

High Level Synthesis aims at transforming a generic input algorithm into a Register Transfer Level (RTL) architecture for a given target technology. This allows obtaining better productivity

when compared to classical implementation approaches using direct RTL design in languages such as VHDL or Verilog, as the designer can work on a higher level language which is much easier to maintain.

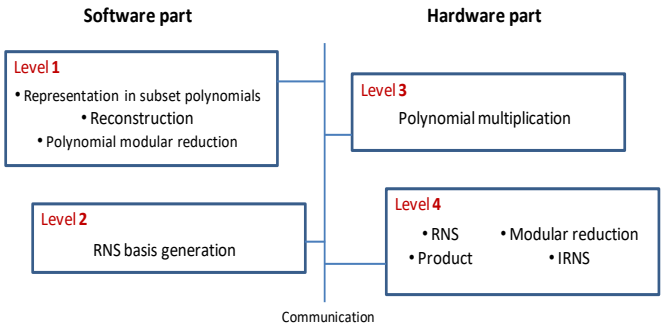


Figure 7. Modular polynomial multiplier: SW/HW partitioning overview

Several HLS tools exist; in this work, we use AUGH [17] since it is open, it may support different targets through the corresponding tool chain, and it can provide early estimations about the performance of the final result that can guide the designer in rapid refinement loops. AUGH is an autonomous HLS tool: it generates RTL descriptions quickly, under only global resource and frequency constraints [18]. This is achieved by performing incremental transformations of the input design description. The small complexity of the design space exploration (DSE) algorithm and the efficient use of all internal circuit structure constraints make this HLS tool very fast and able to generate pertinent solutions.

In this work, we will target two different FPGA technologies and a SoC (System on Chip) from Xilinx, but the methodology can be applied to other targets as well, provided that the corresponding flow is available. In our context, specifying new polynomials just implies to modify the input algorithms described in C language and let the HLS tool produce the RTL descriptions for every new specification (see Figure 8). Changing other parameters is fast and simple as well, as the designer can modify directly the high level description of the algorithm. Similar interventions on a RTL description of the design would take much more time and would be much more prone to errors due to a more complex description code.

AUGH provides different techniques for design optimizations, including unrolling and pipelining for the loops, wiring for the branch conditions and using maximum operator sharing [18]. Each RTL generation is analyzed by the user who can then command the HLS tool (with these directives) to converge towards a better solution in the next trial. DSE process detects possible transformations of the design that bring more parallelism, and applies these transformations until the user resource constraints are reached. Then, the RTL design description is generated.

5. Implementations, comparisons and discussions

5.1. Comparison with the state of the art

In order to provide a fair comparison of our results with the state of the art, we configure our design with parameters as close as possible to [10], [11] and [12] and perform the synthesis on similar targets.

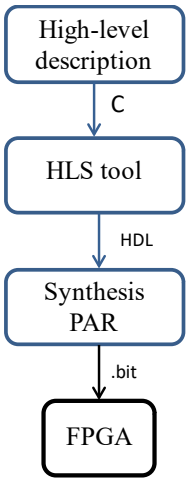


Figure 8. HLS based approach: tool chain

We set parameters for multiplication of polynomials of degree 512, 1024 and 32768 and with coefficients sizes of 26 and 32 bits. Under these configurations and for the smallest degrees, we do not divide inputs into smaller polynomials and we feed our basic block of level 3 directly with the corresponding polynomials.

Table 1 demonstrates that our approach, though about twice slower compared to manually optimized designs, consumes much fewer resources than the two designs reported in [11]. We obtain a reduction by a factor of 9 on average. This significant difference can be explained by the choice of the polynomial multiplication algorithm. For the FTT- and NTT-based algorithms (as in [11]), the pre- and post-computation steps are complex and require storing additional parameters. For the Schoolbook method, no pre- and post-computations are needed and we only have to store the input and output coefficients. Thus, FTT and NTT implementations have lower latency than our accelerator thanks to their lower complexity and their hand-crafted design but have several constraints and require a large amount of hardware resources.

For comparison with [12], where the authors implement the Schoolbook algorithm, we present two solutions in Table 2. In order to exploit parallelism and to achieve the minimal latency, we apply an optimization directive of AUGH not allowing operator sharing. Doing so, we instruct the HLS tool to use as many DSPs as possible: in this case only LUTs are used to store the coefficients and a maximum number of DSPs can be easily parallelized (first design in Table 2). In the other solution (second design in Table 2) BRAMs are used to store information, which limits the number of DSPs that can be used efficiently. It must be stressed, however, that our results are after placement and routing, while [12] only gives results after synthesis. Depending on the choice made with the HLS tool, the latency may be only slightly augmented (+28%) compared with the noticeable gain in resources (more than 35x). Resources can be further reduced up to a factor of 41 with the second solution, but leading to a loss of performance by a factor of 12.6.

For polynomials with degree 32768 used in [10], we divide the inputs into polynomials of degree 8192 and we apply our

approach first without any optimization. In this case (first design in Table 3), the hardware resources required by our design are nearly 28 times less than [10] on average; on the other hand, our design has a latency of 41 ms, while the authors of [10] report a latency of about 9.5 ms. Since our accelerator requires relatively few hardware resources, we can compute in parallel on 4 instances: with this optimization, the second design in Table 3 has a latency comparable to [10], while still maintaining a large advantage in resources.

Globally, these examples of results for our designs show that they consume fewer resources than the state of the art thanks to the proposed approach, while achieved performances can be close to hand-made designs. Using fewer resources also leads to possible parallelization of several instances while meeting the resource constraints of a given FPGA; in that case, even better performances can be reached. The approach offers at the same time a large degree of configurability, allowing the designer to change several important parameters on-the-fly. In addition, our approach is highly flexible, since the same generic high-level description can be used to produce a new circuit with different area/performance trade-offs i.e., we can generate a very cheap (but slow) polynomial multiplier, or a fast but more resource consuming one. Other possible optimizations can be applied, as it will be shown in the next sections.

5.1. Design space exploration

With our flexible design, we can instantiate our architecture with different coefficients size and different degrees, targeting several platforms. Larger parameters imply large values of n and q that grant high security levels and significant multiplicative depths.

When we handle large coefficients, our approach proposes to transform them into their respective RNS representation. The choice of the RNS basis (size and co-prime modulus) is primarily based on the coefficients size of the input polynomials and the available resources on the target device. The key idea is to take advantage of the parallelism offered by the RNS representation and speed up our computations. If we ever need to increase the supported size, it is sufficient to extend the RNS basis without need to change the underlying architecture.

The RNS basis size and modulus are kept as parameters as well as the degree K of the basic block performing the Schoolbook multiplication. Let us now fix for example the coefficients size at 64 bits and vary the degree from 8192 to 32768 to cover a new range of parameters. For each degree, we choose a different degree K of the basic block. When the resources of our target are sufficient, we perform computations in parallel on several instances.

Figure 9 shows that some configurations are more efficient than others. In fact, when computations are running on P instances in parallel, we roughly multiply the resources by P and divide the latency by P (illustrated by the cases of n=8192 and n=32768, figure 9). But, setting for example K=1024 to calculate the multiplication of polynomials with a degree 16384 is not an appropriate configuration. The performance loss is essentially due to 16^2 calls to the basic block *Polynomial_K*. When changing K and/or applying an optimization, the latency is also affected by the cost of the data transfers and therefore some configurations are not suitable.

Table 1. Comparison on Spartan 6 with polynomials of degree 1024 and 26-bit coefficients

References	**Resources**				
	Slice LUT	Slice Register	DSP	BRAM	Latency (μs)
[11](1)	10801	3176	0	0	40.98
[11](2)	2464	915	16	14	32.28
Our work	182	114	3	10	69.1

(1) Multipliers are built by pure LUTs (2) Use DSPs and Brams

Table 2.Comparison on Virtex 7 with polynomials of degree 512 and 32-bit coefficients

References	**Resources**				
	Slice LUT	Slice Register	DSP	BRAM	Latency (μs)
[12]*	252341	130826	512	2048	4.11
Our work (1)	7032	920	368	0	5.27
Our work (2)	171	102	3	3	66.41

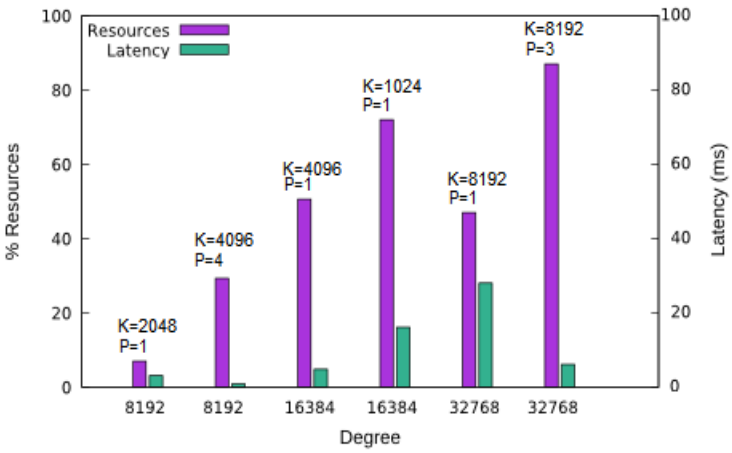
(1) Data is stored in Lutrams (2) Data is stored in Brams *Synthesis results

Table 3.Comparison on Virtex7 with polynomials of degree 32786 and 32-bit coefficients

References	**Resources**				
	Slice LUT	Slice Register	DSP	BRAM	Latency (μs)
[10]	219192	90789	139	768	9.51×10^3
Our work (1)	3392	1920	48	792	41.12×10^3
Our work (2)	13568	7680	192	792	10.28×10^3

(1) Our approach without optimizations (2) Several instances of the basic block running in parallel

Figure 9. Performance of our designs on polynomials with 64-bit coefficients and



different configurations (degree K, parallelization, degree P)

Several multipliers with different area/performance ratios can be generated by modifying the RNS basis and the degree K of the basic block, and also by the parallelization of the computations on many instances of the basic block. Thanks to the HLS design flow, rapid feedback on circuit characteristics is used to evaluate deep architectural changes in short time and pick up the more suitable parameter sets. Then, the designer can select the design satisfying his constraints among the set of generated circuits. The development timeline of a new solution is about 3 hours on average, which is a very short time in comparison with a hand-made design.

5.2. Our system performances

In this section, we evaluate the performance of our hybrid modular polynomial multiplier with respect to the communication times between hardware and software. We run computations on polynomials of degree 512 with 32-bit coefficients and we pick two choices for the modular polynomial:$f(x) = x^{512} + 1$, and a general cyclotomic polynomial of degree 512. This choice is motivated by the possibility, allowed by a generic polynomial, of applying the batching optimization [19].This technique can be used in order to pack multiple messages into one single ciphertext, therefore allowing parallel homomorphic evaluations. Hence, it permits great versatility in the computations and improves performance. This technique is based on the CRT theorem and requires that the polynomial f(x) is different from $x^n + 1$.

To transfer data, we consider an AXI interconnection between the hardware and software part. Our platform is the Zybo Zynq-7000. In our case, we send to the FPGA two input polynomials of 32*512 bits each and we receive one polynomial of 73*1024 bits. We use one high performance AXI bus to send the input coefficients and two high performance AXI buses to receive the output polynomials: this is due to the fact that the polynomial reduction is made in software, and the multiplication result is thus (almost) twice larger than the operands. With this solution, we can enhance the performance of our system and have the final result faster: a speed up factor of about 2.98 is obtained compared with a solution with only one bus, as the hardware acceleration is mitigated by the data transfers.

Our architecture reports a global latency, including transfers, of 0.75 ms when $f(x) = x^{512} + 1$ and of 1.89 ms when f(x) has a general form, which is still very fast compared to pure software implementations.

When compared with the state of the art, we have shown that our implementations consume few resources but report smaller speed-up factors. To counter this, we have proposed to run computations on many instances of the basic block in parallel. Figure 10 illustrates an optimized proposition when performing computations on 4 instances of blocks with degree 512 in parallel.

5.3. System evaluation on larger parameters

In this section, we evaluate our approach on designs with larger parameters. For each configuration, we choose two forms of the irreducible polynomial f(x). The first one corresponds to the popular choice $x^n + 1$. The second one has a general form and allows optimizations in the homomorphic context. Tables 4 and 5 provide implementation results and comparisons with software implementations we developed with Sage and run on a Intel Core i5-2450M (2.5 GHz). We decided to make such a comparison in order to get the same parameters configuration of parameters and because published works do not cover such parameters. Our reference embedded platform (the Zybo board) has processing power and memory that are not even comparable to desktop or server CPUs. Nonetheless, and withstanding the overhead of the data transfers through the AXI bus, the results obtained are quite interesting.

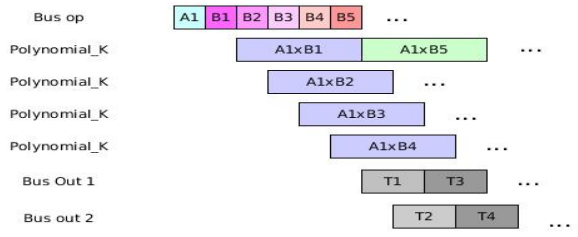


Figure 10. Parallelization of the computations on many instances of the basic block *Polynomial_K* [1].

Table 4. Implementation results (ms) when $f(x) = x^n + 1$

n	q	Software implementation	Our design
4096	32	32.6	2.1
8192	64	76.7	5.8
16384	128	162.8	11.7

Table 5. Implementation results (ms) when f(x) has a general form

n	q	Software implementation	Our design
4096	32	41.8	4.1
8192	64	88.5	12.7
16384	128	202.9	21.8

Results show that our approach reports significant speed-up factors when compared with pure software implementations, in spite of the large parameter values. The latencies when the

polynomial reduction is computed modulo $x^n + 1$ are better than the reduction modulo a polynomial with general form. This is due to the complex form of the polynomial in the second case (only 2 coefficients vs. (n+1) coefficients) and to the number of iterations required to get the final results. For a fair comparison, it must be stated that such a general form is not covered by most works in the literature, especially when hardware implementations are considered.

Several software implementations and libraries dedicated to homomorphic cryptography exist today [20], **Error! Reference source not found.**[21], which achieve very interesting performance when executing on high-end processors. These implementations are usually based on the NTT-based algorithm in order to speed up the modular multiplication on very large operands. This algorithm, however, has very large requirements in terms of memory usage, it constrains the choice of parameters, and it is usually optimized by exploiting the advanced instruction set available on modern processors, such as SSE and AVX. When targeting limited devices and/or current client-server frameworks, those implementations cannot be used directly because of the memory cost, or do not perform well enough due to missing advanced instructions. For this reason, a comparison with these works is difficult, since different application domains and platforms are targeted.

6. Conclusion and future works

We present a flexible and generic SW/HW co-design for the modular polynomial multiplication, the most computationally intensive operation in homomorphic cryptosystems based on the RLWE problem. Given the large parameters required in such schemes, we propose an RNS based algorithm and an efficient decomposition of the large input polynomials. Our design can be easily configured thanks to a HLS approach and sets no restrictions on the parameters that define the RLWE problem leading to high security levels and large multiplicative depths when necessary. Our approach can also optimize the accelerator for applications requiring small parameters. Our architecture can be instantiated to accelerate any RLWE-based scheme; additionally, even if the proposed methodology has been illustrated only on the polynomial multiplication, it can be used to implement and accelerate other primitives required by homomorphic schemes.

Future works include more complex communication schemes, such as using two AXI High Performance input ports and thus increase the number of instances performing in parallel. As the memory is one major bottleneck, we can reduce the amount of memory by controlling and scheduling the loading of sub-polynomials $A_i(x)$ and $B_j(x)$. It may also be interesting to evaluate the benefits of a hardware implementation of the polynomial reduction. Other primitives used in the FHE schemes may also be implemented in hardware using the same methodology. Finally, these accelerators will be integrated to evaluate a full homomorphic scheme on an FPGA/processor platform targeting embedded applications, and the performance will be compared to state-of-art software libraries ported to the same environment.

References

[1] A. Mkhinini, P. Maistri, R. Leveugle and R. Tourki, HLS design of a hardware accelerator for Homomorphic Encryption, IEEE 20th International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS'17), pp. 178-183, Dresden, Allemagne, 19 au 21 avril 2017.

[2] C. Gentry, A fully homomorphic encryption scheme, Ph.D. dissertation Stanford University, 2009.

[3] Joppe W. Bos, Kristin Lauter, Jake Loftus, and Michael Naehrig. Improved security for a ring-based fully homomorphic encryption scheme. In Martijn Stam, editor, Cryptography and Coding, volume 8308 of LNCS, pages 45–64. Springer Berlin Heidelberg, 2013.

[4] Junfeng Fan and Frederik Vercauteren. Somewhat practical fully homomorphic encryption. Cryptology ePrint Archive, Report 2012/144, 2012. http://eprint.iacr.org/.

[5] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (leveled) fully homomorphic encryption without bootstrapping. In ITCS 2012, pages 309–325, Cambridge, Massachusetts, 2012. ACM.

[6] Tancrède Lepoint and Michael Naehrig. A comparison of the homomorphic encryption schemes FV and YASHE. In Progress in Cryptology - AFRICACRYPT 2014,7th International Conference on Cryptology in Africa, Marrakesh, Morocco, May 28-30, 2014. Proceedings, pages 318-335, 2014.

[7] Zvika Brakerski. Fully homomorphic encryption without modulus switching from classical GapSVP. In Advances in Cryptology - Crypto 2012, volume 7417 of LNCS, pages 868–886. Springer, 2012.

[8] Shai Halevi and Victor Shoup. Algorithms in helib. In Advances in Cryptology-CRYPTO 2014 - 34th Annual Cryptology Conference, Santa Barbara, CA, USA, August 17-21, 2014, Proceedings, Part I, pages 554-571, 2014.

[9] Kim Laine and Rachel Player. Simple encrypted arithmetic library - seal (v2.0). Technical report, Microsoft Research, September 2016.

[10] Y. Doröz, E. Öztürk, E. Savas, B. Sunar , "Accelerating LTV Based Homomorphic Encryption in Reconfigurable Hardware",CHES '15, 185-204.

[11] D. D. Chen, N. Mentens, F. Vercauteren, S. S. Roy, R. C. C. Cheung, D. Pao, and I. Verbauwhede, "High-speed polynomial multiplication architecture for ring-lwe and SHE cryptosystems," IEEE Trans. on Circuits and Systems, vol. 62-I, no. 1, pp. 157–166, 2015.

[12] Cedric Jayet-Griffon, M.-A. Cornelie, P. Maistri, Ph. Elbaz-Vincent and R. Leveugle. Polynomial Multipliers for Fully Homomorphic Encryption on FPGA. In 2015 International Conference on ReConFigurable Computing and FPGAs (ReConFig), Mexico City, 2015, pages 1-6.

[13] S. Sinha Roy, K. Järvinen, F. Vercauteren, V. Dimitrov, and I. Verbauwhede, Modular hardware architecture for somewhat homomorphic function evaluation, in Proc. of Cryptographic Hardware and Embedded Systems – CHES 2015. Springer, pp. 164–184.

[14] V. Migliore; M. Mendez Real, V. Lapotre, A. Tisserand, C. Fontaine, G. Gogniat. Hardware/Software co-Design of an Accelerator for FV Homomorphic Encryption Scheme using Karatsuba Algorithm. In IEEE Transactions on Computers, vol.PP, no.99, pp.1-1.

[15] H. L. Garner, The residue number system, IRE Transactions on Electronic Computers, vol. EC-8, no. 2, pp. 140–147, Jun. 1959.

[16] A. Mkhinini, P. Maistri, R. Leveugle, R.Tourki and M. Machhout, A Flexible RNS based Large Polynomial Multiplier for Fully Homomorphic Encryption, 11th IEEE International Design & Test Symposium (IDT), Hammamet, Tunisia, December 18-20, 2016, pp. 131-136.

[17] A. Prost-Boucle, "Augh project" 2016, [Online]. Available: http://tima.imag.fr/sls/research-projects/augh/

[18] A. Prost-Boucle, O. Muller, and F. Rousseau, Fast and standalone design space exploration for high-level synthesis under resource constraints, Journal of Systems Architecture, vol. 60, n. 1, 79–93, 2014.

[19] N. P. Smart, F. Vercauteren, Fully homomorphic SIMD operations, Des. Codes Cryptography 71(1): 57-81 (2014).

[20] Seal : Simple Encrypted Arithmetic Library, https://sealcrypto.codeplex.com/

[21] Helib library, https://github.com/shaih/HElib

[22] NFLlib library, https://github.com/CryptoExperts/FV-NFLlib

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 434-442 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Smart Mobile Application for Assisting Parents in Anti-Drug Support

Tsz Hei Yeung, Chi Kit Ng, Vincent Ng*

Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Article history:
Received: 01 December, 2017
Accepted: 15 January, 2018
Online: 10 February, 2018

Keywords :
Drug Abuse
Mobile Application
Location Check-in
Behaviour analysis
Instant Messaging

ABSTRACT

Drug abuse problem is not a new problem, but it is still serious in Hong Kong. The emergence of hidden drug abuse youth has brought more difficulties to them and their parents. The usual ways of providing anti-drug abuse assistance are not sufficient to these types of young people. Intelligent methods have been developed to assist anti-drug workers. In this paper, we present three mobile applications with a set of intelligent tools for parents and social workers to help our young people. The applications have an alerting tool to detect the possibility of drug abuse of a young student, and to alert corresponding parents and social workers for remedial actions. Another function is the nearby notification tool to help social workers in identifying people who may need assistance. For the alerting function, a number of data mining and text analysis techniques have been adopted for behavior analysis. A supporting instant messaging has been co-developed for communicating alerting messages.

1. Introduction

Drug abuse problem is not a new problem, but it is still a big problem in Hong Kong Society. Even there is a tendency of decreasing in drug abusing, the government report indicated that the age of first drug taking is decreasing. This shows that the problem is in serious with depth (age) but not the width(quantity). Besides, from the perspective of teenagers, they do not have much knowledge on how to get the treatment service from different organizations.

Furthermore, Hong Kong has another problem which is the hidden youths or "socially withdrawal". The term "hidden youths" is originated from Japanese word "Hikikomori" and it means that a person is scheduled at home by at least six months [1]. There is one more level under the "hidden youths" which is "Otaku" in terms of Japanese. There are different characteristics for "Otaku", but the commonly use is that "Otaku" always stay at home for activities and avoid social activities. A study found that these people are satisfied with their quality of life [2]. Also, it is difficult for other people to interact with them normally. And once they are drug abuser, they become the riskiest group of people as people cannot discover them easily and help them return to life. In Hong Kong, it has been reported that, since 2010, 80% young drug abusers take drugs at home.

*Vincent Ng, Department of Computing, The Hong Kong Polytechnic University, 852-2766-7242, Email: cstyng@comp.polyu.edu.hk

According to Narcotics Division, the median drug abuse history of newly reported abusers (i.e. the time for abusers to be reported to the CRDA by reporting agencies from their first drug abuse) remained between 1.7 and 2.1 years during 2007 to 2009. It then increased continuously to 5.9 years in 2015, and dropped to 4.6 years in 2016 [3]. One reason is because psychotropic substance abuse is more popular among young drug abusers [4]. Unlike traditional drugs such as heroin, psychotropic substance abuse is more "hidden" in nature. Young drug abusers tend to take psychotropic substance at home. It is hard for us to bring the young drug abusers to light.

Smartphones have become popular among young people. There are close to 5 million Hong Kong persons aged 10 and over had smartphones in 2014 and the smartphone penetration rate has exceeded 77.2% [5]. Developing anti-drug mobile applications can be an effective way to pass healthy information to young people as well as parents and social workers. Hence, in 2014, we submitted a proposal to the Beat Drugs Fund in HKSAR to develop a series of mobile applications for students, parents and social workers with a backend platform to support information sharing.

In our design, we aim to find out the hidden or potential drug abusers, and to warn their family and anti-drug workers about the problem. Another important function is to support the nearby function for social workers to access young people. For the rest of the paper, Section 2 provides the background review. Next, in Section 3 and 4, we describe our application design and

implementation approach. Section 5 concludes our present work in this paper.

2. Background Review

2.1. Anti-Drugs and Behavior Analysis Mobile Apps

Dealing with the drug abuse problem, different parties in Hong Kong society have developed different mobile applications to link with the youths. There are two mobile applications recently developed in HK and they are “Anti-Drug Information Center” and “毒癮無可忍” to deal with the drug abuse problem. The “Anti-drug Information Center” (禁毒資訊站) developed in Macau provides the drug information as shown in Figure 1 and can support parents as a simple drug abuse detection tool [6].



Figure 1. Home screen of the Anti-drug Information Center

This mobile application adopts an expert system approach which is based on the anti-drug resource kit designed by the T.W.G.Hs. CROSS Centre [7]. However, there were criticisms on the test because the behavior criteria could have different interpretations, not only for detecting drug abuse issues. For example, it simply identifies “emotionally unstable” and “performance mysterious contact with friends” as symptoms of drug abuse, but these symptoms can also be characterized adolescent.

For the ”毒癮無可忍” application, it was developed by a Christian organization. It also provides the drug related information and detection test. The information it provided are the real cases sharing, information of referral center and some encouragement message from the bible. The detection test is rule-based and has a number of questions as shown in Figure 2. That is, when your answer meets all criteria of detection, response will then show up. The responses are video formats so as to arise interest and attention of the young people.

Drugout is an iOS Apps aims to passing knowledge about anti-drug [8]. In the application, players will try to help a game avatar to fight with his drug abuse issues as shown in Figure 3. Score will be accumulated during the actions, and players can get bonus score if they invite others playing the application. Finally, there will be real prizes given to the high score players. Overall, it is an interesting game that can attract students to play and be able to deliver the anti-drug information to the players. The bonus scoring can encourage players to invite more people play the game and players can experience what a drug abuser will suffer from the drugs. Yet, it only provides information without any communication platform. Also, its use is for the youth and parent may not get help from this application.



Figure 2. Addiction testing



Figure 3. Dragout

Another similar mobile application is called Detoxification [9]. Besides anti-drug information, it also provides interesting game and situational questions which raise awareness of drug abuse as shown in Figure 4. Yet, the detoxification application has not been completed and it also lacks of resource links to support youths when they encounter a problem.

Another type of mobile applications is for behavioral analysis, such as the “Pro-sociality Behavior Test”. It contains a self-assessment to test pro-sociality behavior tendency. Its problem is having too many questions for an assessment and users often feel tired when doing it. There is a similar mobile application, “Depression Test”, for depression testing [10]. It provides a nice historical review and users can see the changes of their depression levels. The application provides no connection between users and social service organizations. Users cannot find an easy way to seek for help if they encounter a problem.

Although there are many mobile applications about behavior analysis in the app stores, most of them are not related to anti-drug.

This motivates us to develop a suitable solution with design objectives below:

- Provide an accurate prediction based on behavioral information provided by users.
- Able to detect potential drug abuse with a few guided questions with their answers.
- Able to provide connection and interaction channels between users and anti-drug social service organization.

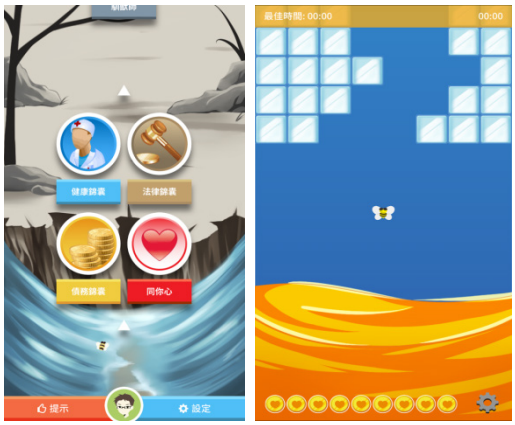


Figure 4. Detoxification

2.2. Alerting Systems

Alerting systems have been applied in different service domains. The Brigham Integrated Computing System (BICS) provided a clinical alerting system at Brigham and Women's Hospital (BWH) since June 1994 [11]. It used an event monitor to determine whether a new clinical data warrants an alert. The notification program would send an automated e-mail message messages to a duty physician, or through other means of communication. The physician could then log on any computer to review the alert and take immediate therapeutic action. The alerting system reduced the time it took for physicians to respond to critical laboratory results. Physicians thought that the communicated information was important and the service can help clinicians to be well informed of important abnormalities in the flood of data they have continually receiving.

Alerting system is also essential for providing the location and information about vehicles to passengers, owners or users [12]. Bus transportation system has been proposed to track a large number of buses simultaneously and detects the routes and directions automatically through GPS satellite [9]. It used a web-based, mobile communication and SMS platform for communicating bus arrival times and predicted the real-time inter-station speed.

Email alerting has been proposed with education, health care and other life applications. They can be used for two-way knowledge translation, which involves receiving case situation to professionals and then sending their evidence-based recommendations. In [13], it is reported that professionals were familiar with email communication. While many felt comfortable and liked using emails in their professional life, yet some others felt overwhelmed by the volume of emails they receive. The study also suggests that email alerting can contain valid and trustworthy information as an easy way to disseminate information to multiple recipients. This would give the reader an option to read emails immediately or later.

Instant messaging alerting is another frequent channel to support alerting messages. The major advantage of instant messaging alerting is real time or instantly delivered. When a receiver is online, the message will be received as soon as the user sends it. Also, instant messaging alerting tends to be confined to a single line and easy to read [14]. However, instant messaging systems are often of proprietary platforms and cross platform support is difficult. Communicating via instant messaging requires that both parties have accounts with the same instant messaging service.

With the above reviews, we have proposed to use multiple channels for sending our altering messages, both email and instant messaging. As social workers are familiar with both channels, this multi-channel platform can effectively work for their clients.

2.3. Location Sharing

Check-in provides location awareness support so that it can act as a channel to share, comment and rate any spots in different areas. It aims to establish a location-based social network (LBSN) for users. All users included students, parents, social workers can publicize their current locations to others. Through this check-in step, users can contact their friends nearby and chat with instant messaging. A special function for social workers is that they can be notified the youths' current locations if youths choose to open the information.

In Google Map, users can leave comments and rate places. The comments and ratings are public to everyone in the world. More than comments and rating, users can also upload and share photos about the spot. Sharing the location with user's friends through other applications like Facebook and Whatsapp, is available in Google Map [15]. A good example is an application developed by Apple Inc., FindMyFriends. It allows users to follow and track friends with their mobile device. Users can also choose to share their location with others. The app determines users' location using GPS in the device. If the user is requested by his friends to see their current location, users can contact their friends on the map immediately with the contacts in their device [16].

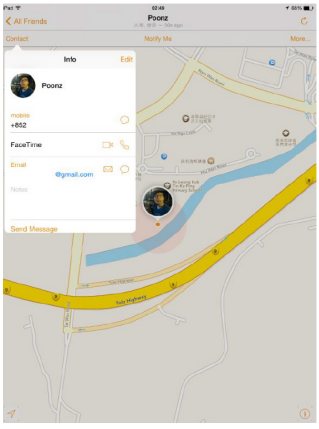


Figure 5. FindMyFriends

Another similar application is Swarm which is developed by Foursquare. It is a companion application of Foursquare. Users are allowed to share their locations within their social network. A check-in can be broadcasted to through other channels like Facebook or Twitter. User can easily see nearby friends, make future plans as well as checking in a location. Figure 6 shows a check-in page of Swarm.

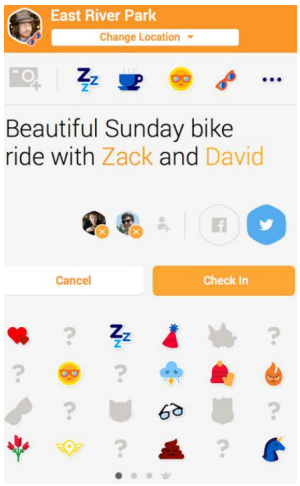


Figure 6. Swarm in Foursquare

Swarm has all the required functions for our objectives in the aspect of location sharing. From "check-in" and "check-out" to chat with friends. However, because our target users require a high level of confidentiality, only secured sharing platform is allowed.

3. Design Methodology

After the review the current problems in the previous sections, a set of mobile application is decided to develop to suit the needs from the different users. For parents, the mainly functions will help them on drug abuse detection of their kids and they can use the application to seek for professional help and build a community between other parents and social workers. For students, the application will try to collect their data that will be useful to the project team, also the application will help the youth have more connection with their parents, friends and social workers by instant messaging or share their location information. For social worker, they should able to use the application to upload the anti-drug information and give help to the youths and their parent. With the types of users, 3 different mobile applications have been designed and developed and they are called Aurora, Giai and Theoi, respectively.

In the rest of this section, the overall design and three functional components are presented. The first component is on the youth behavioral analysis for users to predict a potential drug abusing case. The second component is to support the alerting function when such case is detected. It can effectively work for their clients. The last is about the location sharing and near-by search support.

3.1. Overall Design

At the beginning of our work, teachers, principals, students, social workers and anti-drug experts have been consulted and surveyed on the suitable functions of the three applications. For the student mobile application, it includes the following functions:

a) Head News Center: An information receiving channels for youths to receive anti-drug information, such as events, activities. Some of the news can be in a form of alerting and pop-up such that youths would receive something interesting to consider. It is a way to encourage youths to get more positively involved in the community. An interest profile can be set by the young user for customized information.

b) Q&A Section: a resourceful function for youths to receive answers of common questions, such as where and how they can seek help.

c) Survey (game corner): it is useful to collect youth's current status from time to time. They can provide some of their own data on a voluntarily basis with a bonus point encouragement. The bonus point scheme can be supported by small souvenirs (produced by the project). The survey is in the form of a small game and its data collected can be made available to the social workers and the project team.

d) Searching: a searching function for allowing the youths to quickly identify the information from (a), (b) or (e).

e) Dashboard: a sharing space for users to post/upload their digital products (games, videos, art works, etc.) for sharing. After attending training workshops available from the project, these products can be shared and users can rate them. After uploading, other users can access these products.

f) Quick Contact: this function supports a user to contact a social worker for assistance quickly. The user can enter free text or answer a number of pre-set questions and a social worker would be alerted and try to arrange help.

g) Instant Messaging: with different groups (among youths, parents, and social workers) formed, members within each group can do online chat and share information.

h) Check-In/Location Information: the function is to support users in contacting people nearby and using it as a social media platform for attracting youths. It allows a user to release his/her current location information to other users. It is a powerful tool so that a user can know any other users in the same area. Users can provide comments of the check-in locations. It is a voluntary basis and the access of the location information is limited to the people that the user allows. Some initial designs have been done as show in the next page.

The second application, Giai which focuses on parental support, is developed for parents with the functions below. For similar functions in Aurora, the functional descriptions are skipped.

a) Head News Centre

b) Searching

c) Dashboard

d) Quick Contact

e) Instant Messaging

f) Friend List: a function which allows a user to add a friend or create a new user group (limited to some types of users).

g) InfoSeek: it is a searching function similar to the Location Information function in the Youth Apps

h) Family Center: this is the intelligent alerting tool. It enables parents to input their observations of their children on a regular basis. The input is semi-guided which can help parents to enter structured information with certain set of unusual behaviors. Once parents' observations are entered, the second function is to perform analysis with the knowledge base of an expert system available in the backend server. The expert system is designed and developed by the project team with the help of anti-drug experts and social workers. In case of parent users do not how to start, a series of prompts (similar to a quiz/survey) can be used to collect data. The analysis result would be in form of hints or advice to parents, suggesting where to seek help or acquire further information. This function can work with (f) above to ease the worries of parents.

i) Parental Corner: it is a special sharing space for parental groups and social workers to share information in addition to Dashboard and Head News Center. Parents/social workers can post articles or provide comments/advice of any articles. This corner can access special cases, emergency procedures and FAQs that are stored in a knowledge repository developed by antidrug experts and/or social workers. A searching feature similar to (b) is available also. This function is aimed to encourage parents to form mutual supporting groups for helping each other.

j) Referral Center: the function is acting a help-contact service which allows a parent to have an initial contact with a social worker. This is a resourceful function for youths to receive answers of common questions, such as where and how they can seek help.

The third application, Theoi, is developed for social workers with the functions below. For similar functions in the other applications, the functional descriptions are skipped.

a) Head News Center

b) Searching

c) Dashboard

d) Instant Messaging

e) Friend List

f) InfoSeek

g) P-Action+: Parents are able to request help from social workers with the altering tool, an email with test records will be generated automatically and sent through their own email accounts which they used in registration.

h) Y-Action+: It is similar to P-Action+ but the corresponding contacts are youths.

i) Enrich+: According to the feedbacks from social workers, a web base platform is developed for this function, instead of included in the apps. It allows social workers to post and share information, and update the FAQ database.

Figure 7 shows the software architecture of the system in which different modules are labeled with different border colors. Modules in the red rectangles mean these are shared by all 3 applications. Modules in blue, green and purple rectangle represent the functions only serve students, parents and social workers respectively.

3.2. Behavior Analysis

For the behavior analysis component, there are 4 core modules as the training, classifying, calculating and collecting. In training module, we have chosen 2 techniques, decision tree (see Figure 9) and logistic regression, to build a classifier with the given training data. The calculating module is to estimate an index for possible explanation of the recommendation. The collecting module will receive user's feedback and store them for the enhancement of our analysis (i.e. re-training with additional data). The 4 modules will work together to analyze behavior of youths as shown in Figure 7.

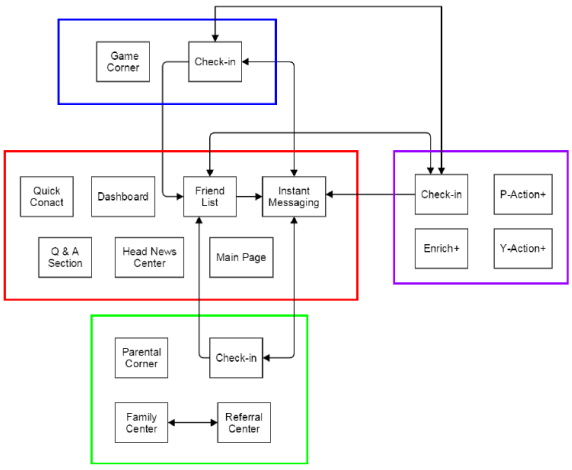


Figure 7. System modular design

The training module will run every time when the additional data is available. It will retrieve the training data and then apply a set of weight behavior criteria with the construction of a decision tree and the execution of logistic regression algorithm independently. The training results will pass to the classifying module and the calculating module afterwards. As the current available is small, the overhead of re-constructing the prediction model is acceptable. Yet, when the dataset size becomes larger, an incremental technique should be considered.

The classifying module is going to utilize the decision tree. A user is required to complete a set of guided questions which represent the behaviors of a potential drug abuser. Then the survey answers will be input to the decision tree. If it is classified as a non-drug abuser, it will bypass the calculating module and return the index as 0. Otherwise, it will go on the calculating module.

The calculating module is going to use the results of logistic regression. It is used for selecting a explanation description by estimating an index, which refers to the probability of drug abuse.

The determined class and index will return to the user. The collecting module is an optional module. A user can return expected status of the predicted drug abuse case to the server. This user feedback and the results of the previous survey will be added to the training data for later prediction improvement.

Input :
X = Training set,
A = each behavioral criteria scale for each training data set from database (training data)
Y = class for each training data set from database (training data)
B = each behaviora criterial scale from a user (according to survey)

Output: Decision tree

// Stage 1 – Training function (construct tree)
Create a new tree T with a single root node.
if stoppingCriterion(X) then
    Mark T as a leaf with the most common value in X as a label.
else
    Find a that obtain the best splittingCriterion(ai,X; y).
    Label n with a.
    for each outcome vi of a do
        Set subtree ti = treeGrowing(a=viX; A; y).
        Connect the root node of nt to the subtree ti with an edge that is labelled as vi.
    return T

Formula for spliting:

$$Q_{left}(\theta) = (x, y)|x_j <= t_m$$
$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta)$$
$$G(Q, \theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$$
$$\theta^* = \operatorname{argmin}_\theta G(Q, \theta)$$

// Stage 2 – Classifying function
    SearchTree(B, T)

Figure 8. Decision Algorithm

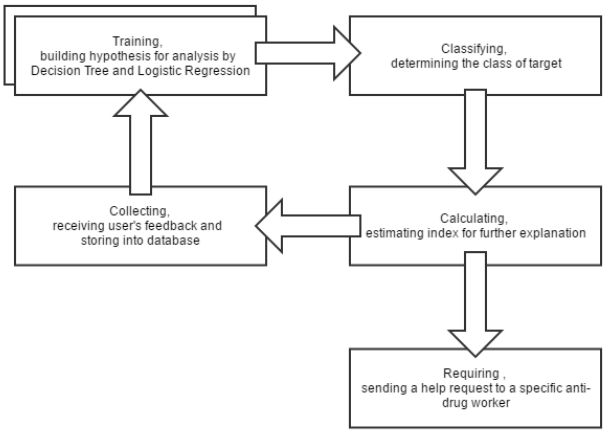


Figure 9. Behavior Analysis Framework

3.3. Altering

For the alerting component, there are 3 modules as confirmation, generation, and sending (as shown in Figure 10). In the confirmation module, we collect the information on the results of behavior analysis, user's demographic data, and the social worker whom users have chosen. Information would then be passed to the generation module. The generation module is used for generating email and instant message. They are generated by two servers in parallel. After that, the sending module will send the messages out.

3.1. Location Sharing and NearBy Support

Check-in support is a major component of our applications. Through the establishment of a healthy location-based social network (LBSN), it does not only share location-embedded information to an existing social network, but also create a new social structure by connecting individuals. With the location information, a person-to-person communication function can be enable, which is called the Friends Nearby Notification in our applications.

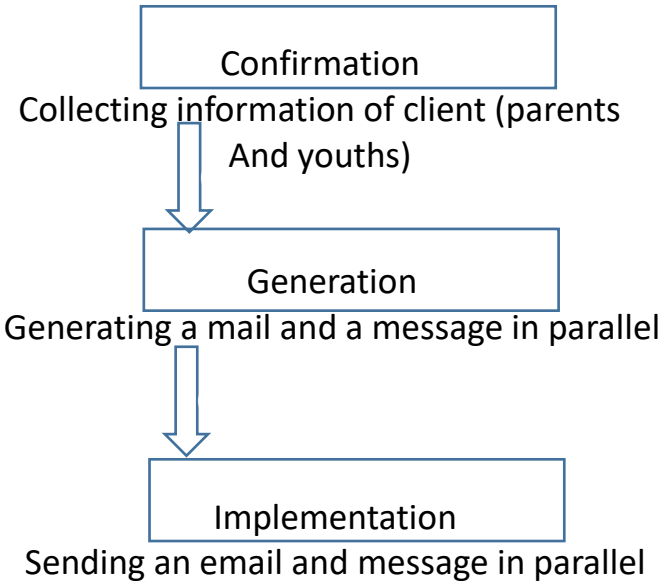


Figure 10. Alerting component

When a user would like to view others' comments to his current location, he can get them from a safe and secured database. Users can chat with their friends or other user after reading their comments. They can be redirected to Instant Messaging engine. In order to find their friends for having a face-to-face talk, the find route function is provided. The default origin will be the user's current location, and after user selecting a destination, data will be sent to the server. A string in JSON format will be returned and transformed into a polyline shown on the map which represents the route between two places. In our design, the Friends Nearby helps users to identify all nearby friends and accessibilities with the following features:

- Notify users when friends are in proximity
- Check nearby friends or accessibilities in regular time intervals
- Provide setting to turn off notification to support privacy
- Provide setting to hide from a particular friend.
- Show the nearby friends and accessibilities details, such as distance away from it, and contact info.

- Call or send short messages to nearby friends.

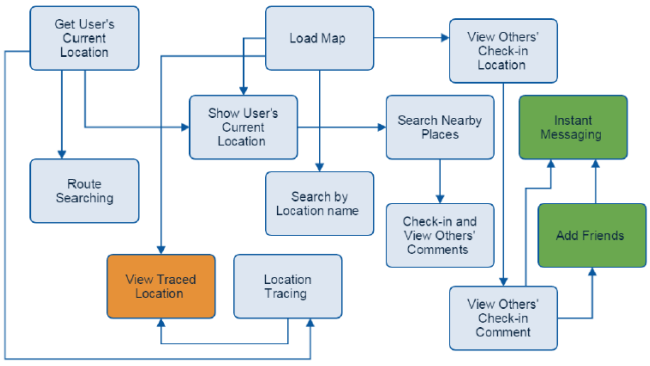


Figure 11. Check-in function

Figure 11 shows how the check-in function is designed and the starting module is the top left box in the figure, 'Get User Current Location.' The boxes in green are the functions that linked with other parts of the platform while the orange box is the function only available for social workers to trace locations of young people, provided with their permission. The locations of these young people will be uploaded to a central database for every 10 minutes. Social workers can view the students' location on the map by retrieving location information from the database.

4. Implementation

The mobile applications and its backend support are designed as a 3-tier client server model as shown in Figure 12. When a user uses a smartphone to access the system, requests will be sent to middle tier, i.e. web server, through HTTP request first. Then, the web server will translate the request into SQL queries and sent it to a database server to retrieve data. The return data will be transformed into JSON response and sent to the client device. All functions are based on different modules as shown in Figure 12.

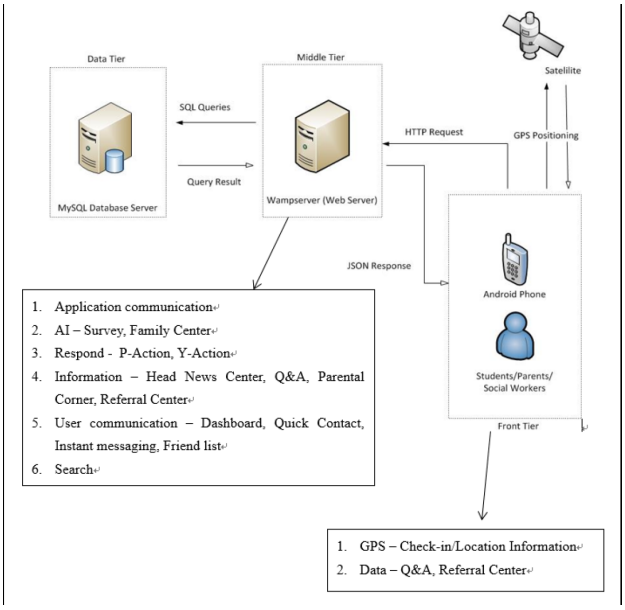


Figure12. Three tier implementation

In the current design, check-in requires only 2 relational tables. The first one stores basic user information, latitude and longitude, which are automatically uploaded by the application. The later one stores all check-in records. We do not archive all user location information since the amount of data is too large to be stored locally. For finding out a user's location, there are two common methods GPS and A-GPS. After balancing the advantages and disadvantages, we are using GPS to locate a user if it is available. If not, A-GPS will be applied. Figure 13 shows a location detection of a user.

Without GPS assistance, a user can check-in manually by entering his nearby location. Google Places API Web Service is used here since it can save device's resources. Several parameters including location, radius and API Access Key are required to retrieve the nearby locations. Location information in the format of latitude and longitude of user's current location can be gotten automatically by the system. Radius means the distance in meters within the returned place results. Check-in sets it as 100 meters since the known maximum error of A-GPS is 100 meters. After preparing these parameters properly, an HTTP request will be sent to Google. Maximum 20 locations will be returned by Google in JSON format. The returned results will contain sort of information about the location. Only place name will be shown to users in a list as shown in Figure 14.



Figure13. Location detection – Check in

Check-in provides location awareness support so that it can act as a channel to share, comment and rate any spots in different areas. Users can share their current location with others. Only 4 data values are required to "check-in", i.e. username, check-in's location, comment and rating. Username will automatically be gotten by the system. Check-in location is stored after user finds the nearby locations. A user can also enter comments and ratings about the location in Check-in page as shown in Figure 15. An error will be shown if the length of a comment is more than 160 English characters and 80 Chinese characters. A limit is set with the same idea from the famous social network like Twitter and Weibo. They have a special feature that the post length is limited. According to [17], 160 characters provide plenty of space to express most of their thoughts with friends. Chinese messages can usually present more information or feelings with fewer words.

Thus, 80 characters should be a proper length constrain in a check-in comment.



Figure14. A list of NearBy locations is returned



Figure15. Comments for a checked location

The behavior analysis module is implemented with Python 2.7 at the server end. We have selected the scikit-learn toolkit to support the decision tree and regression analysis [18]. The CART (Classification and Regression Trees) is used for building a model to determine if a person is a potential drug abuser. The same toolkit contains the logistic regression function in its 'liblinear' library and it is adopted.

The Drug Abuse Screening Test (DAST-10) is a 10-item brief screening tool formed the basis of the data being used in the behaviors modules [19]. Other data attributes, such as sleeping patterns, school academic performance, washroom visiting frequency and periods being alone, are included. During of the development of the module, the project team could not acquire the live data because of privacy issues. Instead, a pseudo dataset of 250 records (cases) is generated randomly with 20 attributes per case. Anti-drug experts were then invited to label the cases with different level of drug abuse (where 0 means nil and 5 means high potential). The first 200 records have been used for the decision tree and regression analysis in the scikit-learn toolkit. Cross-validation was then used to verify the accuracy and the results were in the range of 70-80%. The last 50 records were used as additional data for the continuous training. Yet, the results have not been improved.

The email alerting method is mainly based on PHPMailer. It is an open source class with methods to support HTML-based email and attachments. The messaging alerting method is mainly based on Messaging Queuing Telemetry Transport (MQTT) [20, 21]. MQTT is standardized by OASIS as standards to transport data. In MQTT, there is no message formats specialized in it and standard used is depending on developers. But it has advantages over using HTTP and XMPP, which it is low-weight in transmission as lower size in overhead and more energy saving [22, 23].

With the above set up environment, three mobile applications have been developed. The mobile apps for parents contain the alerting tool. It has a main page which displays Drug Abuse Prediction and or Result Feedback (see Figure 16).

In the behavior analysis, the parents act as observers to record their children's behaviors. When the parent starts the drug abuse prediction test, a series of questions on youths' behavior has to be answered in order to get the prediction results. The estimated probability of drug abuse is returned to the screen of the mobile device as shown in Figure 16. When a warning or alerting message is received, parents can seek help from social workers who have joined the platform according to their regions. In seeking the help, parents can provide more specific information about themselves and their children so that suitable social workers in the district can be located faster. They can also describe the situation in order to shorten the remedial time. When a help request is sent, the selected social workers will receive a notification from our platform as a reminder and email message describing parent's situation.

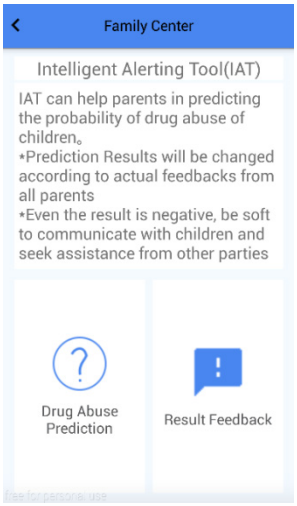


Figure 16a Main Page of behavior analysis



Figure 16b Result of probability of drug abuse

For result feedback, parents are required to answer "positive result" or "negative result" to confirm the drug abuse problem of their children (see Figure 17b). And the feedback is limited to previous detection, which any new prediction will overwrite the old predictions for feedback.

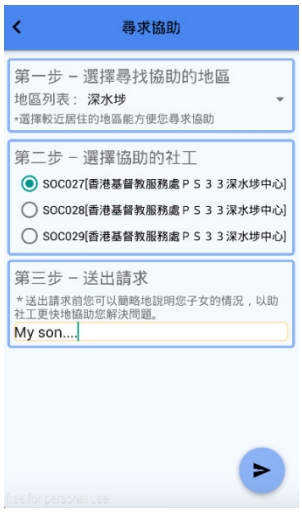


Figure 17a Seek help from social worker in different district

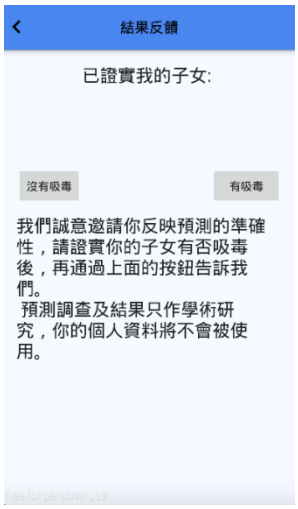


Figure 17b Result feedback

5. Conclusions

In this paper, we have presented three mobile applications with a set of intelligent tools for parents and social workers to help our young people as shown in Figure 18. The application for parents has a tool to analyze drug abuser behavior and send alert to them. The applications have been placed in GoogleStore for download availability.

The work is funded by the Beat Drugs Fund of HKSAR. We have invited a number of secondary schools, community centers, anti-drug experts and social workers to try the applications. All data is collected with a centralized server with confidentiality protection and agreement from users. Users can have options to select if any personal data can be used. Towards the end of the project, positive feedbacks from different user groups have been received. Many parents are interested in the alerting tool while expecting more rules and advice can be embedded.



Figure18. Three mobile applications for anti-drug abuse

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Isobe, 'On publication of denominational precious volumes (Jiao pai xi Pao juan) in Late Ming and Early Ch'ing as seen in, "Niwatazumi".', Studies of Publishing Culture in East Asia, vol. 8, pp. 187-226, 2004.

[2] H. Chan and T. Lo, 'Quality of Life of the Hidden Youth in Hong Kong', Applied Research Quality Life, vol. 9, no. 4, pp. 951-969, 2013.

[3] The Central Registry of Drug Abuse, "Central Registry of Drug Abuse Sixty-sixth Report," Narcotics Division, Hong Kong, 2016.

[4] Policy 21 Limited, "A review of estimation methods on prevalence of drug abuse Final Report," Narcotics Division, 2013.

[5] Social Surveys Section, "Thematic Household Survey Report No. 54 Information technology usage and penetration," Census and Statistics Department HKSAR, 2015.

[6] 禁毒資訊站 [Online]. Available: https://apkpure.com/%E7%A6%81%E6%AF%92%E8%B3%87%E8%A8%8A%E7%AB%99/com.antidrugs

[7] 許寶強, "校園驗毒：「認真」調查結果 馬虎禁毒措施," 明報, 2010.

[8] DrugOut [Online]. Available: https://itunes.apple.com/ao/app/dugout/id1163815036?mt=8

[9] Dihua Sun, Hong Luo, Liping Fu, Weining Liu, Xiaoyong Liao, and Min Zhao, "Predicting Bus Arrival Time on the Basis of Global Positioning System Data", Transportation Research Record: Journal of the Transportation Research Board,No. 2034, D.C., 2007, pp. 62–72.

[10] Depression Test [Online]. Available: https://www.psychologytoday.com/tests/health/depression-test

[11] " Cloud Messaging | Google Developers," Google Developers, 2015. [Online]. Available: https://developers.google.com/cloud-messaging/?hl=zh-tw.

[12] "Messenger for Android - Documentation - Facebook for Developers," Facebook Developers, [Online]. Available: https://developers.facebook.com/docs/messenger/android.

[13] Hani Badran, Pierre Pluye, Roland Grad, "Advantages and Disadvantages of Educational Email Alerts for Family Physicians", J Med Internet Res 2015;17(2):e49.

[14] BradleyTony. (2012). Email vs. IM vs. SMS: Choosing the Right One. PCWorld.

[15] Google Maps will let you share your location with friends and family for a specific period of time [Online]. Available:,https://techcrunch.com/2017/03/22/google-maps-now-lets-you-share-your-location-with-friends-and-family-for-a-specific-period-of-time/

[16] "Google Map," Google, [Online]. Available: https://www.google.com.hk/maps?source=tldsi&hl=zh-TW.

[17] "Find My Friends," Apple Inc., [Online]. Available: https://www.apple.com/apps/find-my-friends/.

[18] Scikit learn [Online]. Available: http://scikit-learn.org/stable/

[19] The Drug Abuse Screening Test (DAST-10) [Online]. Available: https://www.drugabuse.gov/sites/default/files/dast-10.pdf

[20] "MQTT," Mqtt.org, 2015. [Online]. Available: http://mqtt.org.

[21] Y. Zhu, "Mobile Application Message Push and MQTT Protocol," in Wireless Internet Technology, vol. 8, 2015, pp. 1-3.

[22] "The XMPP Standards Foundation," Xmpp.org, 2015. [Online]. Available: http://xmpp.org.

[23] W. Huang, "Design of Android Instant Messaging system based on XMPP agreement," in International Electronic Elements , vol. 19, no. 8, 2011, pp. 57-59.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 443-450 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Actuator Fault Reconstruction based Adaptive Polytopic Observer for a Class of Continuous-Time LPV Systems

Radhia Houimli[*,1,2], Neila Bedioui[1], Mongi Besbes[2]

[1]*Robotics, Informatics and Complex Systems (RISC), ENIT, University of Tunis Al Manar, Institute, 1002, Tunisia*

[2]*Higher Institute of Information and Communication Technologies, University of Carthage, 1164, Tunisia.*

A R T I C L E I N F O

Article history:
Received: 04 December, 2017
Accepted: 07 January, 2018
Online: 10 February, 2018

Keywords :
Polytopic Linear Parameter-Varying (LPV) system
Adaptive observer
LMI

A B S T R A C T

This paper is an extension of work originally presented in conference name. The goal is to propose new fault detection and fault isolation techniques for a polytypic linear parameter-varying system (LPV). In this work, an adaptive observer design is formulated for a given polyquadratic Lyapunov function. Subsequently, new sufficient conditions are given in terms of Linear Matrix Inequalities (LMIs).
To show the effectiveness of the proposed algorithm, an illustrative example is included.

1. Introduction

Research on fault detection, from a theoretical and experimental point of view, has been intensively developed during the last decades. All physical systems must operate normally and without anomaly. However, some conditions cause one or more faults in the process and interrupt this operation. So fault detection is an essential task to avoid degradation of system performance or even its damage [1].

The synthesis of observers for uncertain systems is based on the asymptotic stability of the error estimation equations or on the eliminating influence of uncertain perturbations and measurement of error on errors estimation. Linear Parameter varying system (also known as LPV system) is a special class of system that includes parameters-varying equations and parameters-varying state-space equations. These uncertain systems can be considered as a linearization of state-space nonlinear systems.

The polytopic LPV form is a special class of LPV systems. Indeed, it is a description of the system as a convex combination of sub-models defined by the vertices of a convex polytope [2] [3]. Subsequently, these sub-models are combined by convex weighting functions which give a global model. Similarly, [4] proposes a linear piecewise interpolation model of a diesel engine. The nonlinear model of the machine was transformed into an LPV model.

The synthesis of LPV observers is a direct extension of the LTI control methodologies [5]. The LPV theory has allowed to extend linear methods to nonlinear domains [6]. LPV modeling is used to study nonlinear systems, multiple models or switched models [7].

In this paper, the main contribution is a generalization of the obtained results for actuator fault reconstruction in [1] for LTI systems to continuous-time LPV systems. It was proved that linear methods could be extended to nonlinear domains [8] [9]. In this direction, the actuator fault reconstruction problem is articulated as an LMI feasibility problem. The existence of polyquadratic Lyapunov function could insure the stability of the error estimation [10].

Due to the varying parameter, the polyquadratic approach considers that the Lyapunov function depends on the parameters associated with the description of the polytope.

The paper is organized as follows: In section 2 a model of linear polytopic time-varying (LPV) system is presented. An adaptive observer for fault detection is described in section 3. In section 4, we introduce the polyquadratic adaptive observer for the polytopic LPV system which leads to less conservative conditions on terms of LMI. The simulation result illustrates the effectiveness of our contribution.

Notation. For conciseness the following notations are used: $sym\ (A) = A + A^T$, $\begin{bmatrix} A & B \\ \bullet & C \end{bmatrix} =, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$

[*]Radhia Houimli, BP 8, Face Street, 2021, Tunisia. radhia.houimli@gmail.com

2. Problem Statements and Preliminaries

A continuous-time LPV system in presence of fault can be described by the state-space equations in the following form:

$$\begin{cases} \dot{x}(t) = A(\theta(t))x(t) + B(\theta(t))u(t) + E(\theta(t))f(t) \\ y(t) = Cx(t) \end{cases} \quad (1)$$

Where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$ and $y \in \mathbb{R}^p$ are, respectively, the state space, the input and the output vectors of the system. Variable $f(t)$ represents actuator fault.

The scheduling θ is a set of varying parameters evaluated in hypercube domain Θ such as:

$$\Theta = \left\{ \theta(t) \in \mathbb{R}^p \mid \theta_1 \in \left[\theta_1^{\min}, \theta_1^{\max}\right], ..., \theta_p \in \left[\theta_p^{\min}, \theta_p^{\max}\right] \right\} \quad (2)$$

Where $\theta_i^{\min}$ and $\theta_i^{\max}$, $i = 1,...,p$ are the lower and upper bounds of the parameter.

The parameter dependent Lyapunov function is assumed to be measurable [11] :

$$\theta(t) = \left[\theta_1(t) \quad \theta_2(t) \quad \cdots \quad \theta_r(t)\right]^T \in \mathbb{R}^r \quad (3)$$

Furthermore, LPV system (1) can be defined via barycentric combination of a matrix polytope described by N vertices [12].

$$\begin{aligned} A(\theta(t)) &= \sum_{i=0}^{N} \rho_i(\theta(t)) A_i \\ B(\theta(t)) &= \sum_{i=0}^{N} \rho_i(\theta(t)) B_i, \\ E(\theta(t)) &= \sum_{i=0}^{N} \rho_i(\theta(t)) E_i \end{aligned} \quad (4a)$$

$$\begin{aligned} &\sum_{i=0}^{N} \rho_i(\theta(t)) = 1 \\ &\rho_i \geq 0 \end{aligned} \quad (4b)$$

Where $A_i \in \mathbb{R}^{n \times n}$, $B_i \in \mathbb{R}^{n \times m}$, $E_i \in \mathbb{R}^{n \times r}$ are time invariant matrices defined for the i^{th} ertex of the hypercube and $\rho_i\left((\theta(t))\right) = \rho(\theta_i^{min}, \theta_i^{max}, \theta, t)$. The weighting function $\rho_i\left((\theta(t))\right)$ define the relative contribution of each vertices (A_i, B_i, E_i) to build the system described by $(A(\theta), B(\theta), E(\theta))$.

2.1. Assumption 1 [11]:

The state-space matrices $\left(A\left(\rho(\theta(t))\right), B\left(\rho(\theta(t))\right)\right)$ are continuous and bounded functions and depend on $\rho(\theta(t))$.

2.2. Assumption 2 [11]:

The real parameters $\rho(\theta(t))$ that can be known by on-line measurement values exist in LPV system and vary in a polytope Θ as:

$$\begin{aligned} &\rho(t) \in \Theta \\ &\Theta = \left\{ \sum_{i=1}^{N} \alpha_i(t) w_i : \alpha_i(t) \geq 0, \sum_{i=1}^{N} \alpha_i(t) = 1, N = 2^r \right\} \end{aligned} \quad (5)$$

And the rate of variation $\dot{\rho}(\theta(t))$ are well defined at all times and vary in a polytope $\Theta_v$as:

$$\dot{\rho}(t) \in \Theta_v \quad (6)$$

With

$$\Theta_v = \left\{ \sum_{k=1}^{N} \beta_k(t) v_k : \beta_k(t) \geq 0, \sum_{k=1}^{N} \beta_k(t) = 1, N = 2^r \right\} \quad (7)$$

The following assumption is made:

2.3. Assumption 3:

Without loss of generality matrix C is considered full row rank.

2.4. Assumption 4:

$$rank(CE_i) = rank(E_i) = p \qquad \forall i = [1,...,N] \quad (8)$$

2.5. Assumption 5:

The triple matrix (A_i, E_i, C) is observable.

For simplicity, the time variable t of $\theta(t)$ will be omitted if no confusion is caused.

3. Adaptive Observer Design

For polytopic LPV system (1), an adaptive observer is described by the following state representation:

$$\begin{cases} \dot{\hat{x}}(t) = A(\theta(t))\hat{x}(t) + B(\theta(t))u(t) \\ \qquad + E(\theta(t))\hat{f}(t) - L(\theta(t))(y(t) - \hat{y}(t)) \\ \hat{y}(t) = C\hat{x}(t) \end{cases} \quad (9)$$

Where $\hat{x}(t)$, $\hat{y}(t)$ are the state and outputs estimated vectors and $\hat{f}(t)$is the fault estimation.

In this case the gain matrix is given by the following polytopic form:

$$L(\theta) = \sum_{i=0}^{N} \rho_i(\theta) L_i \quad (10)$$

Where L_i represents the gain of i^{th}vertex.

Remark 1: Since it has been assumed that the pair (A_i, C) is observable, the gain matrices L_i can be selected such that $(A_i - L_i C)$ is stable.

Denote $e_x(t)$, $e_y(t)$, $e_f(t)$ are respectively state ,output and fault estimations errors :

$$\begin{aligned} e_x(t) &= \hat{x}(t) - x(t) \\ e_y(t) &= \hat{y}(t) - y(t) \\ e_f(t) &= \hat{f}(t) - f(t) \end{aligned} \quad (11)$$

Then, the error dynamics are expressed as follows:

$$\dot{e}_x(t) = \sum_{i=1}^{N} \rho_i(\theta(t))\left((A_i - L_i C)e_x(t) + E_i e_f(t)\right) \quad (12)$$

$$\dot{e}_y(t) = C\dot{e}_x(t) \quad (13)$$

The default $f(t)$ is constant, hence $\dot{f}(t) = 0$ [13], consequently the derivate of $e_f(t)$ with respect to time can be written as:

$$\dot{e}_f(t) = \dot{\hat{f}}(t) \tag{14}$$

The state observer (9) is combined with the law for the fault estimation updating of the form [1]

$$\dot{\hat{f}}(t) = -\Gamma F e_y(t) \tag{15}$$

Where $F \in \mathbb{R}^{r\times p}$ and $\Gamma \in \mathbb{R}^{r\times r}$ is the learning rate.

It has been to note a modification of (13) is presented in [1], [6] for time varying $f(t)$ in the form:

$$\dot{\hat{f}}(t) = -\Gamma F(\theta)\left(\dot{e}_y(t) + \sigma e_y(t)\right) \tag{14}$$

Where $\sigma \in \mathbb{R}$ is a positive scalar and can guaranty $\lim_{t\to\infty} e_x(t) = 0$ and $\lim_{t\to\infty} e_f(t) = 0$.

4. Main Result

Consider the LPV system described by (1) with an additive fault. The choice of the parameter-dependent Lyapunov functions for polytopic systems is a dilemma in the literature. The rate of change of the scheduling parameter could be represented in diverse methods. [14] proposed a rate of change that cannot be physically justified. Furthermore, in the expression given by [13], the derivative of the uncertain parameter does not impose special conditions.

In this section, we develop a new adaptive observer for LPV polytopic system. Before, we introduce some instrumental tools which will be used in the proof of characterization of this observer.

4.1. Lemma 1 [1]:

Given scalar $\mu > 0$ and symmetric positive definite matrix $P\left(\rho(\theta(t))\right)$, the following inequality holds:

$$2x^T y \le \frac{1}{\mu} x^T P(\theta) x + \mu y^T P(\theta)^{-1} y \qquad x, y \in \mathbb{R}^n \tag{17}$$

4.2. Lemma 2 *[15]:*

Given a symmetric matrix $\psi \in \mathbb{R}^{n\times n}$, and two matrices P, Q of column dimensions n, there exists X such that the following LMI holds:

$$\psi + sym(P^T X^T Q) < 0 \tag{18}$$

If and only if the projection inequalities with respect to X are satisfied:

$$\mathcal{N}_P \psi \mathcal{N}_P^T < 0, \quad \mathcal{N}_Q^T \psi \mathcal{N}_Q < 0 \tag{19}$$

Where $\mathcal{N}_p$ and $\mathcal{N}_Q$ denote arbitrary bases of the null spaces of P and Q respectively.

Proof. See [15]. ■

4.3. Lemma 3:

Let Φ a symmetric matrix and N, J matrices of appropriate dimensions. The following statements are equivalent:

- $\Phi < 0$ and $\Phi + NJ^T + JN^T < 0$.
- There exists a matrix X such that:

$$\begin{bmatrix} \Phi & J + NX \\ J^T + X^T N^T & -X - X^T \end{bmatrix} < 0 \tag{20}$$

Proof: The proof is obtained remarking that (19) can be developed as follows:

$$\begin{bmatrix} \Phi & J + NX \\ J^T + X^T N^T & -X - X^T \end{bmatrix} = \begin{bmatrix} \Phi & J \\ J^T & 0 \end{bmatrix} + sym\left\{ \begin{bmatrix} 0 \\ I \end{bmatrix} X^T \begin{bmatrix} N^T & -I \end{bmatrix} \right\} < 0 \tag{21}$$

and by applying Lemma 2 . ■

In this part, we consider the case such as the fault is time-varying, which implies $\dot{f}(t) \neq 0$, and the derivate of $e_f(t)$ with respect to time is:

$$\dot{e}_f(t) = \dot{\hat{f}}(t) - \dot{f}(t) \tag{22}$$

The objective of this section is to propose an approach to design a new adaptive observer for polytopic LPV system (1). So, we propose the following Theorem 2.

Theorem 2. Under the assumptions 1, 2 and 3, the system (9) is an adaptive observer for the system (1) if, for a given scalars $\sigma > 0$, $\mu > 0$, $\alpha > 0$ and $\nu > 0$, there exists, for each vertex, asymmetric positive definite matrix $P_i \in \mathbb{R}^{n\times n}$, $L_i \in \mathbb{R}^{n\times r}$, $G_i \in \mathbb{R}^{r\times r}$ and $X_1 \in \mathbb{R}^{n\times n}$ such that the following conditions hold:

$$\begin{bmatrix} \nu P_i - 2P_i + sym(P_i A_j) & -\frac{2}{\sigma} A_j^T P_i E_k & P_i + X_1 & P_i - \alpha C^T L_i^T \\ \bullet & -\frac{2}{\sigma} E_k^T P_i E_j + \frac{1}{2\mu\sigma} G_i & 0 & -\frac{2}{\sigma} E_j^T P_i \\ \bullet & \bullet & -X_1 - X_1^T & 0 \\ \bullet & \bullet & \bullet & -2\alpha I \end{bmatrix} < 0 \tag{23}$$

Under constraint:

$$E_j^T P_i = F_i C \tag{24}$$

The gain observer matrix of the system (1) is given by the following polytopic form:

$$L(\theta) = \sum_{i=0}^{N} \rho_i(\theta) L_i \tag{25}$$

Remark 3: The principle of the polytopic formulation is based on the fact that the system and stability conditions (here in a LMI form) have affine dependence on the parameters. If, for some reason, the affine dependence is lost, the stability of the system is not equivalent (or even implied only) to the feasibility of the LMI at each vertex [16], [17], [18].

Proof: With respecting to the system parameter, it is clear that $e_x(t)$ is linear. Thereby, consider the Lyapunov polytopic function defined by:

$$V\left(e_x(t),\right) e_f(t)) = e_x^T(t) P(\theta) e_x(t) + e_f^T(t) \Gamma^{-1} e_f(t) \tag{26}$$

Where $P\left(\rho(\theta(t))\right) > 0$ is a symmetric positive defined matrix.

Then, the derivative of (26) with respect to t is:

$$\dot{V}\left(e_x(t), e_f(t)\right) = \dot{V}_1\left(e_x(t), e_f(t)\right) + \dot{V}_2\left(e_x(t), e_f(t)\right) < 0 \quad (27)$$

Where :

$$\begin{aligned}\dot{V}_1\left(e_x(t), e_f(t)\right) &= \dot{e}_x^T(t)P(\theta)e_x(t) \\ &+ e_x^T(t)\dot{P}(\theta)e_x(t) + e_x^T(t)P(\theta)\dot{e}_x(t) \\ &= e_x^T(t)\Big[\dot{P}(\theta) + \left(A(\theta) - L(\theta)C\right)^T P(\theta) \\ &+ P(\theta)\left(A(\theta) - L(\theta)C\right)\Big]e_x(t) \\ &+ e_f^T(t)E^T(\theta)P(\theta)e_x(t) \\ &+ e_x^T(t)P(\theta)E(\theta)e_f(t)\end{aligned} \quad (28)$$

And

$$\begin{aligned}\dot{V}_2\left(e_x(t), e_f(t)\right) &= \frac{1}{\sigma}\dot{e}_f^T(t)\Gamma^{-1}e_f(t) + \frac{1}{\sigma}e_f^T(t)\Gamma^{-1}\dot{e}_f(t) \\ &= \frac{1}{\sigma}\left(\dot{\hat{f}}(t) - \dot{f}(t)\right)^T \Gamma^{-1}e_f(t) \\ &+ \frac{1}{\sigma}e_f^T(t)\Gamma^{-1}\left(\dot{\hat{f}}(t) - \dot{f}(t)\right)\end{aligned} \quad (29)$$

Substituting (16) into (28) leads to :

$$\begin{aligned}\dot{V}_2\left(e_x(t), e_f(t)\right) &= -\frac{2}{\sigma}e_f^T(t)\Gamma^{-1}\dot{f}(t) \\ &- \frac{2}{\sigma}e_f^T(t)\Gamma^{-1}\Gamma\Big[F(\theta)\left(\dot{e}_y(t) + \sigma e_y(t)\right)\Big] \\ &= -\frac{2}{\sigma}e_f^T(t)\Gamma^{-1}\dot{f}(t) - \frac{2}{\sigma}e_f^T(t)F(\theta)C\dot{e}_x(t) \\ &- e_f^T(t)F(\theta)Ce_x(t) - e_x^T(t)C^T F^T(\theta)e_f(t)\end{aligned} \quad (30)$$

Then, substituting (12) and (13) into (30), the following inequality is hold:

$$\begin{aligned}\dot{V}\left(e_x(t), e_f(t)\right) &= e_x^T(t)\Big[\dot{P}(\theta) + \left(A(\theta) - L(\theta)C\right)^T P(\theta) \\ &+ P(\theta)\left(A(\theta) - L(\theta)C\right)\Big]e_x(t) \\ &+ e_f^T(t)E^T(\theta)P(\theta)e_x(t) \\ &+ e_x^T(t)P(\theta)E(\theta)e_f(t) \\ &- e_f^T(t)F(\theta)Ce_x(t) \\ &- e_x^T(t)C^T F^T(\theta)e_f(t) \\ &- \frac{2}{\sigma}e_f^T(t)\Gamma^{-1}\dot{f}(t) - \frac{2}{\sigma}e_f^T(t)F(\theta)C\dot{e}_x(t)\end{aligned} \quad (31)$$

If the following condition is introduced:

$$\begin{aligned}&e_f^T(t)\Big[E^T(\theta)P(\theta) - F(\theta)C\Big]e_x(t) \\ &+ e_x^T(t)\Big[P(\theta)E(\theta) - C^T F^T(\theta)\Big]e_f(t) = 0\end{aligned} \quad (32)$$

This implies that:

$$E^T(\theta)P(\theta) = F(\theta)C \quad (33)$$

The inequality (31) becomes by using (33) and :

$$\begin{aligned}\dot{V}\left(e_x(t), e_f(t)\right) &= e_x^T(t)\Big[\dot{P}(\theta) + \left(A(\theta) - L(\theta)C\right)^T P(\theta) \\ &+ P(\theta)\left(A(\theta) - L(\theta)C\right)\Big]e_x(t) \\ &- \frac{2}{\sigma}e_f^T(t)E^T(\theta)P(\theta)\left(A(\theta) - L(\theta)C\right)e_x(t) \\ &- \frac{2}{\sigma}e_f^T(t)E^T(\theta)P(\theta)E(\theta)e_f(t) \\ &- \frac{2}{\sigma}e_f^T(t)\Gamma^{-1}\dot{f}(t)\end{aligned} \quad (34)$$

From Lemme 1, we can suppose that:

$$\begin{aligned}&2\left(-\frac{1}{2\sigma}e_f(t)\right)^T\left(\Gamma^{-1}\dot{f}(t)\right) \\ &\le \frac{1}{2\mu\sigma}e_f^T(t)Ge_f(t) + \frac{\mu}{2\sigma}f_1^2\lambda_{\max}\left(\Gamma^{-1}G^{-1}\Gamma^{-1}\right)\end{aligned} \quad (35)$$

Then, subsisting (35) in (34), the following inequality is done:

$$\begin{bmatrix} a_{11} & * \\ a_{21} & a_{22}\end{bmatrix} < 0 \quad (36)$$

With:

$$a_{11} = \dot{P}(\theta) + sym\left(\left(A(\theta) - L(\theta)C\right)^T P(\rho)\right)$$

$$a_{21} = -\frac{2}{\sigma}E^T(\theta)P(\theta)\left(A(\theta) - L(\theta)C\right)$$

$$a_{22} = -\frac{2}{\sigma}E^T(\theta)P(\theta)E(\theta) + \frac{1}{2\mu\sigma}G(\theta)$$

The derivate of the Lyapunov function is defined as follows:

$$dP(\theta)/dt = \sum_{k=1}^{N}\beta_k(t)Pt(v_k) = \sum_{k=1}^{N}\beta_k(t)\left(P(v_k) - \hat{P}_0\right) \quad (37)$$

$$\dot{P}(\theta) = \sum_{i=1}^{N}\dot{\rho}_i(\theta)P_i \quad (38)$$

$$\sum_{i=1}^{N}\dot{\rho}_i = 0 \quad (39)$$

The rate $\dot{\rho}(t)$ can be represented in several ways. In fact most of the time, it is difficult to give adequate modeling of it. For LPV system, the derived parameter does not vanish as in the LTI case.

In our case, we suppose that [14]:

$$\dot{\rho}(t) < \upsilon\rho(t) \quad (40)$$

$$\dot{P}(\theta) < \upsilon P(\theta) \quad (41)$$

Unfortunately, (41) is not convex in P and L, and cannot be solved by the LMI tools.

We can introduce some transformations to simplify the product term $(P(\theta)L(\theta)C)$ of the inequality (41). In fact, in this solution

we introduce an additive variable in order to allow the decoupling between the Lyapunov matrix and the observer gain in one side and to preserve a general structure to the Lyapunov matrix in the other side.

We suppose that:

$$\Phi=\begin{bmatrix} \upsilon P(\theta)-2P(\theta)+sym\left(P(\theta)A(\theta)\right) & \\ -\frac{2}{\sigma}E^T(\theta)P(\theta)A(\theta) & \\ * & \\ -\frac{2}{\sigma}E^T(\theta)P(\theta)E(\theta)+\frac{1}{2\mu\sigma}G(\theta) \end{bmatrix} \quad (42)$$

$$N^T=\begin{bmatrix} I & 0 \\ -L(\theta)C & 0 \end{bmatrix} \quad (43)$$

$$J=\begin{bmatrix} P(\theta) & P(\theta) \\ 0 & -\frac{2}{\sigma}E^T(\theta)P(\theta) \end{bmatrix} \quad (44)$$

By lemma 2 with (42), (43) and (44), there exists a matrix X of appropriate dimensions such that inequality (45) is satisfied.

$$\begin{bmatrix} M_1 & M_2 & M_3 & M_4 \\ * & M_5 & 0 & M_6 \\ * & * & -X_1-X_1^T & 0 \\ * & * & 0 & -2\alpha I \end{bmatrix}<0 \quad (45)$$

Where

$$\mathrm{X}=\begin{bmatrix} X_1 & 0 \\ 0 & \alpha I \end{bmatrix} \quad (46)$$

$$M_1=\upsilon P(\theta)-2P(\theta)+sym(P(\theta)A(\theta))$$

$$M_2=-\frac{2}{\sigma}A(\theta)P(\theta)E(\theta)$$

$$M_3=P(\theta)+X_1$$

$$M_4=P(\theta)-\alpha C^T L^T(\theta)$$

$$M_5=-\frac{2}{\sigma}E^T(\theta)P(\theta)E(\theta)+\frac{1}{2\mu\sigma}G(\theta)$$

$$M_6=-\frac{2}{\sigma}E^T(\theta)P(\theta)$$

Remark 4: The main advantage of problem (45) will appear when dealing with poly-quadratic observer. In that case, we will see that it theoretically improves the obtained results.

5. Numerical example

The above-described algorithm was applied to reconstruct the fault applied to the following LPV system described in [5] as:

$$A(\theta)=\begin{bmatrix} -1.75+\theta_2 & 1 & 0 & 0 \\ 1 & -1+\theta_1 & 0 & 0 \\ -1.8 & -1 & -0.75+\theta_1 & 0 \\ -1 & 0 & 0 & -1-\theta_2 \end{bmatrix}$$

$$B(\theta)=\begin{bmatrix} 1+\theta_1 & 1 \\ 1 & 0.5+\theta_2 \\ 1 & 0 \\ \theta_2 & 0 \end{bmatrix},\ C=\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$E(\theta)=\begin{bmatrix} 0 \\ 0.6+\theta_1 \\ 0 \\ 1 \end{bmatrix}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\rho_i(\theta)\rho_j(\theta)\rho_k(\theta)\begin{bmatrix} vP_i-2P_i+sym(P_iA_k) & \cdots & -\frac{2}{\sigma}A_jP_iE_k & \cdots \\ * & \cdots & -\frac{2}{\sigma}E_k^TP_iE_j+\frac{1}{2\mu\sigma}G_i & \cdots \\ * & \cdots & * & \cdots \\ * & \cdots & * & \cdots \end{bmatrix}$$

$$\begin{bmatrix} P_i+X_1 & P_i-\alpha C^TL_i^T \\ 0 & -\frac{2}{\sigma}E_j^TP_i \\ -X_1-X_1^T & 0 \\ 0 & -2\alpha I \end{bmatrix}<0 \quad (47)$$

The gain scheduling vector is defined as:

$$\theta=\begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}^T \quad (48)$$

Where

$$\theta_1\in\begin{bmatrix} -0.05 & 0.05 \end{bmatrix} \quad (49)$$

$$\theta_2\in\begin{bmatrix} -0.1 & 0.1 \end{bmatrix} \quad (50)$$

The system has four vertices and evolves in a hypercube. The weighting functions, which verified (4b), are computed as follows:

$$\begin{aligned} \rho_1(\theta)&=\frac{\theta_1-\underline{\theta_1}}{\overline{\theta_1}-\underline{\theta_1}}\frac{\theta_2-\underline{\theta_2}}{\overline{\theta_2}-\underline{\theta_2}}=\frac{(\theta_1+0.05)(\theta_2+0.1)}{0.02} \\ \rho_2(\theta)&=\frac{\theta_1-\underline{\theta_1}}{\overline{\theta_1}-\underline{\theta_1}}\frac{\overline{\theta_2}-\theta_2}{\overline{\theta_2}-\underline{\theta_2}}=\frac{(\theta_1+0.05)(0.1-\theta_2)}{0.02} \\ \rho_3(\theta)&=\frac{\overline{\theta_1}-\theta_1}{\overline{\theta_1}-\underline{\theta_1}}\frac{\theta_2-\underline{\theta_2}}{\overline{\theta_2}-\underline{\theta_2}}=\frac{(0.05-\theta_1)(\theta_2+0.1)}{0.02} \\ \rho_4(\theta)&=\frac{\overline{\theta_1}-\theta_1}{\overline{\theta_1}-\underline{\theta_1}}\frac{\overline{\theta_2}-\theta_2}{\overline{\theta_2}-\underline{\theta_2}}=\frac{(0.05-\theta_1)(0.1-\theta_2)}{0.2} \end{aligned} \quad (51)$$

The four local models represented the LPV system are calculated as the following:

$$A_1=\begin{bmatrix} -1.85 & 1 & 0 & 0 \\ -1 & -1.05 & 0 & 0 \\ -1.8 & -1 & -0.8 & 0 \\ -1 & 0 & 0 & -1.1 \end{bmatrix} \quad (52)$$

$$A_2 = \begin{bmatrix} -1.65 & 1 & 0 & 0 \\ -1 & -1.05 & 0 & 0 \\ -1.8 & -1 & -0.8 & 0 \\ -1 & 0 & 0 & -0.9 \end{bmatrix} \quad (53)$$

$$A_3 = \begin{bmatrix} -1.85 & 1 & 0 & 0 \\ -1 & -0.95 & 0 & 0 \\ -1.8 & -1 & -0.7 & 0 \\ -1 & 0 & 0 & -1.1 \end{bmatrix} \quad (54)$$

$$A_4 = \begin{bmatrix} -1.65 & 1 & 0 & 0 \\ -1 & -0.95 & 0 & 0 \\ -1.8 & -1 & -0.7 & 0 \\ -1 & 0 & 0 & -0.9 \end{bmatrix} \quad (55)$$

$$B_1 = \begin{bmatrix} 0.95 & 1 \\ 1 & 0.4 \\ 1 & 0 \\ -0.1 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0.95 & 1 \\ 1 & 0.6 \\ 1 & 0 \\ 0.1 & 0 \end{bmatrix}, \quad (56)$$

$$B_3 = \begin{bmatrix} 1.05 & 1 \\ 1 & 0.4 \\ 1 & 0 \\ -0.1 & 0 \end{bmatrix}, \quad B_4 = \begin{bmatrix} 1.05 & 1 \\ 1 & 0.6 \\ 1 & 0 \\ 0.1 & 0 \end{bmatrix} \quad (57)$$

$$E_1 = E_2 = \begin{bmatrix} 0 \\ 0.55 \\ 0 \\ 1 \end{bmatrix}, \quad E_3 = E_4 = \begin{bmatrix} 0 \\ 0.65 \\ 0 \\ 1 \end{bmatrix} \quad (58)$$

By applying algorithm (9), gains matrices are as the following:

$$L_1 = 10^{-7} \times \begin{bmatrix} 0.8103 & 0.1578 & -0.489 \\ 0.1433 & 0.0788 & -0.0832 \\ 0.3573 & 0.0560 & -0.1950 \\ -0.3956 & 0.0074 & 0.2464 \end{bmatrix}$$

$$L_2 = 10^{-7} \times \begin{bmatrix} 0.9953 & 0.2005 & -0.6049 \\ 0.1606 & 0.0981 & -0.0962 \\ 0.4209 & 0.0641 & -0.2329 \\ -0.4902 & 0.0192 & 0.3089 \end{bmatrix}$$

$$L_3 = 10^{-7} \times \begin{bmatrix} 0.7851 & 0.1489 & -0.4744 \\ 0.1376 & 0.0795 & -0.0798 \\ 0.3467 & 0.0535 & -0.1886 \\ -0.3845 & 0.0104 & 0.2398 \end{bmatrix}$$

$$L_4 = 10^{-7} \times \begin{bmatrix} 0.8443 & 0.1676 & -0.5104 \\ 0.1469 & 0.0780 & -0.0860 \\ 0.3687 & 0.0576 & -0.2022 \\ -0.4115 & 0.0053 & 0.2563 \end{bmatrix}$$

$$P_1 = 10^{-7} \times \begin{bmatrix} 0.1529 & 0.0034 & 0.0439 & -0.0329 \\ 0.0034 & 0.0098 & 0.0042 & -0.0004 \\ 0.0439 & 0.0042 & 0.0175 & -0.0138 \\ -0.0329 & -0.0004 & -0.0138 & 0.0254 \end{bmatrix}$$

$$P_2 = 10^{-7} \times \begin{bmatrix} 0.1877 & 0.0023 & 0.0523 & -0.0407 \\ 0.0023 & 0.0126 & 0.0044 & 0.0002 \\ 0.0523 & 0.0044 & 0.0203 & -0.0169 \\ -0.0407 & 0.0002 & -0.0169 & 0.0328 \end{bmatrix}$$

$$P_3 = 10^{-7} \times \begin{bmatrix} 0.1500 & 0.0032 & 0.0429 & -0.0319 \\ 0.0032 & 0.0097 & 0.0040 & -0.0000 \\ 0.0429 & 0.0040 & 0.0172 & -0.0134 \\ -0.0319 & -0.0000 & -0.0134 & 0.0250 \end{bmatrix}$$

$$P_4 = 10^{-7} \times \begin{bmatrix} 0.1562 & 0.0033 & 0.0448 & -0.0343 \\ 0.0033 & 0.0099 & 0.0043 & -0.0008 \\ 0.0448 & 0.0043 & 0.0179 & -0.0144 \\ -0.0343 & -0.0008 & -0.0144 & 0.0261 \end{bmatrix}$$

The synthesis of the LPV observer gains L_i for each vertex is achieved with LMI Toolbox of Matlab. The observer gain L of the system described by (58) is determined offline using (9) in the different simulation case bellow. The parameters of simulation are fixed as $\sigma = 10^{10}, \Gamma = 120 \, and \, \mu = 0.2$.

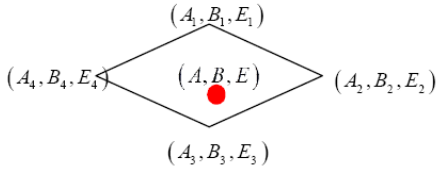
To illustrate the effectiveness of our algorithm, we choose a particular system matrix (A, B) computed using (4) for arbitraries values of $\theta_1 = 0.03$ and $\theta_2 = 0$:

$$A = \begin{bmatrix} -1.756 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ -1.8 & -1 & -0.75 & 0 \\ -1 & 0 & 0 & -1.006 \end{bmatrix} \quad (59)$$

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 0.494 \\ 1 & 0 \\ -0.006 & 0 \end{bmatrix} \quad (60)$$

$$E = \begin{bmatrix} 0 \\ 0.6 \\ 0 \\ 1 \end{bmatrix} \quad (61)$$

The given LPV system is defined in the vertex as shown in the figure above:



A. Case of constant fault signal

In the first case, consider the constant fault described by the following equation:

$$f(t)=\begin{cases}0 & t>0\\ 2 & 51\le t\le 20s\\ 0 & 101\le t\le 150\\ 2 & other\end{cases} \quad (62)$$

Figure 1 and Figure 2 show the estimation of the fault applied to the system described by matrix (59), (60) and (61).

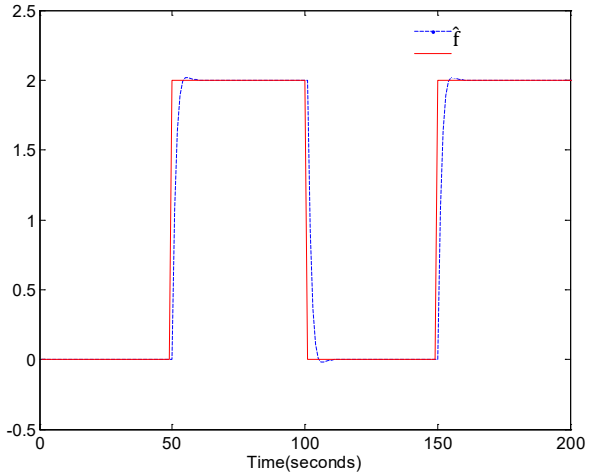


Fig. 1. f and its estimation $\hat{f}$.

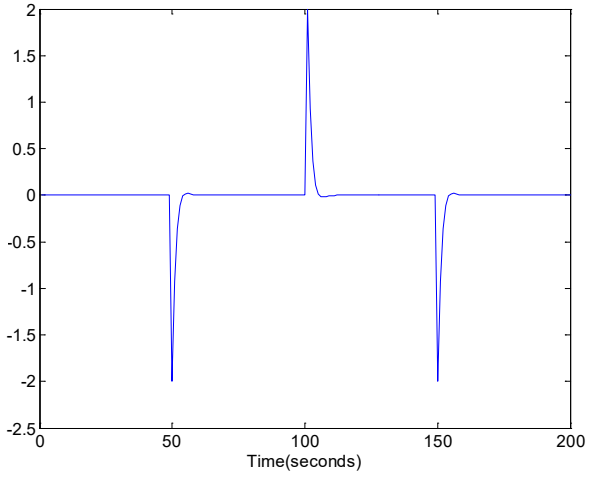


Fig. 2. The error $e_f(t)$ of the fault estimation

B. Case of variable fault signal

The actuator fault is described by:

$$f_{a2}(t)=\begin{cases}0.1\sin 5t+0.04\cos 3t\\ \quad +0.06\sin t+0.05\\ 0 \quad other\end{cases} \quad 7s\le t\le 12s \quad (63)$$

For the arbitrary values of $\rho(\theta)$, the simulation results are as the following:

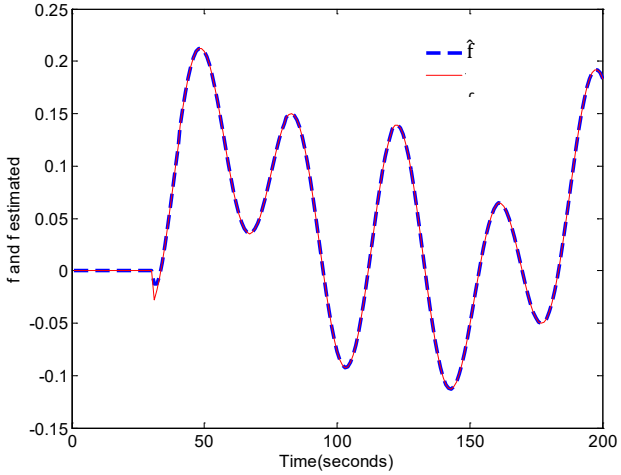


Fig. 3. f and its estimation $\hat{f}$.

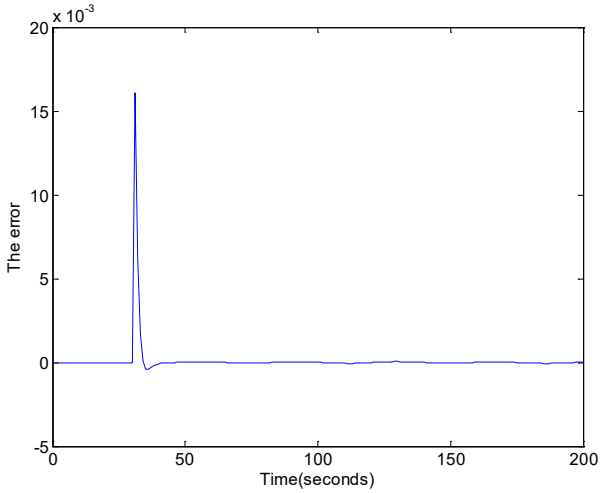


Fig. 4. The error $e_f(t)$ of the fault estimation

For the different types of fault signals considered above, Figure 2 and Figure 4 show the evolution of the error estimation. As can be observed, the error estimation converges asymptotic to zero even in the presence of disturbances. The real and the estimated constant and variable faults are displayed in Figure 1 and Figure 3 respectively. Summarizing, this approach can estimate the states and the fault functions with good performance and small error.

6. Conclusion

In this paper, an adaptive LPV observer using LPV approach has been presented for a polytopic system. A constant fault and a variable fault were considered. A polyquadratic Lyapunov function was used to perform the stability analysis. The problem was formulated in terms of linear matrix inequalities to develop the observer. The simulation results show the performances of the proposed observer. The main advantage of this representation is that it doesn't depend directly on the varying parameter. Moreover, this representation is defined as a difference between two parameters that evolve in two known and defined polytopes.

References

[1] K. Zhang, B. Jiang and V. Cocquempot, "Adaptive Observer-based Fast Fault Estimation" International Journal of Control, Automation, and Systems, **6**(3), 320-326, 2008. https://doi.org/10.1109/ChiCC.2015.7260636

[2] M. A. Montes de Oca, D. Aydın, T. Stutzle, "An Incremental Particle Swarm for Large-Scale Continuous Optimization Problems: An Example of Tuning-in-the-loop (Re) Design of Optimization Algorithms". Soft Computing, **15**(11):2233–2255. https://doi.org/10.1007/s00500-010-0649-0

[3] S. Grenaille, D. Henry and A. Zolghadri, "A method for designing fault diagnosis filters for LPV polytopic systems Journal of Control Science and Engineering. Volume 2008 (2008) : 1–11. http://dx.doi.org/10.1155/2008/231697

[4] Changhui W, Zhiyuan L., "A LPV adaptive observer approach to calibrate MAF sensor map in diesel engine". In: 54th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Hangzhou, China, pp.1375–1380. Piscataway, NJ: IEEE, 2015. https://doi.org/10.1109/SICE.2015.7285416

[5] C. Hoffmann and H. Werner, "A Survey of Linear Parameter-Varying Control Applications Validated by Experiments or High-Fidelity Simulations", IEEE Transactions on Control Systems Technology, Volume: **23**, Issue: 2, pp. 416 – 433, 2015. https://doi.org/10.1109/TCST.2014.2327584

[6] F. Bruzelius. "Linear Parameter-Varying Systems an approach to gain scheduling". PhD thesis, Department Of Signals And Systems, Chalmers University of Technology, 2004.

[7] I. Szászi, B. Kulcsár, G. J. Balas and J. Bokor, "Design of FDI filter for an aircraft control system", Proceedings of the American Control Conference Anchorage, AK May 8-10, 2002. https://doi.org/10.1109/ACC.2002.1024596

[8] F. Bruzelius, "Linear Parameter-Varying Systems: an approach to gain scheduling", Phd thesis, University of Technology of Göteborg, February 2004.

[9] F. Bruzelius, S. Pettersson and C. Breitholtz, "Linear parameter varying descriptions of nonlinear systems", in Proceeding of the IEEE American Control Conference (ACC), pages 1374–1379, Boston, Massachusetts, July 2004.

[10] O. Sename, P. Gáspár and J. Bokor, Robust Control and Linear Parameter Varying Approaches: Application to Vehicle Dynamics. Library of Congress Control, Number: 2012955946. Springer-Verlag Berlin Heidelberg, 2013.

[11] W. Xie, "H2 gain scheduled state feedback for LPV system with new LMI formulation", IEEE Proc.-Control Theory Appl., **152(**6), November 2005.

[12] D. Ichalal and S. Mammar, "On Unknown Input Observers for LPV Systems", IEEE Transactions on Industrial Electronics, 2015. https://doi.org/10.1109/TIE.2015.2448055

[13] J. Geromel and P. Colaneri, "Robust stability of time varying polytopic systems", Systems and control letters, **55**:81–85. 2006. https://doi.org/10.1016/j.automatica.2006.08.024

[14] Y. Caoand Z. Lin, "A Descriptor System Approach to Robust Stability Analysis and Controller Synthesis", IEEE Transaction on automatic control, Vol. 49, No. 11, November, 2004. https://doi.org/10.1109/TAC.2004.837749

[15] S.P. Boyd, L.E Ghaoui, E. Feron and V. Balakrishnam, Linear Matrix Inequalities in System and Control Theory, the Society for Industrial and Applied Mathematics, 3600 University City Science Center, Philadelphia, Pennsylvania 19104-2688, vol. **15**, 1994.

[16] P. Apkarian and R.J. Adams. "Advanced gain-scheduling techniques for uncertain systems". IEEE Transactions on Automatic Control, **6**:21–32, 1998. https://doi.org/10.1109/87.654874

[17] M. Jungers, P.L.D. Peres, E.B. Castelan, E.R. De Pieri, and H. Abou-Kandil. "Nash strategy parameter dependent control for polytopic systems". In 3rd IFAC Symposium on Systems, Strcuture and Control, Brazil, 2007. https://doi.org/10.3182/20071017-3-BR-2923.00097

[18] R. C. L. F. Oliveira, V. F. Montagner, P. L. D. Peres, and P.-A.Bliman. "LMI relaxations for H∞ control of time-varying polytopic systems by means of parameter dependent quadratically stabilizing gains". In 3rd IFAC Symposium on System, Structure and Control, Foz do Iguassu, Brasil, 2007. https://doi.org/10.3182/20071017-3-BR-2923.00099

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 451-459 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Innovative Research on the Development of Game-based Tourism Information Services Using Component-based Software Engineering

Wei-Hsin Huang[*,1], Huei-Ming Chiao[1], Wei-Hsin Huang[2]

[1]*Graduate Institute of Design Science, Tatung University & Department of Digital Game and Animation Design Taipei University of Maritime Technology, Taipei, Taiwan, R.O.C.*

[2]*Department of Media Design, Tatung University, Taipei, Taiwan, R.O.C.*

ARTICLE INFO

Article history:
Received: 14 December, 2017
Accepted: 18 January, 2018
Online: 10 February, 2018

Keywords:
Digital Guide
Game-Based Learning
Component-based Software Engineering

ABSTRACT

In recent years, a number of studies have been conducted exploring the potential of digital tour guides, that is, multimedia components (e.g., 2D graphic, 3D models, and sound effects) that can be integrated into digital storytelling with location-based services. This study uses component-based software engineering to develop the content of game-based tourism information services. The results of this study are combined with 3D VR/AR technology to implement the digital 2D/3D interactive tour guide and show all the attractions' information on a service platform for the gamification of cultural tourism. Nine kinds of game templates have been built in the component module. Five locations have completed indoor or external 3D VR real scenes and provide online visitors with a virtual tour of the indoor or outdoor attractions. The AR interactive work has three logos. The interactive digital guide includes animated tour guides, interactive guided tours, directions and interactive guides. Based on the usage analysis of the component databases built by this study, VR game types are suited to object-oriented game templates, such as the puzzle game template and the treasure hunt game template. Background music is the database component required for each game. The icons and cue tones are the most commonly used components in 2D graphics and sound effects, but the icons are gathered in different directions to approximate the shape of the component to be consistent. This study built a vivid story of a scene tour for online visitors to enhance the interactive digital guide. However, the developer can rapidly build new digital guides by rearranging the components of the modules to shorten the development time by taking advantage of the usage frequency of various databases that have been built by this study to effectively continue to build and expand the database components. Therefore, more game-based digital tour guides can be created to make better defined high-quality heritage attractions understood.

1. Introduction

In the United States, museums and historical sites generated 15 billion U.S. dollars in revenue in 2017 [1]. With the rapid development in information technology, the diversification and popularization of mobile carriers have made for rich digital content development. The Taiwan Tourism Bureau survey analysis also shows that cultural tourism has become an important tourism trend. The integration of multi-material, cross-platform and cross-industry values has become another focus of the content industry. Digital guides and online tourism maps of museums and historical sites' interactive tourist information services have gradually changed reading habits in recent years, replacing the hard copy of tourism guide maps. Since the development of the global digital economy, the output of digital content industries in Taiwan [2, 3] has reached over a 10% annual growth rate in recent years and a value of about 32 billion U.S. dollars in 2015. Based on the advantage of many excellent manufactories of information and communication technology in Taiwan, virtual reality technology has been regarded as an important driver of the digital economy. Applications such as commercial games, knowledge learning, and museum guides have increasingly introduced virtual reality, augmented reality, and mixed reality in Taiwan. For example, HTC and Foxconn Technology Group announced the first

*Corresponding Author: Wei-Hsin Huang, Tatung University, Taipei, Taiwan, R.O.C., Email: wshuang@ttu.edu.tw

VIVELAND™, which includes many virtual reality games, in Taiwan [4]. The Institute of Transportation adopted an interactive mixed reality platform [5] to develop a tool to assist teachers to teach children to walk safely around their schools [6]. The Shihsanhang Museum of Archaeology invites visitors to use a virtual reality tool to explore an important Taiwanese underwater cultural asset, the British steamship SS Bokhara, which sank near Taiwan in 1982 [7].

1.1. Digital Tour Guiding

Brown in [8] examined earlier applications that demonstrated location‐aware guiding by mobile devices, such as Cyberguide, which enabled a visitor to locate nearby bars, and GUIDE, which provided visitors with a means of personalizing information to suit their own interests and the environmental parameters. To increase the interactive elements, Bellotti in [9] built VeGame to operate on a cellular phone-mediated wireless connection between a pocket PC and the server, which communicated with the pocket PC via Bluetooth. VeGame was designed to enhance tourists' experience of art and history through a pleasant and challenging interaction with the heritage and people of Venice. When mobile devices became more widely used, Ballagas in [10] designed REXplorer, a device consisting of a Nokia mobile phone, a global positioning system (GPS) receiver, a camera, and a stretchable textile overlay with a zipper on the back that transformed a standard phone keypad into an 8-key game interface. Tourists could use REXplorer with gestures to interact with the significant buildings in Regensburg, Germany. Regarding indoor tour guiding, in addition to radio frequency identification (RFID) technology for positioning [11], Tsai in [12] proposed a position estimation method to design a location-aware tour guide system to be used with a personal digital assistant (PDA) by visitors to the National Palace Museum of Taiwan. As the user moved closer to items in the collection, the size of the materials became larger on the screen. Steiniger and Edwardes [13] indicated that location-based services (LBS) combined with GPS, geographic information systems (GIS), quick response (QR) codes, mobile facilities and scene-point tour databases are an important future direction of such technology. Currently, advances in mobile technology make it feasible to use virtual and augmented reality technology for learning [14]. Doong in [15] investigated the chocolate-related knowledge learning effect of a game-based learning system developed with a cross-platform of LBS and mixed reality technologies. The virtual treasure-hunt game world was the real location of a chocolate factory that the participants had to find on Google Maps.

1.2. Agile Development and Component-Based Software Engineering (CBSE)

Developing game-based digital content is an expensive and risky activity. Using CBSE methods allows one to design digital content quickly by reusing existing functions and models to produce a system [16]. Folmer and Mehm's overview of developing games with components presents a reference architecture that outlines the relevant areas of reuse and signifies some of the problems with developing components unique to the domain of gaming [17, 18]. Elements characteristic of games (good competition, character development, improvement, inspiring and creative challenges, and activity) can easily be adapted to the needs arising from cultural settings. The results would be an increase in motivation and the development of positive relations with the cultural subject. Based on the spirit of the agile development method, Wu Junyue (2013) devised a set of development processes for mobile customer service that are best for the mobile service application [19]. Wu Junyue's development process f reduces the output of unnecessary documents and uses repetitive development processes to shorten the development schedule. For these reasons, a challenge for the scientific community is to find new ways to visualize and disclose 3D digital contents to achieve better access to and communication of cultural heritage information [20]. Therefore, the purpose of this study is to determine how to develop digital guides rapidly and make them suitable for more kinds of handheld equipment.

1.3. 3D Virtual Game-based Environment

Effective learning is situated, active, and problem-based and requires immediate feedback [21]. Gaming elements provide an instructional environment with stimulation and learning motivation [22, 23]. A well-designed virtual educational game provides complex holistic problem-based environments, making it possible to develop situated understanding [24] and thus supporting effective learning. Virtual game-based environment designs follow the rules of game design. Nevertheless, the purpose is not only to entertain [25] but also to use the characteristics of video and computer games to create attractive and immersive learning experiences to achieve specific learning objectives [26]. In addition to the four important features (gameplay, feedback, interface, and challenge) of a good game [25], the design of a game-based learning system must further consider the realism of the game, the opportunities to explore and obtain new information, and the meaning of learning controls [26]

1.4. Agile Development and Digital Content Development

The effectiveness of the selection of the game template components and the game design is also the most thought-out part of the process of implementation. For example, Teng Feng Fishball Museum (TFFM) at Tamsui is a food tourism location where the TFFM combines the history and cultural characteristics of the Tamsui. The TFFM is focused on individuals or family members and is supported by a rich history and cultural knowledge. The TFFM can present more profound educational material and understanding with 3D VR game types. However, this kind of innovative food museum can also present a variety of innovative food products by focusing on entertaining 2D game templates. Both digital guides focus on increasing the pleasure of the user experience, allowing the user to learn more about tourist attractions, or increasing tourists' pleasure in traveling on their own. Digital guides also meet educational needs and draw visitors into visiting the museum.

2. Methods

2.1. The Game-based Tourism Information Service System (GTISS)

The Game-based Tourism Information Service System (GTISS) was developed for the purpose of breaking geographical and temporal restrictions. 3D virtual and game-based tourist attractions were created using Unity software. Three major elements are including in the GTISS: itinerary planner, games and

virtual reality design, and cultural tourism features. The concept of CBSE was adopted in the development of the GTISS by agile methods.

The GTISS is designed to break geographical and temporal restrictions by creating 3D virtual and game-based tourist attractions using Unity software. There are three major elements in the GTISS, including an itinerary planner, virtual game-based design, and cultural tourism features. The concept of CBSE was adopted in the development of the GTISS by agile methods. Employing CBSE in the design of a tourism information platform expedites the process, that is, multimedia components (e.g., game plans, 2D icons and pictures, music, sound effects, 3D models, and programs) in a previously created database can be used and reused for various projects, which lowers the threshold for developing a tourism information platform and adds value to existing tourism materials. Cultural tourism is often full of story-like elements; thus, to enhance the attractiveness of a cultural tourism information platform, it is necessary to improve its storytelling ability and incorporate the LBS and game elements into the platform [27]. Web3D, GIS, GPS, and Unity3D VR software were used in the present study. The goal of CBSE is to establish a digital content development platform consisting of reusable components; this platform can shorten the time for developing a project. Drawing on the concept of CBSE, the multimedia component database allows developers who do not have enough game development background to design the content of the platform easily. Finally, all the components in the database were created to be compatible with the Unity 3D engine.

To develop the GTISS, first, information content about tourist attractions was collected. Then, based on the features and characteristics of each tourist attraction and the corresponding game plans, appropriate components in the multimedia component database were searched to design a game that presents the information about each tourist attraction. LBS integrated the QR code to enhance the contextualized interaction. Tiered with levels of difficulty, it features requirements, such as time limits and penalties, that add excitement and competition to the game. The accomplishment of finding a destination involves answering questions or solving problems in some well-known scenic spots. Moreover, for users, the platform provides travel guides to the tourist attractions by showing available public transport through dynamic maps and consecutive pictures of real street views. In addition, by using Google Maps application program interfaces (APIs), the platform shows the GPS locations of the tourist attractions on Google Maps, which is familiar to many users.

2.2. Agile Development and Digital Content Development

A number of tourist attractions that the government actively promotes were selected for information collection. Each digital guide for the tourist attractions provides public transport information integrated into the digital map to promote a Low Carbon Tour. Applying the appropriate game plan for an attraction's information confirms that the characteristics of the attraction's content are sufficient for the resources required for the game plan. For example, if there is not enough information about the attraction, then the use of the real tour will lead to digital guide content that is not attractive. Finally, we applied the modular component and game design principles of the plan to archive the digital guides. With the Unity 3D engine, the completed digital guide can be built into different carrier formats (such as html, apk, and exe). The process of component-based agile game and digital content development used by this study is shown in Figure 1.

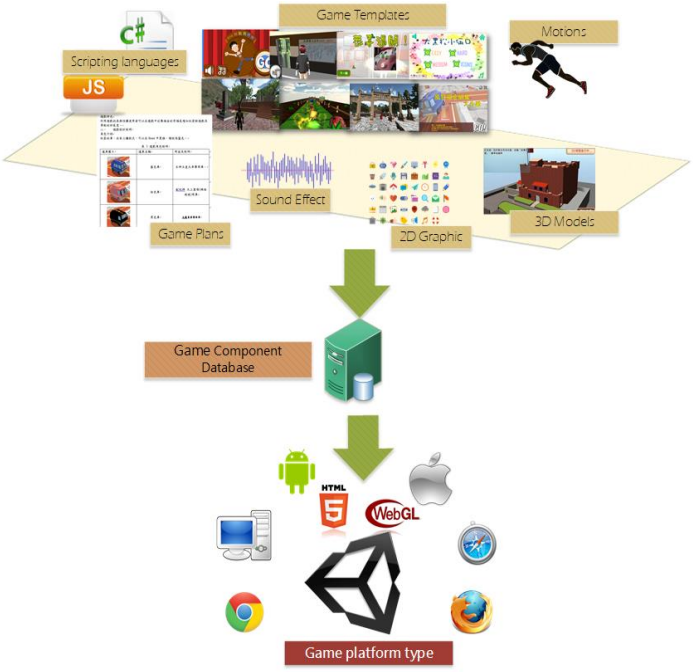


Figure 1 The process of use agile design method to make digital content

The effectiveness of the selection of the game template components and the game design is also the most thought-out part of the process of implementation. For example, Teng Feng Fishball Museum (TFFM) at Tamsui is a food tourism location where the TFFM combines the history and cultural characteristics of the Tamsui. The TFFM is focused on individuals or family members and is supported by a rich history and cultural knowledge. The TFFM can present more profound educational material and understanding with 3D VR game types. However, this kind of innovative food museum can also present a variety of innovative food products by focusing on entertaining 2D game templates. Both digital guides focus on increasing the pleasure of the user experience, allowing the user to learn more about tourist attractions, or increasing tourists' pleasure in traveling on their own. Digital guides also meet educational needs and draw visitors into visiting the museum.

To rapidly develop a game-based digital guide, component-based templates were used to develop a digital guide of tourist attractions. The construction of the modular component database replaced the homemade or commercially available development resource modularization database. Seven component databases were established and used to develop a game-based digital guide that can be viewed on webpages or handheld devices. The component databases are shown in Table 1. A motion database and a game template database were recently completed. After developing several games, the humanoid character can be used in a cross-game motion template. We expanded our motion database

by using x-box to scan simple motions that the game needed and binding the motion to the characters.

Table 1: Component-Database of this study

Database Name	Content
3D Model	The 3D model library contains models such as character, buildings, and objects that built from a variety of modeling software, such as 3D Max, Maya, and SketchUp. All models are finally imported in the development engine with .fbx format
2D Graphic	This database contains illustrations for animations and interface designs such as picture or icon. The 2D material in database licensed under creative commons. Some of these art files may contain .jpg, .png, and so on in different file formats.
Scripting language	This study aimed at scripting and sorting all kinds of shareable functions in game templates. For example, the function of these feet for the game interface to return and explain the function, the game time count/pause/stop, props configuration, character action control. All scripts has been documentation for re-use by non-game developers.
Game Plan	The game template of this study will be self-development and collect from web resource database. We write planning documents for each free or paid game template for non-game developer to use.
Sound Effect	Collect, organize and use the sound studio or online recording software produced by the various sound effects database. All content licensed under creative commons, and contains include .mp3 or .wmv and other file formats of sound.
Motion	In addition to collecting free motion files available on the network, find out how to organize the motion files from ready-made game templates into other game templates. This study also used X-Box for Motion capture to record the movements of real people and mapping the movements onto the CG character.
Game Template	By implement the game plan into real game, there are eight game template had been build. There are nine game template in our database.

In addition to the integration of models, animations and sound effects in the game development process, the development of game script language is also very important. However, the number of script resources that can be used directly in the game is less than the number of models or sound effects. As shown in Figure 2, if building a script language database can be organized, then the time that programmer used to rewrite the code can be saved. The developer simply confirms the required functionality, finds the appropriate script resource in the database, and imports it into the Unity game engine. Fine-tuning allows game developers to quickly develop game features and add new script into the script database.

2.3. 360-Degree Virtual Tour Guide

To increase the sense of participation, an attempt was made to imitate the indoor and outdoor feeling of the attractions that visitors would experience. Some of the attractions are not well defined by their location or are connected to a factory, and the user cannot find the entry.

This study uses Google Street View's screenshots to build a 360-degree VR scene. The application of this 360-degree scene setting in this platform mainly provides an indication of the user's arrival status at the attraction. Using the 360-degree circular guide at the entrance increases the user's understanding of the geographical location of the scenery. The design of this scene can be applied to the design of future AR scenes. The scene of this study is based on the use of Google Street View and real indoor photos made during field trips to build outdoor and indoor scenes. In the future, the user may use the camera lens of the mobile device as a source, and the digital guide instructs the user to provide information or tips based on all the current locations.

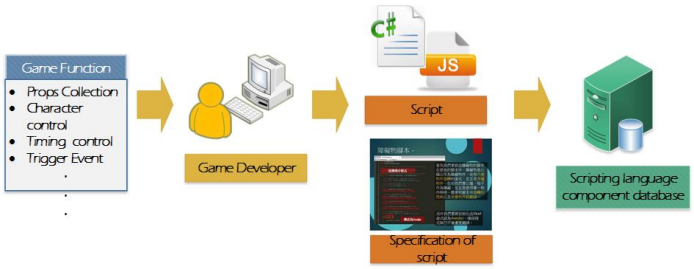


Figure 2 The process of build Scripting language component database

3. Results

3.1. Game-based Component Database

This study built seven component databases, as described in Table 1. Figure 3, Figure 4 and Table 2 are the samples of the 3D models, different motions and sound effects collected in this study in the present component databases. Regarding the 3D character models, for example, there are currently ten 3D character models in the database. As shown in Figure 3, the same character can be used in different games quickly with various motions. The movement of running includes running forward, turning left, and turning right. The motions of touching the head and waiting are bound to form the other motion. Some motion files for men and female are different. In the different games shown in Figure 3, it is possible not only to reuse the 3D character model in the game template but also to use 3D models of terrain and trees. User interface (UI) templates in the 2D graphic database can be created for different attractions easily, quickly and repeatedly. As shown in Figure 4(a), the same UI template was used in the National Dr. Sun Yat-sen Memorial Hall and the Vigor Kobo Dream Museum. Through a color change, the same UI component can quickly adapt to each game template. Similarly, by using consistent icons, the user can understand the meaning of the graphic on the screen. As shown in Figure 4 (b), the icons on the screen are the indispensable elements that guide the user to operate the icon related functions. During the preliminary stage of design, it took about a month to accomplish the construction of a scenic spot whereas only five working days were needed in the later stage for the creation and modification of an attraction, which is the most time-consuming of the entire project.



(a) Turn-left running motion on female (b) Turn-left running motion on man

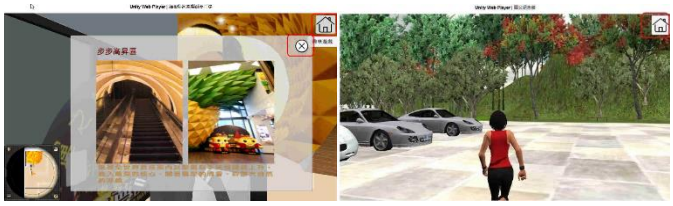


(c) The waiting motion (d) Touch head motion and two character

Figure 3 The 3D character models and motions in different game



(a) The UI template used by two attraction



(b) The icon used by different game

Figure 4 The 2D UI template and close icon be reuse in our project for saving the time of development

3.2. The Tourism Digital Guiding Platform (TDGP)

In the present study, the GTISS was designed and shown as in Figure 5. All attraction locations were labeled in Google Maps and categorized into six types: food, shopping, heritage, transportation, education, and entertainment. User-friendly interfaces enable visitors to find interesting locations easily. Fifteen cultural locations were chosen and designed in the TDGP. Figure 6 shows the template for planning the route to the destination (e.g., from Taipei Main Station to the Taiwan Socks Museum). Users can choose to board at the mass rapid transport (MRT) station. Then, according to the routing algorithm, the user can achieve the minimum number of transfers or the shortest distance calculated from the number of stops. Both the interactive display technology of the LBS geographic location and the 3D VR/AR are important trends in the digital content industry.

The interactive traffic information guideline was provided by the destination websites. Combined with the street/attractions and

Table 2 Component-Database of this study

Attractions	BGM
Anping Castle	Path to Follow.mp3
Republic of Chocolate	The_Messenger.mp3
National Dr. Sun Yat-sen Memorial Hall	Clouds.mp3
Longshan Temple	Clouds.mp3
Taiwan's Socks Museum	On the Bach.mp3
W & W Museum of Jewelry	On the Bach.mp3
Vigor Kobo Dream Museum	Festival.mp3
National Palace Museum	The_Engagement.mp3
Lanyang Museum	Retreat.mp3
Dali Tian Gong Temple	Succotash.mp3
Kuo Yuan Ye Museum of Cake and Pastry	Reasons_to_Smile.mp3
Taiwan Nougat Creativity Museum	Morning Walk.mp3
KONG YEN Yi Shou Dou Cultural Museum	Microchip.mp3
Yumeeriren maternity tour factory	Keith.mp3
Shu Shin Bou Wagashi Museum	Morning Walk.mp3



Figure 5 The tour platform use Google Maps API to presentation



Figure 6 Dynamic route planning to Taiwan's Socks Museum(

outdoor/indoor navigation technology of an attraction's features, users can have an in-depth understanding of the tourist information and be impressed by the digital tour guides.

The completion of each level involves answering questions and solving problems. Answers to the questions can only be found by asking locals questions, exploring the surroundings, or using certain services specific to a location. At the outset, the QR code of a zone must be photographed, and when all the questions in that zone are answered or when the problems are solved, the user can then advance to the next level of his/her choice.

Figure 6 shows useful travel guides included in the tourist attractions with available public transport route information through dynamic maps. By using the 360-degree circular guide, consecutive pictures of real street views can be provided to increase the visualization and understanding of those locations (shown in Figure 7). The platform shows the GPS locations of the tourist attractions on Google Maps, and more LBS information can be added to the system.

The 2D/3D game-based learning content of fifteen attractions are shown in Figure 8. These are famous attractions in Taiwan and are promoted by the Taiwan Tourism Bureau. The development of a digital tour guide is mainly based on 3D game types. However, the architecture is the main part of building the 3D models of the attractions, and it requires more time and work to adjust during the development process. To attain the purpose of the agile development of the new digital guide, this study also built some interesting 2D game templates for our database to save time and reduce the needed process in developing digital tour guides, as shown in Figure 8.

3.3. 3D AR / VR Game-based Tourism Service

Seven databases were used in designing the 15 attractions for the effective and efficient development of the game. Table 3 shows the number of 2D graphics with the game templates applied in the 15 attractions. Figure 9 shows several 2D graphics used in the game templates. Most reusable images are icon-type graphics, which can change color, size, and text. In addition, each game has background music, which was chosen from the background music (BGM) database based on rhythm and types. Table 4 shows all the music applied in the game templates for the study.

To create a 3D VR game template, a goal-oriented puzzle or treasure hunt was created by interacting with the object in the scene (Table 5). The Unity3D game engine was used to build indoor and outside VR panoramas of attractions in this tourism platform. Google street view was used for outdoor scenes, and 3D models with real indoor photos of the museums were created in the present study. Through the implementation of an online VR tour, users can have pleasure and further understand the details of the attractions.

There are nine kinds of game templates built in the component modules. By using these component modules, a digital guide with 3D VR real scenes were provided to visitors for touring both the indoor and outdoor attractions (as is shown in Figure 10). This study organized the game-based component templates and created vivid scenarios of scene tours for visitors to enhance the interaction with the digital guide. Nine kinds of game templates with three game types are shown in Table 5.

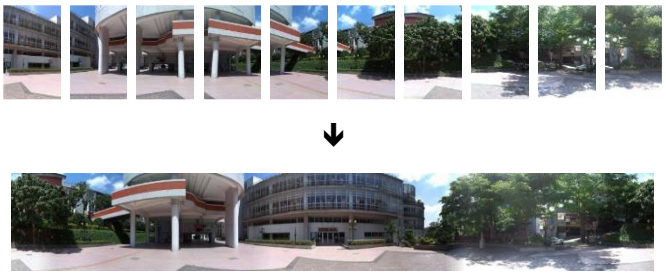
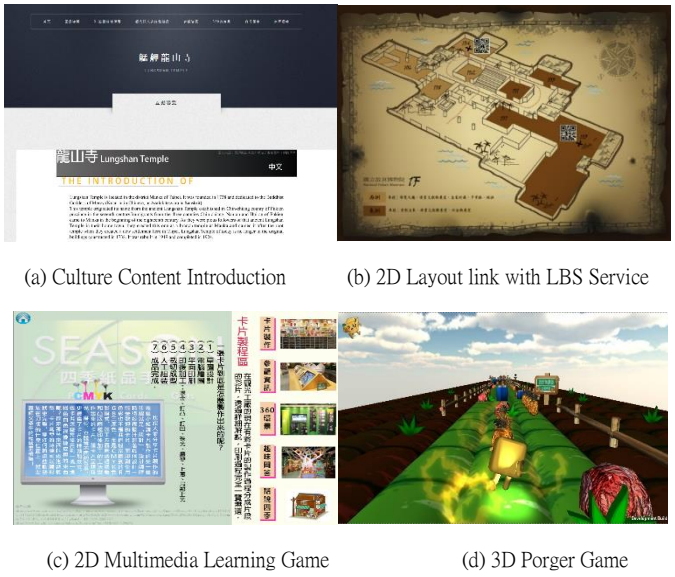


Figure 7 This study use Google Street View's screenshots to build up a 360-degree VR scene



(a) Culture Content Introduction (b) 2D Layout link with LBS Service

(c) 2D Multimedia Learning Game (d) 3D Porger Game

Figure 8 The Cultural Tourism Digital Guiding Platform

Table 3 Number of Sound used by 15 Attractions with Game type

Sound	Game Type			Total
	2D	3D	3D VR	
BGM	**5**	**5**	**5**	**15**
Bird		1		1
Cannon sound			1	1
Cue Tone -Button Click	4			4
Cue Tone -Notice		1	4	5
Cue Tone -Pass	1	1		2
Cue Tone -Pick up Item	1	2		3
Cue Tone -Wrong	**4**	**1**	**4**	**9**
Cue Tone-Car Crash		1		1
Cue Tone-Loss Health points		2		2
Cue Tone-Miss Hit	1			1
Cue Tone-Most Time Up	4			4
Cue Tone-Notice		1		1
Cue Tone-Right	**4**	**1**	**4**	**9**
Cue Tone-Time Up	4			4
Explosion		1		1
Mining sound			1	1
Racing Sound		1		1
Shooting sound	1	1		2
Train arrive		1		1
Typing Effect		1		1
Water of fountain		1		1
Wheel Rolling Sound	1			1
Wind sound			1	1
Total	30	22	20	72

Table 4 Number of Attractions with Game template and Game type

Game Template	Game Type			Number of Attractions
	2D	3D	3D VR	
Book Template		1		1
t Wheel Game Template	1			1
t Personal Shooting Game		1		1
den Object Game Template	2			2
stery Jigsaw Game Template	2			2
kour games Template		1		1
zle Game Template			2	2
ing Game Template		1		1
asure hunt game Template		1	3	4
al	5	5	5	15

Table 5 Number of 2D Graphic used by 15 Attractions with Game type

2D Graphic	2D			2D Total	3D					3D Total	3D VR	
	A.	B.	C.		D.	E.	F.	G.	H.	I.		J.	K
Entry Graphic	1		2	3		1					1		
Exit Icon	**1**	**2**	**2**	**5**			**1**	**1**	**1**		**3**		
Game Over	**1**	**2**	**2**	**5**			**1**	**1**	**1**		**3**		
Home Icon	**1**	**2**	**2**	**5**		**1**			**1**	**1**	**3**	**2**	
Level-Hard		2		2									
Level-Midium		2		2									
Level-Simple		2		2									
Man-Waiter	1			1									
Next icon			2	2									
Next-Page						1					1		
Pause								1			1		
Pre-Page						1					1		
Replay Icon	**1**	**2**	**2**	**5**			**1**	**1**	**1**		**3**		
Right						1				1	2	2	
Setting Icon	1	2		3					1		1		
Store									1		1		
Success								1		1	2	2	
Treasure Map													
Try Again										1	1		
UI Background	1	2	2	5									
Wrong						1					1	2	
總計	8	18	14	40		6	3	5	6	4	24	8	

Notice the Symbol of Game Template :

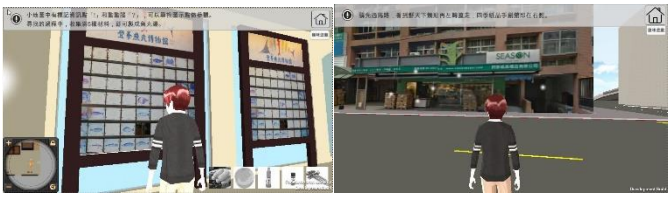
A. Dart Wheel Game Template
B. Hidden Object Game Template
C. Mystery Jigsaw Game Template
D. 3D Book Templete
E. First Personal Shooting Game
F. Parkour games Template
G. Racing Game Template
H. Treasure hunt game Template
I. Puzzle Game Template
J. Treasure hunt game Template



Figure 9 The 2D Gameful Design Digital Guide of this research



(a) The 3D Geometry-base Model VR Digital Guide



(b) The Image-base Background with 3D Geometry-base Model VR Digital Guide

Figure 10 The Digital Guide of this research

During the 2D game development period, a navigation e-book was also built to be merged with the 3D model. The design of the navigation e-book has been improved from the original 2D material to the 3D model library by adding a UI interface for a more interesting and clear introduction. To upgrade the function of the navigation e-book content, we incorporated 2D text descriptions and dynamic 3D guidance to enhance the entire navigation e-book'tals readability. This navigation e-book is released on a Unity platform. There is a PC version, a web format, and an .apk format for other handheld devices.

Animations and interactive elements were added to this navigation e-book to increase the interactivity between the digital content and the users. Using Lan Yang Museum's 3D navigation e-book as an example, on the floor map page, by clicking on the 3D text on the left and the red indicator on the right side of the page, it will move to the floor on the map (Figure 11(a)). In the traffic guide page, the 3D models of moving trains and buses, as shown in Figure 11(b), were also added to improve the appeal of the e-book as well as the enjoyment of the visitors.

To further expand our VR games into AR games. Three museums' logo images were chosen as the AR image-based



(a) PC version 2D e-book model (b) 2D e-book with UI



(c) E-book with 2D map and 3D words (d) E-book With dynamic 3D model (Bus and Train)

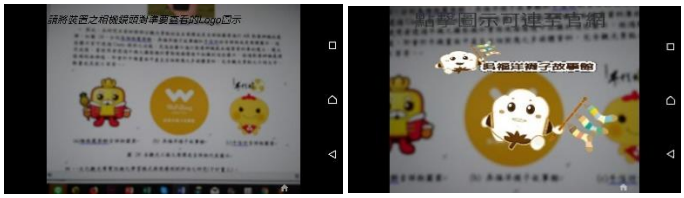
Figure 11 The screenshots of Lan Yang Museum's navigation e-book (apk)

identification, which can generate four identifications supported by the Unity software: single picture identification, cuboid identification, cylinder identification and 3D object identification. When the AR glass lens captures the correct image object, other instructions from the weblink or hyperlink to the museum website are triggered (as shown in Figure 12).

4. Discussion



(a) The screenshot of mobile device that hint user how to use AR

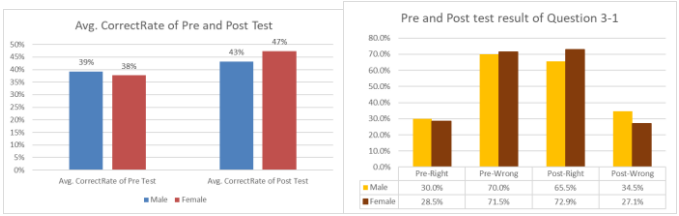


(b) The logo photo capture by mobile device (c) Show the Mascot's Animation

Figure 12 Screenshot of AR identification in this research

It is important to design and develop game-based digital tour guides using agile development and component databases in the future. Game-based digital guide learning easily increases not only the entertainment effect but also the understanding and recollection of the cultural attractions. However, to accomplish this complex work (drawing the 3D model, writing the coding, producing the music, etc.) requires a long development process. Developers with experience in game design need at least a month to finish a game project without using a component database. Using the game-based component database, intern students with no experience in game development were trained to complete the first digital game guide within two weeks. After completing the first project, four intern students found new attractions and used game templates to develop new game-based digital guides on their own. The intern students can not only develop independently but also shorten the design period to within a week. In the future, non-game professional tour staff can complete their own interesting digital tour guides easily and quickly.

After the platform is established, the attention of younger users can be captured by measuring the accuracy of the questions that users answer before and after using the platform. If the digital tour guide of this platform is sufficient to attract users to browse, read and memorize the content, the accuracy of the answers in post-test results should be improved. There were 477 student participants. Their ages ranged from 18 to 22 years old, and the number of female students (338 or 70.9%) was higher than the number of male students (139 or 29.1%). This study used a pre-test and post-test on knowledge content design issues for the platform attractions. The pre-and post test aimed to examine whether students enhanced their understanding of the attractions and the online travel planning. The items include an itinerary planner, games and virtual reality design, cultural features of the attractions, and Tourism English. Students did the pre-test before the experiment and the post test after it. The experiment lasted for one hour. The purpose of the system design was explained and the operational functions were demonstrated before the experiment. Pre- and post-tests were created based on the content of the system to examine student learning on the platform. As shown in Figure 13 (a), both males and females increased their correct answer rate after using the platform. Figure 13 (b) shows the pre- and post-test correct answer rates for question 3-1 in the questionnaire. The post-test correct answer rates for both males and females were 10% or higher than the pre-test. These results are sufficient to initially demonstrate the validity of the digital tourism content on the platform established by this study. The game-based digital tour guides with locality-specific information and culture in mind can make not only real hands-on experiences but also experiences that are memorable and worth sharing.



(a) Avg. CorrectRate of Pre and post test (b) Pre and Post test result of Question 3-1

Figure 13 Two result of pre and post test

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is supported by the Ministry of Science and Technology, Taiwan, R.O.C. under Grant no. MOST 104-2632-H-262 -001 -MY2.

This paper was originally published in the 2017 IEEE International Conference on Applied System Innovation(ICASI), May 13-17, 2017, Hotel Emisia, Sapporo, Japan.

References

[1] Courtney Tucci. Statistics and Facts on Museums. Retrieved from https://www.statista.com/topics/1509/museums/ (2015).
[2] [2] Ministry of Culture (2015). Taiwan Cultural Creative Industries Annual Report 2016 (in Chinese). Ministry of Culture.
[3] [3] Institute for Information Industry (2016). 2015 Digital Content Industry in Taiwan (in Chinese). Ministry of Economic Affairs, Industrial Development Bureau.
[4] [4] HTC. VIVELAND (https://www.vive.com/tw/viveland/).
[5] [5] Ji-Liang Doong Ching-Huei Lai, Kai-Hsiang Chuang, Chun-Chia Hsu (2015, Jul). The Development of Location-based Mobile Augmented Reality (LoMAR). The 3rd International Conference on Hospitality, Leisure, Sports, and Tourism. July 23-24, 2015, Waseda University, Tokyo, Japan .
[6] [6] Chun-Chia Hsu, Chih-Yung Lin, Yu-Li Chen, Wei-Shin Huang, Kai-Hsiang Chuang, Huei-Ming Chiao in (2017) Tools development of investigating roads and using roads for pedestrians and bicyclists (1/3) – Applications of road safety audit (in Chinese). Institute of Transportation, Department of Transportation and Communications.
[7] [7] The Shihsanhang Museum of Archaeology. Virtual reality exploring the S. S. Bokhara (http://www.sshm.ntpc.gov.tw).
[8] [8] Brown, Elizabeth; Brner, Dirk; Sharples, Mike; Glahn, Christian; de Jong, Tim and Specht, Marcus (2010). Location-based and contextual mobile learning. A STELLAR Small-Scale Study. STELLAR European Network of Excellence in TEL (EU).
[9] Bellotti, F., Berta, R., De Gloria, A., Ferretti, E., & Margarone, M. (2003). VeGame: exploring art and history in Venice. Computer 36(9), 48-55
[10] Ballagas, R., Kuntze, A. and Walz, S. P. Gaming Tourism: Lessons from Evaluating REXplorer - a Pervasive Game for Tourists. In: J. Indulska in (Eds.): Pervasive 2008, LNCS 5013, pp. 244–261, 2008. Springer.
[11] Wang, Y., Yang, D., Liu, S., Wang, R. and Meng, X. A RFID & Handheld Device-Based Museum Guide System. in Proceedings of the International Conference on Pervasive Computing and Applications (ICPCA), 2007, pp.308–313.
[12] Tsai C Y, Chou S Y, Lin S W. Location-aware tour guide systems in museum. Scientific Research and Essays, 2010, 5(8): 714-720.
[13] Steiniger, S., Neun, M., & Edwardes, A. (2006). Foundations of Location Based Services: University of Zurich.
[14] Specht, M., Ternier, S., & Greller, W. (2011). Dimensions of mobile augmented reality for learning: a first inventory. Journal of the Research Center for Educational Technology, 7(1), 117–127.
[15] Ji-Liang Doong, Ching-Huei Lai, Kai-Hsiang Chuang, Chun-Chia Hsu, Learning Effects of Location Based Mixed Reality Game: A Pilot Study, Procedia Manufacturing, Volume 3, 2015, Pages 1603-1607.World Tourism Organization. (2017).UNWTO Tourism Highlights: 2017 Edition - UNWTO Elibrary
[16] Bei-lin Lin, “On the rebuild-of the XBRL Demosite by using component-based design.”, Master Thesis, National Central University, 2011
[17] Folmer, E., Component based game development : A solution to escalating costs and expanding deadlines. CBSE'07 Proceedings of the 10th international conference on Component-Based Software Engineering(pp. 66-73) (ISBN: 978-3-540-73550-2) (2007)
[18] Florian Mehm, Stefan Gobel, Ralf Steinmetz. Introducing component-based templates into a game authoring tool. In Dimitris Gouscos and Michalis Meimaria (ed.) 5th European Conference on Games Based Learning (pp. 395–403)., Reading, UK: Academic Conferences Limited. (2011)
[19] JYUN-YUE WU, "Applying Agile Method to Building Mobile Application Software and Services", Master thesis, Shih Hsin University,2013
[20] Poria Y, Bulter R, Airey D. The core of heritage tourism. Annals of Tourism Research, 2003, 30(1): 238-254.
[21] Boyle, E., Connolly, T. M., & Hainey, T. (2011). The role of psychology in under- standing the impact of computer games. Entertainment Computing, 2(2), 69–74.
[22] Din H W H. Play to learn: Exploring online educational games in museums. In The 33rd International Conference and Exhibition on Computer Graphics and Interaction Techniques, Boston, MA, 2006.
[23] Sørensen B H, Meyer B. Serious games in language learning and teaching - A theorectical perspective. In Proceedings of 3rd International Conference of the Digital Games Research Association, 2007, pp. 559-566.
[24] Annetta L A, Folta E, Klesath, M. V-Learning: Distance Education in the 21st Century through 3D Virtual Learning Environments. Springer, New York, 2010.
[25] Johnson, L.W., Vilhjalmsson, H., Marsella, S. in: Artificial Intelligence in Education, IOS Press, Amsterdam, Netherlands, 2005, pp. 306-313.
[26] de Freitas, S. Learning in immersive worlds, 2006. (http://www.jisc.ac.uk/media/documents/programmes/elearninginnova-tion/gamingreport _v3.pdf).
[27] Ghiani G, Paternò F, Santoro C, Spano L D. UbiCicero: A location-aware, multi-device museum guide. Interacting with Computers, 2009, 21(4): 288-303.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 460-466 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Performance Analysis of Regenerative Braking in Permanent Magnet Synchronous Motor Drives

Andrew Adib*, Rached Dhaouadi

College of Engineering, American University of Sharjah, P.O. Box 26666 Sharjah, UAE

ARTICLE INFO

Article history:
Received: 11 November, 2017
Accepted: 09 January, 2018
Online: 10 February, 2018

Keywords:
Permanent Magnet Synchronous Motor
Regenerative Braking
DC-DC Converter
Ultracapacitor
Pulse Width Modulation
Energy Harvesting

ABSTRACT

This paper describes the design and analysis of a regenerative braking system for a permanent magnet synchronous motor (PMSM) drive for electric vehicle (EV) applications. First studied is the principle for electric braking control of a PMSM motor under field-oriented control (FOC). Next, the maximum braking torque in the regeneration mode as well as the braking torque for the maximum regeneration power, respectively, are deduced. Additionally, an optimum switching scheme for the inverter is developed with the objective of maximizing energy recovery during regenerative braking to the DC-bus capacitor. The integration of an ultracapacitor module with the battery allows for the efficient and high power transfer under regenerative braking. It was important to manage the power flow to the DC-bus as this is a key issue that affects the efficiency of the overall system. Finally, the amounts of braking energy that can be recovered, and the efficiency with which it can be returned to the battery/ultracapacitor, is analyzed for a PMSM coupled with a DC motor as the load. The results of the analysis are validated through experimentation.

1. Introduction

Interest in regenerative braking is growing drastically nowadays; as the market is slowly transitioning to electric vehicles (EV) instead of the traditional vehicles that run on fossil fuels. Regenerative braking (RB) utilizes the kinetic energy generated by the motor during the deceleration, or braking, process. Therefore, recovering the braking energy is an effective approach for improving the driving range of an EV [1]-[4]. Usually, within traditional vehicles, all of the braking energy is lost in the form of heat due to friction losses. In RB, the motor acts as a generator and the kinetic energy is harvested by applying the proper switching schemes to the power converter switches. This harvested energy can be used to charge the vehicle's battery, or stored in an ultracapacitor bank [5]-[8]. A hybrid energy storage system can be used to alternate power generation and storage between an ultracapacitor and a battery, depending on the required power. Knowing the parameters of the system is essential to building a clear idea regarding the amount of energy harvested as opposed to that being generated.

The permanent-magnet synchronous motor (PMSM) is widely adopted as the traction motor in electric vehicles (EV) due to its high efficiency and high torque density. Vector control, also called field-oriented control (FOC), is a popular and powerful method in electrical drive applications. This control strategy is used to effectively control the PMSM motor torque and flux in order to force the motor to accurately track the command trajectory regardless of machine and load parameter variations, or any other external disturbances [9]. Electric braking control of the PMSM based on FOC is realized by requesting a negative q-axis current according to the braking torque demanded. The maximum amount of current produced while braking is calculated depending on many variables, including the motor speed and input voltage [10].

This paper is an extension of work originally presented in ICMSAO'17, where regenerative braking was analyzed for a DC motor with battery/supercapacitor energy storage [11]. The maximum amount of current produced by the DC motor while braking was calculated depending on the system variables such as motor speed, armature resistance, and input voltage. The theoretical analysis was next validated by experimental results. The effect of varying the duty cycle of the braking signal was also studied to find the optimal duty cycle that gives the best efficiency in regenerative energy harvesting.

In this paper, electric braking is first analyzed for the PMSM under Field Oriented Control (FOC). Next, a dedicated maximum Energy Recovery Switching Scheme (MERSS) is developed to control the inverter switches during regenerative braking to maximize energy recovery. Regenerative braking energy calculations are confirmed by experimental results on a prototype

*Corresponding Author: Andrew Adib, Email: b00062741@aus.edu

PMSM motor drive system. The harvested energy is stored into the DC-bus capacitor. The efficiency with which energy can be returned to the ultracapacitor is analyzed for the FOC scheme and the MERSS. The analysis, benefits, limitations and experimental results of each control strategy are provided.

This paper is divided into 5 sections; section 2 presents the concept of regenerative braking under FOC and provides some theoretical analysis of the required motor current during this mode of operation. Section 3 presents the maximum energy recovery switching scheme. Section 4 presents the experimental results and discussion when braking the PMSM under FOC and MERSS. Finally, the conclusion is presented in Section 5.

2. Regenerative Braking of the PMSM Under FOC

In conventional passive dynamic braking the kinetic energy of electric machines is dissipated through the armature coils and the additional braking resistors. The principle of the regenerative braking with the voltage-source inverter is similar to that of the conventional dynamic braking, however; the additional resistors are eliminated and no additional power switches are used.

According to the driving conditions, the braking process can be classified as either constant speed braking or variable speed braking. The constant speed braking is usually required during the downhill road driving, while the variable speed braking often occurs during the general deceleration process.

2.1. System Setup

Figure 1 shows the general topology used in electric vehicle applications with hybrid energy storage, where the mechanical load is typically coupled to a Permanent-Magnet Synchronous Machine supplied by a battery source through an inverter. An ultracapacitor module is used as an auxiliary power source connected through a bi-directional DC-DC converter to the DC-link, thus making it possible to obtain an optimized charge/discharge operation mode.

2.2. Regenerative Braking principle of the PMSM

The FOC scheme for the PMSM is shown in Fig. 2. In FOC, the stator phase currents are measured and converted into a complex vector. This current vector is then transformed to a coordinate system rotating with the rotor of the machine.

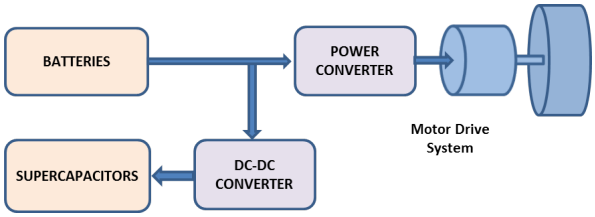


Figure 1: Topology of the Battery-Ultracapacitor Energy Storage System

The real x-axis component of the stator current vector i_d, in this rotor flux-oriented coordinate system, is used to control the rotor flux linkage. The imaginary y-axis component i_q is used to control the motor torque. For the PMSM, maximum torque–current control can be achieved by holding the d-axis current at zero ($i_d = 0$).

Electric braking control, based on FOC, is realized by requesting a negative q-axis current according to the braking torque demanded, or by controlling the speed to follow a ramp reference and gradually approach zero [6].

The input power and the electromagnetic power can be expressed as

$$P_{in} = V_d i_d + V_q i_q, \quad (1)$$

$$P_{em} = T_{em}\omega\,, \quad (2)$$

$$T_{em} = \frac{3}{2}\frac{p}{2}\lambda_{PM} i_q\,. \quad (3)$$

where, i_d and i_q are the d-axis and q-axis current components, V_d and V_q are the d-axis and q-axis voltage components, ω is the motor speed, T_{em} is the electromagnetic torque, λ_{PM} is the permanent magnet flux, and p is the number of poles.

The electric machine power losses are given as

$$P_{loss} = i_d^2 R + i_q^2 R. \quad (4)$$

where, R is the per-phase stator resistance. By balancing the input and output power

$$P_{in} = P_{em} + P_{loss}, \quad (5)$$

$$P_{in} = \frac{3}{2}\frac{p}{2}\lambda_{PM} i_q \omega + R(i_d^2 + i_q^2). \quad (6)$$

To find the regions of regenerative braking, P_{in} is set to 0. This means that no power is being drawn from the DC source.

Next, an expression for i_q in terms of i_d can be found

$$i_q = \frac{-\frac{3p}{22}\lambda_{PM}\omega \pm \sqrt{\left(\frac{3p}{22}\lambda_{PM}\omega\right)^2 - 4R^2 i_d^2}}{2R}. \quad (7)$$

The equivalent electromagnetic torque is given as

$$T_{em} = \frac{3}{2}\frac{p}{2}\lambda_{PM}\left(\frac{-\frac{3p}{22}\lambda_{PM}\omega \pm \sqrt{\left(\frac{3p}{22}\lambda_{PM}\omega\right)^2 - 4R^2 i_d^2}}{2R}\right). \quad (8)$$

This equation describes the electromagnetic torque inside the regenerative braking region. To find the boundaries of this region, the above equation is solved for the maximum and minimum braking torque.

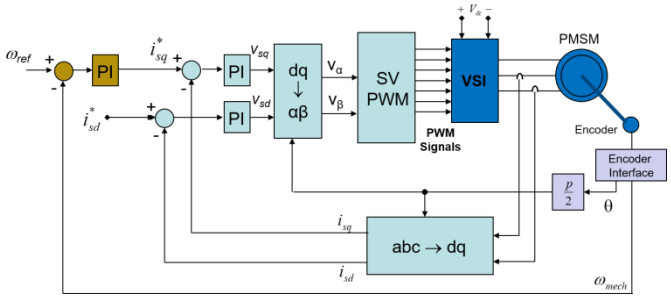


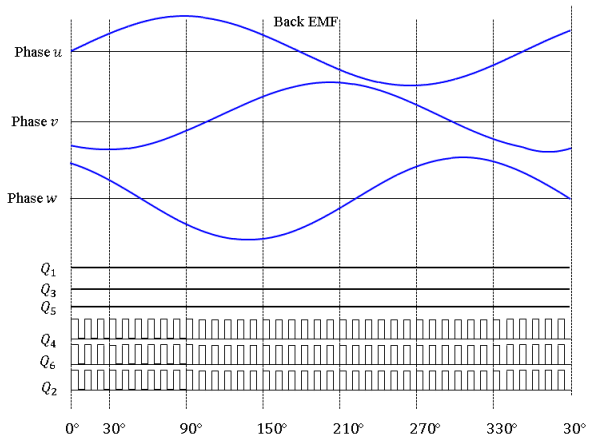
Figure 2: Field Oriented Control Scheme of the PMSM

$$\frac{dT_{em}}{di_d} = 0. \quad (9)$$

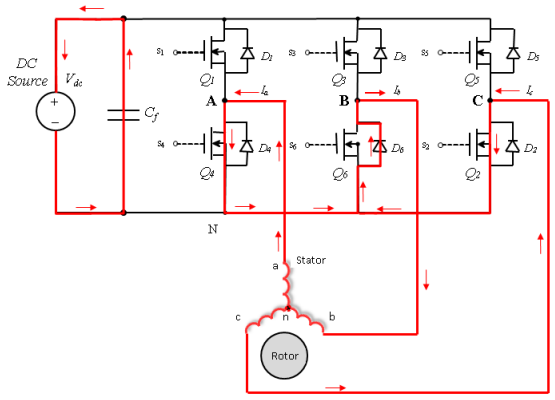
Therefore,

$$i_d = 0. \quad (10)$$

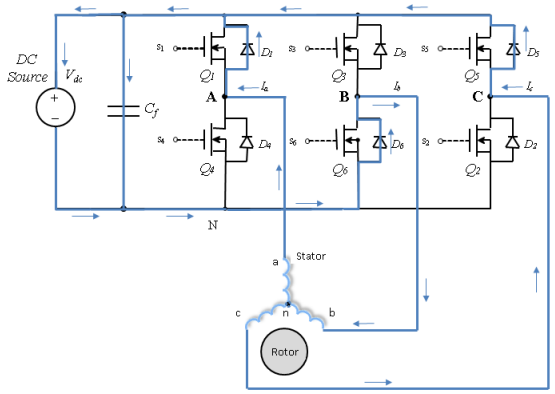
Substituting this result in equation (7) gives



(a)



(b)



(c)

Figure 3: Regenerative braking scheme. (a) Emf and PWM switching signals, (b) Current flow during ON-Time and 0-30° period. (c) Current flow during OFF-Time and 0-30° period

$$i_q = -\frac{3P\lambda_{pm}}{4R}\omega. \tag{11}$$

The equivalent electromagnetic torque is then given by

$$T_{em} = -\frac{\left(\frac{3}{4}p\lambda_{pm}\right)^2}{R}\omega. \tag{12}$$

This equation defines the minimum electromagnetic torque needed by the PMSM to operate the machine in the regenerative braking region.

Next, to find the maximum regenerative braking current absorbed, the input power is minimized.

$$\nabla P_{in} = \begin{bmatrix} \frac{\partial P_{in}}{\partial i_d} \\ \frac{\partial P_{in}}{\partial i_q} \end{bmatrix} = \begin{bmatrix} 2Ri_d \\ \frac{3}{2}\frac{p}{2}\lambda_{pm}\omega + 2Ri_q \end{bmatrix}. \tag{13}$$

The minimum power is obtained by setting this gradient to zero and solving for both variables, i_d and i_q.

$$i_d = 0, \tag{14}$$

$$i_q = -\frac{3p\lambda_{pm}}{8R}\omega. \tag{15}$$

The equivalent electromagnet torque generated by these current commands is given by

$$T_{em} = -\frac{9p^2\lambda_{pm}^2}{32R}\omega. \tag{16}$$

This torque guarantees maximum absorbed current by the DC source during regenerative braking.

3. Regenerative Braking with the Maximum Energy Recovery Switching Scheme (MERSS)

The switching scheme of the inverter is developed with the objective of maximizing energy recovery during regenerative braking. The idea is to utilize the motor phase inductors along with the inverter switches, functioning as a boost converter and allowing the phase currents to reverse their direction to flow back to the DC-bus capacitor [12], [14]. Regenerative braking is achieved by controlling only the lower switches (Q_2, Q_4, Q_6) through PWM and switching OFF all of the upper switches (Q_1, Q_3, Q_5) as shown in Fig. 3. In this mode of operation, the three lower switches are controlled with the same PWM command signal. During the ON-time, a path is provided for the phase current to flow in the negative direction through the closed switch, or in the positive direction through the free-wheeling diode depending on the polarity of the back emf.

Figure 3a shows the phase relationship between the back EMF, armature current, and the switching signals. Figure 3b and Figure 3c both show the closed loop path of the 3-phase currents during the 0-30s time interval of each cycle. During the OFF-time, the phase currents must maintain their direction and are therefore forced to flow through an alternate path created by the upper free-wheeling diodes, $D_1 - D_3$ and then back to the DC-bus capacitor. In this time period, regenerative braking is achieved and the capacitor is charged by the recovered electrical energy.

4. Experimental Testing and Discussions

In this section, experimental results of the regenerative braking process are analyzed to confirm the theoretical analysis. The energy recovered by the DC-bus capacitor is compared to the motor's mechanical energy to evaluate the efficiency of the

regenerative braking process. Figure 4 shows the topology of the PMSM experimental setup. The system consists of a PMSM, which is controlled using a MYWAY MWINV-9R144 inverter. The inverter switches are controlled using the dSPACE 1103 board. The PMSM is coupled with a Bühler DC Motor through flexible couplings and additional disc inertia mounted on the same shaft. The DC motor acts as a mechanical load and is controlled using a DC-DC converter, which in turn is controlled using a dSPACE 1104 board. Two encoders are used: the first is an incremental encoder directly connected to the DC motor side, and the second is a sine/cosine encoder connected to the PMSM side. The system parameters are listed in Table 1.

In the following sections, regenerative braking is implemented on the PMSM machine controlled with FOC. Two methods are implemented to realize regenerative braking. The first method is based on speed control and uses a ramp speed reference with negative slope to brake the motor. This linear deceleration simulates the braking of an EV, where the speed decreases gradually until it reaches 0. The braking time is adjusted by controlling the slope of the reference speed.

The second method operates the PMSM in torque control mode. The speed control loop is disconnected from the vector control scheme and the current commands, i_d^* and i_q^*, are generated directly. The i_q^* command will generate a negative torque that brings the motor to a stop. As discussed in section II-B, generating this torque guarantees maximum current absorption by the DC source, allowing for the returned current to achieve a higher amplitude and therefore the recovered energy is maximized.

Table 1: Motor drive system parameters

Parameter	*Value*
Resistance (R_{p-p})	6.8 Ω
Total inertia (J)	0.00315 Kg.m^2/s^2
Back EMF constant (K_b)	98 $V/krpm$
Torque constant (K_t)	1.6 Nm/A
Rated Torque	3.9 Nm
Stall Current	2.7 A
Inductance (L_{p-p})	24.3 mH
Damping coefficient (B)	4.741x10^{-4} Nm/(rad/s)
Coulomb friction (τ_c)	0.1343 Nm

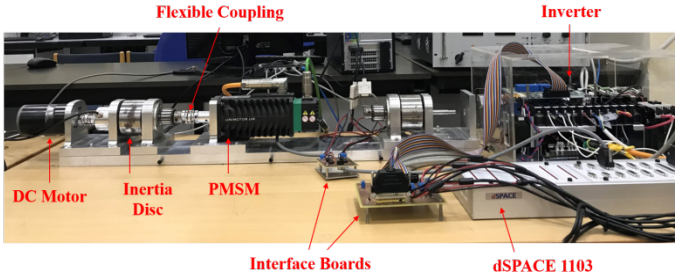
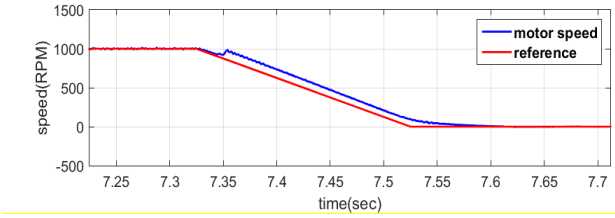


Figure 4: PMSM drive system

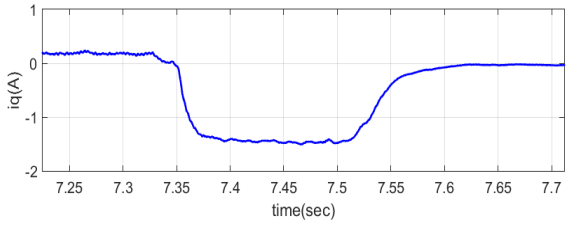
In order to observe the voltage increase during regenerative braking, the three-phase line is disconnected from the inverter and the DC-link voltage is maintained only by the DC-bus capacitor. As a result, the voltage begins decreasing. In this period of time, the motor is operating under constant speed. Once the DC-link voltage reaches 300V, the regenerative braking command signal is triggered and the motor operates under braking mode.

4.1. Regenerative braking using speed control mode

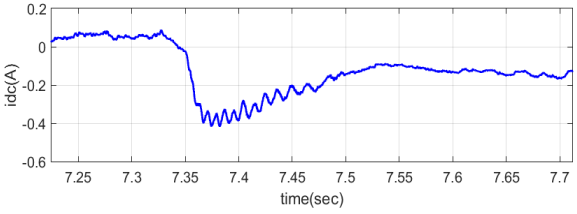
Figure 5a shows the reference speed and actual speed of the PMSM before and after regenerative braking is activated. The motor decelerates gradually to zero with a breaking time set to 0.2s. During the regenerative braking period, a negative q-axis current, i_q^*, is generated by the FOC controller as shown in Figure 5b. The motor mechanical energy is supplied back to the capacitor. Figure 5c shows a negative DC-link current indicating that energy is flowing from the motor to the capacitor. Figure 5d shows that the capacitor voltage is being charged during this period.



(a) Reference speed and actual speed



(b) q-axis current



(c) DC-link current

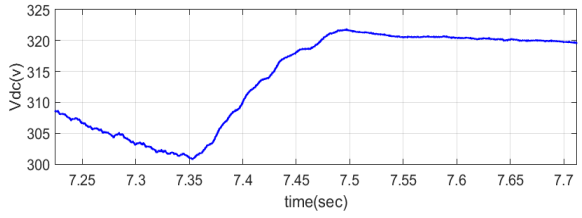


Figure 5: PMSM transient response during regenerative braking under the speed control mode of operation

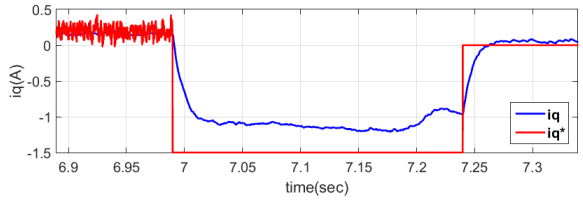
Table 2 gives a summary of the mode of operation for this form of regenerative braking. The total recovered energy is 70.496% of the motor energy.

Table 2: Summary of the regenerative braking performance for the speed control mode of operation

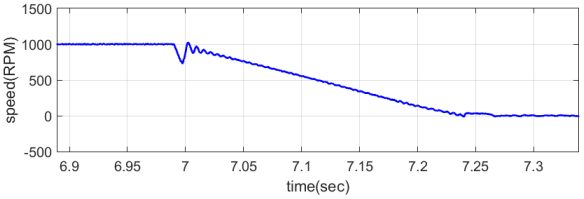
ω_0 (rpm)	Max i_{dc} (A)	ΔV_{dc} (V)	Recovery time (t_r)	Recovered Energy (J)	Braking Power (W)	Max i_q (A)	Mech. Energy (J)	Efficiency (%)
1000	−0.419	21.4	0.145	12.18	128.5	−1.508	17.27	70.50

Table 3: Summary of the regenerative braking statistics of the torque control mode of operation

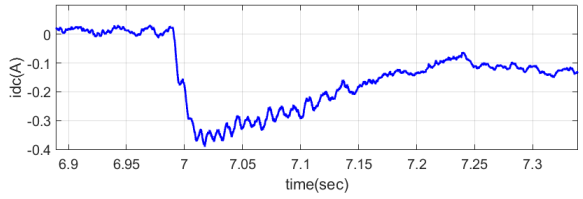
ω_0 (rpm)	Max i_{dc} (A)	ΔV_{dc} (V)	Recovery time (t_r)	Recovered Energy (J)	Braking Power (W)	Max i_q (A)	Mech. Energy (J)	Efficiency (%)
1000	−0.391	21.6	0.242	16.48	119.3	−1.218	17.27	95.43



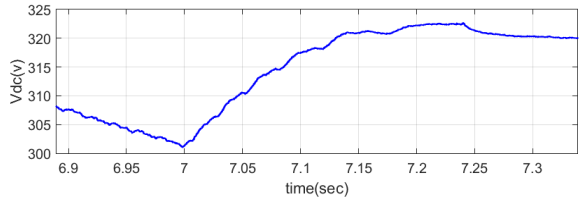
(a) Actual and command q-axis current



(b) Motor speed



(c) DC-link current



(d) DC-bus voltage

Figure 6: PMSM transient response during regenerative braking under the torque control mode of operation

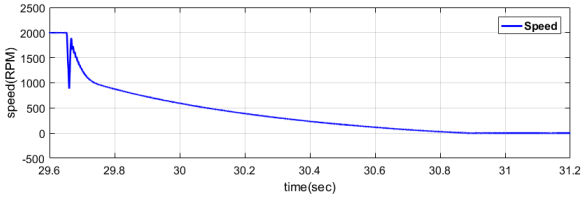
4.2. Regenerative braking using torque control

Figure 6 shows the motor variables during the torque mode of operation. Initially, the motor speed is regulated by FOC to the desired reference value. Once regenerative braking is initiated, the speed controller is disabled and the motor torque is controlled through the q-axis current. Figure 6a shows the reference current, i_q^*, and the generated q-axis current, i_q.

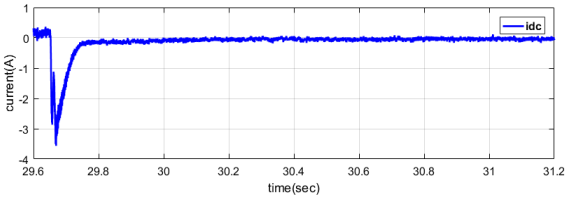
The braking time must be tuned in order for the speed to reach zero at the end of the breaking time. Table 3 summarizes the regenerative braking event results under the torque control mode of operation. The results indicate that greater amounts of energy are harvested with the torque control operational mode in comparison to the case of the speed controlled braking.

4.3. Regenerative braking using the Maximum Energy Regeneration switching scheme (MERSS)

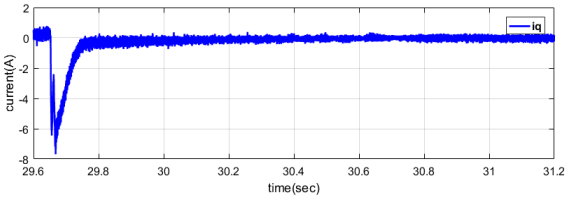
To evaluate the PMSM performance with MERSS, the motor is initially powered by the DC-bus capacitor to operate the drive system in driving mode. Next, a break command is activated to operate the system in regenerative braking mode.



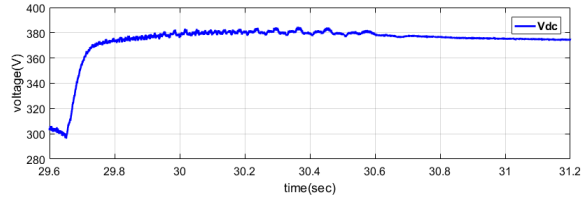
(a) Motor speed



(b) q-axis current



(c) DC-link current



(d) DC-bus voltage

Figure 7: PMSM transient response during regenerative braking under MERSS mode of operation. ω_0=2000 rpm, d=0.7

As a result, the phase currents reverse their direction and provide energy back to the DC-bus capacitor. The DC-bus voltage increases allowing the DC-bus capacitor to be charged.

Figure 7 shows the transient response of motor speed, q-axis motor current, DC-bus current, and DC-bus voltage. The motor is initially running at constant speed and the DC-bus capacitor is supplying power. When the break command is received, the motor starts decelerating until it comes to rest. The regenerative braking region is a subset of this period where the kinetic energy of the

rotor is used to generate electrical energy, and the motor acts as a generator. Figure 8 shows the electric power and energy recovered by the DC-bus capacitor. Following the braking command, energy recovery is activated through the MERSS and the capacitor voltage increases as current is absorbed by the capacitor. Regenerative breaking stops when the motor current reaches the minimum level set by the speed, as given through equation (11).

This process was repeated for different motor speeds and different PWM duty cycles. Varying the duty cycle of the brake command will change the ON-OFF times of the lower inverter switches, as explained previously in Figure 3. Increasing the ON-time period will charge the motor's inductance for a longer time allowing for the storage of additional energy. This process is similar to the operation of a boost converter. However, if it the ON-period is heavily increased, energy will not be fully recovered due to a short OFF-time period, and will eventually be lost in the switching process. For this reason, the duty cycle must be carefully selected for optimum energy recovery. Figure 9 shows the regenerated power and energy as a function of duty cycle for different motor speeds.

The maximum regenerated energy is compared with the mechanical energy to evaluate the efficiency of the system under MERSS. The energy of the motor and the energy stored in the ultracapacitor during braking are given by:

$$\mathrm{E_{mech}} = \frac{1}{2}\mathrm{J}\omega_0^2, \tag{17}$$

$$E_{cap}(k) = E_{cap}(k-1) + \frac{T_s}{2}[P_{cap}(k) + P_{cap}(k-1)], \tag{18}$$

where, ω_0 is the initial speed of the motor when the braking process starts. Therefore, the efficiency of the braking process takes into consideration the power losses during switching.

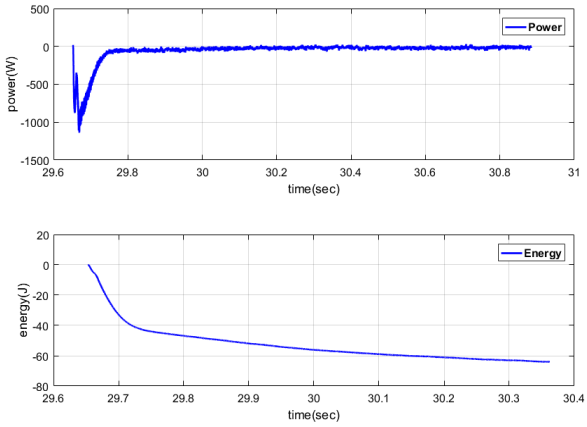


Figure 8: Electric power and energy recovered by the DC bus with MERSS. ω_0=2000 rpm, d=0.7

Table 4 summarizes the results from this experiment and displays the efficiency, as well as the energy harvested. It can be observed that the optimum duty cycle that yields maximum energy recovery is a function of the speed.

The performance of the MERSS is next compared to FOC by conducting the regenerative braking experiments with FOC under the same operating conditions. Table 5 shows a summary of the results. It can be observed that MERSS yields always a better efficiency than FOC.

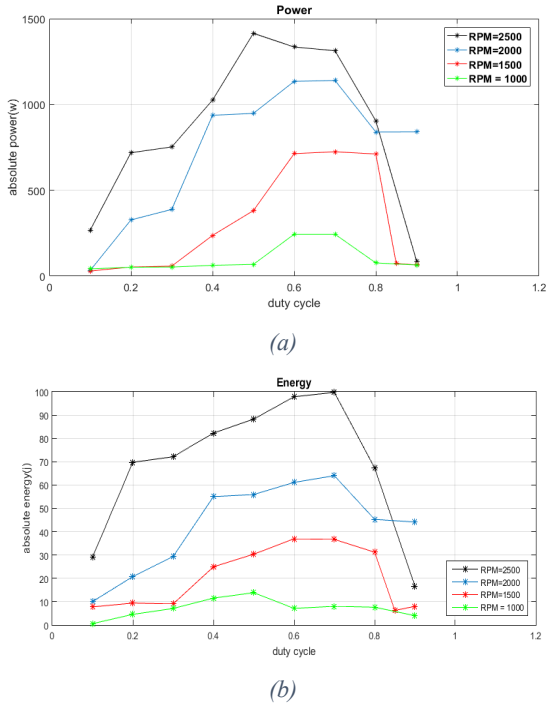


Figure 9: Electric power and energy recovered on the DC bus with MERSS

$$\eta = \frac{E_{cap}}{E_{mech}}. \tag{19}$$

Table 4: Maximum efficiency of braking at different motor speeds with MERSS.

ω_0 (rpm)	Duty Cycle	$E_{mech}(J)$	E_{cap} (J)	η (%)
1000	0.5	17.27	13.84	80.14
1500	0.6	38.86	36.96	95.11
2000	0.7	69.09	64.00	92.64
2500	0.7	107.95	99.74	92.40

Table 5: Comparative analysis between FOC and MERSS

	ω_0 (rpm)	Max i_{dc} (A)	ΔV_{dc} (V)	Max i_q (A)	E_{mech} (J)	E_{cap} (J)	Efficiency (%)
FOC	1000	-0.419	21.4	-1.508	17.27	12.18	70.50
	1500	-1.005	54.9	-2.312	38.86	27.73	71.36
	2000	-1.575	96.8	-2.958	69.09	50.62	73.27
MERSS	1000	-0.2203	16.6	-0.740	17.27	13.84	80.14
	1500	-2.13	52.0	-4.968	38.86	36.96	95.11
	2000	-3.564	85.9	-7.683	69.09	64.00	92.64

5. Conclusion

This paper discusses regenerative braking in a PMSM drive system. A new maximum energy recovery switching scheme is developed and compared to FOC. Analysis is carried out on an experimental setup to confirm the effectiveness of regenerative braking under the new scheme. Experimental results show that the maximum current absorbed by the ultracapacitor does not exceed a set limit which is in turn dependent on the ultracapacitor's voltage, internal resistances of the system and the motor speed. The effect of varying the duty cycle is also studied to uncover the

highest efficiency duty cycle for optimum harvesting of regenerated energy.

References

[1] M. K. Yoong, Y. H. Gang, G. D. Gan, C. K. Leong, Z. Y. Phuan, B. K. Cheah, and K. W. Chew, "Studies of regenerative braking in electric vehicle" in 2010 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT), Petaling Jaya, Malaysia, 2010. https://doi.org/10.1109/STUDENT.2010.5686984.

[2] M.-Ji Yang, H.-L. Jhou, B.-Y. Ma, and K.-K. Shyu, "A Cost-Effective Method of Electric Brake With Energy Regeneration for Electric Vehicles" Ming-Ji Yang, Hong-Lin Jhou, Bin-Yen Ma, and Kuo-Kai Shyu, Member, IEEE Trans. Ind. Electronics, 56(6), 2203 - 2212, June 2009. https://doi.org/10.1109/TIE.2009.2015356.

[3] D. Lu, M. Ouyang, J. Gu, and J. Li, "Instantaneous optimal regenerative braking control for a permanent-magnet synchronous motor in a four-wheel-drive electric vehicle" Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 228(8), 894-908, July 2014. https://doi.org/10.1177/0954407014521173.

[4] C.-H. Chen, W.-C. Chi, and M.-Y. Cheng, "Regenerative braking control for light electric vehicles" in IEEE 9th International Conference on Power Electronics and Drive Systems (PEDS), Singapore, 2011. http:// https://doi.org/10.1109/PEDS.2011.6147317.

[5] F. Naseri; E. Farjah; T. Ghanbari, "An Efficient Regenerative Braking System Based on Battery/Ultracapacitor for Electric, Hybrid and Plug-In Hybrid Electric Vehicles with BLDC Motor" IEEE Trans. Vehicular Technology, 66(5), 3724–3738, 2017. https://doi.org/10.1109/TVT.2016.2611655.

[6] Y. Bian, L. Zhu. H. Lan, A, Li. Anhu, and X. Xu, "Regenerative Braking Strategy for Motor Hoist by Ultracapacitor" Chin. J. Mech. Eng., 25(2), 377–384, March 2012. https://doi.org/10.3901/CJME.2012.02.377.

[7] Z. Zhang, X. Zhang, W. Chen, Y. Rasim, W. Salman, H. Pan, Y. Yuan, and C. Wang, "A high-efficiency energy regenerative shock absorber using supercapacitors for renewable energy applications in range extended electric vehicle" Applied Energy, 178, 177–188, 2016. https://doi.org/10.1016/j.apenergy.2016.06.054.

[8] S. Ding, M. Cheng, Chao Hu, Guishu Zhao and Wei Wang, "An energy recovery system of regenerative braking based permanent magnet synchronous motor for electric vehicles" in 2013 International Conference on Electrical Machines and Systems (ICEMS), Busan, South Korea, 2013. https://doi.org/10.1109/ICEMS.2013.6754468.

[9] Ned Mohan, Advanced Electric Drives: Analysis, Control and Modeling Using Simulink, Wiley 2014.

[10] A. Samba Murthy, "Analysis of regenerative braking in electric machines," MSc Thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, May 2013.

[11] A. Adib and R. Dhaouadi, "Modeling and analysis of a regenerative braking system with a battery-supercapacitor energy storage" in 7th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO), Sharjah, UAE, 2017. https://doi.org/10.1109/ICMSAO.2017.7934897.

[12] A. S. Murthy, D. P. Magee and D. G. Taylor, "Vehicle braking strategies based on regenerative braking boundaries of electric machines" in 2015 IEEE Transportation Electrification Conference and Expo (ITEC), Dearborn, MI, 2015. https://doi.org/10.1109/ITEC.2015.7165809.

[13] T.-H. Kim, J.-H. Lee, and C.-Y. Won, "Design and control methods of bidirectional DC-DC converter for the optimal DC link voltage of PMSM drive" Journal of Electrical Engineering and Technology, 9(6), 1944-1953, 2014. http://dx.doi.org/10.5370/JEET.2014.9.6.1944.

[14] X. Jiaqun and C. Haotian, "Regenerative brake of brushless DC motor for light electric vehicle" in 2015 18th International Conference on Electrical Machines and Systems (ICEMS), Pattaya, Thailand, 2015. https://doi.org/10.1109/ICEMS.2015.7385262.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 467-471 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Tracking and Detecting moving weak Targets

Naima Amrouche[*,1,2], Ali Khenchaf[2], Daoud Berkani[1]

[1]*Ecole Nationale Polytechnique d'Alger, 10 Rue Frère Oudek, Elharrach, Algérie, 16200*

[2]*Lab-STICC CNRS UMR 6285, ENSTA Bretagne, 02 Rue François Verny, Brest Cedex 09, France, 29806*

ARTICLE INFO

Article history
Received: 13 December, 2017
Accepted: 24 January, 2018
Online: 10 February, 2018

Keywords :
Target Tracking
Track Before Detect
Detection
Particle Filter

ABSTRACT

Detect and tracking of moving weak targets is a complicated dynamic state estimation problem whose difficulty is increased in case of high clutter conditions or low signal to noise ratio (SNR). In this case, the track-before-detect filter (TBDF) that uses unthresholded measurements considers as an effective method for detecting and tracking a single target under low SNR conditions. In this paper, a particle filter based track-before-detect (PF-TBD) method is proposed to address the problem of detection and tracking with unthersholded data and a binary variable of the existence of the target for two motion models. Simulation results using image measurements based on TBD scenarios are also presented to demonstrate the capability of the proposed approach.

1. Introduction

The classical approach to target tracking is based on target measurements (position, range rate, and so forth) that are extracted by thresholding the output of a signal processing unit of a surveillance sensor [1].The primary role of thresholding is to reduce the data flow and thus simplify tracking. For a target of a certain signal-to-noise ratio (SNR), the choice of the detection threshold determines the probability of target detection and the density of false alarms. The false alarm rate, on the other hand, affects the complexity of the data association problem in the tracking system. In general, higher densities of false alarms require more sophisticated data association algorithms.

The undesirable effect of thresholding the sensor data, however, is that in restricting the data flow, it also throws away potentially useful information. For high SNR targets this loss of information is of little concern because one can achieve good probability of detection with a small false alarm rate. Recent developments of stealthy military aircraft and cruise missiles have emphasized the need to detect and track low SNR targets. For these dim (stealthy) targets, there is a considerable advantage in using the unthresholded data for simultaneous detection and track initiation [2], [3]. Depending on the type of sensor in use, the unthresholded data can be a sequence of range-Doppler maps, bearing-frequency distributions.

[*]Naima Amrouche, 02 rue François Verny, Tel : +33 2 98 34 88 00 &
Email: naima.amrouche@ensta-bretagne.org

The concept of simultaneous detection and tracking using unthresholded data is known in literatures as track-before-detect (TBD) approach. Typically TBD is implemented as a batch algorithm using the Hough transform [4], dynamic programming [2] [3] or maximum likelihood estimation [5].

TBD algorithms based on the Hough transformation, dynamic programming or maximum likelihood methods are generally computationally intensive [6]. With recent advancement in Sequential Monte Carlo techniques, TBD algorithms implemented using PF are now computationally feasible [7] [8].

In this paper we also develop a recursive Bayesian TBD estimator; however, our formulation and implementation are based on the particle filter [7] [9]. The PF based TBD incorporates unthresholded data and a binary target existence variable into the target state estimation process. The presence and absence of target are explicitly modelled [10] [11]. This concept, allows us to calculate the probability of existence of the target directly from the filter.

The paper is organized as follows: section 2 the system dynamics and measurement model, are introduced for the TBD application. In section 3 formulates the TBD approach as a nonlinear filtering problem and describes the conceptual recursive Bayesian solution. The implementation of this solution using a particle filter is presented in section 4. In section 5 collects our simulations and results. Finally, we report our conclusions and direction for future research in section 6.

2. Formulation Problem

2.1. Dynamic Model

We assume that want to track a target moving in a 2-D plane with an unknown state vector s_k at time step k. We consider the state model given by:

$$s_{k+1} = Fs_k + v_k \quad (1)$$

Where F is the state transition matrix, assuming constant velocity motion and coordinate turn motion respectively, k is the discrete-time index, v_k is the process noise sequence, and s_k is the state vector defined as:

$$s_k = [x_k \quad \dot{x}_k \quad y_k \quad \dot{y}_k \quad I_k] \quad (2)$$

Here $(x_k, y_k), (\dot{x}_k, \dot{y}_k)$ and I_k denote the position, velocity, and the intensity of the target, respectively.

2.2. Transition Matrix

A target can be present or absent from the surveillance region at a discrete-time k. Target presence variable E_k is modelled by a two-state Markov chain, that is $E_k = \{0,1\}$. Here 0 denotes the event that a target is not present, while 1 denotes the opposite. Furthermore, we assume that transitional probabilities of target "birth" P_b and "death" P_d, defined as:

$$P_b \triangleq P\{E_k = 1 | E_{k-1} = 0\} \quad (3)$$

$$P_d \triangleq P\{E_k = 0 | E_{k-1} = 1\} \quad (4)$$

Are known, the other two transitional probabilities of this Markov chain, the probability of staying alive $1 - P_d$ and the probability of remaining absent $1 - P_b$ respectively, are given by:

$$1 - P_d \triangleq P\{E_k = 1 | E_{k-1} = 1\} \quad (5)$$

$$1 - P_b \triangleq P\{E_k = 0 | E_{k-1} = 0\} \quad (6)$$

The corresponding transition matrix for the Markov process is:

$$\Pi = \begin{bmatrix} 1 - P_b & P_b \\ P_d & 1 - P_d \end{bmatrix} \quad (7)$$

2.3. Sensor Model

The sensor provides a sequence of two-dimensional images (frames) of the surveillance region, each image consisting of $(n \times m)$ resolution cells (pixels). A resolution cell corresponds to a rectangular region of dimensions $\triangle_x \times \Delta_y$ so that the center of each cell (i, j) is defined to be at $(i\Delta_x \times j\Delta_y)$ for $i = 1, \dots n$ and $= 1, \dots m$.

At each resolution cell (i, j) the measured intensity is denoted as $z_k^{(i,j)}$ and modeled as:

$$z_k^{(i,j)} = \begin{cases} h_k^{(i,j)}(s_k) + \omega_k^{(i,j)} & if\ target\ present \\ \omega_k^{(i,j)} & if\ target\ absent \end{cases} \quad (8)$$

Where $h_k^{(i,j)}(s_k)$ is the target contribution to intensity level in the resolution cell (i, j) and $\omega_k^{(i,j)}$ is measurement noise in the resolution cell (i, j), assumed to be independent from pixel to pixel and from frame to frame. Thus for a point target of intensity I_k at position(x_k, y_k), the contribution to pixel(i, j) is approximated as:

$$h_k^{(i,j)}(s_k) \approx \frac{\Delta_x \Delta_y I_k}{2\pi\Sigma^2} \exp\left\{-\frac{(i\Delta_x - x_k)^2 + (j\Delta_y - y_k)^2}{2\Sigma^2}\right\} \quad (9)$$

Where Σ is the amount of blurring introduced by the sensor. The complete measurements recorded at time k a $n \times m$ matrix denoted as:

$$z_k = \left\{z_k^{(i,j)} : i = 1, \dots, n, j = 1, \dots, m\right\} \quad (10)$$

While the set of complete measurements collected up to time k is denoted as usual: $Z_k = \{z_i, i = 1, \dots k\}$.

3. Bayesian Solution to TBD Filtering

The formal recursive Bayesian solution can be presented as a two-step procedure, consisting of prediction and update.

3.1. Prediction

The predicted target state can be written in terms of the target state and existence at the previous time, giving

$$\begin{gathered} p(s_k, E_k = 1 | Z_{k-1}) = (1 - \\ P_d) \int p(s_k | s_{k-1}, E_k = 1, E_{k-1} = 1) \times \\ p(s_{k-1}, E_{k-1} = 1 | Z_{k-1}) ds_{k-1} + \\ P_b \int p_b(s_k) p(s_{k-1}, E_{k-1} = 0 | Z_{k-1}) ds_{k-1} \end{gathered} \quad (11)$$

The pdf $p_b(s_k)$ denotes the initial target density on its appearance.

3.2. Update

The update equation in the Bayesian framework is given by:

$$\begin{gathered} p(s_k, E_k = 1 | Z_k) = \\ \frac{p(z_k | s_k, E_k = 1) p(s_k, E_k = 1 | Z_{k-1})}{p(z_k | Z_{k-1})} \end{gathered} \quad (12)$$

Where prediction density $p(s_k, E_k = 1 | Z_{k-1})$ is given by (11) and $p(z_k | s_k, E_k)$ is the likelihood function given by:

$$\begin{gathered} p(z_k | s_k, E_k) = \\ \begin{cases} \prod_{i=1}^{n} \prod_{j=1}^{m} p_{S+N}\left(z_k^{(i,j)} \middle| s_k\right), & for \quad E_k = 1 \\ \prod_{i=1}^{n} \prod_{j=1}^{m} p_N\left(z_k^{(i,j)}\right), & for \quad E_k = 0 \end{cases} \end{gathered} \quad (13)$$

Here $p_N\left(z_k^{(i,j)}\right)$ is the probability density function of background noise in pixel (i, j), while $p_{S+N}\left(z_k^{(i,j)} \middle| s_k\right)$ is the likelihood of target signal plus noise in pixel (i, j),given that the target is in state s_k, This two probability density function can be further expressed as:

$$p_N\left(z_k^{(i,j)}\right) = \mathcal{N}\left(z_k^{(i,j)}, 0, \sigma^2\right) \quad (14)$$

$$p_{S+N}\left(z_k^{(i,j)}\middle|s_k\right)=\mathcal{N}\left(z_k^{(i,j)},h_k^{(i,j)},\sigma^2\right) \quad (15)$$

Since the target (if present) will affect only the pixels in the vicinity of its location (x_k,y_k), the expression for $p(z_k|s_k,E_k=1)$ can be approximated as follows:

$$p(z_k|s_k,E_k=1)\approx \prod_{i\in C_i(s_k)}\prod_{j\in C_j(s_k)} p_{S+N}\left(z_k^{(i,j)}\middle|s_k\right)\cdot \prod_{i\notin C_i(s_k)}\prod_{j\notin C_j(s_k)} p_N\left(z_k^{(i,j)}\right) \quad (16)$$

Where $C_i(s_k)$ and $C_j(s_k)$ are the sets of subscripts i and j, respectively, corresponding to pixels affected by the target.

4. A Particle Filter for Track Before Detect (PF-TBD)

The recursive Bayesian solution of the track problem described in the previous section can be implemented using a particle filter [7] [9] [12] [13]has some similarities to the MMPF [14]. In this case we introduce the augmented state vector to include the existence variable. $y_k=[\mathrm{s}_k^T \quad E_k]^T$. Let us denote a random measure that characterizes the posterior probability density function at $k-1$, namely $p(y_{k-1}|Z_{k-1})$, by $\{y_{k-1}^n,\omega_{k-1}^n\}_{n=1}^N$.As usual, N is the number of particles, while y_k^n consists of s_k^n and E_k^n.The pseudocode of a single cycle of the PF developed for the TBD problem is presented in Table 1. The next step is the prediction of particle target states; this is done, however, only for those particles that are characterized by $E_k^n=1$.For remaining particles (with $E_k^n=0$), the target state components are undefined. There are two possible cases here:

4.1. Newborne Particles

This group of predicted particles is characterized by the transition from $E_{k-1}^n=0$ to $E_k^n=1$.The target state particles are uniformly drawn at time step k based on some a priori information on the minimum and maximum possible values on the target state.

4.2. Existing Particles

This group of particles that continues to stay "alive", with $E_{k-1}^n=1$ to $E_k^n=1$.The state transition model in equation (1) is used to update the target state particles.

The importance weights are computed next. For this purpose we need to introduce the likelihood ratio in pixel (i,j) for a target in state s_k^n, defined as:

$$\ell\left(z_k^{(i,j)}\middle|s_k^n\right)\triangleq\frac{p_{S+N}\left(z_k^{(i,j)}\middle|\mathrm{s}_k^n\right)}{p_n\left(z_k^{(i,j)}\right)} \quad (17)$$

$$\ell\left(z_k^{(i,j)}\middle|s_k^n\right)=\exp\left\{-\frac{h_k^{(i,j)}\left(h_k^{(i,j)}-2z_k^{(i,j)}\right)}{2\sigma^2}\right\} \quad (18)$$

Where $h_k^{(i,j)}$ was defined in (9). Equation (18) follows from (14), (15), and (11). The importance weights (up normalizing constant) are now given by [7]:

$$\tilde{\omega}_k^n=\begin{cases}\prod_{i\in C_i(s_k^n)}\prod_{j\in C_j(s_k^n)}\ell\left(z_k^{(i,j)}\middle|\mathrm{s}_k^n\right) & if \quad E_k^n=1\\ 1 & if \quad E_k^n=0\end{cases} \quad (19)$$

$[\{y_k^n\}_{n=1}^N]$ =TBD-PF$\left[\{y_{k-1}^n\}_{i=1}^N,z_k\right]$

- Target existence transitions using the Regime Transition Algorithm given in [6]

$[\{E_k^n\}_{n=1}^N]$ =RT$[\{E_{k-1}^n\}_{n=1}^N,\Pi]$

- FOR $i=1:N$
  - IF a newborn particle ($E_{k-1}^n=0$ and $E_k^n=1$) Draw $s_k^n\sim q_b(s_k|z_k)$
  - IF an existing particle ($E_{k-1}^n=1$ and $E_k^n=1$) Draw $s_k^n\sim q(s_k|s_{k-1}^n,z_k)$
  - Evaluate importance weight using (13)
- END FOR
- Calculate total weight: t =SUM$[\{\tilde{\omega}_k^n\}_{n=1}^N]$
- FOR $n=1:N$
  - Normalize: $\omega_k^n=t^{-1}\tilde{\omega}_k^n$
- END FOR
- Resample using systematic resampling algorithm given in [6]

$[\{y_k^n,-,-\}_{n=1}^N]$ =RESAMPLE$[\{y_k^n,\omega_k^n\}_{n=1}^N]$

The PF for track-before-detect performs target detection using the estimate of the posterior probability of target existence. This estimate is computed as:

$$\hat{P}_k=\frac{\sum_{n=1}^N E_k^n}{N} \quad (20)$$

And satisfies $0<\hat{P}_k<1$. Target presence is declared if $\hat{P}_k$ is above a certain threshold value. This declaration can then trigger the initialization of a track based on the estimated target state

$$\hat{s}_{k/k}=\frac{\sum_{n=1}^N s_k^n\cdot E_k^n}{\sum_{n=1}^N E_k^n} \quad (21)$$

5. Simulation and Results

In the simulation we used two scenarios for target motion and random walk model is adopted for target intensity.

5.1. Scenario1

The first model is a nearly constant velocity model is used. The dynamic model [15] for the target can be described by (1).

Where $F=diag[F_1,F_1]$,$Q=diag[Q_1,Q_1,q_2T]$

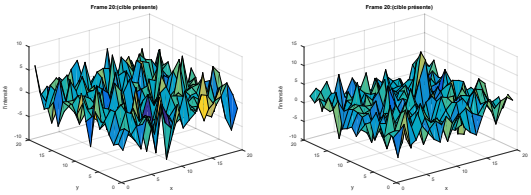
and $F_1=\begin{bmatrix}1 & T\\ 0 & 1\end{bmatrix}$, $Q_1=q_1\cdot\begin{bmatrix}T^3/3 & T^2/2\\ T^2/2 & T\end{bmatrix}$

Where q_1 and q_2 denote the level of process noise in target motion and intensity, respectively. A sequence of 30 frames of data has been generated with the following parameters:

$\Delta_x=\Delta_y=1, n=m=20, T=1s, \sigma=3, \Sigma=0.7$. The target is absent from frame 1 to frame 5 to be introduced in frame 6 with the initial intensity I_0 .The initial state is $[4.2 \quad 0.45 \quad 7.2 \quad 0.25 \quad I_0]$ [6] [16] [17].

The simulations are conducted under an initial intensity $I_0 = 9, 13$ and 25, which corresponds to an SNR of 3.18,6.71, and 12 dB, respectively, according to the calculation equation $SNR = 10log\left[\frac{I\Delta_x\Delta_y/2\pi\Sigma^2}{\sigma}\right]^2$. The target exists until frame 24 and is again absent in frames 25, 26,…, 30.

Figure 1 (a) and (b) show the measurement frame at time step 20 for 6.71dB and 12 dB peak respectively.



(a) (b)

Fig. 1. Measurements Frame (20): (a) for 6.71 dB Peak SNR, (b) for 12 dB Peak SNR for CV model.

The particle filter parameters are selected as follows: transitional probabilities $P_b = P_d = 0.05$;initial existence probability $\mu_1 = 0.05; v_{\min} = -1\ unit/s$; $v_{\max} = 1\ unit/s$; initial intensity range from $I_{min} = I_0 - 5$ to $I_{max} = I_0 + 5$; $p = 2$ and number of particles $N = 2000$.

In figure (2) the probability of presence is shown for a SNR of 6.71 dB. Existence probability remains very stable and above 0.97 until frame 25.Then it drop sharply in frame 26, when the target is disappear.

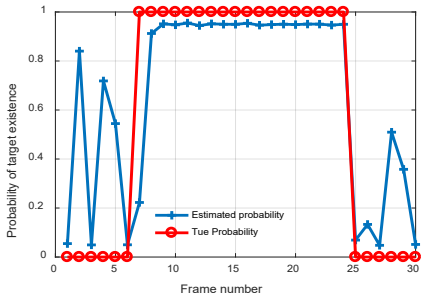


Fig. 2. True and Estimated Target Existence Probability for SNR=6.71 dB

Figure (3) displays the true target path against the track, produced by the filter. Note how the target trajectory deviates slightly from the straight line due to process noise. The PF-TBD tracks the target with a small positional error.

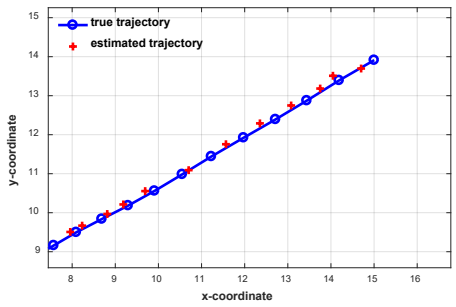


Fig. 3. True and Estimated Target Trajectory for SNR=6.71 dB

Figure (4) shows the position RMSE for three different peak SNR conditions (3.18 dB,6.71 dB, 12 dB). The position error is lower in 6.71 dB than 3.18 dB. As it can be seen, the PF-TBD was able to closely track the target even under low SNR.

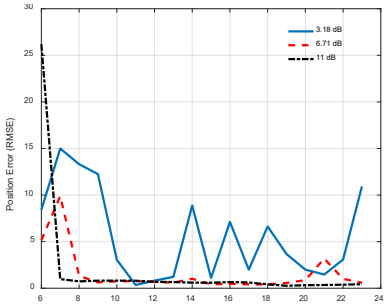


Fig. 4. Position RMSE for the PF-TBD for differnet peak SNR

5.2. Scenario2

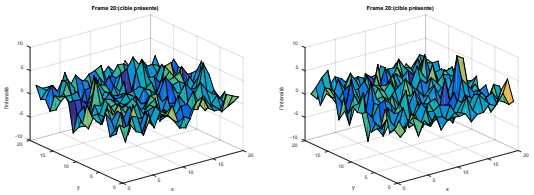
The second model is a Coordinate turn model is used [15]. The dynamic model for the target can be described by (1).

Where:

$$F = \begin{bmatrix} 1 & F_1 & 0 & F_2 & 0 \\ 0 & F_3 & 0 & -F_4 & 0 \\ 0 & -F_2 & 1 & F_1 & 0 \\ 0 & F_4 & 0 & F_3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, Q = \begin{bmatrix} Q_1 & Q_2 & 0 & Q_3 & 0 \\ Q_2 & Q_4 & -Q_3 & 0 & 0 \\ 0 & -Q_3 & Q_1 & Q_2 & 0 \\ Q_3 & 0 & Q_2 & Q_4 & 0 \\ 0 & 0 & 0 & 0 & Q_5 \end{bmatrix}$$

And $F_1 = \frac{sin(\Psi T)}{\Psi}$, $F_2 = \frac{(-cos(\Psi T)+1)}{\Psi}$, $F_3 = \cos(\Psi T)$, $F_4 = \sin(\Psi T)$, $Q_1 = \frac{2(\Psi T - sin(\Psi T))q_1}{\Psi^3}$, $Q_2 = \frac{(1-cos(\Psi T)q_1)}{\Psi^2}$ $Q_3 = \frac{(\Psi T - sin(\Psi T)q_1)}{\Psi^2}$, $Q_4 = q_1 T$, $Q_5 = q_2 T$, $\Psi = 6$ is a constant angular rate.

Figure 5 (a) and (b) show the measurement frame at time step 20 for 6.71dB and 12 dB peak respectively.



(a) (b)

Fig. 5. Measurements Frame (20): (a) for 6.71dB Peak SNR, (b) for 12dB Peak SNR for CT model.

In figure (6) the probability of presence is shown for a SNR of 6.71 dB. Existence probability probability is still increase above frame 7 until frame 17 and still stable until frame 25. Therefore, it drops rapidly following the target disappears from the monitoring region after frame 25.

Figure (7) shows the true and estimated target trajectories for coordinate turn model, the estimated trajectory is very close to the true trajectory.

Figure (8) shows the position RMSE for three different peak SNR conditions (3.18 dB,6.71 dB, 12 dB). The position error is lower in 6.71 dB than 3.18 dB. As it can be seen, the PF-TBD was able to closely track the target even under low SNR.

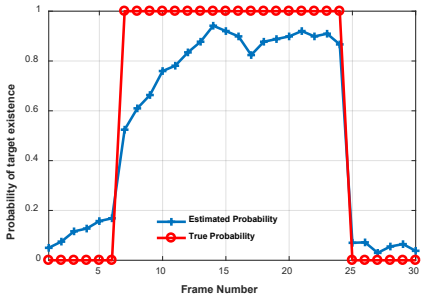


Fig. 6. True and Estimated Target Trajectory for SNR=6.71 dB

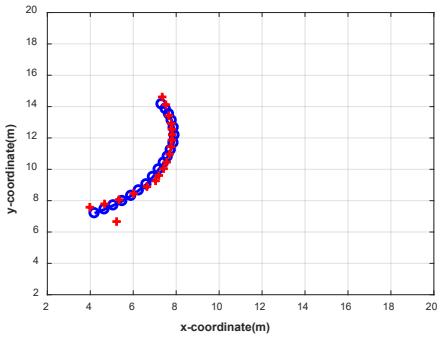


Fig. 7. True and Estimated Target CT Trajectory for SNR=6.71

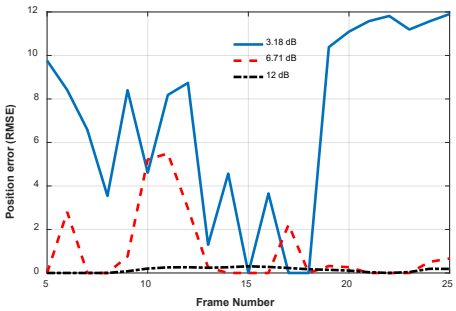


Fig. 8. Position RMSE for the PF-TBD for differnet peak SNR

6. Conclusion

In this paper, to manipulate moving weak targets, the PF-TBD algorithm is proposed for two dynamics models (CV and CT). The major advantage of the track-before-detect approach based on target existence variable and as a result, the developed particle filter can detect and track low SNR maneuvering target. The results from the simulation show that the PF-TBD algorithm has a successfully detection and tracking performance, both for constant velocity and coordinate turn models of moving targets, under severe conditions such as high noise or low SNR. Therefore, further work will mainly concentrate on how to detect and track multiple targets in high noise and high clutter.

References

[1] S. Blackman and R. Popoli, Design and Analysis of Modern Tracking System, Norwood: MA: Artech House, 1999.

[2] Y. Barniv, Dynamic programming algorithm for detecting dim moving targets, in Multitarget Multisensor Tracking: Advanced Application (Y, Bar Shalom, ed),ch 4, Norwood: MA, Artech House, 1990.

[3] J. Arnold, S. Shaw and H. Pasternack, "Efficient target tracking using dynamic programming," *IEEE Trans Aerospace and Electronic Systems,* vol. 29, pp. 44-56, January 1993.

[4] B. Carlson, E. D. Evans and S. L. Wilson, "Search radar detection and track with the Hough transform, part I: System concept," *IEEE Trans Aerospace and Electronic System,* vol. 30, pp. 102-108.

[5] M. Tonissen and Y. Bar-Shalom, "Maximum likelihood track-before-detect with fluctuating target amplitude," *IEEE Trans, Aerospace and Electronic Systems,* vol. 34, pp. 796-809, July 1998.

[6] R. Ristic, S. Arulampalam and N. Gordon, Beyond the Kalman Filter: Particle Filters for Tracking Applications, Boston: MA:Artech House, 2004.

[7] D. Salmond and H. Birth, "A particle filter for track-before-detect," *in Proc, Americain Control Conf,* pp. 3755-3760, June 2001.

[8] Y. Boyers and H. Drissen, "Particle filter based detection for tracking," *In Proceedings of the American Control Conference,,* pp. 4393-4397, June 2001.

[9] M. Rollason and D. Salmond, "A particle filter for track-before-detect of a target with unknown amplitude," *in IEE Int, Seminar Target Tracking: Algorithms and Application,* p. 14, October 2001.

[10] D. B. Colegrove, A. W. Davis and J. K. Alyliffe, "Track initiation and nearest neihbours incorporated into probabilistic data association," *Journal of Electrical and Engineers,* vol. 6, pp. 191-198, September 1986.

[11] D. Musicki, R. Evans and S. Stankovic, "Integrated probabilistic data association," *IEEE Trans. Automatic Control,* vol. 39, pp. 1237-1240, June 1994.

[12] D. J. Ballantyne, H. Y. Chan and M. A. Kouritzin, "A Novel branching particle method for tracking," *in Proc, SPIE, Signal and Data Processing of Small Targets,,* vol. 4048, p. 287, 2000.

[13] M. G. Rutten, B. Ristic and N. J. Gordon, "A Comparison of Particle Filters for Recursive Track-before-detect," in *7th International Conferenceon Information Fusion (FUSION)*, 2005.

[14] S. Mcginnity and G. W. Irwin, "Multiple Bootstrap Filter for Maneuvering Target Tracking," *IEEE Transaction of Aerospace and Electronic systems,* vol. 36, no. 3, July 2000.

[15] Y. Bar-Shalom, X. R. Li and T. Kirubarajan, Estimation with Applications to Tracking and Navigation, New York: Jhone Wiley&Sons, 2001.

[16] E. S. P and -S. A. P, "Generalized Recursive Track-Before-Detect With Proposal Partitioning for Tracking Varying Number of Multiple Targets in Low SNR," *IEEE Transactions on signal processing,* vol. 64, no. 11, 2016.

[17] E. P. S and -S. P. A, "Generalized Recursive Track-Before-Detect With Proposal Partitioning for Tracking Varying Number of Multiple Targets in Low SNR," *IEEE Transactions on signal processing,* vol. 64, no. 11, 2016.

[18] N. Amrouche, A. Khenchaf and D. Berkani, "Detection and Tracking Targets under Low SNR," in *IEEE International Conference on Industrial Technology*, Toronto, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 472-482 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Structure-Preserving Modeling of Safety-Critical Combinational Circuits

Feim Ridvan Rasim*, Sebastian M. Sattler

Chair of Reliable Circuits and Systems, Friedrich-Alexander-University Erlangen-Nuremberg, Paul-Gordan-Str. 5, 91052 Erlangen, Germany

ARTICLEINFO

Article history:
Received: 31 October, 2017
Accepted: 01 February, 2018
Online: 18 February, 2018

Keywords:
Combinational circuit
Structure
modeling
Symbolic analysis
Signal flow graph
Signal flow plan
Boolean function
Resolution method
KV diagram

ABSTRACT

In this work, a representative combinational circuit is visualized in various ways. It is abstracted (concretized) from transistor level to gate level and a structure-preserving transition is carried out into a signal flow graph. For creating a signal flow plan it is necessary to swap the nodes and the edges in the signal flow graph. After having executed this action the result is a signal flow plan. A value table exhibits the coding of the whole circuit. Then the so called module view is used to get the familiar compact and directed display and neighborhood relations are repeated once more, the resolution method is used. It is observed that undefined results can occur in digital circuits. But, these must be avoided in safety critical circuits. These events have to be secured in practice by costly and expensive verification and testing. In order to deal with the problem now, the structure-preserving modeling has to be understood, since this is the only way to achieve a one-purpose, qualitative and cost effective search for errors.

1 Introduction

This paper is an extension of work orginally presented at the 20th IEEE International Symposium on DDECS 2017, Dresden, Germany [1].

In order to ensure the functional safety of circuits or systems which are regarded as critical to safety, the mutual convert of models and functions is of great importance. The inconsistency problem is omnipresent; therefore, the essential claim for conformity with the formal derived function and the function derived from the real structure has a present role [2]. The directed mode of operation of a system should be represented by a circuit or switching table, also called a table of values, one-to-one in the sense that the encoding can be reproduced. In safety-critical circuits it is necessary not defined results, which often occur in complex circuits, to avoid or to monitor. The transferability of circuits into additional and other display possibilities is therefore a necessary property to ensure the functional safety of safety-critical circuits. In this work, a representative combinational circuit is visualized in various ways. In all these representations, however, it should be noted that the "structure-preserving modeling and transfer" is maintained. This means that the formally derived function must consistently match the function derived from the respective representation type. Both functions must in no case have inconsistencies, since only the fault-free function is included in the circuit. Functional safety can be guaranteed by the condition of the *structure-based modeling and transfer.*
To present the application we use an electrical circuit as a use case. And visualize it in various ways like Gate Level (GL), Signal Flow Graph (SFG), Signal Flow Plan (SFP), Module View (MV), Resolution Method (RM) and KV diagram. During creation of different display possibilities, we explain the rules of the structure-based modeling and transfer. In addition, the mathematical axioms, which are based on Propositional Logic (AA), are declared. The advantage of the method is that each type of representation (respectively presentation) has a depth of accuracy, clarity and compactness. The transferability of circuits into other possibilities of directed representation is a necessary property to ensure the functional safety of safety critical circuits.

Organization of the paper: First, the theoretical foundations are briefly explained in Chapter 2. They are regarded as basic knowledge in order to understand this work. Subsequently, the implementation is described in detail by an example in Chapter 3 and visualized by sketches and mod-

*Corresponding Author: Feim Ridvan Rasim, feim.rasim@fau.de

https://dx.doi.org/10.25046/aj030158

els. In the end, the results and the core outline of the work are summarized again and an outlook is given.

2 Theoretical Foundations

2.1 Category level (CL)

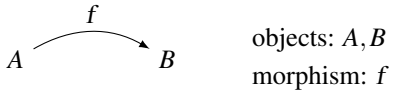


Figure 1: Morphism f from A to B

Morphisms are concretized on category level. A morphism is a directed association between two objects being associative and identitive. In Fig. 1, the morphism f acts on object A and is substituted (=) by object B.

2.2 Value table (VT)

The table of values, also known as truth table or switching sequence table, is a tabular list of the truth value course of a logical statement. In this table, the assignments of the inputs are linked together and the states at the outputs are shown binary. Value Tables (VT) are used for logic operations such as AND, OR, NAND and NOR gates, but also for flip-flops or complex circuits [3], [4]. A value table thus serves to represent the value of a composite statement as a function of the truth values of its partial statements. Tab. 1 shows a truth table for a NOT (left), AND2 (middle) and OR2 (right) function. The input for a NOT function is A and for AND2 and NOR2 the inputs are A and B.

A	NOT A	A	B	A AND B	A	B	A OR B
0	1	0	0	0	0	0	0
1	0	0	1	0	0	1	1
		1	0	0	1	0	1
		1	1	1	1	1	1

Table 1: Value table for NOT (left), AND2 (middle) and OR2 (right) function in (0,1)

2.3 Combinational circuits (CC)

A	B	A NOR B
0	0	1
0	1	0
1	0	0
1	1	0

Table 2: Value table for NOR2 circuit

Under a combinational circuit is a circuit that is realized with simple basic gates such as AND, OR and inverters to understand. It realizes a one-to-one mapping, a function. The outputs of this logic are dependent on the inputs, which means there is no feedback from the outputs to the inputs. The output variable is thus only a function of the input variables [4]. Tab. 2 shows the value table for NOR2 circuit. NOR2 represents a combinational circuit, because the circuit does not include feedback.

2.3.1 Abstraction of a circuit

Under an abstraction of a circuit is a kind of "simplified" representation to understand. A complex, combinational circuit is shown in a different way, in order to make it easier to understand. However, in the various representations, the core message of the output circuit must not be changed. So the function of the circuit must not be changed when transferring it to another representation. The circuit must deliver the same function, no matter in which visual presentation it is shown. For example, an abstraction of a circuit is the transfer of a circuit from the transistor level to the gate level. In addition, a circuit can be presented in its module view (MV) (see section 2.10), which is an isomorphism to the signal flow plan of the circuit. The aim of an abstraction of a circuit is the simplified or clearer visual design or presentation.

2.3.2 Structural changeover and modeling

In order to be able to carry out a structure-faithful modeling, it is to be known that two parallel connected transistors are combined in propositional logic as follows. Fig. 2 shows two transistors connected in parallel. V_{DD} is the operating voltage.
It is a complex gate with two inputs $(\bar{A}, \bar{B})$ and one output (C) between which the logical link "OR" exists. This OR2 outputs "1" at the output when one of the inputs are assigned a "0". This means that if one of the two inputs is assigned a "0", the output creates a "1". The two transistors are connected in parallel, they must be concatenated (addition in field algebra). Furthermore the operating voltage V_{DD} has to be considered. It runs in series with each of the two transistors, it muss be catenated (order, multiplication in field algebra) with two transistors.

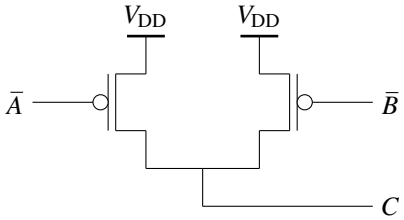


Figure 2: Example of parallel connected transistors

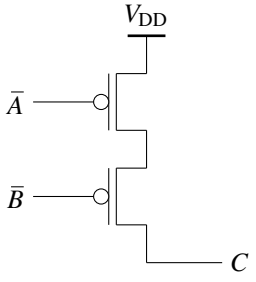


Figure 3: Example of serial connected transistors

If C in Fig. 2 is to be expressed as a function (contrary to logic, only true values can be processed in propositional logic (see section 2.9)), then this is:

$$C = V_{\mathrm{DD}} \cdot (\bar{A} + \bar{B}) \tag{1}$$

In Fig. 3 two transistors are connected in series. It is a complex gate with two inputs $(\bar{A}, \bar{B})$ and one output (C) between which the logical link "AND" exists. This AND2 outputs "1" at the output when both inputs are assigned a "0". This means that if both of the two inputs is assigned a "0", the output creates a "1". Furthermore the operating voltage V_{DD} has to be considered. It runs in series with each of the two transistors, it muss be catenated with two transistors. C can be expressed in propositional logic as follows:

$$C = V_{\mathrm{DD}} \cdot (\bar{A} \cdot \bar{B}) \tag{2}$$

Structure-faithful modeling unites function and structure one-by-one in the sense of a monomorphism injective - that is, the structure has at most one solution (this is the function) - and of a epimorphism surjective - that is, the function has at least one solution (that is the structure). Such a mapping enables a one-to-one (local-bijective) and understandable description of a generating system. During transferring into various presentation possibilities the structure-faithful modeling has a significant role. It is extraordinary important that the formally derived (modeled) function coincide with the function generated by the real structure. Consequently the function has to correspond to reality and shall not exhibit any inconsistencies. Only in this way the functionality of a circuit can be ensured. Structure-faithful therefore means that the relation to reality must never be lost during modeling. Shortly spoken, each pin of the model at gate level (GL) must show up in the real world at transistor level (TL). But, pins have to be correctly labelled in reality at transistor level (TL). This is mandatory.

During the transfer, it is also important that the function generated from the real structure consistently matches the function derived from the signal flow graph or any other type of presentation. Only in this way the functionality of the generating circuit can be ensured. In addition, there is a structure-based transfer only in the absence of inconsistencies. A transfer of the signal flow graph or of the circuit into a value table must also be structure-faithful. Thus the function derived from the evaluation table must correspond to the same function derived from the signal flow graph or the generated circuit. Shortly spoken, each undefined pin of a given reality or not proper assigned signal of a given model has to be shifted to undefined. We consistently use the symbol "$*$".

2.4 RS Buffer

In complex circuits, many structures exist that can create undefined results. These undefined results must not occur in safety-critical circuits, since otherwise the desired function of the circuit cannot be guaranteed. For this reason, the RS buffer structure is established [2]. It can intercept undefined cases in combination with a dual-rail approach. Thus, it is possible to stabilize a complex circuit in its function without glitch. These stabilized states do not produce unpredictable events and can therefore be processed by the circuit without causing errors. Fig. 4 shows the circuit symbol and Fig. 5 shows the circuitry of the RS buffer. The value at the node X in the circuit corresponds to the value at the pin $\bar{Y}$, because of the inverter. Thus, the X is neglected for the sake of clarity in the VT of Tab. 3. On closer examination of Tab. 3, it is noticeable that the RS buffer triggers a switching process only during assignments $(S, \bar{R}) = (1,0)$ and $(S, \bar{R}) = (0,1)$. The old state is retained for assignments $(S, \bar{R}) = (1,1)$ and $(S, \bar{R}) = (0,0)$. The function for the output Y is therefore $Y = R(S \vee Y) \vee S(R \vee Y) = YR \vee YS \vee SR$.

S	$\bar{R}$	Y
0	0	Y
0	1	0
1	0	1
1	1	Y

Table 3: Value table of the RS buffer [2]

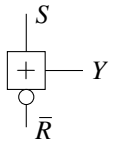


Figure 4: Circuit symbol of the RS buffer

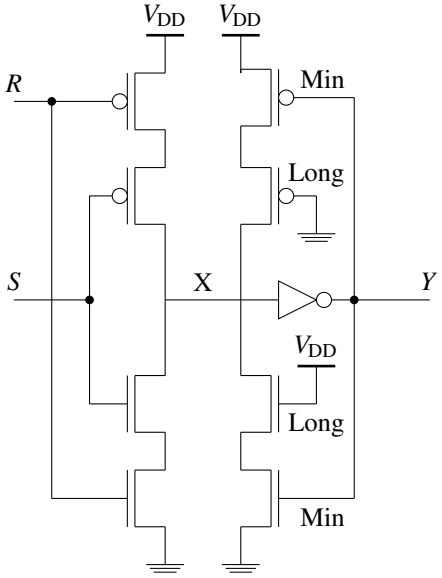


Figure 5: Circuit of the RS buffer [2]

2.5 Signal flow graph (SFG)

The signal flow graph (SFG) is a vividly method to present the internal structure of a system or the interaction of several systems. This presentation allows a better understanding of the function as well as the interrelations of one or more systems. In addition, the signal flow graph is the appropriate tool for abstracting functions or connections to one level above category level (associativity and not identity). The signal flow graph is a directed and weighted graph whose nodes represent objects (sets) and its edges morphisms (functions). The edges of this graph can be understood in a dual view (SFP) as small processing units which process incoming signals (edges) in a particular form and then send the result to all outgoing edges (signals). Signal

flow graphs are formally defined graphs [5]-[9].

2.6 Signal flow plan (SFP)

The signal flow plan (SFP) has a special significance in control engineering and, with the representation method, is based on a block diagram and adds to a further loan of the relationships within a system. A signal flow plan is used to identify the complexity of a system. This contains unidirectional blocks, also called nodes, which transmit incoming signals as small processing units into outgoing signals. In the signal flow graph these are represented by edges. From an SFP, the transfer function of a system can be derived. The signals are redirected to edges, unlike the signal flow graph. In the signal flow graph, they are represented by nodes. Edges in a signal flow plan are also directed connections between two nodes. They illustrate the effect of a signal by its weighting. Furthermore, it is possible to transfer a signal flow graph into a signal flow plan by transferring nodes into edges and edges into nodes [5], [7], [9]. Fig. 6 shows a simple example of a signal flow plan derived from a signal flow graph. The input signals x_0 and x_1 adds up to an output signal y.

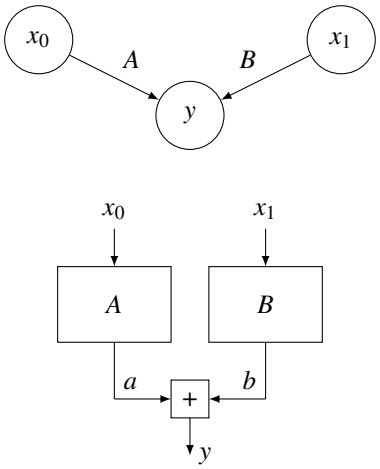


Figure 6: From a signal flow graph (above) to a signal flow plan (below)

2.7 Boolean algebra (BA)

The switching algebra (boolean algebra) is based on decisions and comparisons, so it can explain and visualize logical links very well. Successful results are represented by a "1", unsuccessful results by a "0". These two symbols are complementary to one another. At any time, each pin must be occupied, because only then the system can be a total system and can be calculated by the switching algebra. The switching algebra is not sufficient for a detailed representation of a circuit. However, it is suitable for the functional description without restrictions to the general [10].

2.8 Positive logic (PL)

In the positive logic the symbol "1" stands for a successful event. Unsuccessful events are called undefined. Positive logic is an event that occurs just as it is expected. This means that if a negative event is expected and it occurs, this event is considered successful. This also applies analogously to a positively expected event. If a positive event is expected and it occurs, then this event is also successful. In the positive logic, therefore, only the "1" exists as a value [11]. Here, we need a "0", too. Therefore, the "0" is also a "1" but only a part (child) of the "1".

2.9 Propositional logic (AA)

The propositional logic comes from the formal logic and can be continued into the switching algebra without restrictions to the general. This "Aussagenlogischer Ausdruck (AA)" describes the relationship between statements. Statements can be seen partially in the propositional logic. This means that there should be only one unary statement. The symbols are true (w) and not false ($\bar{\text{f}}$). The following example is intended to illustrate the propositional logic: The statement $Y = (A \wedge \bar{B}) \vee C$ contains two statements and is nevertheless unary, regardless whether is equal to a positive literal or a negative literal. The following statement contains only one statement and is also unary $\bar{Y} = (A \wedge B)$. Thus, in AA a statement is always true, only its complemented content can be interpreted as "false". The logical sign for a true statement is "w " or "$\bar{f}$" [11], [12].

2.10 Module view (MV)

Through a model representation, a real system can be displayed simpler and clearer. The natural section of reality can be simplified by model visualization, and the model representation can also serve as a reference model for further development of a system. In order to represent a real circuit with real electrical components as a model, the structure-faithful transfer is of great importance. In addition, it is necessary that in safety-critical circuits, the modeling is performed structurally, because only then the structure-related functional safety of the circuit is guaranteed. This means that the formally derived and modeled function with which the function created by the real structure must be absolutely identical. Fig. 7 shows a simple example of the asymmetric and irreflexive module view. Due to the directed representation as a module view, a clear visualization is achieved. You can see the input (input vector) X, the output F and the programming vectors Z and Y. By presenting it as a module view, the complexity of the representation can be completely directed.

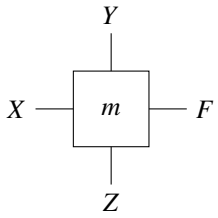


Figure 7: Example of a module view

Only the inputs, outputs and programming vectors are seen. The function, how exactly the interaction of these variables ever is, is not described and explained. The function (transitive closure) is represented by m (model) [5].

2.11 Resolution method (RM)

The resolution method (RM) is a method to check the truth value. It is primarily about being able to prove the satisfiability or unfulfillability. In order to provide this proof, a clause set is extended by new clauses, called resolvents, until an empty clause is generated. If this succeeds, the initial quantity is unsatisfiable. Here, the clauses are conjunctions. A set of clauses, therefore, is a disjunction. It follows that if an empty clause can generated, the initial set is tautological. Resolvent of clauses C_1 and C_2 (according to literal l) is given in [13]. An example of the resolution method is shown in Fig. 8. Here you can see that the clause set can be non-tautologically fulfilled, since in the end no empty clause can be generated. Will be a clause set with the resolvent resulting from their clauses, this results in a logically equivalent set of clauses. This can be reunited with its resolvents without affecting satisfiability, and so on. If finally the empty clause can be formed, a tautological set is proven. In the other case, for example if the last clause is not empty, this is the proof of the satisfiability of the disjunvtive clauses.

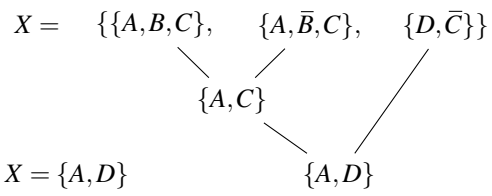


Figure 8: Example of a resolution

The resolution method obeys the following neighborhood relations:

- The resolver $\{y,z\}$ (largest common cover) can be generated from the clauses $\{x,y\}$ and $\{\bar{x},y,z\}$.
- The resolver $\{\}$ can be generated from the clauses $\{x\}$ and $\{\bar{x}\}$.

Definition: Let C_1 be a clause containing the literal l and let C_2 be a clause containing the literal $\bar{l}$. Then the clause is called:

$$C = (C_1 \backslash \{l\}) \cup (C_2 \backslash \{\bar{l}\}) \tag{3}$$

2.12 KV diagram

The Karnaugh-Veitch diagram (KV diagram) is used for the clear presentation and simplification of Boolean functions into a minimal expression. A KV diagram can be used to transform any disjunctive normal form (DNF) into a minimal disjunctive logical expression. The diagram is labeled with the variables at the edges. Each variable occurs in negated form (negative literal) and not negated form (positive literal). The assignment is arbitrary. However, it should be noted that horizontally and vertically adjacent fields may only differ in exactly one variable [4].

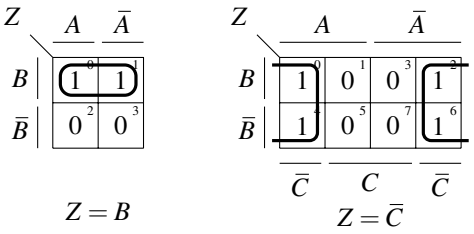


Figure 9: KV diagram examples

Minterm Method:

- Try to group as many horizontally and vertically adjacent fields as possible that contain a "1" (1, 2, 4, 8, 16, 32, ...).
- A block may continue over the right / left or bottom of the chart.
- From these identified blocks so many are to be selected that all "1" fields are covered (Fig. 9).
- Now the conjunct terms are formed.
- These conjunctive terms are linked with an OR and a DNF results.

Maxterm method: This method differs from the Minterm method in the following points:

- Instead of ones, zeros are combined into blocks.
- Subsequently, disjunction terms are formed.
- These disjunction terms are AND, resulting in a conjunctive normal form (CNF).

3 Implementation

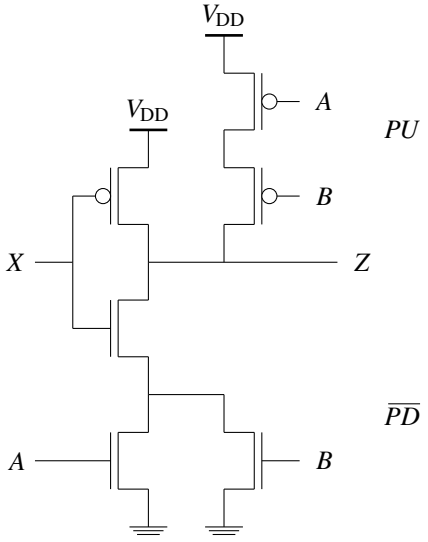


Figure 10: Simple electrical circuit at transistor level (TL)

In this Chapter, an electrical circuit is considered at the transistor level in Fig. 10. Afterwards, the circuit is transfered from the transistor level into the gate level in a structured manner. It should be noted that the analog circuit, that

is, the circuit at transistor level, is described at the gate level in propositional logic. Subsequently, the circuit is converted into a signal flow graph and signal flow plan at the gate level. Various possibilities for the representation of the output circuit, then such as the evaluation table, the module view or the resolves are presented. With all these possibilities of representation, it should be noted that the respective derived function must not have any inconsistencies that means, the formally derived function must agree with a function generated from a real structure (TL). Fig. 10 shows the circuit at the transistor level, this circuitry is a combinational circuit because there are no feedback lines, which is to be transferred to the gate level. It is a complex gate with three inputs and an output.

3.1 Concretization to gate level

The analysis of a circuit at the transistor level is more detailed and more complex than viewing at the gate level, since the representation in gates is only a "model" which allows a simplified and clear view of the circuitry. The transfer of a structure at the transistor level into a structure at the gate level is therefore a concretization (often called abstraction) and serves to increase the clarity and simplify the understanding of the structure. From propositional logic and category point of view, however, transistor level (parent) is the abstraction of the gate level (child). This is important to keep in mind.

In the first step is now the electrical circuit transferred to the gate level. First, the pull-up (PU) and the pull-down $(\overline{PD})$ are viewed separately from each other. When the transistors of the pull-up or pull-down are connected in parallel, they must be concatenated. On the other hand, when the transistors are connected in series, they must be catenated. Furthermore the operating voltage V_{DD} has to be considered in the pull-up. This runs in series with each of the two lines of the pull-up. And the mass potential runs serially to both line of the pull-down.
The transistors in the *PU* are switched with a "0" low-active input and the transistors in the $\overline{PD}$ are switched with a "1" high-active input. *PU* and $\overline{PD}$ are connected by a composition (concatenation). With the composition it is meant that different things are substituted but do preserve their directed manner (the morphisms are preserved). This composition ensures the RS buffer in Fig. 5. In Fig. 5 after the pull-down, a switch "¬" is installed in order to meet the propositional composition (we use switch to replace the two levels "0" and "1", we can switch from one level to another). The function of the RS buffer has already been explained in the basic chapter. The operating voltage is usually designated with V_{DD}, the reference point is the mass with the low-active input $\overline{\mathrm{GND}}$ (since the gate level has been in AA, the ground is written as $\overline{\mathrm{GND}}$). Please keep in mind that pins have to be coded correctly. You have to choose the correct literals, either positive literals or negative literals.

In the second step subcircuits are described as concrete mathematical functions in the propositional domain (in the AA domain). It should also be noted that each partial circuit is basically at least disjunct, in this case even complementary, this mean a pull-up (PU) and a pull-down $(\overline{PD})$. Pull-up means that the output is pulled up to the operating voltage V_{DD}. This part of the circuit is low-active since a logical "0" must be present at the primary input in order to trigger the switching process. The $\overline{PD}$ draws the output to the mass potential. It is referred to as high-active, since a logical "1" must be present at the primary input, so that a switching process is triggered. For example, Tab. 3 shows the functionality of the RS buffer. The operation is assumed to be already known. Thus, the circuit consisting of a *PU* and a $\overline{PD}$ has the following equation:

$$Y = PU + \neg\overline{PD} = S + \neg\bar{R} \tag{4}$$

Eq. 4 will play an important role in the later course of the work. After all steps, as explained above, have been carried out, a structure-preserving model at the gate level results. Important is, that during all steps inconsistencies must not occur. In Fig. 11, the electrical circuit is now displayed in propositional logic at gate level. As described above, the *PU* and $\overline{PD}$ are summarized using the composition. In addition, a switch is built between $\overline{PD}$ and the composition. Its task is also to highlight the low-active input of the RS buffer. It is important that the stars here in the continuation only serve as a "monitor" for checking. If the circuit is designed correctly, each star $(*)$ can only supply a "0".

In the third step, the node *Z* (pin *Z*) is expressed in a function: The *PU* and $\overline{PD}$ is connected to the composition "+". The operating voltage V_{DD} flows in series with the inputs in the *PU*. The mass runs serially to all inputs in the $\overline{PD}$. The switch between composition and pull-down switches the logic value of the last part to a negative literal - the substitute of its complement in Eq. 4. In summary, it should be emphasized that viewing at the gate level (in propositional logic) allows a directed and more simplified view, which makes the derivation of the function at node *Z* easier.
The node *X* is treated as another input. Thus, it must be noted that node *X* affects the *PU* and the $\overline{PD}$. Furthermore, the *PU* and the $\overline{PD}$ are linked by a composition. This composition works like an RS buffer. The inputs are then linked to the logical operators. Not to forget is the switch, which is between $\overline{PD}$ and the composition. This "switches" the part of the function belonging to the $\overline{PD}$. The operating voltage V_{DD} and the mass run serial to the inputs.
The stars $(*)$ also serve (Fig. 11) for to check the correctness of the circuit. For example, each individual star can only assume the value "0" in order to ensure that the circuit is correctness. Fig. 11 shows the structure-faithful modeling of the circuit at gate level (in AA). After the circuit is transferred to gate level, the node *Z* can be expressed as a mathematical function in Eq. 5.

$$Z = V_{\mathrm{DD}} \cdot (\overline{X} + \bar{A} \cdot \bar{B}) + \neg(\overline{\mathrm{GND}} \cdot (A + B) \cdot X) \tag{5}$$

The circuit at gate level in Fig. 11 shows the concreteness of the output circuit (Fig. 10). As described above, it was transferred from the transistor level to the gate level in a structure-preserving manner. The functions derived at the transistor level and the functions derived at the gate level must be identical with respect to their partial order (directed manner), this means the function of the circuit must

not be changed by the transfer. Only then is the transfer a structure-faithful one.

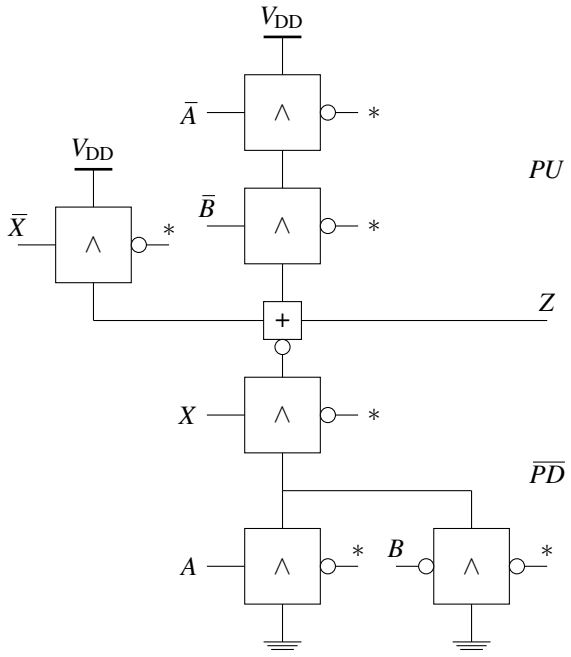


Figure 11: Simple electrical circuit at gate level (in AA)

The gate level can be viewed as a model view. It serves to increase clarity as well as contribute to an understanding of the circuit.

3.2 Transfer to a signal flow graph (SFG)

With the basic knowledge from Chapter 2, the function Z is now transferred step by step into an SFG at Fig. 12.

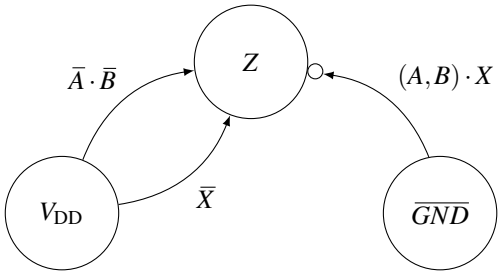


Figure 12: SFG of electrical circuit

For the creation of SFG, it is determined which signals are visualized in a node and which are represented by an edge. The operating voltage V_{DD} and the ground $\overline{GND}$ are visualized by a node, while the inputs, also the X, are used as edge weights for the processing of the signals. The operating voltage is catenated by the inputs $\overline{A}$, $\overline{B}$ and $\overline{X}$. Please remember that we are in propositional logic. The negative literal $\overline{A}$ is not the complement of A but can be substituted by the complement of A, $\overline{A} = \bar{A}$. BA is constituted by $\overline{A} = \neg A$. Thus, V_{DD} represents a node while the inputs with the switch reside on the edges. The third edge has as an edge weight the catenation of the X with the two inputs A and B. The mass is represented as a node. The switch (bubble at node Z) must be considered as shown in Fig. 12. These three edges end at to the node Z. The switch, which must be installed here, is not to be forgotten. These two edges concatenate to the node X. When looking at the Z function it is noticeable that the node X has a major influence on this function:

It is important that the function Z generated from the real structure is consistent (in the direction) with the functions derived at the gate level and the signal flow graph. The SFG allows a further comprehensible and simple visual consideration of the problem. The system is represented simply and visually by weighted, directed graphs. In dual sense (SFP), blocks are small processing units that process incoming signals (that are edges) in a certain form and send the result to all outgoing edges (that are signals). In Fig. 12, the comma "," means in parallel. We use the comma "," to write different things next to each other but preserving the directed manner. Since the directions have to be preserved in the structure faithful modeling and transfer, we announce two edges A and B next to each other. From Fig. 12 follows equation 6:

$$Z = V_{DD} \cdot (\overline{X} + \bar{A} \cdot \bar{B}) + \neg(\overline{GND} \cdot (A + B) \cdot X) \qquad (6)$$

3.3 Transfer to a signal flow plan (SFP)

As already explained, the node can be interpreted in the dual sense as a partition, a signal, and an edge over its weight as processing (operation) of the signal. Thus it generates (substitutes) a new signal. The states of the output circuit are to be found in the nodes. The edges are supplemented with their weights. If the electrical circuit is now considered more closely at gate level, the background knowledge of this work can be used to derive a signal flow plan. It is important to know that the function which is derived from the signal flow graph has to coincide with the function which is derived at the gate level. For only then the modeling has been done in a structured way and the relation to reality has not been lost. In order to be able to transmit a signal flow graph into a signal flow plan, the following must be carried out. By interchanging nodes and edges, a signal flow graph results in a signal flow plan and vice versa.

In order to set up a signal flow plan one has to determine which signals are visualized in a node and which are represented by an edge. The edge is a directed line, which connects two nodes and effects the processing of a signal via its weight. The mass as well as the operating voltage represent the nodes in the SFG. The primary inputs are shown as edges. The signal flow plan (action plan) is used to determine the complexity of a system. The nodes (blocks) in a signal flow plan are small processing units that transmit incoming signals on edges into outgoing signals on edges. Fig. 13 shows the signal flow plan derived from the signal flow graph of the electrical circuit. The signals, in this case the edges (nets) of the signal flow plan, are found on the nodes of the signal flow graph. The nodes (blocks) of the signal flow plan are found in the edges of the signal flow graph.

After the gate level of node Z has been transferred to a signal flow plan, the signal flow plan is displayed in a module view (Fig. 14).

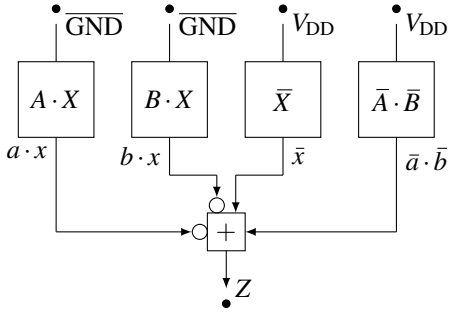


Figure 13: SFP of an electrical circuit

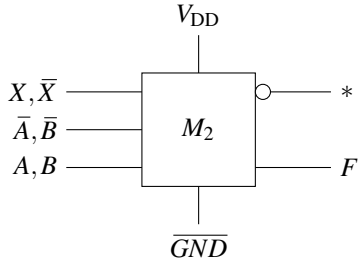


Figure 14: Module-view of an electrical circuit

This digital structural element serves for further simplified viewing of the real system. This further simplifies understanding of the structure. The module-view for node Z has:

- an input vector (inputs): A, B, $\overline{A}$, $\overline{B}$ and $X, \overline{X}$
- a programming vector (states): $\overline{\text{GND}}$ and V_{DD}
- an output vector (output): $Z =: F$

As the MV is using the transitive closure it preserves all pins (that occur in reality). Therefore, it is structure-faithful and a compact presentation of the circuit.
Within M_2, the function is outdated here. We can convert this function from MV to other display posibilities structure-faithful.
At the star $(*)$ we should only observe the "0".

3.4 Evaluation table of the circuit

In the next step the node Z (Eq. 7) of the output circuit (Fig. 11) is shown in the form of a total switching table (in positive logic (Tab. 4), followed by the partial switching table in positive logic (Tab. 5).

$$Z = V_{\text{DD}} \cdot (\overline{X} + \overline{A} \cdot \overline{B}) + \neg(\overline{\text{GND}} \cdot (A + B) \cdot X) \qquad (7)$$

The function in Eq. 7 can be expressed as follows:

$$\begin{aligned} Z &= (Z, \overline{Z}) \\ &= V_{\text{DD}} \cdot (\overline{X} + \overline{A} \cdot \overline{B}), \neg(\overline{\text{GND}} \cdot (A + B) \cdot X) \\ &= V_{\text{DD}} \cdot (\overline{X} + \overline{A} \cdot \overline{B}) + \neg(\overline{\text{GND}} \cdot (A + B) \cdot X) \\ &= PU + \neg\overline{PD} \end{aligned} \qquad (8)$$

In Eq. 7, the node Z is represented in different versions. $Z = (Z, \overline{Z})$ means that Z is a partition of two blocks provided by a positive and a negative literal. The last part of Eq. 7 means that Z consists of two parts, the pull up and the pull down. The relationship between PU and $\overline{PD}$ is the same as already shown in Eq. 4, not to forget the composition "+" between PU and $\overline{PD}$, the RS buffer.
Tab. 3 is also required in order to be able to set up the switching table. The operating voltage V_{DD} depends on the pull-up. It is important to know that if only the operating voltage supplies a "1", a reliable switching process can be present in the PU. The pull-down depends on the mass, $\overline{\text{GND}}$. A switching operation in the $\overline{PD}$ can only take place when the mass is at "0".
Tab. 4 shows the total switching table of node Z in positive logic (PL). On closer examination of the table it can be seen that this table is composed of three divisional tables. The result of the PU and the $\overline{PD}$ and the node Z represent a further table. The coding universe for the PU consists of the operating voltage V_{DD} and $(\overline{A}, \overline{B}, \overline{X})$. For the $\overline{PD}$, the coding universe consists of (A, B, X) and $\overline{\text{GND}}$. Thus the table for the PU and $\overline{PD}$ has a total of $2^4 = 16$ assignments. The output Z represents the respective state for the one-to-one coding of the table.
The lower part of the table (last eight lines) represents assignments which tend not to trigger any switching operations. The reason for this is that during the pull-up the operating voltage V_{DD} can only assume the value "1". The mass does not matter. For pull-down, $\overline{\text{GND}}$ must only have the value "0", so that a switching operation is triggered. The operating voltage does not matter.
The results of the pull-up and the pull-down are calculated by Eq. 7. In order to determine the resulting node Z, Tab. 3 is to be considered. It represents the relationship between PU and $\overline{PD}$.
In the next step, the total switching table Tab. 4 (in PL) is transferred to a partial switching table Tab. 5 (in PL). The correctness of the partial switching table is only given because a structure correct transformation and modeling has taken place. This ensures that the circuit has been designed without errors and thus a partial switching table can be applied without errors.
Tab. 5 shows the partial switching table of node Z. The operating voltage is unsignificant for the $\overline{PD}$, whereas in the PU it can trigger a switching operation only with a "1". The mass is meaningless for the PU, whereas only a "0" the mass in the $\overline{PD}$, triggers a switching process. Thus the last eight lines of the total switching table is lost. For the safety of a defect-free structure it can be firmly assumed that in these assignments, the output Z assumes the states in Tab. 5.
The symbol $*$ represents undefined.

3.5 Resolution method for the circuit

In the last step, the node Z (Eq. 9) of the output circuit (Fig. 11) is entered into a KV diagram.

$$Z = V_{\text{DD}} \cdot (\overline{X} + \overline{A} \cdot \overline{B}) + \neg(\overline{\text{GND}} \cdot (A + B) \cdot X) \qquad (9)$$

PU					$\overline{PD}$							
$\overline{\text{GND}}$	$\overline{X}$	$\overline{A}$	$\overline{B}$	V_{DD}	V_{DD}	X	A	B	$\overline{\text{GND}}$	PU	$\overline{PD}$	Z
$*$	0	0	0	1	$*$	0	0	0	0	1	$*$	1
$*$	0	0	1	1	$*$	0	0	1	0	1	$*$	1
$*$	0	1	0	1	$*$	0	1	0	0	1	$*$	1
$*$	0	1	1	1	$*$	0	1	1	0	1	$*$	1
$*$	1	0	0	1	$*$	1	0	0	0	1	$*$	1
$*$	1	0	1	1	$*$	1	0	1	0	$*$	1	0
$*$	1	1	0	1	$*$	1	1	0	0	$*$	1	0
$*$	1	1	1	1	$*$	1	1	1	0	$*$	1	0
$*$	0	0	0	0	$*$	0	0	0	1	$*$	$*$	$*$
$*$	0	0	1	0	$*$	0	0	1	1	$*$	$*$	$*$
$*$	0	1	0	0	$*$	0	1	0	1	$*$	$*$	$*$
$*$	0	1	1	0	$*$	0	1	1	1	$*$	$*$	$*$
$*$	1	0	0	0	$*$	1	0	0	1	$*$	$*$	$*$
$*$	1	0	1	0	$*$	1	0	1	1	$*$	$*$	$*$
$*$	1	1	0	0	$*$	1	1	0	1	$*$	$*$	$*$
$*$	1	1	1	0	$*$	1	1	1	1	$*$	$*$	$*$

Table 4: Total switching table (in PL): PU, $\overline{PD}$ and Z

$\overline{X}$ for PU X for $\overline{PD}$	$\overline{A}$ for PU A for $\overline{PD}$	$\overline{B}$ for PU B for $\overline{PD}$	V_{DD}	$\overline{\text{GND}}$	PU	$\overline{PD}$	Z
0	0	0	1	0	1	$*$	1
0	0	1	1	0	1	$*$	1
0	1	0	1	0	1	$*$	1
0	1	1	1	0	1	$*$	1
1	0	0	1	0	1	$*$	1
1	0	1	1	0	$*$	1	0
1	1	0	1	0	$*$	1	0
1	1	1	1	0	$*$	1	0

Table 5: Partial switching table (in PL): PU, $\overline{PD}$ and Z

For this purpose, Eq. 9 must be converted into a DNF system, as can be seen in the following.

$$Z = V_{\text{DD}} \cdot \overline{X} + V_{\text{DD}} \cdot \overline{A} \cdot \overline{B} + \neg(\overline{\text{GND}} \cdot A \cdot X) + \neg(\overline{\text{GND}} \cdot B \cdot X) \quad (10)$$

Due to the different terms of the Eq. 10, a better understanding of the KV diagram is to be developed in Fig. 15. It is noticeable that some fields overlap in the diagram. In particular, this fields are occupied by "1" and "¬1". The question is how this field actually looks in the KV diagram in AA. Now Eq. 10 is a concrete equation (Category Level). The KV diagram for this equation is to be represented in AA. In this case, the switch "¬" is substituted by complement since the KV diagram is in the Aussagenlogik (AA), the propositional logic, the logic of statements. The KV diagram now looks as Fig. 16. The blocks of the KV diagram (Fig. 16) are now determined from Eq. 10 as follows:

$$(\overline{A} \wedge \overline{B} \wedge V_{\text{DD}}) \vee (X \wedge A \wedge \overline{\text{GND}}) \vee (B \wedge X \wedge \overline{\text{GND}}) \vee (\overline{X} \wedge V_{\text{DD}}) \vee (V_{\text{DD}} \wedge \overline{\text{GND}}) \quad (11)$$

Therefore in Eq. 11, $V_{\text{DD}} \wedge \overline{\text{GND}}$ represents a redundant block, a redundant prime implicant. This case is visualized by dashed lines. This part of the equation is now used for resolving. Fig. 17 shows the resolution method for node Z. The clauses ($\{\overline{A}, \overline{B}, V_{DD}\}$, $\{X, A, \overline{GND}\}$, $\{X, B, \overline{GND}\}$, $\{\overline{X}, V_{DD}\}$) are derived from Eq. 10. In order to be able to determine the clauses, the equation was first transformed into a DNF system in order to connect the switch neglected. Only by neglecting the switch can a resolvent be formed. In order to determine the folowing clauses ($\{\overline{A}, \overline{B}, V_{DD}, \overline{GND}\}$, $\{\overline{A}, \overline{B}, V_{DD}, GND\}$, $\{X, A, GND, \overline{V}_{DD}\}$, $\{X, A, \overline{GND}, V_{DD}\}$, $\{X, B, \overline{GND}, \overline{V}_{DD}\}$, $\{X, B, \overline{GND}, V_{DD}\}$, $\{\overline{X}, V_{DD}, GND\}$, $\{\overline{X}, V_{DD}, \overline{GND}\}$), the ones in the KV diagram (Fig. 16) have to be viewed individually. $\{V_{DD}, \overline{GND}\}$ comes from prime implicant (dashed lines) of Fig. 16.

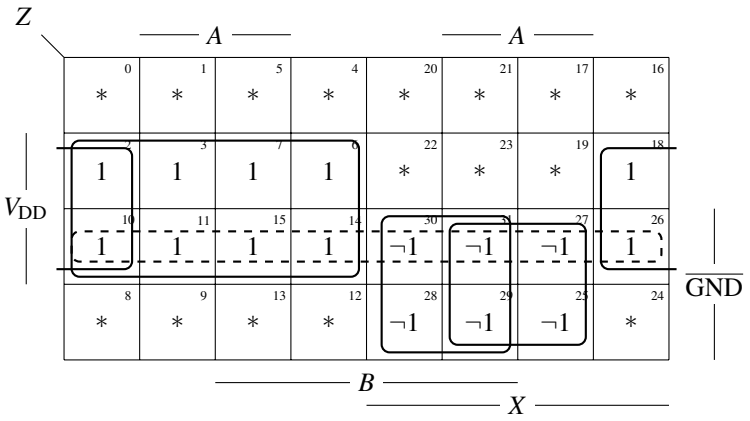


Figure 15: KV diagram for Z (electrical circuit in PL)

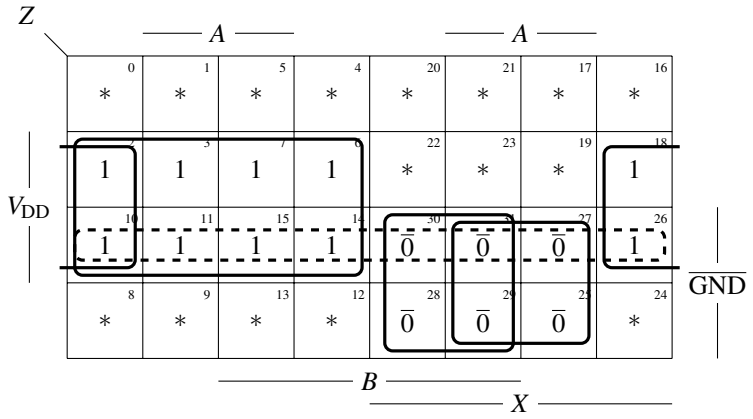


Figure 16: KV diagram for Z (electrical circuit in AA)

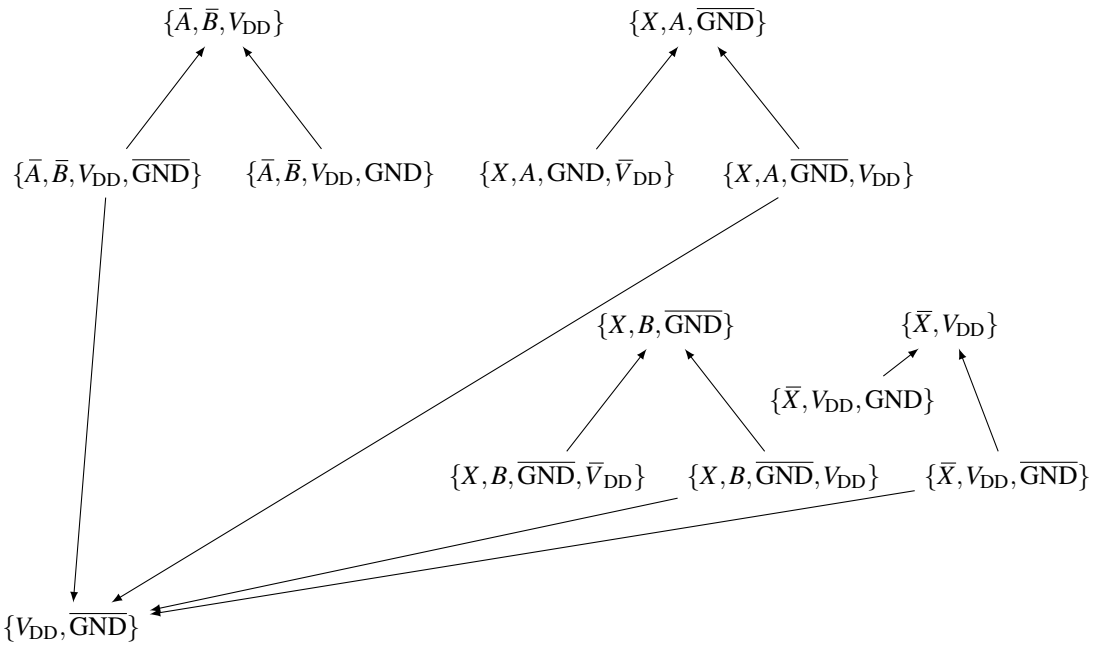


Figure 17: Resolution method for the electrical circuit in PL

So it is noticeable that the KV diagram shows additional information when looking at the ones. The ones that can be useful in resolving are selected. Resolves can be generated only from certain clauses. As already explained in Chapter 2, the following holds: The resolver $\{y,z\}$ can be generated from the clauses $\{x,y\}$ and $\{\bar{x},y,z\}$. The resolver $\{\}$ can be generated from the clauses $\{x\}$ and $\{\bar{x}\}$. The prime implicant (dashed) in the KV diagram is made up of this clause: V_{DD}, $\overline{\mathrm{GND}}$.
This clause resolves from three different clauses, as shown in Fig. 17.

4 Conclusion

The transferability of circuits into other possibilities of representation is a necessary property to ensure the functional safety of safety critical circuits. In this work an output circuit has been visualized in various display possibilities. Each type of presentation has its advantages and disadvantages. Moreover, each type of representation has a depth of accuracy, clarity and compactness. However, all of these representations are common in that their "structure-faithful modeling and transition" must be preserved. This means that a formally derived function has to match consistently with the function derived from the respective representation type. Both functions must in no case have inconsistencies, because only then the fault-free function of the circuit can be maintained. Shortly spoken, each pin of the model at gate level must show up in the real world at transistor level. But, pins have to be correctly labelled in reality. This is mandatory.

References

[1] F. R. Rasim, C. Kocar, S. M. Sattler: Structure-Preserving Modeling of Safety-Critical Combinational Circuits, 20th IEEE International Symposium on DDECS 2017, Dresden

[2] G. Uygur, S. M. Sattler: Structure-Preserving Modeling for Safety Critical Systems. Mixed-Signal Testing Workshop (IMSTW), 2015

[3] M. M. Mano, R. K. Charles: Logic and Computer Design Fundamentals, Third Edition. Prentice Hall, 2004

[4] P. Horowitz, W. Hill: The Art of Electronics, Second Edition. Cambridge Press, 1989

[5] R. C. Dorf, R. H. Bishop, Modern Control Systems Solution Manual, Pearson Studium, 2011

[6] S. J. Mason, Feedback Theory - Further Properties of Signal Flow Graphs, IEEE, vol. 44, pp. 920-926, 1956

[7] W. S. Levine, The Control Handbook, CRC and IEEE Pres, 1996

[8] J. A. Brzozowski, E. J. McCluskey, Signal Flow Graph Techniques for Sequential Circuit State Diagrams. IEEE vol. EC-12, pp. 67-76,1963

[9] L. M. Horowitz, Syntehsis of Feedback Systems, Academic Press INC. London, 2013

[10] R. S. Stankovic, J. Astola: From Boolean Logic to Switching Circuits and Automata, Springer, 2011

[11] L. H. Hackstaff: Systems of Formal Logic, Springer, 2012

[12] Ch. Posthoff, B. Steinbach: Logic Functions and Equations, Springer, 2009

[13] A. Leitsch: The Resolution Calculus, Springer, 1997

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 483-495 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Hardware Acceleration on Cloud Services: The use of Restricted Boltzmann Machines on Handwritten Digits Recognition

Eleni Bougioukou, Nikolaos Toulgaridis, Maria Varsamou, Theodore Antonakopoulos *

Department of Electrical and Computer Engineering, University of Patras, 26504 Rio - Patras, Greece

ARTICLEINFO

Article history:
Received: 30 November, 2017
Accepted: 07 January, 2018
Online: 18 February, 2018

Keywords:
Neural networks
Handwritten Digits Recognition
Cloud servers
Hardware accelerators

ABSTRACT

Cloud computing allows users and enterprises to process their data in high performance servers, thus reducing the need for advanced hardware at the client side. Although local processing is viable in many cases, collecting data from multiple clients and processing them in a server gives the best possible performance in terms of processing rate. In this work, the implementation of a high performance cloud computing engine for recognizing handwritten digits is presented. The engine exploits the benefits of cloud and uses a powerful hardware accelerator in order to classify the images received concurrently from multiple clients. The accelerator implements a number of neural networks, operating in parallel, resulting to a processing rate of more than 10 MImages/sec.

1 Introduction

We live in the era of 'Big Data', where a vast amount of structured, semi-structured and unstructured data are being generated at an ever-accelerating pace and can be mined to obtain valuable information. The most commonly used approach to process this kind of data is to aggregate raw data into large datasets, probably extend them with metadata, and then apply machine learning and/or artificial intelligence algorithms in order to identify repeatable patterns [1]. Artificial neural networks are a rapidly developing category of machine learning structures that give computers the capability to learn without being explicitly programmed to perform specific tasks. They consist of different layers for analyzing and learning data. Each layer consists of a large number of highly interconnected processing elements (neurons), working together to learn from previous data in order to solve specific problems by making proper decisions.

Boltzmann Machine (BM) is a typical example of a neural network structure. BMs are probabilistic Markov Random Field models that use a layer of hidden variables to model a distribution over input variables, called visible variables. In general, learning a Boltzmann machine is a computationally demanding process. However, the learning problem can be simplified by imposing restrictions on the network topology, which leads to Restricted Boltzmann Machines (RBMs)[2]. RBMs are structured as bipartite undirected graphs, which results to efficient inference implementation, and are particularly capable of learning complex features. The last few years, many applications based on RBMs have been developed to cover a large variety of learning problems, such as image classification, speech recognition, collaborative filtering and so on. One such example application is the recognition of handwritten digits. Learning an RBM corresponds to fitting its parameters, so that the distribution represented by the RBM models the distribution underlying the training data, handwritten digits in this case [3]. The storage resources and the time required not only to train an RBM but also to make real-time predictions on new coming data increases exponentially with the number of parameters. Thus the development of a handwritten digit recognition application on a single user machine is a non-trivial task.

Nowadays, cloud computing solutions provide new capabilities to users and enterprises for processing their data remotely in high performance servers. Users do not have to invest in information technology infrastructure, reducing the need for

*Theodore Antonakopoulos, University of Patras, Patras 26504, Greece, Email: antonako@upatras.gr, Tel.: +30(2610)996487

https://dx.doi.org/10.25046/aj030159

advanced hardware at the user side. In addition, cloud providers specialized in a particular area, such as image processing, can bring advanced services to a single user, complex services that are not easily afforded by individual users. Another important benefit is the high processing rate that can be achieved when a high performance server is used for processing requests from multiple independent users.

In this work, which is an extension of the work originally presented in [4] and [5], we exploit the vast amount of resources provided by cloud computing along with the high computation capabilities offered by hardware accelerators in order to build a complete cloud-based engine that can be used in real-time handwritten digits recognition (HDR). The engine collects images from multiple sources over the cloud and processes them as fast as possible resulting in high processing rate. The high processing rate is achieved by using a powerful hardware accelerator that implements a number of neural networks operating in parallel. A major advantage of the proposed engine is that the training of the neural networks is not only performed once during initialization of the system, but it is also fed with new images periodically, thus improving the accuracy of the prediction results. The engine is presented in detail in the sections that follow.

Section 2 gives an overview of existing implementations, especially on handwritten digits recognition. Section 3 analyzes Restricted Boltzmann Machines and how they can be modified in order to be used to solve classification problems, such as image recognition. Section 4 presents the architecture and the functionality of the proposed cloud-based computing server. Section 5 highlights the communication interface between the server and the hardware accelerator. Section 6 describes in detail the implementation of the neural networks, with emphasis on the architecture of the dedicated hardware accelerator. A complete system prototype was developed, which is presented in Section 7. Experimental results that demonstrate the system performance in terms of processing rate for both implementations are also presented.

2 State of The Art

The scientific area of automatic handwriting recognition is of great interest for both academia and industry. Existing algorithms are so efficient in learning to recognize handwritten digits that they are used, for example, by post offices to sort letters and banks to read personal checks. MNIST is the most widely used dataset for studying handwritten digit recognition [6]. State-of-the-art models that present accuracy results in the range of 0.35% down to 0.23% error rates are based on large Convolutional Neural Networks (CNNs), either in the form of a deep single network optimized with various training techniques or as a committee of many smaller networks [7], [8], [9]. The best results so far, 0.21% error rate, have been claimed by an approach based on regularization of neural networks using DropConnect [10].There is a listing of the state-of-the-art results and links to the relevant papers on the MNIST and other datasets collected by Rodrigo Benenson [11].

The disadvantage of CNNs is that they are very resource-demanding and their training and inference procedures are extremely time-consuming. As the amount of data increases, machine learning moves to the cloud and big clusters of high-performance servers are used for providing real-time results. Most works mainly deal with implementing the training procedure by using distributed servers over the cloud ([12], [13]) or by using higly scalable FPGA implementations ([14], [15]). In this work, a high performance computing engine that accepts images from multiple clients over the cloud and classifies them with high accuracy is presented.

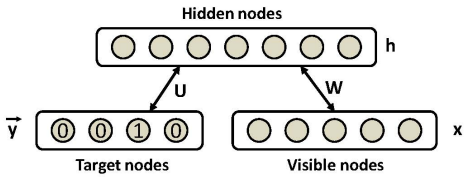


Figure 1: Discriminative RBM modeling the joint distribution of inputs and target classes.

3 Classification Restricted Boltzmann Machines

Restricted Boltzmann Machines are usually used as feature extractors for other learning algorithms or as initializers for deep feedforward neural network classifiers, not as standalone classifiers. However, authors in [16] have proposed a discriminative variant of RBMs that can be used autonomously in classification tasks offering good performance results. The bipartite undirected graph of such an RBM is illustrated in Figure 1. Given a training set $D_{train} = \{(x_1, y_1), ..., (x_i, y_i), ..., (x_D, y_D)\}$, where i denotes the i-th example of the set consisting of an input vector x_i and the corresponding target class $y_i \in \{1, ..., C\}$, we use the specific RBM to model the joint distribution between a layer of N hidden variables $\mathbf{h} = (h_1, ..., h_N)$, usually referred as features, and the observed variables $(\mathbf{x}, \mathbf{y})$. It is a parametric model where the parameters Θ = **(W,b,c,d,U)** represent the following:

- **W:** Weights matrix between **x** and **h**
- **U:** Weights matrix between $\mathbf{e}_y$ and **h**
- **b,c,d:** Respective biases of **x**, **h** and $\mathbf{e}_y$

and $\mathbf{e}_y = (1_{i=y})$ for $i \in \{1, ..., C\}$ the 'one out of C' vector representation of y.

We consider the binary version of the model where each node, hidden or visible, may be in one state, **ON** or **OFF**. A node adopts a new state as a probabilistic

function of the states of its neighboring nodes and the weights on its links to them. It stands:

$$p(h_j = 1|y,x) = sigm(c_j + u_{jy} + \sum_i w_{ji}x_i) \quad (1)$$

$$p(x_i = 1|h) = sigm(b_i + \sum_j w_{ji}h_j) \quad (2)$$

$$p(y|h) = \frac{\exp(d_y + \sum_j u_{jy}h_j)}{\sum_{\text{for all y}} \exp(d_y + \sum_j u_{jy}h_j)} \quad (3)$$

where $sigm(x) = 1/(1 + e^{-x})$ is the logistic sigmoid function.

After the model has been trained, the conditional probability $p(y|x)$ is used for classification. The conditional probability $p(y|x)$, can be computed using $p(y|x) = argmin(F(y,x))$:

$$F(y,h) = -d_y - \sum_j \log(1 + e^{(c_j + u_{jy} + \sum_i w_{ji}x_i)}) \quad (4)$$

where $F(y,h)$ is called *free energy*.

In order to train an RBM to solve a particular classification problem, an objective has to be defined that the learning procedure will try to minimize for all examples in the dataset D_{train}. It is possible to choose among various different objective functions, but generally the following three are used [17]:

- Generative Training Objective
- Discriminative Training Objective
- Hybrid Training Objective

3.1 Generative Training Objective

Given that the model defines a value for the joint probability $p(x,y)$, a natural choice for a training objective is the generative objective:

$$L_{gen}(D_{train}) = -\sum_{i=1}^{|D_{train}|} \log p(y_i, x_i) \quad (5)$$

Computing $\log p(y_i, x_i)$ and its gradient with respect to any RBM parameter Θ is intractable. Fortunately it has been shown that the gradient can be well approximated using the Contrastive Divergence estimator. The analytical computation is replaced by an estimate at a sample generated after a limited number of Gibbs sampling steps, with the sampler's initial state for the visible variables set at the training example (x_i, y_i). A single Gibbs sampling iteration is usually sufficient to learn a meaningful representation of the data [18]. Then, this gradient estimate can be used in a stochastic gradient descent procedure for training. A pseudocode of the procedure is given in Algorithm 1, where γ is the learning rate. Usually, the weights **(W,U)** are initialized using small random values, while the biases **(b,c,d)** are initially zero. Ideally, RBMs require parameter updating after each single example, but mini-batch updating can be also used. However, in order to ensure fast model convergence, the batch size should remain relatively small.

Algorithm 1 RBM Training over (x,y) using 1-step Contrastive Divergence

Input: w_{ij}

for all training samples (x,y) **do**

1. Calculate $p(h = 1|x,y)$ for all hidden nodes
   Sample the hidden distribution $< h_0 >$
2. Perform Gibbs sampling for k steps (k=1):
   Calculate $p(x|h_0)$ for all visible nodes and $p(y|h_0)$ for all target nodes
   Sample the visible distribution $< x_k >$ and the target distribution $< y_k >$
   Calculate $p(h = 1|x_k, y_k)$ for all hidden nodes
   Sample the hidden distribution $< h_k >$
3. Calculate gradients:
   $gW =< h_k > * < x_k > - < h_0 > *x$
   $gU =< h_k > * < y_k > - < h_0 > *y$
   $gb =< x_k > -x$
   $gc =< y_k > -y$
   $gd =< h_k > - < h_0 >$
4. Update weights and biases:
   $W' = W - \gamma * gW$
   $U' = U - \gamma * gU$
   $b' = b - \gamma * gb$
   $c' = c - \gamma * gc$
   $d' = d - \gamma * gd$

end for

3.2 Discriminative Training Objective

The generative training objective can be decomposed as follows:

$$L_{gen}(D_{train}) = -\sum_{i=1}^{|D_{train}|} \log p(y_i|x_i) - \sum_{i=1}^{|D_{train}|} \log p(x_i) \quad (6)$$

This means that the RBM classifier will dedicate some of its capacity at modeling the marginal distribution of the input only. Since classification is a supervised learning task and we are only interested in obtaining a good prediction of the target given the input, we can ignore the unsupervised part of the generative objective and focus on the supervised part. So, the discriminative training objective is defined as:

$$L_{disc}(D_{train}) = -\sum_{i=1}^{|D_{train}|} \log p(y_i|x_i) \quad (7)$$

The most important advantage using this training objective is that it is possible to compute exactly its

gradient with respect to the RBMs parameters for each example (x_i, y_i). For the various model parameters it stands:

- For Θ = **(c,W,U)**:

$$\frac{\vartheta \log p(y_i|x_i)}{\vartheta \Theta} = \sum_j sigm(O_{y_i j}(x_i)) \frac{\vartheta O_{y_i j}(x_i)}{\vartheta \Theta} - \sum_{j,forally} sigm(O_{yj}(x_i)) p(y|x_i) \frac{\vartheta O_{yj}(x_i)}{\vartheta \Theta} \quad (8)$$

where $O_{yj}(x) = c_j + u_{jy} + \sum_i w_{ji} x_i$ for hidden node j.

- For Θ = **(d)**:

$$\frac{\vartheta \log p(y_i|x_i)}{\vartheta d_y} = 1_{y=y_i} - p(y|x_i), \forall y \in \{1, ..., C\} \quad (9)$$

- For Θ = **(b)** the gradient is zero, since the input biases are not involved in the computation of $p(y|x)$.

3.3 Hybrid Training Objective

The effectiveness of both generative and discriminative approaches on various problems has been studied and it has been shown that they have quite different properties. For classification tasks, adding the generative training objective to the discriminative training objective is a way to regularize the second one. To adapt the amount of regularization, the Hybrid Training Objective can be used:

$$L_{hybrid}(D_{train}) = L_{disc}(D_{train}) - \alpha L_{gen}(D_{train}) \quad (10)$$

where the weight α of the generative part can be adjusted based on the performance of the model on a validation set.

3.4 Training using MNIST

In this work, RBMs were applied on a classic classification problem, handwritten digit recognition using the MNIST dataset [6]. MNIST is a large database of handwritten digits, commonly used for training various image processing systems. The database is also widely used for machine learning and pattern recognition methods. The original MNIST dataset is used here, which contains 60,000 training and 10,000 test examples with 28x28 grey-scale images corresponding to all 0-9 digits. This is a multiclass classification problem, where the number of target classes is 10. Before final integration, the best parameters for the RBM model should be selected. These parameters include the number of hidden nodes N, the learning rate γ, which training objective is minimized, the generative weight α, as well as the batch size bs. For that reason, a complete Matlab simulation model was developed and the effect of the different parameters was studied using a validation set. It should be noted that for the specific problem, Hybrid Training led to faster convergence.

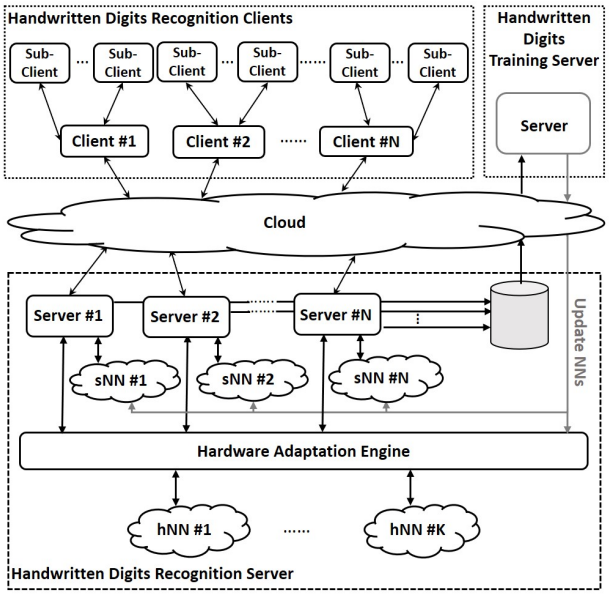


Figure 2: The HDR Computing Engine Architecture

4 HDR computing engine architecture

Based on Classification RBMs, a high performance computing engine able to serve a large amount of real-time requests for detecting the correct values of handwritten digits was implemented. This is a complete infrastructure with multiple entities specially designed either for training of already accumulated data or for real-time classification of multiple new images. Concerning the neural networks implementation, both software modules and dedicated hardware accelerators were developed.

The architecture of the proposed HDR computing engine is shown in Figure 2. This engine accepts requests from various clients from the cloud and processes them in real-time. The system is composed of three basic entities, which are a) Handwritten Digits Recognition clients, b) Handwritten Digits Recognition servers and c) Handwritten Digits (HD) training server. The HDR clients can either be individual clients or sets of sub-clients that are serviced by an HDR client that combines their requests in a single data stream. Each HDR client is associated with a dedicated HDR server. The number of HDR servers that are executed at any given time is variable and is determined by the number of HDR clients that are supported. The main advantage of the proposed architecture is that it does not use a set of predefined parameters but it uses the information of a large number of clients for continuously updating the weights and biases of the HDR algorithm, achieving

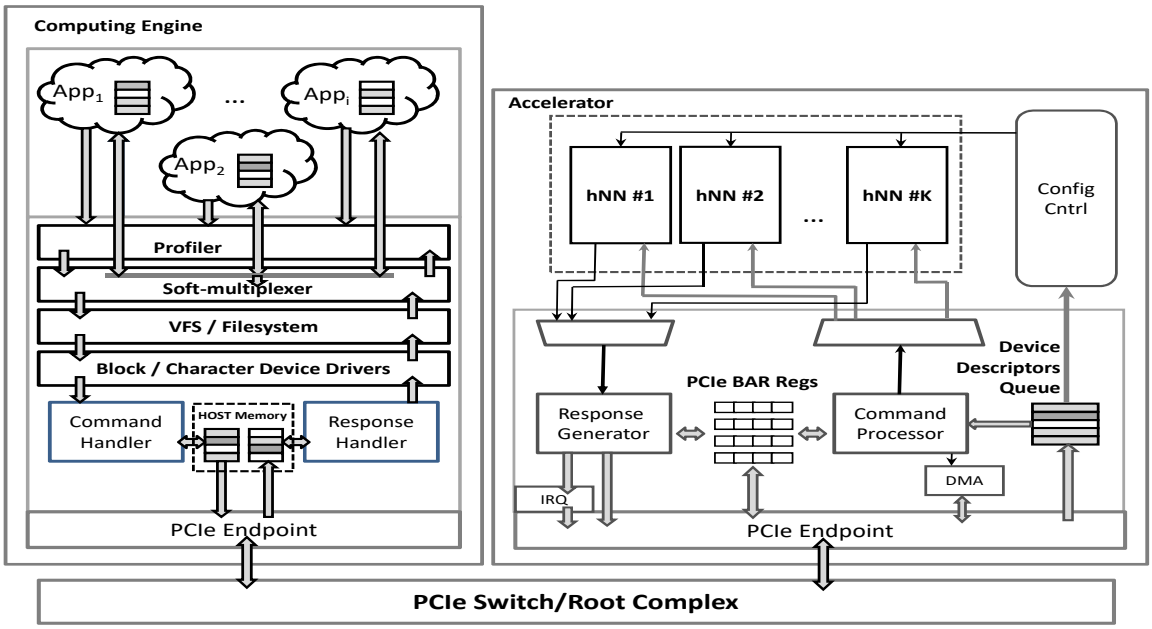


Figure 3: Hardware Adaption Engine - Interfacing with the HDR hardware accelerator

high accuracy for a given complexity. The HD training server is responsible for updating these parameters. More specifically:

HDR Clients: A two-way TCP/IP connection [19] between every HDR client and a dedicated HDR server is established. They exchange variable size data blocks, serviced either partially for better latency or as a whole block for better processing rate.

HDR Servers: When a request for a new connection arrives, a new HDR server thread is activated. Each thread receives bunches of images by its client and stores them temporarily to its local memory. Each HDR server may accept different types of requests from its client. The requests may differ at the number of images per block, the image size, i.e. 16x16 or 28x28 pixels, and pixel information coding (1 bit per pixel, 8 bits per pixel etc.). The server uses either its built-in processing capability to serve the requests, which means that the neural network is implemented in high-performance software, or forwards them to a dedicated hardware accelerator.

HD Training Server: The HD training server receives data and performs the training of the RBM neural networks. The data are forwarded to the HD training server either during initialization of the system or periodically during normal operation. In the latter case, the HDR servers send a random subset of images along with the information regarding the digit recognition. After training, the HD training server is responsible for sending the updated parameters, weights and offsets, back to all neural networks.

Concerning the neural network implementation there are three possible configurations.

(i) *Software-only Configuration:* a software NN (sNN) is initialized per HDR thread. Each NN accepts images by its corresponding thread, classifies them and returns the predicted digits.

(ii) *Hardware-only Configuration:* a small number of NNs implemented in hardware (hNN) is available to all server threads and is shared among them. The execution time is much smaller compared to the software implementation and that results to much higher processing rate. The hardware accelerator is based on a powerful FPGA board attached to the computing engine's CPU using a high-speed interface, either native PCIe Gen 3.0 with more than 1 GBps useful transfer rate [20] or the Coherent Accelerator Processor Interface (CAPI) [21]. An entity called Hardware Adaptation Engine (HAE) provides a seamless interface between the server threads and the accelerator. This entity is described analytically in the next section. Each HDR server does not use a specific hNN but whichever is available. The hardware accelerator may consist of multiple FPGA boards and/or GPUs. Although this architecture is significantly more efficient, in the case of a heavy workload, each hNN may have to process a huge bunch of images. In this case, a hybrid configuration is most preferable.

(iii) *Hybrid Configuration:* the images are still processed by the hardware accelerator (hNNs) but when the processing delay exceeds the software execution time, then the sNNs start receiving some of the images for classification.

5 Hardware Adaptation Engine

The HDR servers receive bunches of images from their clients and store them in a FIFO at the input

of HAE. Each HDR server has its own dedicated FIFO. HAE is responsible for forwarding these images for recognition to the hardware accelerator. HAE consists of multiple functional units that are shown in Figure 3. Initially, the Profiler decides the next block of images that has to be forwarded to the Soft-multiplexer. The number of images per block is selected so that the total system performance is maximized. The Soft-multiplexer interfaces with the hardware accelerator through Block/Character device drivers. The drivers and the accelerator communicate through shared host memory areas. When the images have been processed by the accelerator, a response with the digits values is fed back to the proper HDR server. A more detailed description of the HAE functional units follows.

5.1 Profiler

It is responsible for selecting the next block of images that has to be handled by the Soft-multiplexer. The Profiler decisions depend on the number of pending commands and the total response statistics (execution time etc.). The HDR servers inform the Profiler about their pending commands, while the Soft-multiplexer provides timing information regarding command execution. Initially, the Profiler selects the commands of the next block using either a round-robin or a static algorithm with fixed priorities. When enough statistics have been collected, dynamic allocation that takes into consideration the current load is feasible.

5.2 Soft-multiplexer

It makes the basic calls for interfacing with the Character/Block device drivers (e.g. open, close, pwrite, pread, ioctl etc). The Soft-multiplexer processes the selected block, extracts the commands and forwards them to the drivers. After receiving a notification for the completion of a command, the Soft-multiplexer forwards a response to the corresponding server thread. The soft-multiplexer also informs the Profiler about the execution time of each bunch of images.

5.3 Character/Block Device Driver

It is responsible for transferring the data from user space to kernel space and vice versa, for address translation and interrupt handling. The commands are associated with a data structure, in the form of a descriptor, which is the basic information provided to the hardware accelerator. The Device Driver activates the Command and Response Handler for forwarding commands and receiving the respective responses from the accelerator. The Command Handler accepts commands, creates the descriptors and stores them in a shared host memory area (Host Descriptors Queue). Each descriptor may contain one or more commands, associated with data from the same or different HDR servers. When a block of descriptors has been processed and their responses are ready, the accelerator sends an MSI-X Interrupt to notify the Response Handler that responses are stored in another shared host memory area (Responses Queue). The Response Handler decodes each response and forwards the corresponding results to the Soft-multiplexer.

At the accelerator's side, the Command Processor is responsible for accessing the descriptors at the host's main memory and transferring them into accelerator's local memory (Device Descriptors Queue). It decodes the descriptors, transfers the requested images using a DMA engine and feeds any available hNN. When the image processing has been completed the Response Generator prepares the responses that contain the classified digits, stores them in the Responses Queue and sends an interrupt to the Device Driver.

6 Neural Network Implementation

For better exploitation of the hardware resources, images of 16x16 pixels, with 1 bit/pixel, are used. This can be achieved by proper prefiltering and scaling, without affecting significantly the total system's accuracy. For this image size each hNN uses less hardware resources, thus enabling more hNNs to be integrated in the given accelerator resources, while the accuracy is only slightly reduced. These images are generated by the original ones after proper filtering and optimum threshold application. The pre-processing takes places at the client side before transmission. Before moving to the actual neural network implementation, a theoretical analysis of certain parameters regarding the prediction procedure is necessary.

6.1 Theoretical Analysis

As aforementioned, classification is based on the calculation of *free energy* [22], given by (4). For implementation purposes this equation is rearranged as follows:

$$S_m = \sum_{i=1}^{N_v} (v_i \cdot WV_{i,m}) + HB_m \tag{11}$$

$$LE_{k,m} = \log(1 + \exp(S_m + WT_{k,m})) \tag{12}$$

$$Fe_k = -(\sum_{m=1}^{N_h} LE_{k,m} + TB_k) \tag{13}$$

$$t_k = 1 : min(Fe_k) \tag{14}$$

where N_h is the number of hidden nodes, N_t the number of target nodes, $m \in [1, N_h]$, $k \in [1, N_t]$, v_i is visible node i, h_m is hidden node m and t_k is target node k. Fe_k is the *free energy* of target node k. Regarding the weights and biases, VB_i are the biases of visible nodes, TB_k are the biases of target

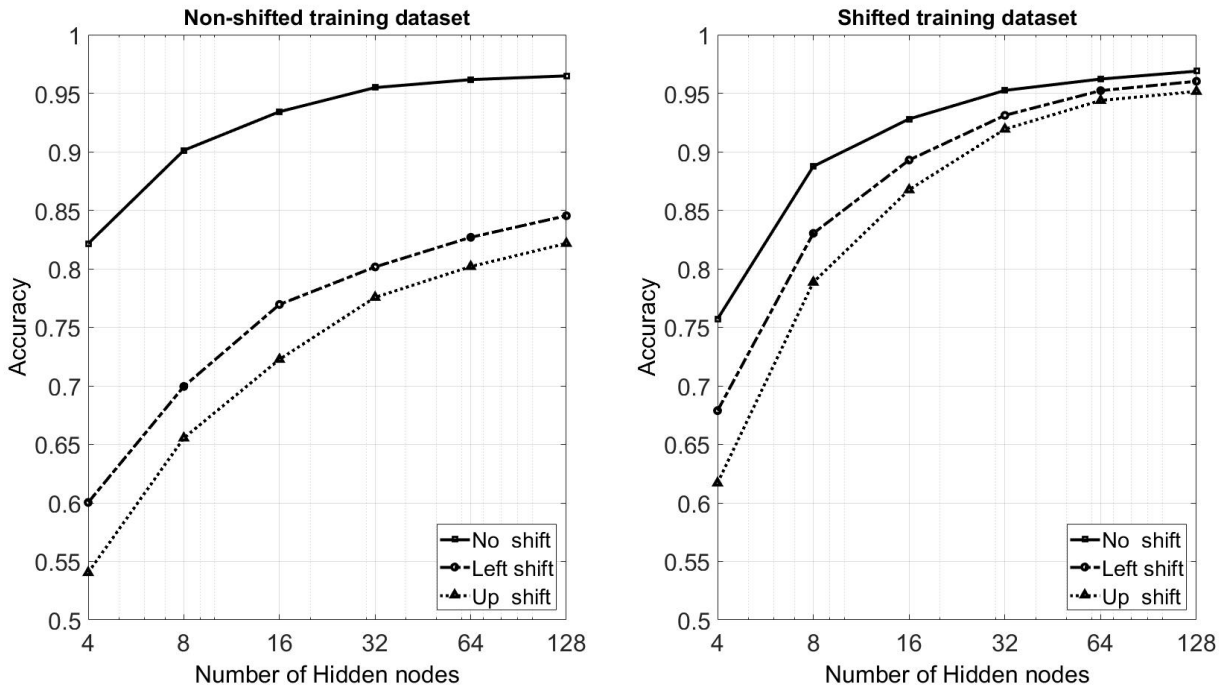


Figure 4: The influence of different training datasets on classification' s accuracy.

nodes, HB_m are the biases of hidden nodes, WV are the weights between visible and hidden nodes and WT are the weights between target and hidden nodes.

In order to study more thoroughly the accuracy of the proposed implementation and validate its effectiveness under a broader set of images, the initial dataset was extended with shifted versions of the original images. More specifically, four additional datasets were created, each one being a shifted copy of the initial one. The images were shifted to the left, right, up or down by a single pixel. In this way, the final dataset is a superset of the original one. Training was performed using both datasets, the original non-shifted one and, to keep the dimensions the same, a randomly selected collection of 60,000 images from the shifted dataset.

Figure 4 shows the effect of the training dataset on the classification accuracy of the model, when it is applied on three individual datasets, the original test dataset with the non-shifted images, the test dataset with the images that have been shifted left by one pixel and the test dataset with the images that have been shifted up by one pixel. As expected, the use of the shifted version for training leads to better overall accuracy, except in the case of the original dataset and for a small number of hidden nodes. This is explained by the fact that an RBM can model fewer dependencies when it includes a small number of hidden nodes, thus being subject to overfitting on the training dataset. Similar results can be taken for the datasets with the one-pixel right and down shifting.

Figure 4 also helps to specify the number of hidden nodes that will be used in both the software and hardware implementations of the neural network. It is obvious that the choice of 32 hidden nodes is a satisfactory trade-off between accuracy and complexity. By choosing 64 hidden nodes the accuracy improves only 2.2%, while the hardware complexity increases almost 8 times.

6.2 Software only Configuration

As described in Section 4, each HDR server receives requests from multiple clients. In software implementation each HDR server implements one RBM neural network. In order to have multiple sNNs operating in parallel, the server application is multithreaded, one thread per neural network. When a high performance CPU engine is used, multithreading results to better exploitation of the available cores. Software multithreading is mostly effective on multiprocessor or multicore systems, where actual parallel or distributed processing is feasible.

Each thread is responsible for classifying the incoming images, by calculating their *free energy* (4) with respect to each one of the ten possible classes. Therefore, vector and matrix operations dominate the computations. To achieve high performance, the threads use off-the-shelf highly-optimized libraries that exist for a variety of computer architectures.

6.3 Hardware only Configuration

Regarding the implementation of the hardware accelerator, a major consideration involves the arithmetic, fixed or floating-point, that will be

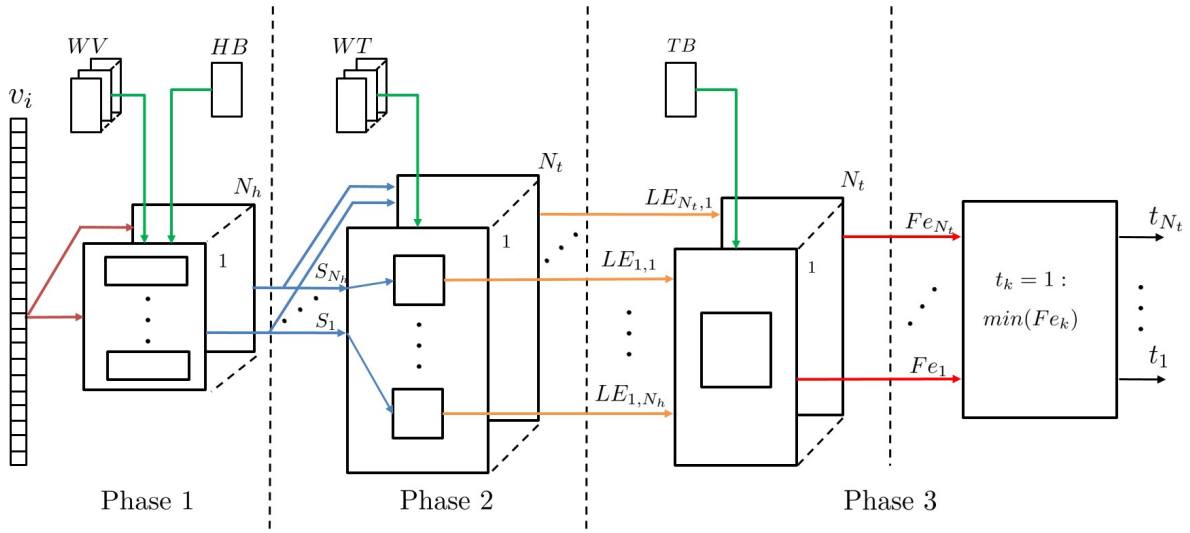


Figure 5: Accelerator Architecture

used for the calculations. The proposed design uses fixed-point arithmetic for all linear functions (*add*, *mul*, *cmp*) and single-precision floating-point for non-linear functions, like *log* and *exp* in (12). To specify the range of the fixed-point arithmetic, simulations were run and statistics were collected based on all available training and test data patterns.

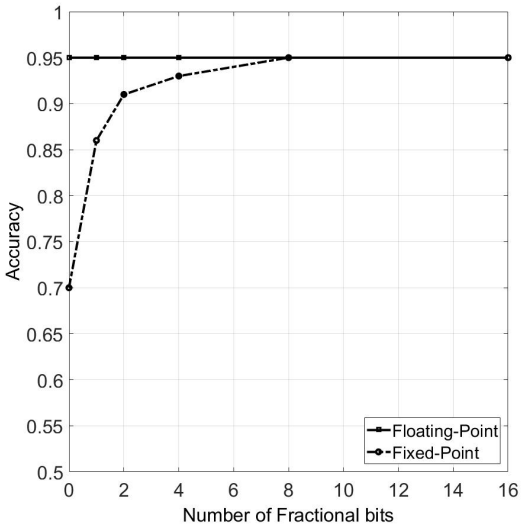


Figure 6: Accuracy vs Fractional bits ($N_h = 32$)

Figure 6 shows the accuracy results for various fixed-point number configurations. It can be seen that by using 16-bits fixed-point numbers, with the first 8 bits being used for the sign bit and the integer part and the remaining 8 bits for the fraction, a good classification accuracy of 95% can be achieved.

The proposed architecture is based on equations (11), (12) and (14) and is organized into three separate phases, shown in Figure 5. The time required to execute each phase determines the performance increase that can be achieved by proper pipeline. Whenever needed, the results of each phase are stored in shared memories and/or cascaded registers. The weights and biases are stored in dual-port RAMs and/or FIFOs, so that they are initialized/updated during system operation when new training data have been used.

A detailed description of the three phases comprising the accelerator architecture follows:

(i) *Phase 1:* Each image consists of a set of pixels, which are represented as binary values. This simplifies the multiplication part of (11), since instead of arithmetic multiplications, low complexity multiplexing functions can be used. The incoming image determines the weights that have to be used in the sum of products. Figure 7 shows a detailed scheduling of the operations implemented in this phase. Each *load* operation refers to reading WV values from memory. The *sel* of each multiplexer is connected to the corresponding pixel of the input image. This module needs 2 clocks to perform *load* and *mux* operations in order to feed the adder trees. After this point all additions are completed in 6 clocks. A total of 4 such modules is used.

Each of these modules includes 256 adders and is responsible for calculating 8 S_m results. Equation (11) determines that a total of 32 S_m results is needed. Every module operates in a pipeline manner. Although each S_m needs 8 clocks to be calculated, by using pipeline a total latency of 37 clocks is achieved.

(ii) *Phase 2:* Apart from the first addition, phase 2 is implemented using floating-point arithmetic. It was designed using the functionalities of Xilinx's Floating-point IP core [23]. The architecture of phase 2 and the respective timing diagram is given in Figure 8. Equation (12) is implemented in two similar and with same latency stages: $exp(a + b)$ and $ln(a + b)$.

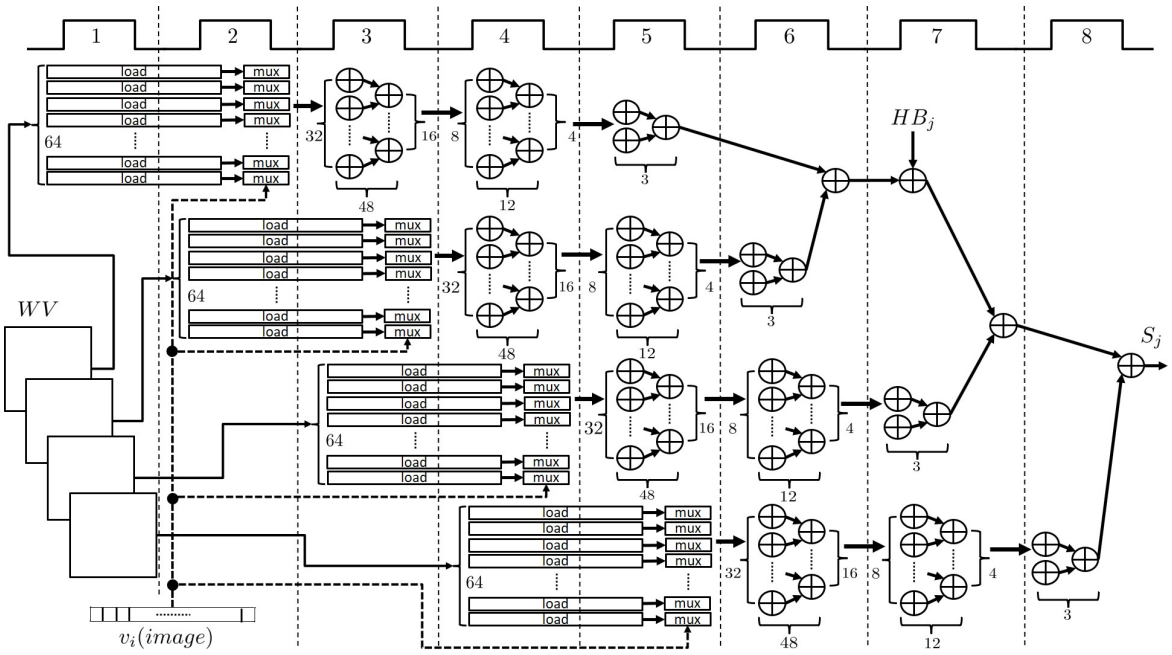


Figure 7: Phase 1 of Accelerator Architecture: Addition and Multiplexing Scheduling

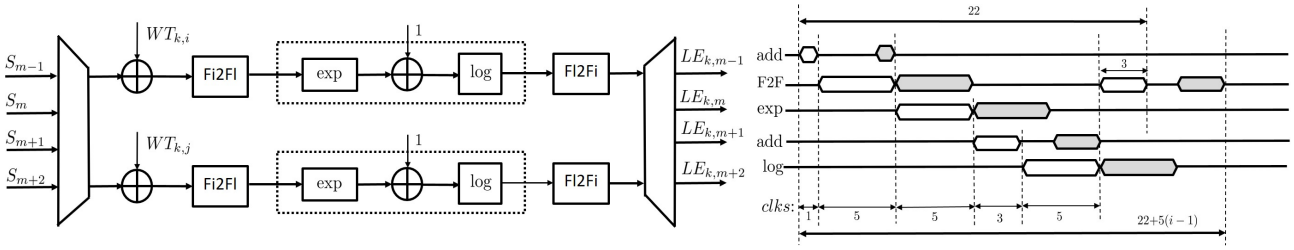


Figure 8: Phase 2 of Accelerator Architecture: Block Design (left) and Timing Diagram (right)

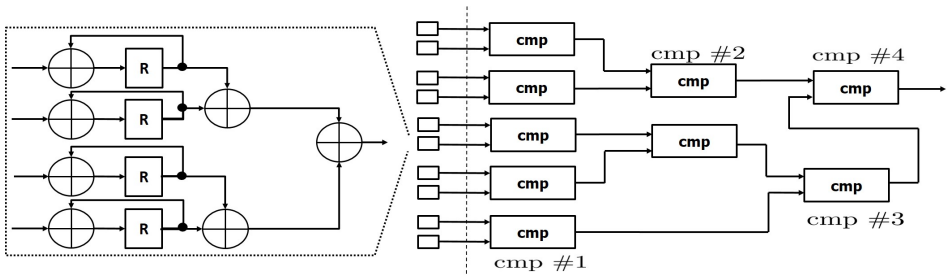


Figure 9: Phase 3 of Accelerator Architecture: Block Design

Therefore, two LE values can be calculated simultaneously using pipeline. For optimum performance, multiplexing of the incoming S_m values is performed, thus reducing the hardware complexity without decreasing the processing rate, and then fixed-to-floating point transformation is applied. Since the Fixed-point addition with the Fixed-to-Floating ($F2F$) operation lasts for less than the duration of each of the aforementioned processing stages, the same floating-point circuit (indicated with a dotted line) can be used for processing continuously a stream of S_m values. Since a total of $N_h x N_t$ LE values have to be calculated, the number of such circuits is determined by the multiplexing/demultiplexing used. $Fl2Fi$ in this figure corresponds to floating-to-fixed transformation. Phase 2 is the slowest phase of the whole accelerator. Reducing its latency would lead to improvement of the whole processing rate of the accelerator. To achieve this, sets of four LE values are calculated simultaneously. However, the required DSP resources are doubled.

(iii) *Phase 3:* The last phase is divided into two distinct sub-phases. The first accumulates the $LE_{k,j}$ inputs from Phase 2 to produce the Fe_k

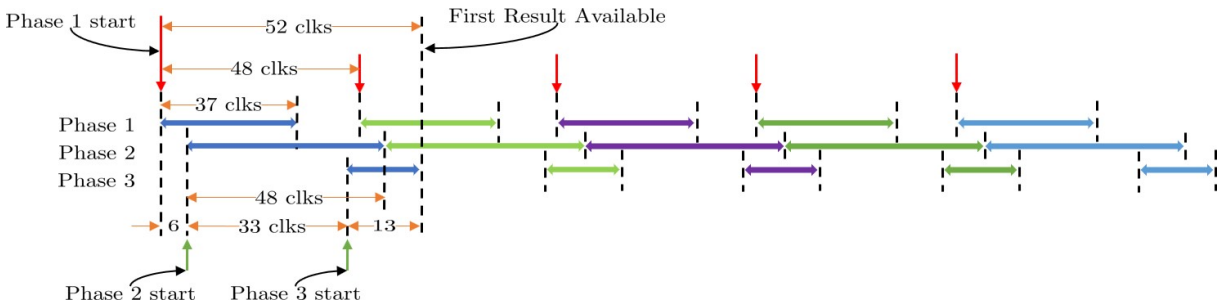


Figure 10: Pipeline strategy applied between processing phases

results (13). The second sub-phase uses the N_t computed Fe values in parallel and in a few clock cycles determines the digit (14) that corresponds to the class of the input image. Figure 9 presents the architecture of Phase 3.

Since all these phases have different processing times and do not require all the results of the previous phase to be available in order to proceed, pipeline can be used for increasing the whole system's processing rate. The latency of phase 1 is 37 clocks. In phase 2 the latency is 48 clocks and 10 such circuits are used in parallel. The combined latency for phase 3 is 10 clocks and 10 circuits implementing the accumulation process are used. Therefore the total latency per hNN without pipeline is 95 clock cycles. A pipeline of two is used and so the module is ready to receive new images every 48 clock cycles, with a latency of 52 clock cycles, as shown in Figure 10.

7 System Prototyping and Performance Results

The above described HDR Cloud Service has been implemented and tested in Xeon and Power8 servers (Table 1) while the hardware accelerator has been implemented in a Virtex-7 FPGA. For generating various workloads, an i7 server has been used, where various clients were implemented with user-defined workload patterns. The HDR clients send bunches of images over 1 Gbps ethernet link.

7.1 Software only Configuration

The applications of clients and servers were developed in C. Provided that HDR servers are multi-threaded, a very popular API is used for threading an application, known as POSIX Thread [24]. Also, CBLAS library is used for performing all the necessary vector and matrix operations. The OS of all platforms is Linux ubuntu 15.10.

In order to validate the efficiency of this configuration, the data rate achieved (in KImages/secs) for various computing servers (Table 1) and for various bunches of images (KImages/req) was measured. To preserve consistency with the hardware implementation, the number of hidden nodes used in sNNs is 32.

Table 1: Clients and Server Platforms

Platfrom	CPU		Memory		
	Cores	GHz	GB	Type	MHz
Xeon	6	1.60	16	DDR3	2133
Power8	4	3.00	64	DDR4	1600/1333
Power8	8	3.32	64	DDR4	1600/1333

When the number of images per request is small the total performance drops because the time between consecutive requests is much higher compared to the processing time of each request. On the other hand, the number of images per request does not affect the performance when it is more than a few hundreds since the overhead is absorbed. This can be seen in Tables 2 and 3.

Table 2: Processing rate [KImgs/sec] using software HDR implementation (N_t =32 and 1KImgs/req)

	Number of HDR Clients			
Servers	1	4	16	64
Intel Xeon	13.20	55.01	82.31	83.10
Power8	16.12	50.52	95.01	98.14
Power8	14.39	46.92	149.03	171.86

Table 3: Processing rate [KImgs/sec] using software HDR implementation (N_t =32 and 10KImgs/req)

	Number of HDR Clients			
Servers	1	4	16	64
Intel Xeon	13.31	55.36	82.11	82.77
Power8	16.20	59.82	98.13	100.09
Power8	14.48	54.88	155.33	171.20

Based on Tables 2 and 3, it is evident that the system performance has a linear relation with the number of HDR clients. For a small number of clients, the system is not fully utilized due to the communication overhead. Each client waits for the completion of a request before sending a new one. As the number of clients increases, this overhead is amortized and full system utilization is achieved. So, the performance increases with the increase in the number of clients. The performance reaches different maximum values, depending on the server platform

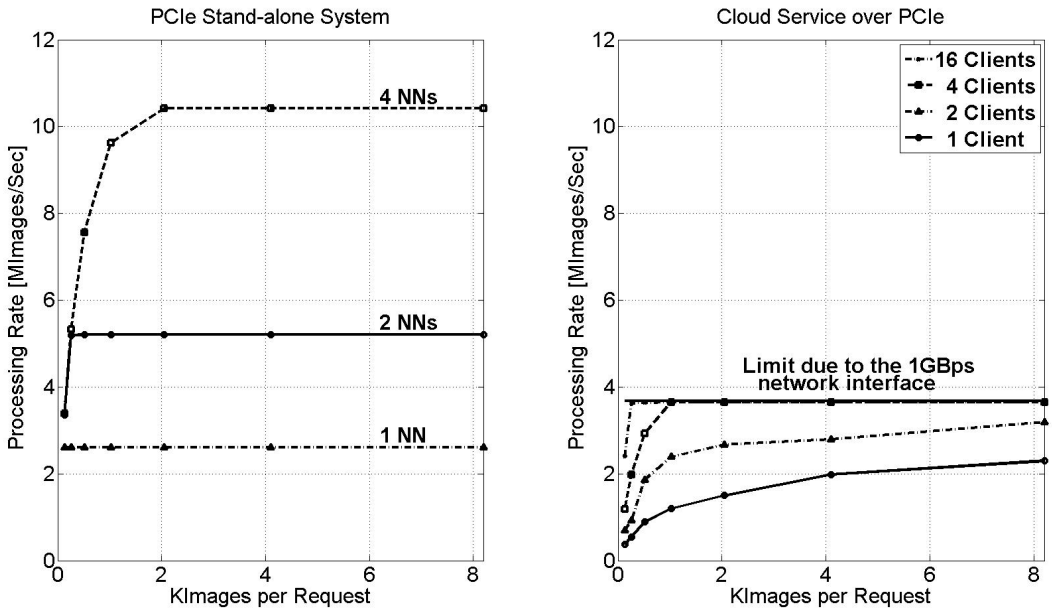


Figure 11: Processing rate using hardware HDR implementation over PCIe

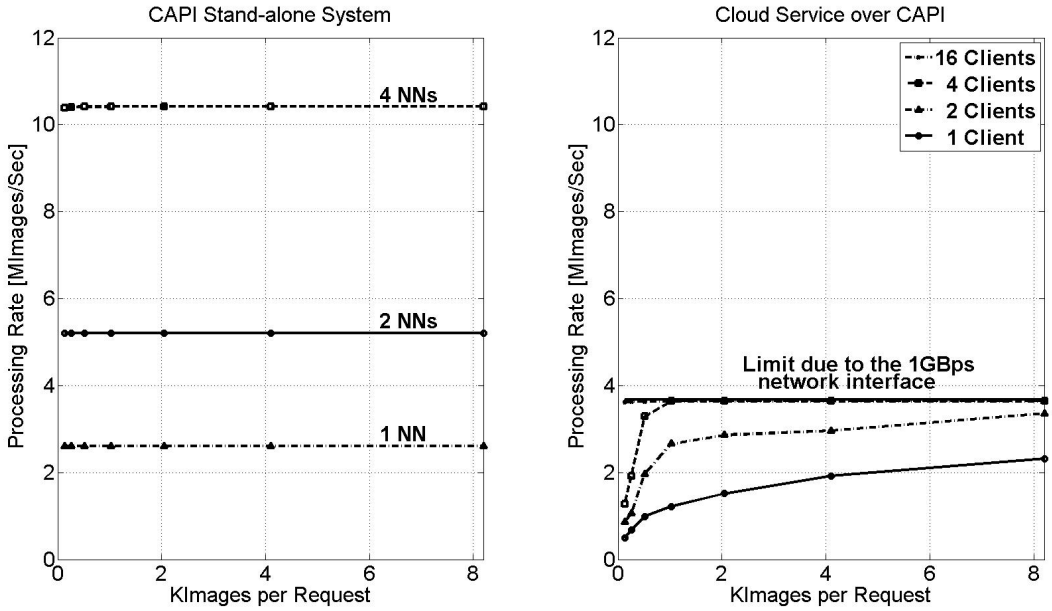


Figure 12: Processing rate using hardware HDR implementation over CAPI

used. This is mainly due to the fact that each server has different number of CPU cores (Table 1). As mentioned, when multiple threads are running, they are distributed to different cores and exploit better the capabilities of each CPU.

7.2 Hardware only Configuration

As far as the accelerator is concerned, based on the latency numbers provided above, a processing rate of 2.6 MImages/sec, at 125 MHz, is achieved by each hNN module. To maximize the total accelerator performance, four hNN modules operate in parallel. That results to a total processing rate of 10.4 MImages/sec, at 125 MHz. The limitation of four modules is introduced by the available resources of the specific Virtex-7 FPGA board (Table 4).

The accelerator supports two interfaces, PCIe Gen 3.0 with 8 lanes and Coherent Accelerator Processor Interface (CAPI). In both cases, a databus of 1024 bits is used that provides images up to four hNNs simultaneously.

The accelerator was attached to a Power8 server with 8 cores (Table 1). For each interface (PCIe, CAPI) two different sets of experiments were conducted in order to measure the performance of the hardware configuration. The results concerning the PCIe are shown in Figure 11. The left part illustrates the

accelerator's processing rate by varying the number of images per request as well as the number of hNNs implemented in the accelerator. In this case, the images are stored locally to the server and not received from multiple clients over the network. The number of images per request does not affect the performance when 1 or 2 hNNs are implemented in the accelerator. The rate is 2.6 MImages/sec and 5.2 MImages/sec correspondingly. When 4 hNNs are integrated, the maximum rate (10.4 MImages/sec) is achieved when each request contains at least 2K images.

Table 4: Implementation parameters of a neural network (N_t =32, XC7VX690T-2 Virtex-7 FPGA)

	Phase 1	Phase 2	Phases 3-4	Total
BRAMs	4 %	1 %	1 %	6 %
DSP	-	13 %	-	13 %
FFs	1 %	3 %	1 %	5 %
LUTs	7 %	8 %	1 %	16 %
Slices	-	-	-	18 %

The right part of Figure 11 illustrates the processing rate, when 4 hNNs operate in parallel, varying the number of connected clients over the network. It is obvious that the maximum rate of this set-up can be achieved even for a small number of active clients. Although in the case of 4 hNNs the maximum achievable rate is 10.4 MImages/sec it can be seen that the maximum processing rate of the whole set-up is less due to the used network interface (934.4 Mbps data rate over a 1 Gbps Ethernet link), since the communication interface becomes the systems bottleneck. Nevertheless, the system with 1 Gbps network interface and hardware acceleration is up to 21 times faster compared to the software implementation.

The achieved processing rates when CAPI is used are shown in Figure 12. The performance is slightly better when 4 hNNs are implemented in the accelerator as it can be seen in the left part of the figure. In this case, the accelerator reaches its maximum value even for a small number of images per request. So, the configuration with CAPI outperforms the configuration with native PCIe. Traditional I/O attachment protocols, like PCIe, introduce significant device driver and operating system latencies, since an application calls the device driver to access the accelerator and the device driver performs a memory mapping operation. With CAPI instead, the accelerator is attached as a coherent CPU peer over the I/O physical interface.

8 Conclusions

The design and implementation of a computing engine for handwritten digits recognition was presented. This engine can be used for cloud applications and achieves high performance, in terms of processing rate, when a hardware accelerator with multiple neural networks is used, as demonstrated by experimental results. Details of neural networks implementation on reprogrammable logic have also been described. The architecture can be parameterized in order to achieve the best compromise between hardware complexity and processing performance.

References

[1] A. Smola and S. VishwanathanKevin, *Introduction to Machine Learning*. Cambridge, UK: Cambridge University Press, 2008.

[2] A. Fischer and C. Igel, "An introduction to restricted boltzmann machines.," in *CIARP* (L. Alvarez, M. Mejail, L. G. Deniz, and J. C. Jacobo, eds.), vol. 7441 of *Lecture Notes in Computer Science*, pp. 14–36, Springer, 2012.

[3] A. Fischer and C. Igel, "Training restricted boltzmann machines," *Pattern Recogn.*, vol. 47, pp. 25–39, Jan. 2014.

[4] E. Bougioukou, N. Toulgaridis, and T. Antonakopoulos, "Cloud services using hardware accelerators: The case of handwritten digits recognition," in *Proceedings of the 6th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, (Thessaloniki), 4-6 May 2017.

[5] N. Toulgaridis, E. Bougioukou, and T. Antonakopoulos, "Architecture and implementation of a restricted boltzmann machine for handwritten digit recognition," in *Proceedings of the 6th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, (Thessaloniki), 4-6 May 2017.

[6] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," 2010. `http://yann.lecun.com/exdb/mnist/`.

[7] D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pp. 1237–1242, 2011.

[8] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012*, pp. 3642–3649, 2012.

[9] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," *ArXiv e-prints*, Sept. 2015.

[10] L. Wan, M. Zeiler, S. Zhang, Y. LeCun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, 2013.

[11] R. Benenson, "Classification datasets results." `https://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html`.

[12] H.-j. Zhang and N.-f. Xiao, "Parallel implementation of multilayered neural networks based on map-reduce on cloud computing clusters," 02 2015.

[13] S. Teerapittayanon, B. McDanel, and H. T. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," 09 2017.

[14] D. L. Ly and P. Chow, "A high-performance fpga architecture for restricted boltzmann machines," in *Proceedings of the ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, FPGA '09, (New York, NY, USA), pp. 73–82, ACM, 2009.

[15] S. K. Kim, L. C. Mcafee, P. L. Mcmahon, and K. Olukotun, "A highly scalable restricted boltzmann machine fpga implementation," in *Proceedings of the International Conference on Field Programmable Logic and Applications*, pp. 367–372, 08 2009.

[16] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, (Helsinki, Finland), pp. 536–543, 2008.

[17] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted boltzmann machine," *J. Mach. Learn. Res.*, vol. 13, pp. 643–669, Mar. 2012.

[18] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, p. 2002, 2000.

[19] M. J. Donahoo and K. L. Calvert, *TCP/IP Sockets in C: Practical Guide for Programmers*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2nd ed., 2009.

[20] PCI SIG, *PCI Express Base Specification, Revision 2.1*, March 2009.

[21] IBM Systems Magazine, *A Deeper Look at POWER8 CAPI and Data Engine for NoSQL*, May 2015. http://ibmsystemsmag.com/power/businessstrategy/competitiveadvantage/capi-deep-look/.

[22] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," in *Predicting Structured Data* (G. Bakir, T. Hofman, B. Schölkopf, A. Smola, and B. Taskar, eds.), MIT Press, 2006.

[23] Xilinx Inc., *Floating-Point Operator*. https://www.xilinx.com/products/intellectual-property/floating_pt.html.

[24] B. Nichols, D. Buttlar, and J. P. Farrell, *Pthreads Programming*. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 1996.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 496-501(2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

An Analysis of K-means Algorithm Based Network Intrusion Detection System

Yi Yi Aung*, Myat Myat Min

Faculty of Computer Sciences, University of Computer Studies, Mandalay, UCSM, 0000, Myanmar

ARTICLE INFO

Article history:
Received: 30 November, 2017
Accepted: 30 January, 2018
Online: 10 February, 2018

Keywords:
Network Intrusion Detection System
K-means
Random Forest
KDDCup 99

ABSTRACT

In this modern age, information technology (IT) plays a role in a number of different fields. And therefore, the role of security is very important to control and assist the flow of activities over the network. Intrusion detection (ID) is a kind of security management system for computers and networks. There are many approaches and methods used in ID. Each approach has merits and demerits. Therefore this paper highlights the similar distribution of attacks nature by using K-means and also the effective accuracy of Random Forest algorithm in detecting intrusions. This paper describes full pattern recognition and machine learning algorithm performance for the four attack categories, such as Denial-of-Service (DoS) attacks (deny legitimate request to a system), Probing attacks (information gathering attacks), user-to-root (U2R) attacks (unauthorized access to local super-user), and remote-to-local (R2L) attacks (unauthorized local access from a remote machine) shown in the KDD 99 Cup intrusion detection dataset.

1. Introduction

On the Internet, users share valuable information around the world. The internet has created various ways to threaten the stability and security of interrelated systems. Both of these mechanisms are static and dynamic. Static mechanisms like firewalls and software updates provide dynamic security and mechanisms such as intrusion detection systems. Today, security is the most serious problem for getting valuable information. Therefore, static mechanisms or dynamic mechanisms are needed to protect individual information despite the precautionary technology. The intrusion detection system detects not only successful aggression, but also helps monitor and prevent timely action.

The intrusion Detection System (IDS) is a standard component of a security infrastructure that allows network administrators to detect policy violations. Check all incoming and outgoing network activity and determine suspicious patterns that indicate network or system attacks from people trying to break or compromise the system.

A secure network must provide the following:

- Data confidentiality: Data transferred over the network must be accessible only to data that has been approved accordingly.
- Data integrity: Data must maintain integrity from when it is sent when it is received. No damage or loss of data from random events or malicious activities is accepted.
- Data availability: The network must be resistant to service attack denial.

IDS technology based on tracking process can be categorized into two approaches:

Abuse/Signature detection: This technology searches for signature attacks and known signatures in network traffic and are used as a reference to detect future attacks. Regularly updated databases are usually used to store signatures of known attacks. The way this technology controls intrusion detection is similar to antivirus software. The advantage of this type of detection is that it can accurately and efficiently detect known attacks. Anomaly detection: This technology is based on tracking traffic anomalies. The gap between traffic is monitored and regular profiles are measured. Different implementations of this technology have been reserved based on metrics used to measure the deviation of traffic profiles. The advantage of this detection type is that it is well suited to detect unknown attacks.

IDS are divided into two parts based on analysis and retention of audit data:

* Yi Yi Aung, Email: yiyiaungresearch@gmail.com

Host-based IDS (HIDS): HIDS is a home based tracking method that allows the system to collect data in the form of multiple host activity records, such as event logs and system logs. Since everything is in the host, there's no need to install additional hardware or software [1]. The advantages of hosted IDS are to check the success or failure of attacks, monitor system activity, and detect attacks that IDS networks cannot detect, close tracking and real-time responses, are not required.

Network-based IDS (NIDS): NIDS is a network approach that collects data directly from a network monitored as a packet instead of collecting data from a particular host / agent. Most NIDS are a free and easy-to-use operating system [2]. Network-based IDS offers advantages such as low cost of ownership, easier, placement, network attack detection, evidence preservation, real-time tracking and rapid response, and detection of failed attacks.

2. Literature Review

Most of intrusion detection system focused on four major attack categories such as denial of service, probe, user-to-root, and remote-to-local but this author specially emphasized on User-to-Root (U2R) attacks in NSL-KDD dataset by using Weka tool. This paper focused on a comparative study analysis of user-to-root attack, which the attacker tries to access normal user account and gains root access information of the system based on several machine learning techniques such as navie bayes, random forest, J48, etc [3].

This paper analyzed anomaly intrusions detection system by using Random Forest classifier with Principal Component Analysis. The author got experimental results by using simulation connection dataset of NSL-KDD. The performance of the system was measured by using Precision, Recall and F-Measure. And also this paper was specially focused on to detect various attacks present in Denial of Service (DoS) such as Neptune, Smurf, Pod, Teardrop, Land, Back, Apache2, Processtable, Mailbomb [4].

This paper used C4.5, CART (Classification and Regression Trees), Random Forest, and REP (Reduced Error Pruning) Tree to investigate the detection of intrusions contained in KDDCUP 1999 DARPA dataset. And compared the performance of the above algorithms based on the measures such as Accuracy, Learning Time (in seconds) and Size of the Tree. According to the experimental results, Random Forest was better as it correctly identifies more number of instances than other. And the accuracy of the REP Tree was very less than other algorithms but the learning time of REP Tress is very less than other [5].

They used Support Vector Machine with Principal Component Analysis (PCA) to choose the optimum feature subset that was useful in applying for intrusion detection system. To determine the effectiveness and feasibility of the proposed IDS system, they choosed NSL-KDD dataset for simulation their system. They found that PCA algorithm is good to select a best subset of features for classification of intrusions. It can help to speed up the training and testing process of intrusions detection which is important for high-speed network applications [6].

In this proposed paper, several classification techniques and machine learning algorithms have been considered to categorize the network traffic. Out of the classification techniques, they have found nine suitable classifiers like BayesNet, J48, PART, JRip, Random Tree, Random Forest and REPTree. The comparison of these algorithms has been performed using WEKA tool [7].

Security has become a crucial issue for computer systems. IDS can protect to our computer network. Different classification and clustering algorithms have been proposed in recent year for IDS. In this paper, multiple algorithms were analyzed to find the optimal algorithm. At last the optimal algorithms Random Forest and DB Scan were occurred for IDS [8].

The purpose of this survey paper was to describe the methods/ techniques which are being used for Intrusion Detection based on Data mining concepts and the designed frame works. This survey paper stated the methods and techniques of data mining to aid the process of Intrusion Detection and the frameworks [9]. The concept of intercepting these two different fields, gives more scope for the research community to work in this area. New approaches enhanced the existing interference detecting system and it was a stepping stone to build effective and efficient IDS to detect different types of attacks [10].

This paper proposed a novel hybrid model for intrusion detection. The proposed framework in this paper may be expected as another step towards advancement of IDS. The Hybrid framework led to effective, adaptive and intelligent intrusion detection [11].

This paper drew the conclusions on the basis of implementations performed using various data mining algorithms. Combining more than one data mining algorithms had be used to remove disadvantages of one another and lead to a better performance than any single classifier. Different classifiers had different knowledge regarding the problem [12].

3. Methodology

This section consists of the conversation of the two algorithms of data mining classification approaches. These are K-means and Random Forest.

3.1. K-means Clustering Algorithm

Clustering, based on distance measurements performed on objects, and classifying objects (invasions) into clusters. Unlike classification, classification because there is no information about the label of learning data is an unattended learning process. For anomalous detection, we can use welding and in-depth analysis to guide the ID model. Measurement of distance or similarity plays an important role in collecting observations into homogeneous groups. Jacquard affinity measurement, the longest common order scale (LCS), is important that the event is to awaken the size to determine if normal or abnormal. Euclidean distance is approximately two vectors X and Y in space Euclidean n-dimensions, the size of the distance widely used for vector space. Euclidean distance can be defined as the square root of the total difference of the same vector dimension. Finally, grouping and classification algorithms need to be channeled effectively, massively, it possible to handle dimension of network data and heterogeneity [13].

In this paper, we use K-means algorithm to cluster dataset connections. The K-means algorithm is one of the widely recognized clustering tools. K-means groups the data in accordance with their characteristic values into a user-specified number of K distinct clusters. Data categorized into the same cluster have identical feature values. K, the positive integer denoting the number of clusters, needs to be provided in advance. The steps involved in a K-means algorithm are given consequently: [14]

1. K points denoting the data to be clustered are placed into the space. These points denote the primary group centroids.
2. The data are assigned to the group that is adjacent to the centroid.
3. The positions of all the K centroids are recalculated as soon as all the data are assigned.
4. Repeat steps 2 and 3 until the centroid unchanged.

This results in the partition of data into groups. The preprocessed dataset partition is performed using the K-means algorithm with K value as 5. Because we have the dataset that contains normal and 4 attack categories such as DoS, Probe, U2R, R2L.

3.2. Random Forest Algorithm

One of the most popular methods or frameworks used by scientists in the science of data is Random Forest. It is a supervised classification algorithm. It can be seen from its name, which is to create a forest by some way and make it random. There is a direct relationship between the numbers of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach [15].

Random Forests grows many classification trees. Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random – but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number mM is specified such that at each node, m variables are selected at random out of the M and the best split on this m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

There are many of top benefits of Random Forest algorithm. Some of these benefits are as follows:

- Accuracy
- Runs efficiently on large data bases
- Handles thousands of input variables without variable deletion
- Provides effective methods for estimating missing data
- Maintains accuracy when a large proportion of the data are missing

4. KDDCup 99 Dataset

The evaluation of any intrusion detection algorithm on real network data is extremely difficult mainly due to the high cost of obtaining proper labeling of network connections. Due to the real sample cannot be gotten for intrusion detection, the KDDCup'99 dataset is used as the sample to verity the performance of the misuse detection model. The KDDCup'99 dataset, referred by Columbia University, was arranged from intrusions simulated in a military network environment at the DARPA in 1998. It contains network connections obtained from a sniffer that records all network traffic using the TCP dump format. The total simulated period is seven weeks. It was performed in the MIT Lincoln Labs, and then announced on the UCI KDD Cup 1999 Archive [16].

KDDCup'99 dataset have two variations of training dataset; one is a full training set having 5 million connections and the other is 10% of this training set having 494021 connections. Since the whole dataset is huge, the experiment has been performed on its smaller amount of dataset that is 10% of KDD. Additionally, the KDDCup'99 dataset includes many attack behaviors, classified into four groups: Probe, Denial of Service (DoS), User to Root (U2R), and Remote to Local (R2L) [17]. These can be seen in table I. Normal connections are created to profile that expected in a military network. The detailed information of the two variations of training dataset can be seen in table II.

Table I: Various Attacks and Categories

Categories	Attacks Subclass
DoS	back, land, Neptune, pod, smurf, teardrop
Probe	ipsweep, nmap, portsweep, satan
U2R	buffer_overflow, loadmodule, perl, rootkit
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster

Table II: Number of Instances in KDD and 10% KDD

Class	Whole KDD	10 % KDD
DoS	3883370	391458
Probe	41102	4107
U2R	52	52
R2L	1126	1126
Normal	972780	97278
Total	4898430	494021

The data set includes 41 features classifying the data records into normal or a type of attacks. The features consist of 34 types of numeric features and 7 types of symbolic features, according to different properties of attack. The nature of features can be divided into the following groups [18].

- Basic Features: Basic functions can be obtained from the packet header without checking the load.

- Content Features: Domain knowledge is used to assess the original TCP packet load. This includes features such as the number of unsuccessful login attempts.
- Time-based Traffic Features: This function is designed to capture properties in the 2-second window. Examples of such functions are the number of connections to the same host every two seconds.
- Host-based Traffic Features: Use the history window to estimate the number of connections (in the case 100) and not the time. Therefore, host-based functionality is designed to assess attacks that include two or more intervals.

4.1. Pre-Processing

KDDCUP 99 data set is pre-processed in order to make it suitable for the data mining learning algorithm. Pre-processing is performed for the following reasons.

Each record in the dataset consists of categorical as well as numeric features. Textual (plain) data is used for categorical features. K-means algorithm needs numeric data (either discrete or continuous). The first step in pre-processing is to covert this categorical feature attributes to numeric attributes. For converting symbols into numerical form, an integer code is assigned to each symbol. For instance, in the case of protocol type feature, 0 is assigned to tcp, 1 to udp, and 2 to the icmp symbol and so on. The dataset contains three categorical attributes while the rest of the thirty eight attributes are numeric. Every category of an attribute is assigned a specific number.

We have used K-means and Random Forest to define normal and attacks in the system. They need specific format so we have converted the dataset to K-means and Random Forest compatible format.

5. Experimental Results and Discussion

To facilitate the experiments, we used eclipse java and weka tool to implement the algorithms on a PC with 64-bit window 7 operating system, 4GB RAM and a CPU of Intel core i3-4010U CPU with 1.70GHz. Data come from MIT Lincoln laboratory of KDDCup99 data set. The table lists the number of instances available in the whole dataset, 10% of KDDCup'99 dataset.

The analysis is performed by using K-means and Random Forest algorithms. We use K-means algorithm to generate heterogeneous dataset to nearly homogeneous dataset. The clustering results of K-means algorithm are described from table III to table VIII.

Table III: Detailed Information of Attack Categories in Cluster-1

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	107219	107217	2
Probe	1610	1605	5
U2R	0	0	0
R2L	6	3	3
Normal	10	3	7
Total	108845	108828	17

Table IV: Detailed Information of Attack Categories in Cluster-2

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	1067	1065	2
Probe	1221	1207	14
U2R	4	0	4
R2L	1	0	1
Normal	21235	21230	5
Total	23528	23502	26

Table V: Detailed Information of Attack Categories in Cluster-3

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	280782	280782	0
Probe	0	0	0
U2R	0	0	0
R2L	0	0	0
Normal	16	14	2
Total	280798	280796	2

Table VI: Detailed Information of Attack Categories in Cluster-4

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	2203	2202	1
Probe	12	1	11
U2R	46	29	17
R2L	1087	1068	19
Normal	75409	75398	11
Total	78757	78698	59

Table VII: Detailed Information of Attack Categories in Cluster-5

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	187	186	1
Probe	1264	1255	9
U2R	2	0	2
R2L	32	22	10
Normal	608	604	4
Total	2093	2067	26

Table VIII: Detailed Information of Attack Categories with Clustering

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	391458	391452	6
Probe	4107	4068	39
U2R	52	29	23
R2L	1126	1093	33
Normal	97278	97249	29
Total	494021	493891	130

By analyzing the clustering results, the characteristics of Denial of Service (DoS) attacks are mostly related to themselves in cluster-3. And then, it is closely similar to the nature of Probe attacks in cluster-1. Probe attacks are also mostly related to DoS attacks in cluster-1. And then, it is nearly same with the nature of

Normal by looking in cluster-5. Normal is mostly similar nature with User-to-Root attacks and Remote-to-Local attacks by studying in cluster-4. And then, Normal is related to Probe by studying cluster-2 and cluster-5. Normal is related to all attacks by looking in all 5 clusters because attacks mimic to normal behavior in intrusions.

Then we apply Random Forest algorithm to know the intrusions and normal traffic. The performance of attacks categories with Random Forest algorithm in 5 clusters of K-means can be seen from table IX to table XIV. The Precision and Recall of the normal and attacks detection are good and the false positive rate is nearly zero.

Table IX: Performance Analysis of Attack Categories in Cluster-1

Attacks	False Positive Rate	Precision	Recall
DoS	0.00738	0.999888	0.999981
Probe	0.000009	0.999377	0.996894
U2R	0	0	0
R2L	0.000018	0.6	0.5
Normal	0.000018	0.6	0.3

Table X: Performance Analysis of Attack Categories in Cluster-2

Attacks	False Positive Rate	Precision	Recall
DoS	0	1	0.998125
Probe	0.000224	0.995874	0.988533
U2R	0.000042	0	0
R2L	0	0	0
Normal	0.008722	0.999058	0.999764

Table XI: Performance Analysis of Attack Categories in Cluster-3

Attacks	False Positive Rate	Precision	Recall
DoS	0.125	0.999992	1
Probe	0	0	0
U2R	0	0	0
R2L	0	0	0
Normal	0.875	0	1

Table XII: Performance Analysis of Attack Categories in Cluster-4

Attacks	False Positive Rate	Precision	Recall
DoS	0	1	0.999546
Probe	0	0	0.083333
U2R	0.000165	0.690476	0.630434
R2L	0.000077	0.994413	0.98252
Normal	0.000495	0.999483	0.999854

Table XIII: Performance Analysis of Attack Categories in Cluster-5

Attacks	False Positive Rate	Precision	Recall
DoS	0.001049	0.989361	0.994652
Probe	0.00965	0.993665	0.992879
U2R	0.000478	0	0
R2L	0.002911	0.785714	0.6875
Normal	0.00606	0.985318	0.993421

Table XIV: Performance Analysis of Attack Categories with K-means Clustering

Attacks	False Positive Rate	Precision	Recall
DoS	0.000156	0.999959	0.999984
Probe	0.000028	0.99657	0.990504
U2R	0.00003	0.65909	0.557692
R2L	0.000028	0.987353	0.969831
Normal	0.000148	0.99928	0.999701

6. Conclusion

This paper presents a comparative analysis hybrid machine learning technique to detect Denial of Service (DoS) attacks, Probing (Probe) attacks, User-to-Root (U2R) attacks and Remote-to-Local (R2L) attacks. We can know the similar nature of attack group by using K-means algorithm. And then we use Random Forest algorithm to classify normal and attack connections. The experiments show that, KDDCup 99 dataset can be applied as an effective benchmark dataset to help researchers compare different intrusion detection models. Future work includes analyzing with other data mining algorithms to classify attack categories and how it can detect on other real time environment dataset.

References

[1] Aung Yi Yi and Myat Myat Min, "An Analysis of Random Forest Algorithm Based Network Intrusion Detection System", Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNDP), 2017 IEEE/ACIS 18 th International Conference, 2017.

[2] Yan. K.Q., S. C. Wang and C. W. Liu, "A Hybrid Intrusion Detection System of Cluster-based Wireless Sensor Networks", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, March 18-20, 2009, Hong Kong.

[3] S. Revathi and Dr. A. Malathi, "Detecting User-To-Root (U2R) Attacks Based on Various Machine Learning Techniques", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 4, ISSN (Online): 2278-1021, ISSN (Print): 2319-5940, April 2014.

[4] S. Revathi and Dr. A. Malathi, "Detecting Denial of Service Attack Using Principal Component Analysis with Random Forest Classifier", International Journal of Computer Science & Engineering Technology (IJCSET), Vol.5, No. 03, ISSN: 2229-3345, March 2014.

[5] Jayshri R. Patel, "Performance Evaluation of Decision Tree Classifiers for Ranked Features of Intrusion Detection", Journal of Information, Knowledge and Research in Information Technology, Vol. 02, Issue -02, ISSN: 0975-6698, Nov 12 to Oct 13.

[6] Heba F. Eid et al., "Principal Components Analysis and Support Vector Machine based Intrusion Detection System", 10 th International Conference on Intelligent Systems Design and Applications, (IEEE, 2010).

[7] Manzoor, Muhammad Asif, and Yasser Morgan, "Network Intrusion Detection System using Apache Storm" , Special Issue on Recent Advances in Engineering Systems, Advances in Science, Technology and Engineering System Journal (ASTES), Vol. 2, No. 3, 812-818 (2017).

[8] S. Choudhury and A. Bhowal, "Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection", International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015, pp 89-95.

[9] R. Venkatesan, R. Ganesan and A.A.L. Selvakumar, "A Comprehensive Study in Data Mining Frameworks for Intrusion Detection", International Journal of Advanced Computer Research, December-2012, Volume-2 Number-4 Issue-7, ISSN (print): 2249-7277 ISSN (online): 2277-7970.

[10] Somani Manish and Roshni Dubey, "Hybrid Intrusion Detection Model Based on Clustering and Association", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.3, Issue 3, ISSN (Print): 2320-3765, ISSN(Online):2278-8875 March 2014.

[11] M. Dhakar and A. Tiwari, "A Novel Data Mining based Hybrid Intrusion Detection Framework", Journal of Information and Computing Science, 2014, Vol-9 No-1 pp. 037-048, ISSN 1746-7659, England, UK.

[12] TR. Patel, A. Thakkar and A. Ganatra, "A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems", International Journal of Soft Computing and Engineering (IUSCE), March-2012, Vol-2, Issue-1, ISSN: 2231-2307.

[13] Youssef Ahmed and Ahmed Emam, "Network Intrusion Detection Using Data Mining and Network Behavior Analysis", International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011.

[14] X. Wu, V.Kumar, Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg, "Top 10 algorithms in data mining", Survey Paper(2008).

[15] S. Devaraju and S. Ramakrishnam, "Performance Comparison for Intrusion Detection System using Neural Network with KDD Dataset", ICTACT Journal on soft Computing , Vol:04, Issue:03, ISSN: 2229-6956, April 2014.

[16] P. S. Rath, M. Hohanty, S. Acharya and M. Aich, "Optimization of IDS Algorithms Using Data Mining Technique", Proceeding of 53rd IRF International Conference, Pune, India,ISBN 978-93-86083-01-2, 2016.

[17] L.S. Parihar and A. Tiwari, "Survey on Intrusion Detection Usingn Data Mining Methods",IJSART, January-2016, Volume-2 Issue-1 ISSN (online): 2395-1052.

[18] Md.E. Haque and T.M. Alkharobi, "Adaptive Hybrid Model for Network Intrusion Detection and Comparison among Machine Learning Algorithms", International Journal of Machine Learning and Computing, February 2015, Vol-5, No-1.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 502-510 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Computation of Viability Kernels on Grid Computers for Aircraft Control in Windshear

Nikolai Botkin[*,1], Varvara Turova[1], Johannes Diepolder[2], Florian Holzapfel[2]

[1]*Zentrum Mathematik, Modeling M6, Technische Universität München, 85748, Germany*

[2]*Institute of Flight System Dynamics, Technische Universität München, 85748, Germany*

ARTICLEINFO

Article history:
Received: 29 November, 2017
Accepted: 24 January, 2018
Online: 28 February, 2018

Keywords:
Aircraft Cruise Flight Phase
Differential Game
Viability Kernel
Grid Method
Supercomputer

ABSTRACT

This paper is devoted to the analysis of aircraft dynamics in the cruise flight phase under windshear conditions. The study is conducted with reference to a point-mass aircraft model restricted to move in a vertical plane. We formulate the problem as a differential game against the wind disturbances: The first player, autopilot, manages, via additional smoothing filters, the aircraft's angle of attack and power setting. The second player, wind, produces disturbances that are transferred, also via smoothing filters, into most dangerous wind gusts. The state variables of the game are subject to state constraints representing aircraft safety conditions related, for example, to the altitude, path inclination and velocity. Viability theory is used to compute the so-called viability kernels, the maximal subsets of state constraints where an appropriate feedback strategy of the first player can keep aircraft trajectories arbitrary long for all admissible disturbances generated by the second player. A grid method is utilized, and challenging computations in seven dimensions are conducted on a supercomputer system.

1 Introduction

Atmospheric conditions such as windshear continue to be considered as a source of potentially severe consequences. They are dangerous for aircraft during landing or take-off, because the wind gusts can occur at relatively low altitudes. Nevertheless, windshear is also dangerous during the cruise flight phase because it can lead to violation of the prescribed flight level.

In view of threats related to wind disturbances, there is permanent interest in designing robust aircraft guidance and control schemes (possibly for use with autopilots). The related question consists in finding safety domains, i.e. sets of initial states from which the control problem can be solved in the case of worst wind disturbance whose components lie in a known range.

There exist a large number of works devoted to the problem of aircraft control in the presence of severe windshears. In particular, papers [1–6] address the problem of aircraft control during take-off in windshear conditions. In works [1] and [2], the wind velocity field is assumed to be known. It is shown that open loop controls obtained as solutions of appropriate optimization problems provide satisfactory results for rather severe wind disturbances. Nevertheless, it is clear that the spatial distribution of wind velocity cannot be measured with appropriate accuracy, and therefore feedback principles of control design are more realistic. Different types of feedback controls are proposed in papers [3–6]. In [3], the design of a feedback robust control is based on the construction of an appropriate Lyapunov function. Robust control theory is used in [4] to develop feedback controls stabilizing the relative path inclination and (in [5] and [6]) for the design of feedback controls stabilizing the climb rate.

An approach based on differential game theory (see e.g. [7]) is presented in paper [8] in connection with the problem of landing. A high dimensional nonlinear system of dynamic equations is linearized and reduced to a two-dimensional differential game using a transformation of variables. The resulting differential game is numerically solved, and optimal feedback controls are constructed and tested in the nonlinear model against a downburst.

[*]Corresponding Author: Nikolai Botkin, Boltzmannstr. 3, 85748 Garching bei München, Germany, Email: botkin@ma.tum.de

https://dx.doi.org/10.25046/aj030161

Another method, also based on differential game theory, is used in papers [9] and [10] to find feedback controls that are effective against downbursts. This approach assumes the computation of the value function, which is a viscosity solution (see. e.g. [11] and [12]) of an appropriate Hamilton-Jacobi equation. The numerical implementation utilizes dynamic programming techniques described in [13] and [14]. The case of known wind velocity field as well as the case of unknown wind disturbances are considered.

Recent investigations [15] and [16] utilize a method (see [17]) for the fast computation of rough approximations of solvability sets in linear conflict control problems. Using techniques of sequential linearization, this method is applied to nonlinear aircraft dynamics to design an appropriate control for take-off in the presence of downbursts.

The current paper addresses the problem of retaining trajectories in an appropriate flight domain (AFD) corresponding to the cruise flight phase (cf. [18] and [19]). Viability theory (see e.g. [20]) provides numerical methods (see e.g. [21], [22], and the Appendix) for finding the viability kernel, i.e. the largest set of initial states lying in the AFD from which viable trajectories emanate. More precisely, it includes all initial states for which there exists a feedback control that generates trajectories remaining in the viability kernel for all possible admissible wind gusts. In the case where the initial state does not belong to the viability kernel, there exists a method of designing a wind disturbance such that all trajectories violate the AFD for all possible controls.

As for wind conditions, it is assumed that only bounds on the wind velocity components are imposed. The dynamics of the aircraft will be considered as a differential game (cf. [9] and [10]) where the first player chooses control inputs, whereas the second player forms the worst wind disturbance. It is assumed that the first player is able to measure the current state vector, whereas the second player can measure both the current state vector and the current control ("future" values are not available) of the first player. Therefore, the second player may use the so-called feedback counter strategies (see [7]).

The current paper has common features with the open-access publication [23] concerning the model description and solution method. It should be noted that the publication [23] is mainly focused on theoretical fundamentals of the differential game approach. Regarding computational results, paper [23] formulates a problem of constructing the viability kernel in seven dimensions and performs several steps of the algorithm to show its feasibility. In contrast, the current paper addresses aspects of implementation on large scale grid computers and completely solves the above mentioned seven-dimensional problem including simulation of optimal trajectories.

The paper is structured as follows:

Section 2 outlines a point-mass aircraft model describing the vertical motion of a generic modern regional jet transport aircraft. The model is closely related to the one described in paper [23]. The difference consists in a more clear method of deriving the dynamics equations.

In Section 3, state constraints related to the cruise flight phase are formulated, and the corresponding computed viability kernels are demonstrated through their three-dimensional sections. Additionally, trajectories yielded by an optimal feedback strategy, working against an optimal control of the disturbance, are shown. It should be noted that an optimal control of the disturbance can be constructed either as counter or pure feedback strategy because the so called saddle point condition holds for the differential game under consideration.

Section 4 outlines some aspects of parallel implementation of the computational method on a supercomputer system. The parallelization principles and data flow inside and between compute nodes are sketched. The novelty of our approach and comparison with existing software tools are discussed.

Section 6 (Appendix) briefly outlines the concept of differential games and viability kernels. Grid schemes for the computation of them are sketched. The details can be found in [23].

2 Model equations

In this section, a point-mass model representing the vertical motion of a generic modern regional jet transport aircraft is considered. Table 1 introduces euclidean coordinate systems (COS) that are necessary to compute the forces exerted on the aircraft. The origins of COSs are located either at the aircraft gravity center (CG) or at a fixed reference point (O) on the earth surface.

Table 1: Aircraft coordinate systems

COS	**Index**	***x*-axis**	**Origin**
Local	N	Parallel to the Earth's surface (x_N) and upwards (z_N)	O
Kinematic	K	In direction of $\overrightarrow{V_K}$	CG
Aerodynam.	A	In direction of $\overrightarrow{V_A}$	CG
Thrust	P	In positive direction of the symmetry axis of the turbine (aft looking forward)	CG
Body Fixed	B	In direction of the nose and in the symmetry plane of the aircraft	CG

Here, $\vec{V}_K$ and $\vec{V}_A$ are kinematic and aerodynamic aircraft velocities, respectively. The angles defining the relationship between the coordinate systems are the following (see also Figure 1):

α_K the kinematic angle of attack,
α_A the aerodynamic angle of attack,
γ_K the kinematic path inclination angle,
σ the thrust inclination angle.

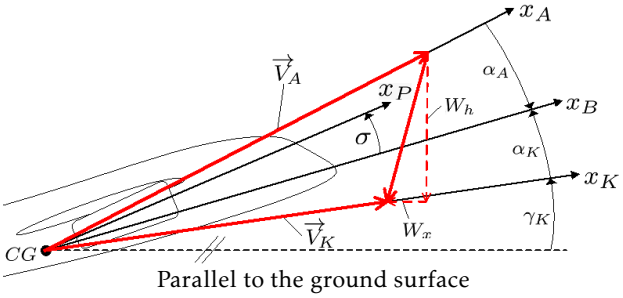


Figure 1: Aircraft coordinate systems and angles

Matrices for the transformations $K \to N$ (Kinematic to Local), $A \to B$ (Aerodynamic to Body), $B \to K$ (Body to Kinematic), and $P \to B$ (Thrust to Body) are defined as follows:

$$M_{NK} = \begin{bmatrix} \cos(\gamma_K) & -\sin(\gamma_K) \\ \sin(\gamma_K) & \cos(\gamma_K) \end{bmatrix},$$

$$M_{BA} = \begin{bmatrix} \cos(\alpha_A) & -\sin(\alpha_A) \\ \sin(\alpha_A) & \cos(\alpha_A) \end{bmatrix},$$

$$M_{KB} = \begin{bmatrix} \cos(\alpha_K) & -\sin(\alpha_K) \\ \sin(\alpha_K) & \cos(\alpha_K) \end{bmatrix},$$

$$M_{BP} = \begin{bmatrix} \cos(\sigma) & -\sin(\sigma) \\ \sin(\sigma) & \cos(\sigma) \end{bmatrix}.$$

The position propagation is given in the local coordinate system (N), whereas the translation dynamics are derived in the kinematic coordinate system (K). The model equations read as follows:

$$\dot{x}_N = V_K \cos(\gamma_K), \quad (1)$$
$$\dot{z}_N = V_K \sin(\gamma_K), \quad (2)$$
$$\dot{V}_K = \frac{X_T}{m}, \quad (3)$$
$$\dot{\gamma}_K = \frac{Z_T}{m V_K}. \quad (4)$$

Here, X_T and Z_T denote the components of the total force $\vec{F}_T$ represented in the kinematic coordinate system (K), and the notation m stands for the aircraft mass. As usually, $\vec{F}_T$ comprises aerodynamic, propulsion, and gravitation forces:

$$\vec{F}_T = \vec{F}_A + \vec{F}_P + \vec{F}_G.$$

Aerodynamic forces. They are defined as follows:

$$\vec{F}_A = M_{KB} M_{BA} \begin{bmatrix} C_D \\ C_L \end{bmatrix} \frac{1}{2} \rho V_A^2 S,$$

where C_D and C_L are the drag and lift coefficients, respectively, $\rho = \rho(h)$ is the air density (depends on the altitude), V_A the aerodynamic velocity, and S the wing reference area.

The lift and drag coefficients $C_D(\alpha_A, M)$ and $C_L(\alpha_A, M)$ are taken in the form:

$$C_D(\alpha_A, M) = c_1^D + c_2^D \alpha_A + c_3^D M + c_4^D \alpha_A^2 + c_5^D \alpha_A M + c_6^D M^2 + c_7^D \alpha_A^3 + c_8^D \alpha_A^2 M + c_9^D \alpha_A M^2, \quad (5)$$

$$C_L(\alpha_A, M) = c_1^L + c_2^L \alpha_A + c_3^L M + c_4^L \alpha_A^2 + c_5^L \alpha_A M + c_6^L M^2 + c_7^L \alpha_A^3 + c_8^L \alpha_A^2 M + c_9^L \alpha_A M^2, \quad (6)$$

where the coefficients c_i^D and c_i^L, $i = 1,...9$ are determined from least square fitting to experimental data.

The absolute value, V_A, of the aerodynamic velocity can be derived using its relation to the kinematic velocity $\vec{V}_K$ in the Local frame (N) and the wind velocities W_x and W_z in the x_N and z_N directions, respectively. Therefore,

$$V_A = \left\| (\vec{V}_K)^N - \begin{bmatrix} W_x \\ W_z \end{bmatrix} \right\|, \quad (7)$$

and finally, using the matrix M_{NK}, this implies the formula

$$V_A^2 = (V_K \cos\gamma_K - W_x)^2 + (V_K \sin\gamma_K - W_z)^2.$$

The aerodynamic angle of attack α_A is computed as follows:

$$\alpha_A = \arctan\left(\frac{w_A}{u_A}\right),$$

where u_A and w_A are x_B and z_B-components of the aerodynamic velocity, respectively.

The Mach number M is defined as follows:

$$M = \frac{V_A}{c}, \quad c = \sqrt{\kappa R T(h)},$$

where c is the speed of sound, κ the adiabatic index for air, R the gas constant for ideal gases, and $T(h)$ the temperature of air at the altitude h. See [23] for more details.

Propulsion Forces. Thrust forces are modeled considering a two-engine setup. Thus,

$$\vec{F}_P = 2 M_{KB} M_{BP} \vec{F}_{P\text{net}}, \quad \vec{F}_{P\text{net}} = \begin{bmatrix} f_V(\delta_T, M) \\ f_\gamma(\delta_T, M) \end{bmatrix},$$

where $\delta_T \in [0,1]$ is the thrust setting, and the functions f_V and f_γ are approximated similar to formulas (5) and (6), with δ_T instead of α_A. See [23] for more details.

Gravitation Force. For a simple gravitational model with constant acceleration g, the corresponding force is computed as:

$$F_G = M_{KN} \begin{bmatrix} 0 \\ -g \end{bmatrix},$$

where M_{KN} is the inverse (transpose) of M_{NK}.

The following two equations are added to the dynamics (1)-(4) to exclude jumps in the controls:

$$\dot{\alpha}_K = \widetilde{\alpha}_K, \quad \dot{\delta}_T = \widetilde{\delta}_T \tag{8}$$

Furthermore, the following equations produce smoothing of wind disturbances:

$$\dot{W}_x = -k_w(W_x - \widetilde{W}_x), \quad \dot{W}_z = -k_w(W_z - \widetilde{W}_z), \tag{9}$$

where the time constant, k_w, is chosen as $k_w = 1\,s^{-1}$.

3 Problem setting and simulation results

Problem. The model consists of equations (2), (3), (4), (8), and (9). Thus, the state vector has seven variables: z_N, V_K, γ_K, α_K, δ_T, W_x, and W_z. The controls are associated with the rate of the angle of attack, $\widetilde{\alpha}_K$, and the rate of the thrust setting, $\widetilde{\delta}_T$. Their instantaneous changes are permitted. The disturbances are now associated with the artificial variables $\widetilde{W}_x$ and $\widetilde{W}_z$ that are inputs of the filters (9). Thus, the physical wind components W_x and W_z do not exhibit instantaneous jumps.

The following constraints on the controls and disturbances are prescribed:

$$\begin{gathered}\widetilde{\alpha}_K \in [-5,5]\,\text{deg/s}, \;\; \widetilde{\delta}_T \in [-0.3,0.3]\,1/\text{s}, \\ |\widetilde{W}_x| \le 5\,\text{m/s}, \;\; |\widetilde{W}_z| \le 5\,\text{m/s},\end{gathered} \tag{10}$$

and the following state constraints are imposed:

$$\begin{aligned}&h_N := z_N - h_0 \in [-90,90]m, \\ &V_K \in [100,200]\,\text{m/s}, \;\; |\gamma_K| \le 10\,\text{deg}, \\ &\alpha_K \in [0,16]\,\text{deg}, \;\; \delta_T \in [0.3,1],\end{aligned} \tag{11}$$

where $h_0 = 10000\,\text{m}$ being the cruise flight altitude.

Additionally, the state constraints $|W_x| \le 5\,\text{m/s}$ and $|W_z| \le 5\,\text{m/s}$ hold automatically because of equations (9) and the constraints (10).

In the numerical construction, the box $[-100,100]\times[90,210]\times[-15,15]\times[-4,20]\times[0.2,1.2]\times[-6,6]\times[-6,6]$ of the space $(h_N, V_K, \gamma_K, \alpha_K, \delta_T, W_x, W_z)$ was divided in $200\times120\times30\times24\times14\times12\times12$ grid cells. The sequence of time steps, $\{\delta_\ell\}$, was chosen as $\delta_\ell \equiv 0.01$, and the computations were performed until $\left|\mathcal{V}^h_{\ell+1} - \mathcal{V}^h_\ell\right| \le 10^{-5}$ for all grid nodes. Totally, 4471 steps of the algorithm (19) were done. The computation has been done on the SuperMUC system at the Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities. The computation was distributed over 100 compute nodes with 16 cores per node, which is regarded as "middle task" on the SuperMUC system. The runtime was about 40 h.

Figures 2-4 show different three-dimensional sections of the seven-dimensional viability kernel. Figures 5-7 respectively demonstrate the same three-dimensional sections and the corresponding projections of two optimal trajectories emanating from points lying in the viability kernel. The start point of trajectory 0 lies near to the boundary of the viability kernel, whereas trajectory 1 starts from a point lying deep inside of the viability kernel. The trajectories are computed when the control uses its optimal feedback strategy, and the disturbance utilizes its optimal counter feedback strategy (see the Appendix). The time step size used in the simulation of the trajectories was equal to 0.01. It is seen that the trajectories go to their attraction cycles and remain there. The simulation time interval is $[0, 15\,\text{min}]$. Figure 8 shows a three-dimensional section of a smaller viability kernel corresponding to the shrinkage of the state constraints (11) by the factor 0.7. The corresponding three-dimensional projection of two optimal trajectories are shown. It is seen that the disturbance can keep trajectory 0 outside the reduced viability kernel because the start point lies outside of it.

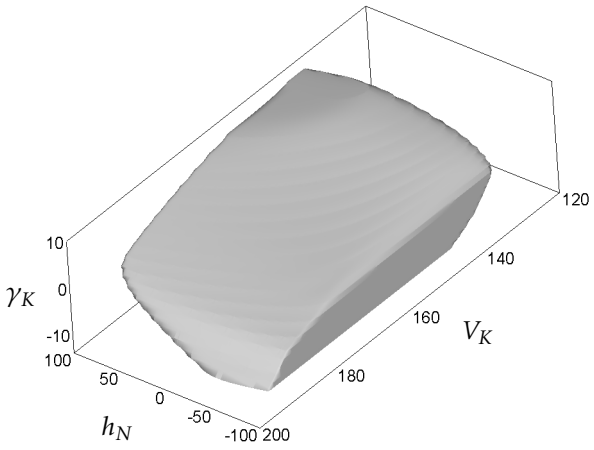


Figure 2: A three-dimensional section of the viability kernel: $\alpha_K = 8$, $\delta_T = 0.65$, $W_x = 0$, and $W_z = 0$.

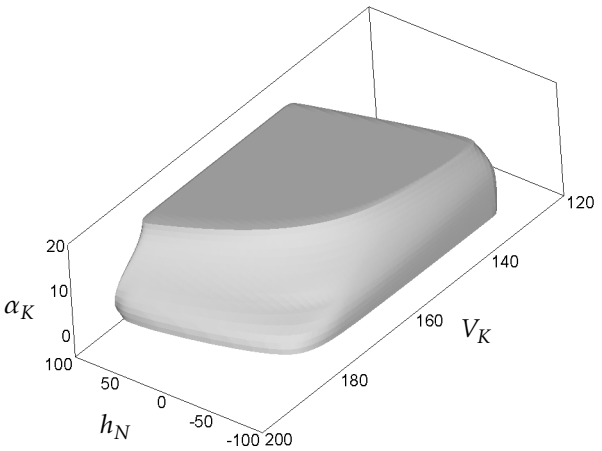


Figure 3: Another three-dimensional section of the viability kernel: $\gamma_K = 0$, $\delta_T = 0.65$, $W_x = 0$, and $W_z = 0$.

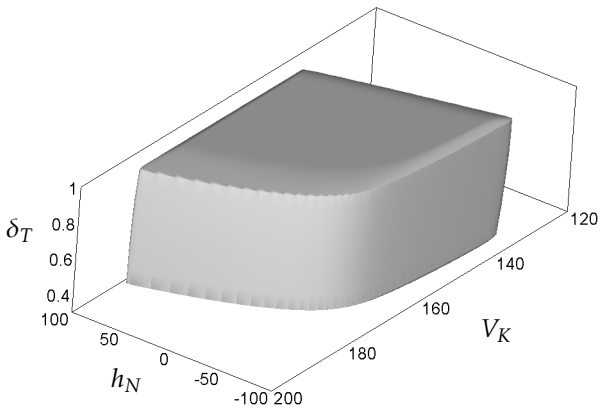


Figure 4: One more three-dimensional section of the viability kernel: $\gamma_K = 0$, $\alpha_K = 8$, $W_x = 0$, and $W_z = 0$.

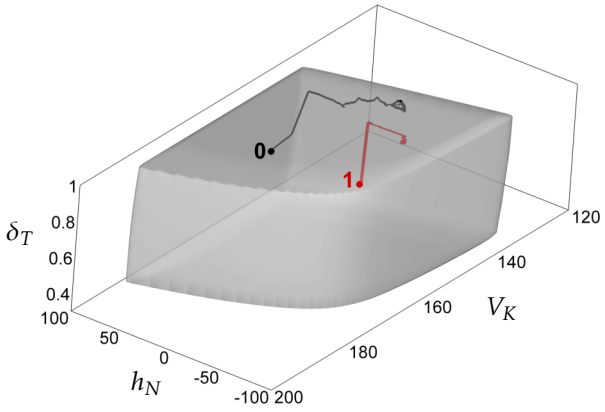


Figure 7: The same section as in Figure 4 and projections of the same two trajectories as in Figure 5.

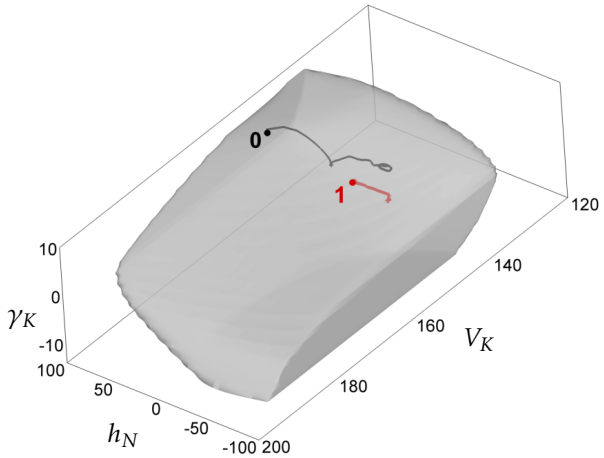


Figure 5: The section from Figure 2 and projections of two trajectories generated by an optimal feedback strategy of the control and an optimal counter feedback strategy of the disturbance. The start points are marked with bullets.

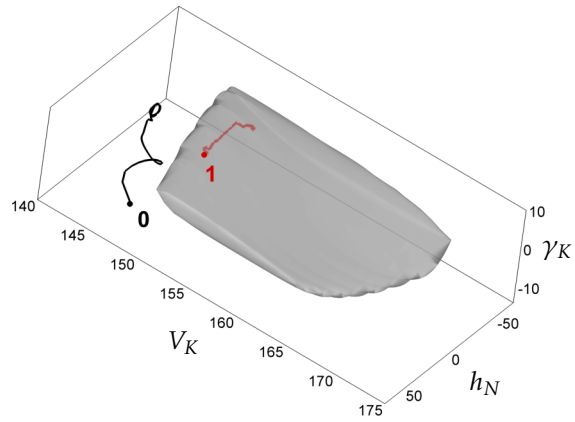


Figure 8: The section ($\alpha_K = 8$, $\delta_T = 0.65$, $W_x = 0$, and $W_z = 0$) of a smaller viability set corresponding to the shrinkage of the state constraints by the factor of 0.7. Projections of the same trajectories as in in Figure 5 are shown. Since the start point 0 does not belong to this viability set, the disturbance can keep the trajectory outside of it.

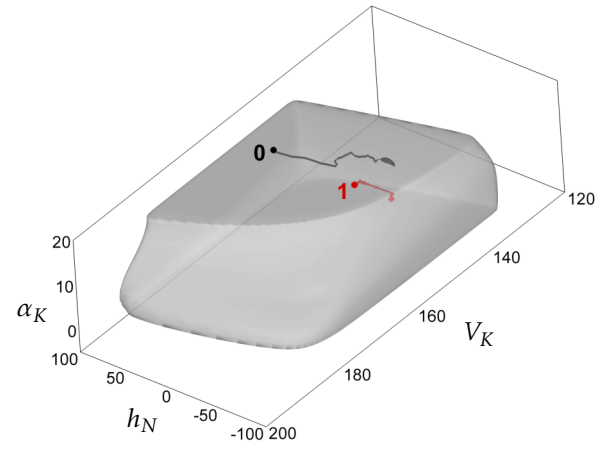


Figure 6: The same section as in Figure 3 and projections of the same two trajectories as in Figure 5.

4 Implementation Aspects

4.1 Grid Computing, Parallelization and Scaling

We use a self developed software code to implement the algorithms of the grid scheme outlined in Appendix 6.3. The code is parallelized using a mixed MPI/OpenMP technique. The first two dimensions of the grid are decomposed, whereas the other dimensions remain unmodified. A compute node cartesian topology (see Figure 9) is then created such that each compute node corresponds to a grid cylinder. Additionally, each grid cylinder is supplied with ghost nodes that allow us to compute divided differences (18) used e.g. in the algorithm (19). In Figure 9, the

lines connecting the compute nodes show the data flow supported by MPI. The parallelization inside of each compute node (i.e. inside a grid cylinder) is supported by OpenMP.

The problem considered in this paper deals with a seven-dimensional grid of 200×120×30×24×14×12×12 cells. Each of the first two dimensions was divided into 10 parts. Thus, there were 100 grid cylinders, each of size $20 \times 12 \times 30 \times 24 \times 14 \times 12 \times 12$, plus the necessary ghost grid nodes. Therefore, the required memory per compute node was equal to about 8 GB.

Our observations show a good scaling behavior (see Table 2). The results were obtained for the problem of computing the viability kernel in five dimensions on a $200 \times 120 \times 30 \times 24 \times 14$ grid. The relative speedup normalized to 32 cores and absolute timing of 30000 steps are shown.

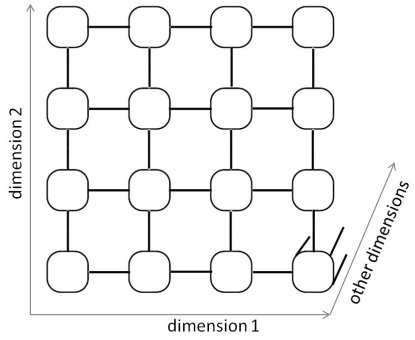


Figure 9: Compute node cartesian topology used in our applications.

Table 2: Scaling behavior of the code

# of cores	Linear prediction of scaling	Observed scaling	Wall time [h]	Performance/core [GFlop/s]
32	1	1	9.7	20
128	4	3.5	2.8	17.5
1600	50	38	0.25	15.2

4.2 Novelty of the method used

This paper utilizes a new method for computing viability kernels, which is based on the results of paper [21]. It is proven in [21] that the viability kernel is the Hausdorff limit of the sets $\{x : V(x,t) \le 0\}$ as $t \to -\infty$, where $V(x,t)$ is the value function (see [7]) of a state constrained differential game .

The numerical implementation of this method requires a theoretical basis and a stable numerical procedure for the treatment of transient Hamilton-Jacobi equations related to differential games with state constraints. Such a theoretical basis is given in paper [13] where the conventional conditions for viscosity solutions are modified to account for state constraints. This theoretical background is numerically implemented in [13] and [14], which results in monotone stable grid scheme (17) and (19). This enables us to perform a large number of time steps, usually several thousands.

Another new feature is that the corresponding software code is deeply parallelized using hybrid MPI/OpenMP techniques and adapted to run on a supercomputer system. Moreover, diverse modified variants of the code, based on sparse representations of grid functions, are tested.

4.3 Comparison with existing software

The well known software for solving Hamilton-Jacobi equations is the Level Set Method Toolbox (LSMT) described in [24]. This tool is really appropriate for solving rather general problems. However, it is not indicated in the manual whether the LSMT can compute viability kernels for differential games with *state constraints*. Moreover, according to the manual, the LSMT is not parallelized, whereas our software runs on a multi-node system.

5 Conclusion

This investigation shows that methods of differential games theory and viability theory can by applied to nonlinear aircraft models to investigate potential control abilities in the presence of wind disturbances. The new feature of our approach is the consideration of viability kernels for differential games. Feedback strategies of the players can be found from limiting grid "value functions" defining viability kernels (cf. Appendix). It is important that the amount of stored data is relatively small, which permits to implement the computed feedback strategies on a flight simulator. Further research will be focused on the treatment of models with more state variables, which will allow us to consider more realistic problems. Moreover, accounting for additional sources of uncertainty such as sensor errors or modeling uncertainties is planned.

Acknowledgment This work was supported by the DFG grants TU427/2-1 and HO4190/8-1. Computer resources for this project have been provided by the Gauss Centre for Supercomputing/Leibniz Supercomputing Centre under grant: pr74lu.

6 Appendix

This section briefly summarizes a method presented in [23] for computing viability kernels. Such a summary should help to provide a self-contained presentation.

6.1 Differential game

Consider a conflict control system with the autonomous dynamics

$$\dot{x} = f(x,u,v),\ x \in R^n,\ u \in P \subset R^p,\ v \in Q \subset R^q. \quad (12)$$

Here; x stands for the state vector; u and v denote control inputs of the first and second players, respectively; and the compact sets P and Q describe constraints imposed on the control and disturbance variables, respectively. Further, it is assumed that all functions of x have global properties. For example, the right-hand side f is supposed to be bounded, continuous, and Lipschitzian in x on $R^n \times P \times Q$.

In the following, it is assumed that the Isaacs saddle point condition

$$\min_{u\in P}\max_{v\in Q}\langle s, f(x,u,v)\rangle = \max_{v\in Q}\min_{u\in P}\langle s, f(x,u,v)\rangle, \quad (13)$$

is true for all $s \in R^n$ and $x \in R^n$. Note that this condition holds for the problem under consideration because controls and disturbances are additively separated in the model equations.

6.2 Viability kernel

For any $v \in Q$, consider the differential inclusion

$$\dot{x} \in F_v(x) = \overline{\text{co}}\{f : f(x,u,v), u \in P\} \quad (14)$$

Let $G \subset R^n$ be a compact set such that $G = \overline{\text{int}\, G}$, this set will play the role of the state constraint. Let T be an arbitrary fixed time instant, and $N = (-\infty, T] \times G$. For any subset $W \subset N$ and any time instant $t \le T$, define the time section of W through the relation $W(t) := \{x \in R^n : (t,x) \in W\}$.

Definition 1 (u-stability property [7]) *A set $W \subset N$ is called u−stable on $(-\infty, T]$ if for any position $(t_*, x_*) \in W$, for any time instant $t^* \in [t_*, T]$, for any fixed $v \in Q$, there exists a solution $x(\cdot)$ of the differential inclusion (14) with the initial condition $x(t_*) = x_*$ such that $(t^*, x(t^*)) \in W$.*

The next proposition is taken as a basis of the definition of viability kernels.

Proposition 1 *[see [21] for the proof] Let W be a maximal u-stable subset of $N = (-\infty, T] \times G$. If $W(t) \neq \varnothing$ for any $t \le T$, then the set*

$$K = \bigcap\nolimits_{t\le T} W(t)$$

is nonempty, and $W(t) \to K$ in the Hausdorff metric as $t \to -\infty$. The set K is called the viability kernel of G and denoted by $Viab(G)$.

Proposition 2 *Let $\bar{t} > 0$ be an arbitrary time instant, and $x_* \in Viab(G)$. Then there exists a feedback strategy $\mathcal{A}(x)$ of the first player such that all trajectories yielded by $\mathcal{A}$ and started at $t = 0$ from x_* remain in the set $Viab(G)$ for all $t \in [0, \bar{t}]$ and any actions of the second player. If $x^* \notin Viab(G)$, then there exists a feedback strategy $\mathcal{B}(x)$ and a time instant t_f such that all trajectories yielded by $\mathcal{B}$ and started at $t = 0$ from x^* violate the state constraint G for $t > t_f$ and any actions of the first player.*

6.3 Grid method for computing viability kernels

For the implementation of numerical method, it is convenient to represent viability kernels as level sets of an appropriate function. Let G_λ be a family of state constraint sets defined by the relation

$$G_\lambda = \{x \in R^n, g(x) \le \lambda\}, \quad (15)$$

where a continuous function g is chosen in such a way that, e.g., G_0 being the desired state constraint. It is necessary to construct a function V representing the viability kernels as follows:

$$Viab(G_\lambda) = \{x \in R^n, V(x) \le \lambda\}. \quad (16)$$

Such a function can be computed as a limiting solution, as $t \to -\infty$, of an appropriate Hamilton-Jacobi equation arising from conflict control problems with state constraints (see [13]). A grid approximation of V is computed as described below (cf. [21], [22], and [23]).

Let $\delta > 0$ be a time step length, and the tuple $h := (h_1, ..., h_n)$ defines space step sizes. Set $|h| := \max\{h_1, ..., h_n\}$ and introduce the following upwind operator defined on grid functions related to the discretization h:

$$\Pi(\phi;\delta,h)(x) = \phi(x) + \delta \min_{u\in P}\max_{v\in Q}\sum_{i=1}^{n}(p_i^{\text{right}} f_i^{+} + p_i^{\text{left}} f_i^{-}), \quad (17)$$

with f_i being the components of $f(x,u,v)$, and

$$\begin{aligned} &a^+ = \max\{a, 0\}, \quad a^- = \min\{a, 0\}, \\ &p_i^{\text{right}} = [\phi(x_1, ..., x_i + h_i, ..., x_n) - \phi(x_1, ..., x_i, ..., x_n)]/h_i, \\ &p_i^{\text{left}} \ = [\phi(x_1, ..., x_i, ..., x_n) - \phi(x_1, ..., x_i - h_i, ..., x_n)]/h_i. \end{aligned} \quad (18)$$

Let $\{\delta_\ell\}$ be a sequence of positive reals such that $\delta_\ell \to 0$ and $\sum_{\ell=0}^{\infty}\delta_\ell = \infty$. Consider the following grid scheme:

$$\mathcal{V}^h_{\ell+1} = \max\left\{\Pi\left(\mathcal{V}^h_\ell; \delta_\ell, h\right), g^h\right\}, \ \mathcal{V}^h_0 = g^h, \quad \ell = 0, 1, \ldots, \quad (19)$$

where g^h is the restriction of g to the grid defined by h.

It can be proven that $\mathcal{V}^h_\ell$ monotonically converges point-wise to a grid function $\mathcal{V}^h$, and this function define approximations of the viability kernels according to formula (16), see [23] for more details.

Remark 1 *In (17), the operation $\min_{u\in P}\max_{v\in Q}$ can be changed for $\max_{v\in Q}\min_{u\in P}$ to obtain almost the same result. The difference tends to zero with $|h|$. The proof follows from the fact that the original operator (17) and the modified one satisfy the same consistency condition (see [13]) involving the following Hamiltonian H:*

$$H(x,p) := \max_{v\in Q}\min_{u\in P}\langle p, f(x,u,v)\rangle = \min_{u\in P}\max_{v\in Q}\langle p, f(x,u,v)\rangle.$$

Numerical computations confirm this observation.

6.4 Control design

This section outlines one of possible methods of control design. Consider the grid scheme (19) assuming that ℓ is large enough so that the required approximation is reached, i.e. $\left|\mathcal{V}^h_{\ell+1} - \mathcal{V}^h_{\ell}\right|_{L_\infty} \le \epsilon$.

The optimal control u and the worst disturbance $v(u)$ at the current state x of the game can be found as solutions of the following program:

$$u, v \to \min_{u\in P} \max_{v\in Q} \mathcal{L}^h\Big[\mathcal{V}^h_\ell\Big]\Big(x + \tau f(x,u,v)\Big). \qquad (20)$$

Here, $\mathcal{L}^h$ is an interpolation operator (see e.g. [14]) defined on the corresponding grid functions, and τ being a parameter which should be several times larger than the time step size of the simulation procedure to provide some stabilization. Note that the function $\mathcal{V}^h_\ell$ can be transferred to a sparse grid (see e.g. [25] and [26]), which may essentially reduce the storage space. The disadvantage of such a technique is a certain loss of accuracy and a slower performance.

Remark 2 *As it is described above, the second player uses the so-called feedback counter strategies, i.e. functions of x and u, where u being the current control action of the first player. Since the saddle point condition (13) holds, the theory of differential games says that the second player can achieve the same result using pure feedback strategies, i.e. functions of x. For example, a near optimal strategy of the second player can be obtained as a solution of the problem*

$$u, v \to \max_{v\in Q} \min_{u\in P} \mathcal{L}^h\Big[\mathcal{V}^h_\ell\Big]\Big(x + \delta f(x,u,v)\Big),$$

see also Remark 1. Thus, the both players achieve optimal results using pure feedback strategies.

Remark 3 *If u and v appear linearly in the right-hand side of (12), then* min *and* max *operations in (17) and (20) can be computed only over extreme points of the sets P and Q respectively. This can be proven using the same arguments as in Remark 1.*

Remark 4 *If x_0 being a start point of the game, then x_0 lies in the approximate viability kernel defined as:*

$$\{x \in R^n : \mathcal{V}^h_\ell(x) \le \mathcal{V}^h_\ell(x_0)\},$$

and all trajectories approximately remain in this set. However, if the second player works non-optimally for a while, then, most likely, $\mathcal{V}^h_\ell(x(\bar{t})) < \mathcal{V}^h_\ell(x_0)$ for some $\bar{t} > 0$, and therefore, the state vector $x(\bar{t})$ lies now in the smaller viability kernel

$$\{x \in R^n : \mathcal{V}^h_\ell(x) \le \mathcal{V}^h_\ell(x(\bar{t}))\}.$$

Thus, faults of the second player improve the result of the first one.

Conflict of Interest The authors declare no conflict of interest.

References

[1] A. Miele, T. Wang, W. W. Melvin, "Optimal take-off trajectories in the presence of windshear" J. Optimiz. Theory App., **49**(1), 1–45, 1986. https://doi.org/10.1007/BF00939246

[2] A. Miele, T. Wang, W. W. Melvin, "Guidance strategies for near-optimum take-off performance in windshear" J. Optimiz. Theory App., **50**(1), 1–47, 1986. https://doi.org/10.1007/BF00938475

[3] Y. H. Chen, S. Pandey, "Robust control strategy for take-off performance in a windshear" Optim. Contr. Appl. Met., **10**(1), 65–79, 1989. doi:10.1002/oca.4660100106

[4] G. Leitmann, S. Pandey, "Aircraft Control under Conditions of Windshear" in C. T. Leondes (ed.) Control and Dynamic Systems, vol. 34, part 1, 1–79, 1990. Academic Press, New York.

[5] G. Leitmann, S. Pandey, "Aircraft control for flight in an uncertain environment: takeoff in windshear" J. Optimiz. Theory App., **70**(1), 25–55, 1991. https://doi.org/10.1007/BF00940503

[6] G. Leitmann, S. Pandey, E. Ryan, "Adaptive control of aircraft in windshear" Int. J. Robust Nonlin., **3**(2), 133–153, 1993. doi:10.1002/rnc.4590030206

[7] N. N. Krasovskii, A. I. Subbotin, Game-Theoretical Control Problems, Springer, 1988.

[8] V. S. Patsko, N. D. Botkin, V. M. Kein, V. L. Turova, M. A. Zarkh, "Control of an aircraft landing in windshear" J. Optimiz. Theory App., **83**(2), 237–267, 1994. https://doi.org/10.1007/BF02190056

[9] N. D. Botkin, V. L. Turova, "Application of dynamic programming approach to aircraft take-off in a windshear" AIP Conference Proceedings, **1479**, 1226–1229, 2012. https://doi.org/10.1063/1.4756373

[10] N. D. Botkin, V. L. Turova, "Dynamic programming approach to aircraft control in a windshear" in V. Křivan, G. Zaccour(eds.) Advances in Dynamic Games. Annals of the International Society of Dynamic Games, vol. 13, 53–69, 2013. Birkhäuser, Cham. https://doi.org/10.1007/978-3-319-02690-9_3

[11] M. G. Crandall, P. L. Lions, "Viscosity solutions of Hamilton-Jacobi equations" T. Am. Math. Soc., **277**(1), 1–47, 1983. https://doi.org/10.1090/S0002-9947-1983-0690039-8

[12] A. I. Subbotin, Generalized Solutions of First Order PDEs: The Dynamical Optimization Perspective, Birkhäuser, 1995.

[13] N. D. Botkin, K.-H. Hoffmann, N. Mayer, V. L. Turova, "Approximation schemes for solving disturbed control problems with non-terminal time and state constraints" Analysis, **31**(4), 355–379, 2011. https://doi.org/10.1524/anly.2011.1122

[14] N. D. Botkin, K.-H. Hoffmann, V. L. Turova, "Stable numerical schemes for solving Hamilton–Jacobi–Bellman–Isaacs equations" SIAM J. Sci. Comput., **33**(2), 992–1007, 2011. https://doi.org/10.1137/100801068

[15] K. Martynov, N. Botkin, V. Turova, J. Diepolder, "Real-Time Control of Aircraft Take-Off in Windshear. Part I: Aircraft Model and Control Schemes" in 25th Mediterranean Conference on Control and Automation (MED), Valletta Malta, 277–284, 2017. IEEE. doi:10.1109/MED.2017.7984131

[16] K. Martynov, N. Botkin, V. Turova, J. Diepolder, "Real-Time Control of Aircraft Take-Off in Windshear. Part II: Simulations and Model Enhancement" in 25th Mediterranean Conference on Control and Automation (MED), Valletta Malta, 285–290, 2017. IEEE. doi:10.1109/MED.2017.7984132

[17] E. K. Kostousova, "On target control synthesis under set-membership uncertainties using polyhedral techniques" in C. Pötzsche, C. Heuberger, B. Kaltenbacher, F. Rendl (eds.) System Modeling and Optimization, vol. 443, 170–180, 2014. Springer-Verlag, Berlin-Heidelberg. https://doi.org/10.1007/978-3-662-45504-3_16

[18] N. Seube, R. Moitie, G. Leitmann, "Aircraft take-off in windshear: a viability approach". Set-Valued Anal., **8**(1-2), 163–180, 2000. https://doi.org/10.1023/A:1008786811464

[19] A. M. Bayen, I. M. Mitchell, M. K. Osihi, C. J. Tomlin, "Aircraft autolander safety analysis through optimal control-based reach set computation" J. Guid. Control Dynam., **30**(1), 68–77, 2007. https://doi.org/10.2514/1.21562

[20] J. P. Aubin, "A survey of viability theory" SIAM J. Control Optim., **28**(4), 749–788, 1990. https://doi.org/10.1137/0328044

[21] N. D. Botkin, V. L. Turova, "Numerical construction of viable sets for autonomous conflict control systems" Mathematics, **2**(2), 68–82, 2014. doi:10.3390/math2020068

[22] N. D. Botkin, V. L. Turova, "Examples of computed viability kernels" Trudy Inst. Mat. i Mekh. UrO RAN, **21**(1), 306–319, 2015. http://mi.mathnet.ru/eng/timm1190

[23] N. Botkin, V. Turova, J. Diepolder, M. Bittner, F. Holzapfel, "Aircraft control during cruise flight in windshear conditions: viability approach" Dyn. Games Appl., **7**(4), 1–15, 2017. https://doi.org/10.1007/s13235-017-0215-9

[24] I. Mitchell, A Toolbox of Level Set Methods, UBC Department of Computer Science Technical Report TR-2007-11, 2007. https://www.cs.ubc.ca/~mitchell/ToolboxLS/toolboxLS-1.1.pdf

[25] C. Zenger, "Sparse grids" in Hackbusch, W. (ed.) Parallel Algorithms for Partial Differential Equations. Notes on Numerical Fluid Mechanics, vol. 31, 241–251, 1991. Vieweg, Braunschweig/Wiesbaden. doi: 10.1002/zamm.19920721115

[26] D. Pflüger, "Spatially Adaptive Sparse Grids for High-Dimensional Problems", Dissertation, Verlag Dr. Hut, München, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 511-520 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Adaptive and Non Adaptive LTE Fractional Frequency Reuse Mechanisms Mobility Performance

Uttara Sawant*, Robert Akl

Department of Computer Science and Engineering, University of North Texas, Denton, Texas 76207, USA

ARTICLEINFO

Article history:
Received: 30 November 2017
Accepted: 5 February 2018
Online: 28 February, 2018

Keywords:
Wireless
Metrics
Optimization
Frequency reuse
Mobility
Femtocells

ABSTRACT

Mobile broadband has gained momentum with the growing demand of user data rates. Long Term Evolution (LTE) technology is the step in mobile communications evolution, developed to satisfy high data rate demand, and meet better spectral efficiency requirements. Effective radio resource management and inter cell interferences are the major challenges. Fractional Frequency Reuse (FFR) is one of the effective interference avoidance mechanisms applied to LTE networks to yield optimal throughput. In this extended paper, we propose a novel performance metric, weighted throughput on user satisfaction, and evaluate an existing adaptation process that dynamically adjusts to optimal network performance determined by FFR mechanism with mobile users. The performance of FFR mechanism with mobility model, adaptation process, and femtocell densification is evaluated and optimized for proposed metric and other metrics. Results optimized by proposed metric show comparatively higher average throughput and lower variance among user throughput.

1 Introduction

Orthogonal Frequency Multiple Access (OFDMA) offers great spectrum efficiency and flexible frequency allocation to users. However, in LTE OFDMA networks the system performance is severely hampered by the Inter-Cell Interference (ICI) due to the frequency reuse, where the cell edge users will experience high interferences from neighboring cells.

FFR is one of the interference management techniques which require minimal or no coordination among the adjacent cells. The basic mechanism of FFR is to partition the macrocell area into spatial regions and each sub-region is assigned different frequency sub-bands for users. In [1, 2], the user satisfaction metric introduced by the authors is evaluated for users' mobility and the performance is compared with other reuse schemes. The authors propose a dynamic mechanism that selects the optimal FFR scheme based on the user satisfaction metric. The proposed mechanism is evaluated through several simulation scenarios that incorporate users' mobility and its selected FFR scheme is compared with other frequency reuse schemes in order to highlight its performance. The research is further enhanced in [3] where cell edge reuse factor is set to 1.5 and results are generated to determine the optimal inner radius and frequency allocation. In [4, 5], a performance study is carried out, where FFR partitions each cell into two regions; inner region and outer region, and allocates different frequency band to each region based on reuse factor. In [6], a frequency reuse technique is proposed which aims at maximizing throughput via combinations of inner cell radius and frequency allocation to the macrocell. In this work, the authors study the interference mitigation techniques in femtocell/macrocell networks and proposed a FFR mechanism that leads to increased overall system performance. Work in [7] proposes a mechanism that selects the optimal FFR scheme based on the user throughput and user satisfaction. In [8, 9], authors present a FFR optimization scheme based on capacity density (bit/s/m^2), which show better performance compared to conventional Reuse-1 and Reuse-3 schemes. The authors formulate an optimization problem and solve it by simulation. Graph theory and similar optimization techniques are presented in [10], a graph-based framework for dynamic FFR in multi-cell OFDMA networks is proposed in this work. The proposed scheme enhances the conventional FFR by

*Corresponding Author Uttara Sawant, University of North Texas, Denton, Texas 76207, +1-940-595-1489 & uttarasawant@my.unt.edu

https://dx.doi.org/10.25046/aj030162

enabling adaptive spectral sharing per cell load conditions. Such adaptation has significant benefits in a practical environment, where traffic load in different cells may be asymmetric and time-varying. Work in [11, 12] provides analysis of the inter cell interference coordination problem in multi-cell OFDMA systems. In order to reduce the interference without losing much frequency resources in each cell, cell users are partitioned into two classes, interior and exterior users. Results determine the optimal configuration of the interior and exterior regions' dimensions as well as the optimal frequency reuse factor.

Authors in [13] use two-stage heuristic approach to find optimal frequency partitioning. The authors intend to propose the FFR scheme by introducing the concept of normalized Spectral Efficiency (nSE). With the optimal Frequency Reuse Factor (FRF), the FFR scheme can maximize the system SE with throughput improvement of cell-edge users. To solve the problem, they divide it into two sub-problems. However, this scheme is asymptotically optimal with the assumption of the uniform distributions of signal and interference in sectors. The simulation results demonstrate the gain of the system SE is about 3% by proposed FFR scheme. The goal in work in [14] is to improve the cell edge throughput as well as the average cell throughput, compared to a network with frequency reuse factor 1. The cell capacity under those reuse schemes is estimated and compared. To attain both high spectral efficiency and good coverage within sectors/beams, a scheme based on coordinated scheduling between sectors of the same site, and the employment of frequency reuse factor above 1 only in outer parts of the sector, is proposed and evaluated. The resulting sector throughput increases with the number of active users. When terminals have one antenna and channels are Rayleigh fading, it results in a sector payload capacity between 1.2 (one user) and 2.1 bits/s/Hz/sector (for 30 users) in an interference-limited environment [15]. In [16], the authors review some of the recent advances in ICI research and discuss the assumptions, advantages, and limitations of the proposed mechanisms. They propose a scheduling algorithm to schedule users based on channel quality and quality of service (QoS) metrics.

Authors in [17] investigate an OFDMA radio resource control (RRC) scheme, where RRC control is exercised at both RNC and base stations. In [18], the authors present an analytical solution to carry out performance study of various frequency reuse schemes in an OFDMA based cellular network. Results of performance study show that results obtained through analytical method are in conformity with those obtained through simulations. In [19], the authors analyze the FFR scheme and propose a dynamic FFR mechanism that selects the optimal frequency allocation based on the cell total throughput and user satisfaction.

Extensive work is done to optimize network performance of FFR mechanism [20]. FFR and exploitation of the channel state information at the transmitter (CSIT) are effective approaches to improving the spectrum efficiency of the outer coverage region. When channels vary within a physical transmission frame, the above improvement is substantially suppressed. To remedy this, work in [21] develops new FFR patterns for OFDMA systems with frequency or time division duplexing (FDD/TDD) in time-varying channels. Simulation results show significant performance gains of the proposed schemes over the existing ones. A semi-centralized joint cell muting and user scheduling scheme for interference coordination in a multi-cell network is proposed under two different temporal fairness criteria. The authors in [22] propose a general pattern set construction algorithm in this paper. Numerical results are provided to validate the effectiveness of the proposed scheme for both criteria. The impact of choice of the cell muting pattern set is also studied through numerical examples for various cellular topologies. Densely deployed cellular wireless networks, which employ small cell technology, are being widely implemented. Authors in [23, 24] aim to maximize the system's throughput through the employment of FFR schemes. The authors derive the optimal configuration of the FFR scheme and evaluate the systems performance behavior under absolute and proportional fairness requirements. Table 1 shows comparison between previous published work and paper contributions.

Table 1: Comparison

Parameter	Previous Work	Current Paper
Throughput variance	High	Low
Femtocells adaptation	Minimal	High

This paper evaluates the performance of the sectorized FFR mechanism with five metrics; four existing and one proposed, which determine the values for inner region radius and frequency allocation for optimal results. For a specific mobility model, an adaptation process is applied to the FFR mechanism optimized for each of the five metrics and its performance is evaluated. Results are also generated and evaluated for non-adaptation process for the same network. Results prove that the proposed metric shows lower variance in user throughput compared to the other metrics from previous published work. The optimized FFR mechanism shows reacting to the adaptation process with user mobility even with the addition of femtocells.

2 FFR Mechanism

The topology of Figure 1 consists of 16 cells with four non-overlapping resource sets. Each cell of the topology is divided into two regions; inner and outer region. The total available bandwidth of the system is split

into four uneven spectrum, denoted by A (blue), B (green), C (red) and D (yellow). Spectrum A, B, and C have equal bandwidth and are allocated in outer regions with Frequency Reuse 3. On the other hand, spectrum D is allocated in all inner regions with Frequency Reuse 1. The frequency resources in all inner regions are universally used, since the inner region users are less exposed to ICI.

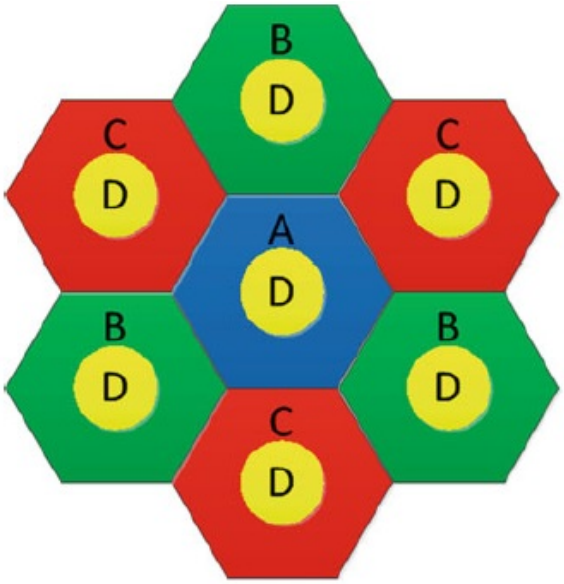


Figure 1: Strict FFR Deployment in LTE Macrocell Deployment [4]

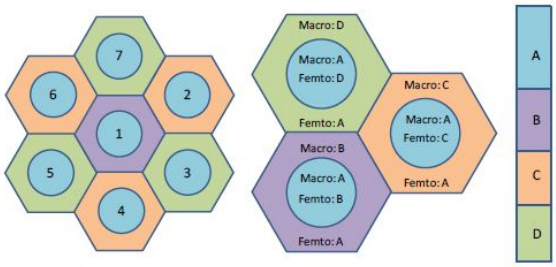


Figure 2: Strict FFR Deployment in LTE Macrocell-Femtocell Deployment [25]

From user's perspective, Integer Frequency Reuse (IFR) can be regarded as a special case of FFR. In IFR, all resource blocks (RBs) allocated to a cell can be used anywhere in the cell without any specification of user's location. For comparison, the FFR scheme that is selected by adaptive mechanism is compared with variations of IFR. The macrocell coverage area is partitioned into center-zone and edge-zone. The entire frequency band is divided into two parts, one part is solely assigned to the center-zone (e.g., sub-band A in Figure 1) and the other part is partitioned into three subbands (e.g., sub-bands B, C, and D) and assigned to the three edge-zones [26]. Figure 2 shows the Strict FFR network layout for integrated macrocell-femtocell deployment. Note that femtocell uses different frequency bands than the overlay macrocell to minimize or eliminate the inter-tier interference in the network.

3 System Model

A set of multicast users are uniformly distributed in the grid of 16 macrocells. In order to find the optimal inner region radius and frequency allocation in the deployment, the mechanism divides each cell into two regions and calculates the total throughput for the following 40 Frequency Allocations (FA), assuming Frequency Reuse 1 and x for inner and the outer regions respectively [26], where x is the frequency reuse factor of 3. Each FA corresponds to paired value of fraction of inner region resource blocks and inner region radius.

- FA1: All 25 resource blocks are allocated in inner region. No resource blocks are allocated to the outer region.
- FA2: 24 resource blocks are allocated in inner region. $1/x$ resource block allocated to the outer region.
  ⋮
- FA39: 1 resource block allocated in inner region. $24/x$ resource blocks allocated to the outer region.
- FA40: No resource blocks are allocated in inner region. $25/x$ resource blocks allocated to the outer region.

For each FA, the mechanism calculates the total throughput, user satisfaction, user fairness, and weighted throughput values based on new metrics. This procedure is repeated for successive inner cell radius (0 to R, where R is the cell radius). The mechanism selects the optimal FFR scheme that maximizes the cell total throughput. This procedure is repeated periodically in order to take into account users' mobility. Therefore, the per-user throughput, the cell total throughput, User Satisfaction (US), and other metrics are calculated in periodic time intervals (the exact time is beyond the scope of this manuscript) and at each time interval, the FFR scheme that maximizes the above parameters is selected. This periodic process is called adaptation [28]. The system model described above can be used for supported LTE bandwidths ranging from 1.4 MHz to 20 MHz.

The signal-to-interference-plus-noise (SINR) for downlink transmission to macro user x on a subcarrier n can be expressed as,

$$SINR_{x,n} = \frac{P_{M,n} G_{x,M,n}}{N_0 \Delta f + \sum_{M'} P_{M',n} G_{x,M',n}} \quad (1)$$

where, $P_{M,n}$ and $P_{M',n}$ is transmit power of serving macrocell M and neighboring macrocell M' on subcarrier n, respectively. $G_{i,M,n}$ is channel gain between macro user i and serving macrocell M on subcarrier n and $G_{i,M',n}$ corresponds to channel gain from neighboring macrocell M'. Finally, N_0 is white noise power spectral density and Δf is subcarrier spacing.

The channel gain G, given by the following equation is dominantly affected by path loss, which is assumed to be modeled based on urban path-loss PL is defined as,

$$G = 10^{-PL/10} \tag{2}$$

Additionally, for the throughput calculation, the capacity of a user i on a specific subcarrier n can be estimated via the $SINR$ from the following equation,

$$C_{i,n} = W \cdot \log_2(1 + \alpha \cdot SINR_{i,n}) \tag{3}$$

where, W denotes the available bandwidth for each subcarrier divided by the number of users that share the specific subcarrier and α is a constant for a target bit error rate (BER) defined by $\alpha = -1.5/ln(5 \cdot BER)$. Here, BER is set to 10^{-6}.

The expression of the total throughput of the serving macrocell M is given as follows,

$$T_i = \sum_n \beta_{i,n} C_{i,n} \tag{4}$$

where, $\beta_{i,n}$ represents the subcarrier assigned to user i. When $\beta_{i,n}$=1, the subcarrier is assigned to user i. Otherwise, $\beta_{i,n}$=0.

4 Mobility Model

This paper evaluates network performance in a scenario with mobile users in the highlighted cell of the topology presented in Figure 3. During the experiment that lasts for 217 seconds, 24 users of the examined cell are moving randomly inside the cell with speed 3 km/h, according to the Pedestrian A channel model [27]. It is assured that all of them remain into the cell's area, ensuring that their total number will remain constant. This corresponds to low mobility scenario with zero to minimal handover.

5 Performance Metrics

The paper evaluates the network performance for adaptive and non-adaptive FFR mechanisms with user mobility. For comparison, the adaptive FFR scheme that is selected by proposed mechanism is compared with three different cases. The first case, where the optimal inner radius and frequency allocation remain constant through the adaptation process, is referred to FFR non-adaptive. The second case, where the cell bandwidth equals the whole network bandwidth, is called IFR with frequency reuse 1 (IFR1). In the third case, the inner region radius is zero and each cell uses one third of the networks bandwidth. This case is called IFR with frequency reuse 3 (IFR3). The difference with the IFR1 case, lies in the fact that only co-channel base station are considered in interference calculation and as a consequence, the interference base station density is divided by 3.

5.1 Existing Metrics

This paper uses a metric US defined by authors in [28]. It is calculated as the sum of the users' throughput divided by the product of the maximum user's throughput and the number of users (X). US ranges between 0 and 1. When US approaches 1, all the users in the corresponding cell experience similar throughput. However, when US approaches 0, there are huge differences in throughput values across the users in the cell. This metric will be utilized in scenarios where fairness to the users is significant such as cell throughput and Jain fairness index.

$$US = \frac{\sum_{x=1}^{X} T_x}{max_user_throughput \cdot X} \tag{5}$$

With metric WT defined by the authors in [28], the aim is to not only generate low variance of the per-user throughput values but also obtain higher values of the cell total throughput.

$$WT = JI \cdot T \tag{6}$$

where T is the cell mean throughput.

To obtain a metric of fairness for performance evaluation, the Jain fairness metric introduced in [29] is used. Assuming the allocated throughput for user i is x_i, Jain fairness index for the cell is defined as,

$$JI = \frac{(\sum_{i=1}^{X} x_i)^2}{X \cdot \sum_{i=1}^{X} x_i^2} \tag{7}$$

This metric is interesting for the evaluation of the proposed method due to its properties. It is scale-independent, applicable for different number of users and it is bounded between [0, 1], where 0 means "total unfairness" and 1 means "total fairness" in terms of throughput division among the users.

5.2 Proposed Metric

This section introduces a new metric, weighted throughput based on user satisfaction WT_{US}, to add weights to the cell throughput corresponding to specific inner radius and inner bandwidth such that the resultant throughput is higher than the corresponding throughput optimized at user satisfaction alone and all the users in the corresponding cell experience similar throughput.

$$WT_{US} = US \cdot T \tag{8}$$

From (6), (8), it is clear that the metrics are cell-based.

6 Performance Evaluation

6.1 Mathematical Model

The product model is applied as mathematical model in the new metric definition and development. Previous metric, weighted throughput, adds the benefits

of the individual component metrics, throughput and fairness index. Similarly, the proposed metric defined by the product model of throughput and user satisfaction, is expected to perform better than the individual component metrics.

6.2 Simulation Parameters

Table 2 are parameters set for simulation.

Table 2: Simulation Parameters

Parameter	Value
System Bandwidth	5 MHz
Subcarriers	300
Subcarrier Bandwidth	180 KHz
Cell Radius	250 m
Inter eNodeB distance	500 m
Noise Power Spectral Density	-174 dBm/Hz
Subcarrier spacing	15 KHz
Channel Model	Typical Urban
Carrier Frequency	2000 MHz
Number of macrocells	16
Macrocell Transmit Power	46 dBm
Path Loss	Cost 231 Hata Model
Users' speed	3 km/h PedA

A sample uniform deployment considered in simulation is shown in Figure 3. Macrocells are located at cell centers. Active users are randomly distributed and are shown with white dots.

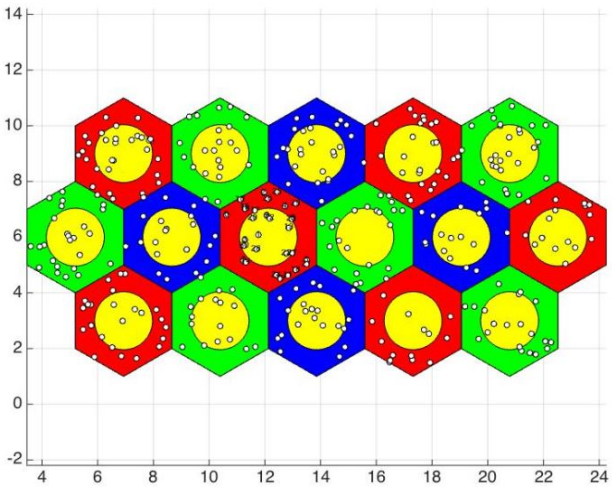


Figure 3: Strict FFR Uniform Deployment for Simulation

6.3 Simulation Guidelines

Note that the simulation is run assuming a stable downlink data traffic, since the simulation framework is analyzing data for downlink traffic. A network snapshot where user association to macrocell remains stable for the duration of the simulation run is considered as the users are associated to the base stations with maximum SINR. A frequent handover will occur in high mobility environment and that is not included in the scope of this paper. The simulation software for our research is based on [30].

6.4 Performance Analysis

6.4.1 Comparison of New Metric to Other Metrics

The new metric is introduced due to the following reasons, better performance in terms of average user throughput and effective user fairness in terms of variance in user throughput. Figure 4 proves the user throughput optimized with new metric shows better performance and low variance.

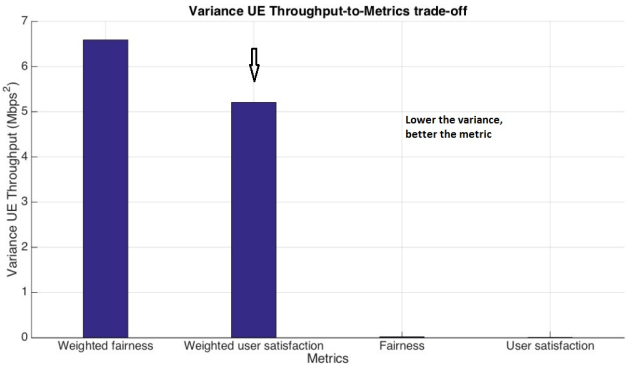


Figure 4: Comparison of Variance in User Throughput

6.4.2 Static Users - Metric Performance

Simulation is carried out by applying the metrics on users with static positions and performance is evaluated with respect to inner region subcarriers and inner region radius. Figures 5 - 8 show the metric performance with respect to inner region subcarriers. Figures 9 - 12 show the metric performance with respect to inner region radius.

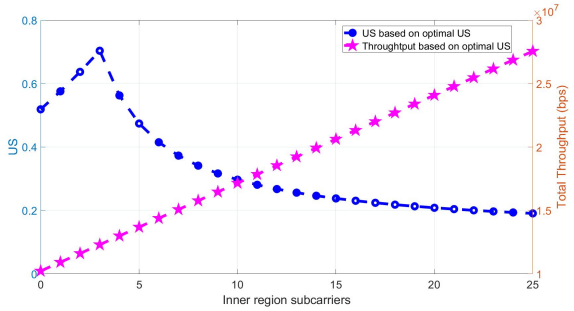


Figure 5: User Satisfaction Metric Performance for Static Users to Inner Region Subcarriers

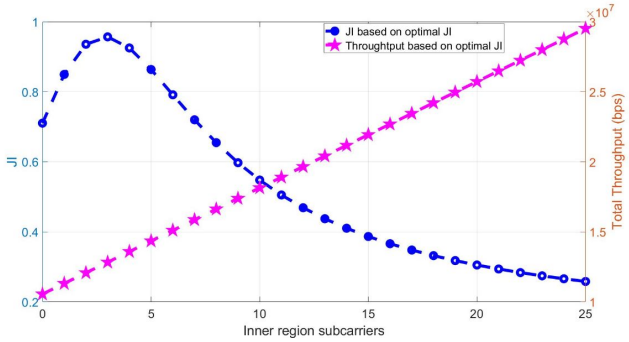


Figure 6: Fairness Index Metric Performance for Static Users to Inner Region Subcarriers

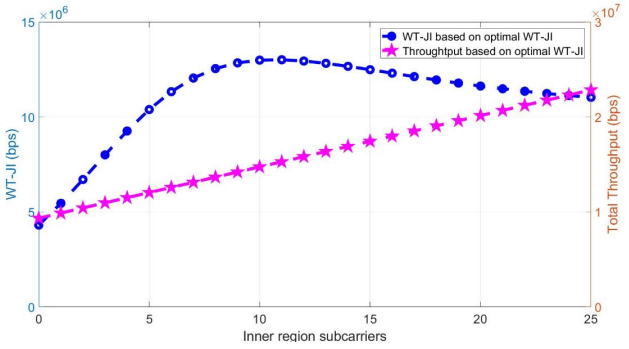


Figure 7: Weighted Fairness Metric Performance for Static Users to Inner Region Subcarriers

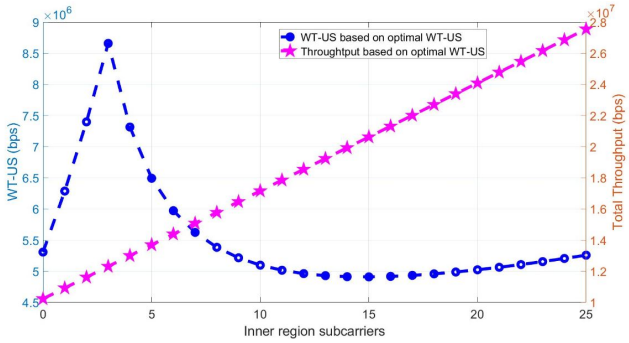


Figure 8: Weighted User Satisfaction Metric Performance for Static Users to Inner Region Subcarriers

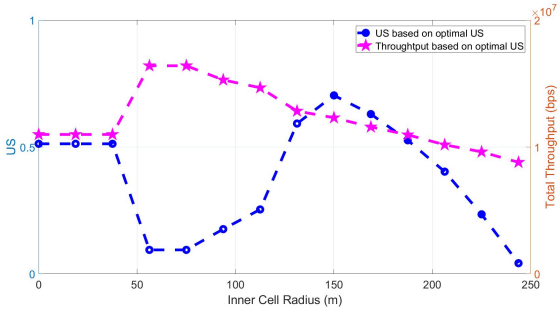


Figure 9: User Satisfaction Metric Performance for Static Users to Inner Region Radius

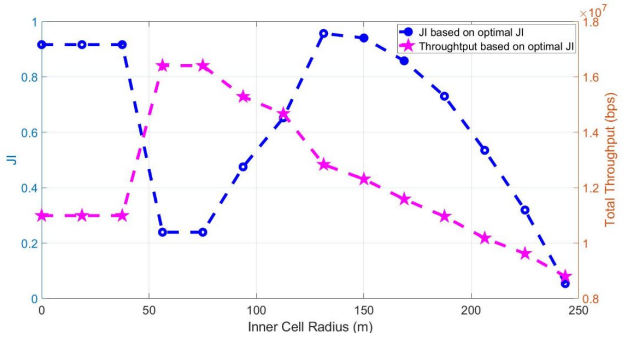


Figure 10: Fairness Index Metric Performance for Static Users to Inner Region Radius

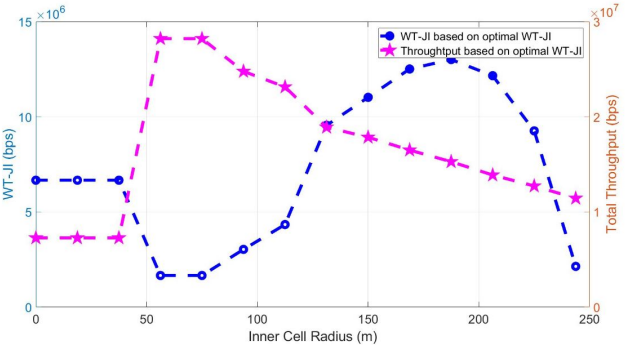


Figure 11: Weighted Fairness Metric Performance for Static Users to Inner Region Radius

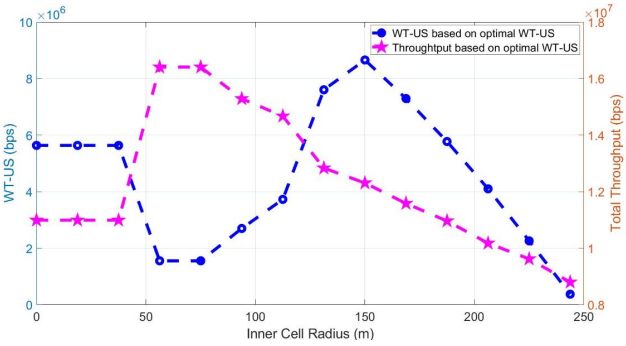


Figure 12: Weighted User Satisfaction Metric Performance for Static Users to Inner Region Radius

6.4.3 Adaptive versus Non-adaptive mechanisms - Metric Performance

As user mobility is introduced and new positions are determined, the simulation calculates metrics and optimal inner radius and subcarrier allocation for adaptive and non-adaptive variations for each FFR mechanism. For non-adaptive FFR mechanism, optimal inner radius and subcarrier allocation remain constant even after the user position changes and performance metrics are calculated. For adaptive FFR mechanism, new values of optimal inner radius and subcarrier allocation are determined based on the new user position and optimal FFR performance. Therefore, the adaptive FFR mechanism reacts to user mobility better than non-adaptive FFR mechanism. Figures 13 - 17 show

FFR performance for all metrics with adaptive and non-adaptive mechanisms applied. In all the adaptive versus non-adaptive mechanisms comparison results, the adaptive mechanism applied on the metrics perform better than non-adaptive mechanism. Particularly, the proposed metric performance Figure 17 shows higher throughput than other metrics and better performance versus non-adaptive and IFR variations. For both weighted user satisfaction (proposed metric) in Figure 17 and weighted fairness (existing metric) in Figure 16, the adaptation process is visible between 50 to 200 seconds. However, despite common benefit of better reaction to the adaptation process, weighted user satisfaction performs better with higher throughput compared to weighted fairness. There is a co-relation between adaptation trend and inner region radius as shown in Section 6.4.4.

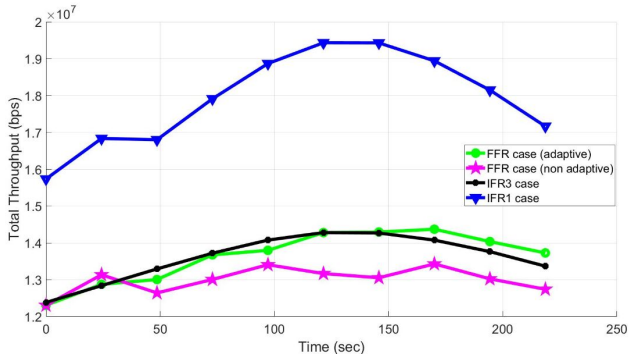


Figure 13: Comparison of adaptive versus non-adaptive mechanisms for Total Throughput

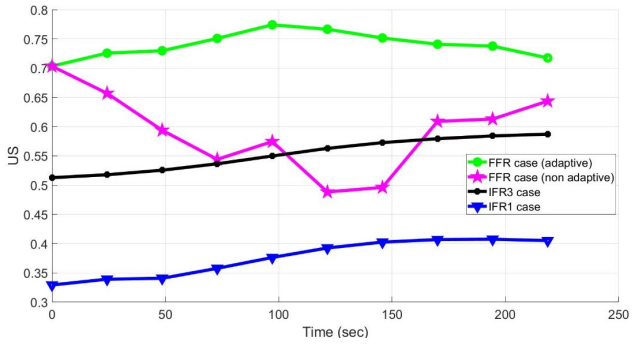


Figure 14: Comparison of adaptive versus non-adaptive mechanisms for User Satisfaction

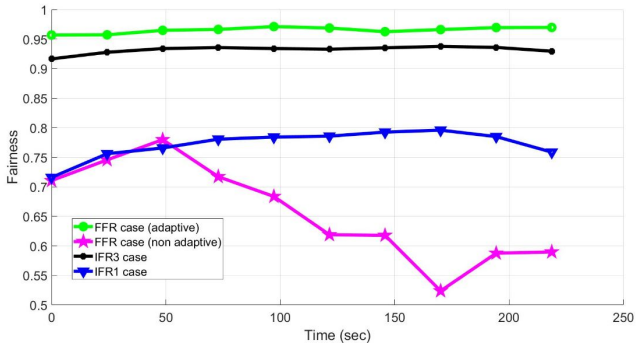


Figure 15: Comparison of adaptive versus non-adaptive mechanisms for Fairness Index

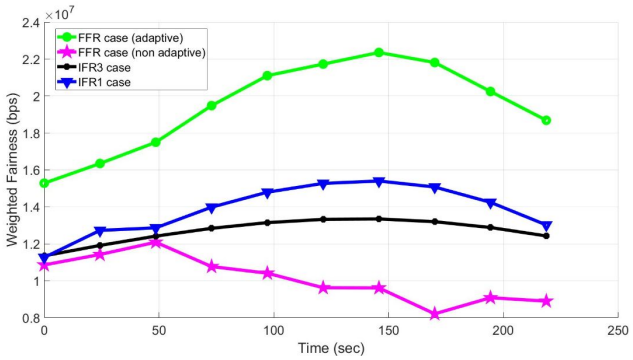


Figure 16: Comparison of adaptive versus non-adaptive mechanisms for Weighted Fairness

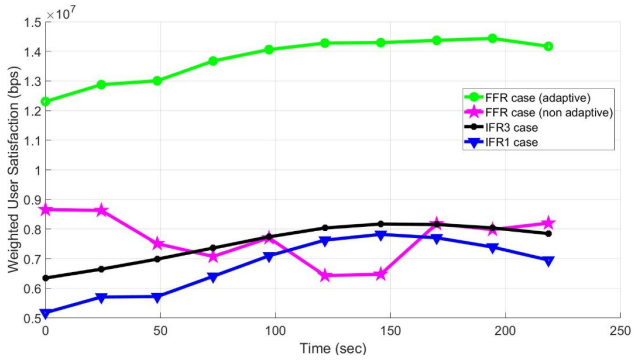


Figure 17: Comparison of adaptive versus non-adaptive mechanisms for Weighted User Satisfaction

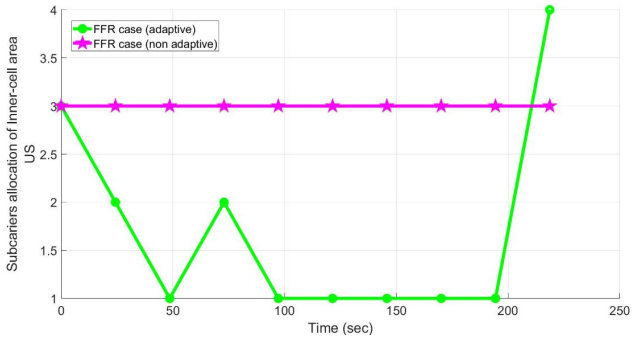


Figure 18: Comparison of Subcarrier Allocation for User Satisfaction

6.4.4 Adaptive versus Non-adaptive mechanisms - Subcarrier Allocation and Inner-Cell Radius

The throughput trend optimized for each metric can be explained with inner region subcarrier and inner region radius results. Figures 22 - 25 show effect of adaptation process on subcarrier allocation for all metrics. Figures 22 - 25 show effect of adaptation process on inner region radius for all metrics. Both subcarrier allocation and inner region radius show active response to the adaptation process applied to the network during the simulation time. For weighted user satisfaction, Figure 25 and weighted fairness Figure 24, the subcarrier allocation reacts positively to the adaptation process between 50 and 200 seconds. High throughput trend during this time interval can be explained due to allocation of all 25 subcarriers. Similarly, the range of

inner region radius changes with the adaptation process during the simulation time compared to its static trend in non-adaptive mechanism as shown in Figures 25, 24.

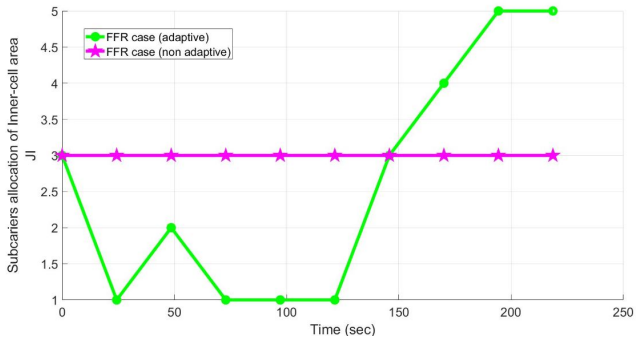


Figure 19: Comparison of Subcarrier Allocation for Fairness Index

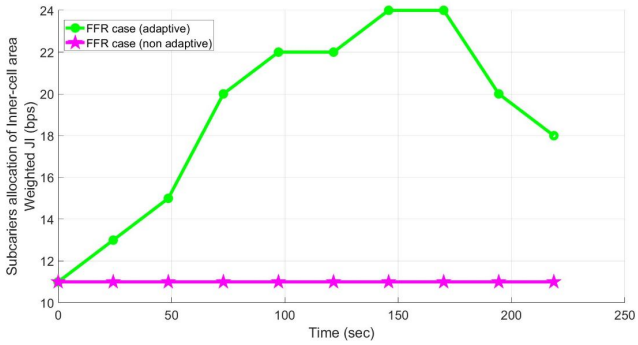


Figure 20: Comparison of Subcarrier Allocation for Weighted Fairness

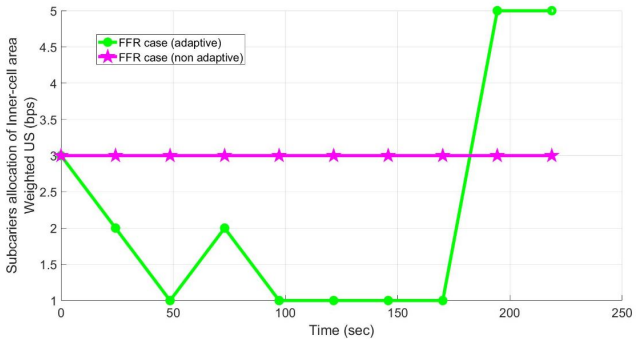


Figure 21: Comparison of Subcarrier Allocation for Weighted User Satisfaction

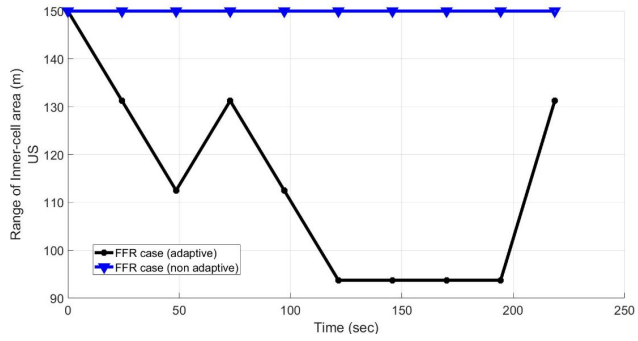


Figure 22: Comparison of Range of Inner-cell for User Satisfaction

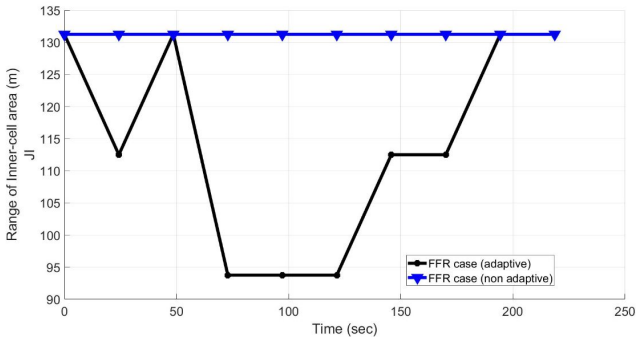


Figure 23: Comparison of Range of Inner-cell for Fairness Index

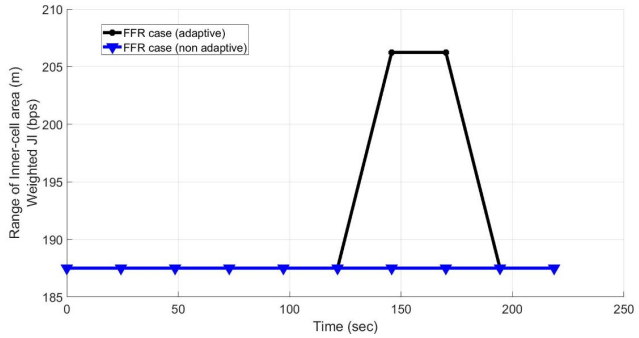


Figure 24: Comparison of Range of Inner-cell for Weighted Fairness

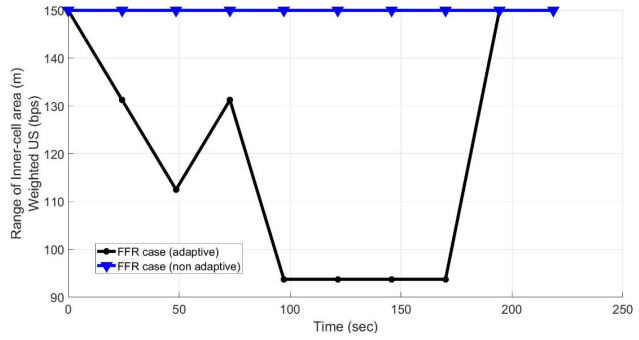


Figure 25: Comparison of Range of Inner-cell for Weighted User Satisfaction

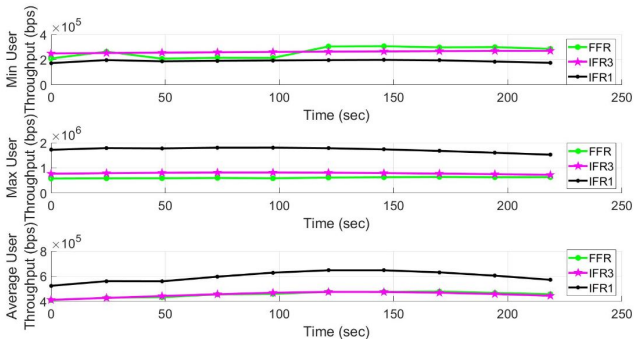


Figure 26: Comparison of Minimum, Maximum, Average for User Satisfaction

6.4.5 Adaptive versus Non-adaptive mechanisms - Minimum, Maximum, and Average Values

FFR mechanism simulation compares their minimum, maximum, and average results over time. Figures 26 - 29 show the metric performance. Maximum and average throughput shows lowest performance for FFR mechanism optimized for weighted user satisfaction metric. This is obvious result since the goal of the new metric is to reduce variance in user throughput and maintain relatively high overall throughput.

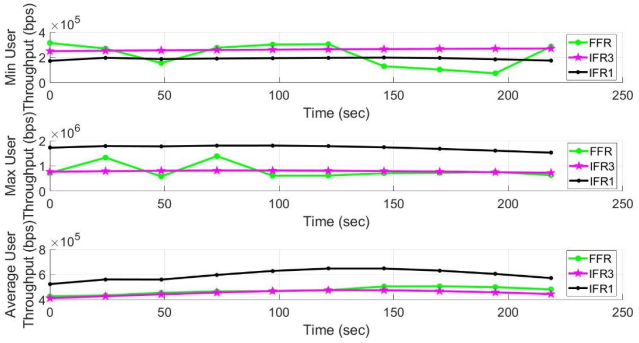


Figure 27: Comparison of Minimum, Maximum, Average for Fairness Index

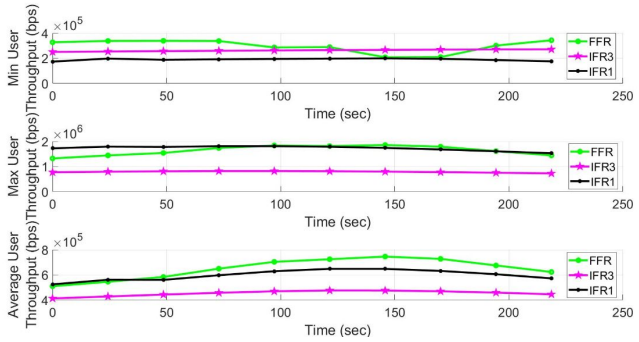


Figure 28: Comparison of Minimum, Maximum, Average for Weighted Fairness

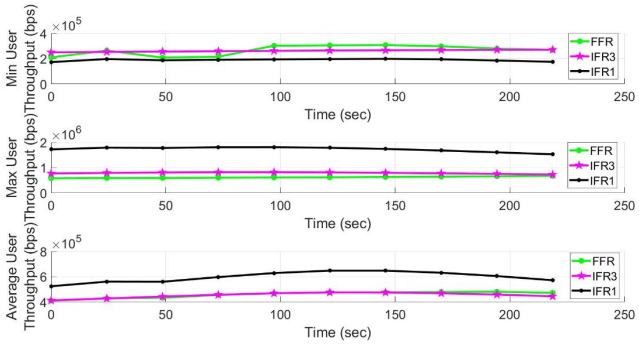


Figure 29: Comparison of Minimum, Maximum, Average for Weighted User Satisfaction

6.4.6 Adaptive versus Non-adaptive mechanisms - Femtocells

LTE femtocells have been developed to increase capacity and raise the throughput of cell-edge users. In this simulation, the LTE network is extended with randomly placed femtocells in each macrocell, assuming an uniform femtocell co-channel deployment. Despite the high throughput benefits, femtocell deployment experience relatively high level of interference from neighboring macrocells and femtocells. FFR frequency deployment is set as indicated in Figure 2. Simulation results in Figure 30 and 31 indicate higher throughput due to addition of femtocell radio resources. Adaptive trend is in line with the previous results without femtocells, where the network throughput adapts to the moving user locations. Network throughput optimized to weighted user satisfaction metric shows higher maximum and average throughput compared to IFR1 and IFR3 due to the presence of additional frequency resources.

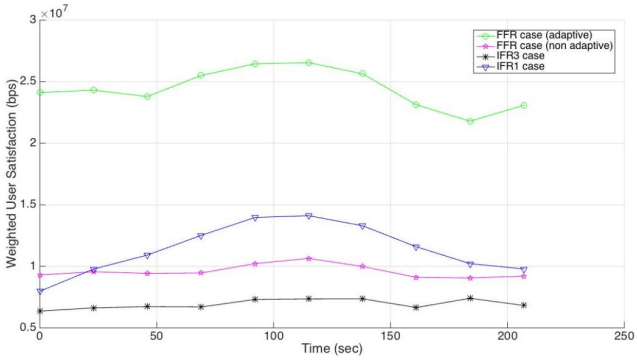


Figure 30: Comparison of adaptive versus non-adaptive mechanisms for Weighted User Satisfaction with Femtocells

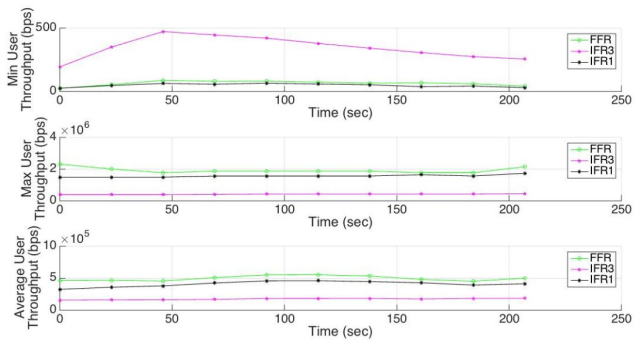


Figure 31: Comparison of Minimum, Maximum, Average for Weighted User Satisfaction with Femtocells

7 Conclusion

In conclusion, the paper evaluated the adaptation process in LTE FFR mechanism. The selected FFR mechanism based on the optimal inner region radius and frequency allocation performed better than other interference co-ordination mechanisms. The FFR mechanism optimized using weighted throughput user satisfaction made a positive trade-off between the existing metrics, as it maintained relatively high total throughput and reduced the variance of per-user throughput. The proposed metric reacted better to the adaptation process in mobility scenarios and generated higher throughput compared to the other metrics. With extreme densification in future mobility, the effect of femtocells on

the network throughput optimized for new metric on moving users was studied. A good research extension would be to review impact of femtocell densification and uniform grid placement on the adaptation process with the proposed metric.

Conflict of Interest The authors declare no conflict of interest.

References

[1] U. Sawant, R. Akl, "Evaluation of Adaptive and Non Adaptive LTE Fractional Frequency Reuse Mechanisms" in The 26th Annual Wireless and Optical Communications Conference, New Jersey, USA, 2017. https://doi.org/10.1109/WOCC.2017.7928998

[2] D. Bilios, C. Bouras, V. Kokkinos, G. Tseliou, A. Papazois, "Selecting the Optimal Fractional Frequency Reuse Scheme in Long Term Evolution Networks" in IEEE Wireless Pers. Communications, 2013. https://doi.org/10.1007/s11277-012-0965-z

[3] M. Yadav, M. Palle, A. Hani, "Performance Analysis of Fractional Frequency Reuse Factor for Interference Suppression in Long Term Evolution" in International Journal of Conceptions on Electronics and Communication Engineering, 2015. https://doi.org/10.1109/PGSRET.2015.7312249

[4] C. Bouras, D. Bilios, V. Kokkinos, A. Papazois, G. Tseliou, "A performance study of Fractional Frequency Reuse in OFDMA networks" in IEEE Wireless Telecommunications Symposium, 2012. https://doi.org/10.1109/WMNC.2012.6416137

[5] C. Bouras, G. Kavourgias, V. Kokkinos, A. Papazois, "Interference management in LTE femtocell systems using an adaptive frequency reuse scheme" in IEEE Wireless and Mobile Networking Conference (WMNC), 2012. https://doi.org/10.1109/PGSRET.2015.7312249

[6] C. Bouras, G. Kavourgias, V. Kokkinos, A. Papazois, "Interference management in LTE femtocell systems using an adaptive frequency reuse scheme" in IEEE Wireless Telecommunications Symposium, 2012. https://doi.org/10.1109/WTS.2012.6266120

[7] D. Bilios, C. Bouras, V. Kokkinos, A. Papazois, G. Tseliou, "Optimization of fractional frequency reuse in Long Term Evolution networks" in IEEE Wireless Communications and Networking Conference (WCNC), 2012. https://doi.org/10.1109/WCNC.2012.6214087

[8] M. Taranetz, J. Ikuno, M. Rupp, "Capacity Density Optimization by Fractional Frequency Partitioning" in IEEE, 2011. https://doi.org/10.1109/ACSSC.2011.6190246

[9] J. Ikuno, M. Taranetz, M. Rupp, "A Fairness-based Performance Evaluation of Fractional Frequency Reuse in LTE" in 2013 17th International ITG Workshop on Smart Antennas (WSA), 2013.

[10] Y. Chang, Z. Tao, J. Zhang, C. Kuo, "A graph approach to dynamic fractional frequency reuse (FFR) in multi-cell OFDMA networks" in In proceedings of IEEE international conference on communications (ICC 2009), 2009. https://doi.org/10.1109/ICC.2009.5198612

[11] M. Assad, "Optimal fractional frequency reuse (FFR) in multicellular OFDMA system" in In proceedings of IEEE 68th Vehicular technology conference (VTC 2008Fall), 2008. https://doi.org/10.1109/VETECF.2008.381

[12] N. Hassan, M. Assad, "Optimal fractional frequency reuse (FFR) and resource allocation in multiuser OFDMA system" in In Proceedings of International Conference on Information and Communication Technologies, (ICICT 2009), 2009. https://doi.org/10.1109/ICICT.2009.5267207

[13] L. Fang, X. Zhang, "Optimal fractional frequency reuse in OFDMA based wireless networks" in Proceedings of 4th international conference on wireless communications, networking and mobile computing, (WiCOM08), 2008. https://doi.org/10.1109/WiCom.2008.166

[14] Y. Xiang, J. Luo, C. Hartmann, "Inter-cell interference mitigation through flexible resource reuse in OFDMA based communication networks" in Proceedings of European Wireless, 2007.

[15] M. Sternad, T. Ottosson, A. Ahlen, A. Svensson, "Attaining both coverage and high spectral efficiency with adaptive OFDM downlinks" in Proceedings of IEEE 58th Vehicular Technology Conference (VTC 2003-Fall), 2003. https://doi.org/10.1109/VETECF.2003.1285981

[16] G. Fodor, C. Koutsimanis, A. Rcz, N. Reider, A. Simonsson, W. Mller, "Intercell interference coordination in OFDMA networks and in the 3GPP long term evolution system" in Journal of Communications, 2009. https://doi.org/10.4304/jcm.4.7.445-453

[17] G. Li, H. Liu, "Downlink radio resource allocation for multi-cell OFDMA system" in IEEE Transactions on Wireless Communications, 2006. https://doi.org/10.1109/TWC.2006.256968

[18] P. Godlewski, M. Maqbool, M. Coupechoux, J. Kelif, "Analytical evaluation of various frequency reuse schemes in cellular OFDMA networks" in Proceedings of 3rd international conference on performance evaluation methodologies and tools (Valuetools 2008), 2008. https://doi.org/10.1016/j.peva.2009.08.001

[19] J. Lim, R. Badlishah, M. Jusoh, "LTE-fractional frequency reuse (FFR) optimization with femtocell network" in 2nd International Conference on Electronic Design (ICED), 2014. https://doi.org/10.1109/ICED.2014.7015863

[20] H. Elfadil, M. Ali, M. Abas, "Fractional frequency reuse in LTE networks" in 2015 2nd World Symposium on Web Applications and Networking (WSWAN), 2015. https://doi.org/10.1109/WSWAN.2015.7210297

[21] P. Yen, Q. Zhan, H. Minn, "New Fractional Frequency Reuse Patterns for Multi-Cell Systems in Time-Varying Channels" in International Journal of Conceptions on Electronics and Communication Engineering, 2015. https://doi.org/10.1109/LWC.2015.2404787

[22] S. Shahsavari, N. Akar, B. Hossein Khalaj, "Joint Cell Muting and User Scheduling in Multi-Cell Networks with Temporal Fairness" in CoRR, 2016.

[23] H. Chang, I. Rubin, "Optimal Downlink and Uplink Fractional Frequency Reuse in Cellular Wireless Networks" in IEEE Transactions on Vehicular Technology, 2015. https://doi.org/10.1109/TVT.2015.2425356

[24] I. Mahmoud, O. Elgzzar, S. Hashima, "An accurate model of worst case signal to interference ratio for frequency reuse cellular systems" in 2016 11th International Conference on Computer Engineering and Systems (ICCES), 2016. https://doi.org/10.1109/ICCES.2016.7822037

[25] N. Saquib, E. Hossain, D. I. Kim, "Fractional frequency reuse for interference management in LTE-Advanced HetNets", in IEEE Wireless Communications, 2013. https://doi.org/10.1109/MWC.2013.6507402

[26] P. Lee, T. Lee, J. Jeong, J. Shin, "Interference Management in LTE Femtocell Systems Using Fractional Frequency Reuse" in The 12th International Conference on Advanced Communication Technology (ICACT), 2010. https://doi.org/10.1109/PGSRET.2015.7312249

[27] Technical Specification Group RAN, "Technical Specification Group RAN, *E-UTRA; LTE RF system scenarios*" in 3GPP Tech. Rep. TS 36.942, 2008. https://doi.org/10.1109/PGSRET.2015.7312249

[28] C. Bouras, V. Kokkinos, A. Papazois, G. Tseliou, "Fractional Frequency Reuse in Integrated Femtocell/Macrocell Environments" in IEEE WWIC, 2013. https://doi.org/10.1007/978-3-642-38401-1_18

[29] R. Jain, D. Chiu, W. Hawe, "A Quantitative Measure of Fairness And Discrimination for Resource Allocation in Shared Computer System" in DEC Technical Report 301, 1984. https://doi.org/10.1109/PGSRET.2015.7312249

[30] Research Unit 6, "FFR Scheme Selection Mechanism", 2013. http://ru6.cti.gr/ru6/FFR_selection_mechanism.zip

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 521-536 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Two-Stage Performance Engineering of Container-based Virtualization

Zheng Li[1,4], Maria Kihl[*,2], Yiqun Chen[3], He Zhang[1]

[1]*Software Institute, Nanjing University, 210008, China*

[2]*Department of Electrical and Information Technology, Lund University, 223 63, Sweden*

[3]*Centre for Spatial Data Infrastructures & Land Administration, University of Melbourne, 3010, Australia*

[4]*Department of Computer Science, University of Concepción, 4070386, Chile*

ARTICLEINFO

Article history:
Received: 14 November, 2017
Accepted: 05 February, 2018
Online: 28 February, 2018

Keywords:
Cloud Computing
Container
Hypervisor
MapReduce
Performance Engineering
Virtualization

ABSTRACT

Cloud computing has become a compelling paradigm built on compute and storage virtualization technologies. The current virtualization solution in the Cloud widely relies on hypervisor-based technologies. Given the recent booming of the container ecosystem, the container-based virtualization starts receiving more attention for being a promising alternative. Although the container technologies are generally considered to be lightweight, no virtualization solution is ideally resource-free, and the corresponding performance overheads will lead to negative impacts on the quality of Cloud services. To facilitate understanding container technologies from the performance engineering's perspective, we conducted two-stage performance investigations into Docker containers as a concrete example. At the first stage, we used a physical machine with "just-enough" resource as a baseline to investigate the performance overhead of a standalone Docker container against a standalone virtual machine (VM). With findings contrary to the related work, our evaluation results show that the virtualization's performance overhead could vary not only on a feature-by-feature basis but also on a job-to-job basis. Moreover, the hypervisor-based technology does not come with higher performance overhead in every case. For example, Docker containers particularly exhibit lower QoS in terms of storage transaction speed. At the ongoing second stage, we employed a physical machine with "fair-enough" resource to implement a container-based MapReduce application and try to optimize its performance. In fact, this machine failed in affording VM-based MapReduce clusters in the same scale. The performance tuning results show that the effects of different optimization strategies could largely be related to the data characteristics. For example, LZO compression can bring the most significant performance improvement when dealing with text data in our case.

1 Introduction

The container technologies have widely been accepted for building next-generation Cloud systems. This paper investigates the performance overhead of container-based virtualization and the performance optimization of a container-based MapReduce application, which is an extension of work originally presented in the 31st IEEE International Conference on Advanced Information Networking and Application (AINA 2017) [1].

The Cloud has been considered to be able to provide computing capacity as the next utility in our modern daily life. In particular, it is the virtualization technologies that enable Cloud computing to be a new paradigm of utility, by playing various vital roles in supporting Cloud services, ranging from resource iso-

[*]Corresponding Author; Address: Ole Römers Väg 3, Lund 223 63, Sweden; Tel: + 46 46 222 9010; Email: maria.kihl@eit.lth.se

https://dx.doi.org/10.25046/aj030163

lation to resource provisioning. The existing virtualization technologies can roughly be distinguished between the hypervisor-based and the container-based solutions. Considering their own resource consumption, both virtualization solutions inevitably introduce performance overheads to running Cloud services, and the performance overheads could then lead to negative impacts to the corresponding quality of service (QoS). Therefore, it would be crucial for both Cloud providers (e.g., for improving infrastructural efficiency) and consumers (e.g., for selecting services wisely) to understand to what extend a candidate virtualization solution incurs influence on the Cloud's QoS.

Recall that hypervisor-driven virtual machines (VMs) require guest operating systems (OS), while containers can share a host OS. Suppose physical machines, VMs and containers are three candidate resource types for a particular Cloud service, a natural hypothesis could be:

> *The physical machine-based service has the best quality among the three resource types, while the container-based service performs better than the hypervisor-based VM service.*

Unfortunately, to the best of our knowledge, there is little quantitative evidence to help test this hypothesis in an "apple-to-apple" manner, except for the similar qualitative discussions. Furthermore, the performance overhead of hypervisor-based and container-based virtualization technologies can even vary in practice depending on different service circumstance (e.g., uncertain workload densities and resource competitions). Therefore, we decided to conduct a twofold investigation into containers from the performance's perspective. Firstly, we used a physical machine with "just-enough" resource as a baseline to quantitatively investigate and compare the performance overheads between the container-based and hypervisor-based virtualizations. In particular, since Docker is currently the most popular container solution [2] and VMWare is one of the leaders in the hypervisor market [3], we chose Docker and VMWare Workstation 12 Pro to represent the two virtualization solutions respectively. Secondly, we implemented a container-based MapReduce cluster on a physical machine with "fair-enough" resource to investigate the performance optimization of our MapReduce application at least in this use case.

According to the clarifications in [4, 5], our qualitative investigations can be regulated by the discipline of experimental computer science (ECS). By employing ECS's recently available Domain Knowledge-driven Methodology (DoKnowMe) [6], we experimentally explored the performance overheads of different virtualization solutions on a feature-by-feature basis, i.e. the communication-, computation-, memory- and storage-related QoS aspects. As for the investigation into performance optimization, we were concerned with the task timeout, out-of-band heartbeat, buffer setting, stream merging, data compression and the cluster size of our container-based MapReduce application.

The experimental results and analyses of performance overhead investigation generally advocate the aforementioned hypothesis. However, the hypothesis is not true in all the cases. For example, we do not see computation performance difference between the three resource types for solving a combinatorially hard chess problem; and the container exhibits even higher storage performance overhead than the VM when reading/writing data byte by byte. Moreover, we find that the remarkable performance loss incurred by both virtualization solutions usually appears in the performance variability.

The performance optimization investigation reveals that various optimization strategies might take different effects due to different data characteristics of a container-based MapReduce application. For example, dealing with text data can significantly benefit from enabling data compression, whereas buffer settings have little effect for dealing with relatively small amount of data.

Overall, our work makes fourfold contributions to the container ecosystem, as specified below.

(1) Our experimental results and analyses can help both researchers and practitioners to better understand the fundamental performance of the present container-based and hypervisor-based virtualization technologies. In fact, the performance evaluation practices in ECS can roughly be distinguished between two stages: the first stage is to reveal the primary performance of specific (system) features, while the second stage is generally based on the first-stage evaluation to investigate real-world application cases. Thus, this work can be viewed as a foundation for more sophisticated evaluation studies in the future.

(2) Our method of calculating performance overhead can easily be applied or adapted to different evaluation scenarios by others. The literature shows that the "performance overhead" has normally been used in the context of qualitative discussions. By quantifying such an indicator, our study essentially provides a concrete lens into the case of performance comparisons.

(3) As a second-stage evaluation work, our case study on the performance optimization of a MapReduce application both demonstrates a practical use case and supplies an easy-to-replicate scenario for engineering performance of container-based applications. In other words, this work essentially proposed a characteristic-consistent data set (i.e. Amazon's spot price history that is open to the public) for future performance engineering studies.

(4) The whole evaluation logic and details reported in this paper can be viewed as a reusable template of evaluating Docker containers. Since the Docker project is still quickly growing [7], the evaluation results could be gradually out of date. Given this template, future evaluations

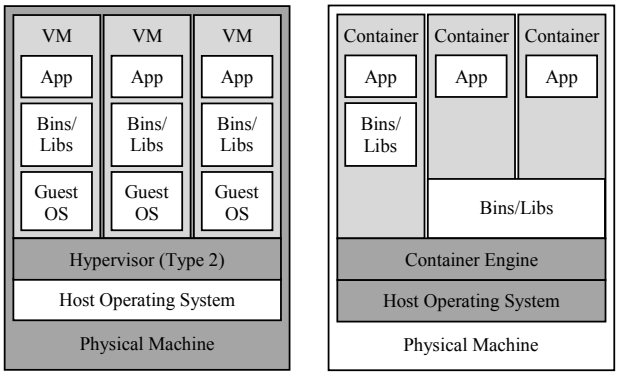
can be conveniently repeated or replicated even by different evaluators at different times and locations. More importantly, by emphasizing the backend logic and evaluation activities, the template-driven evaluation implementations (instead of results only) would be more traceable and comparable.

The remainder of this paper is organized as follows. Section 2 briefly summarizes the background knowledge of container-based and the hypervisor-based virtualization technologies. Section 3 introduces the fundamental performance evaluation of a single container. The detailed performance overhead investigation is divided into two reporting parts, namely pre-experimental activities and experimental results & analyses, and they are correspondingly described into Section 3.2 and 3.3 respectively. Section 4 explains our case study on the performance optimization of a container-based MapReduce application. Section 5 highlights the existing work related to container's performance evaluation. Conclusions and some future work are discussed in Section 6.

2 Hypervisor-based vs. Container-based Virtualization

When it comes to the Cloud virtualization, the de facto solution is to employ the hypervisor-based technologies, and the most representative Cloud service type is offering VMs [8]. In this virtualization solution, the hypervisor manages physical computing resources and makes isolated slices of hardware available for creating VMs [7]. We can further distinguish between two types of hypervisors, namely the bare-metal hypervisor that is installed directly onto the computing hardware, and the hosted hypervisor that requires a host OS. To make a better contrast between the hypervisor-related and container-related concepts, we particularly emphasize the second hypervisor type, as shown in Figure 1a. Since the hypervisor-based virtualization provides access to physical hardware only, each VM needs a complete implementation of a guest OS including the binaries and libraries necessary for applications [9]. As a result, the guest OS will inevitably incur resource competition against the applications running on the VM service, and essentially downgrade the QoS from the application's perspective. Moreover, the performance overhead of the hypervisor would also be passed on to the corresponding Cloud services and lead to negative impacts on the QoS.

To relieve the performance overhead of hypervisor-based virtualization, researchers and practitioners recently started promoting an alternative and lightweight solution, namely container-based virtualization. In fact, the foundation of the container technology can be traced back to the Unix `chroot` command in 1979 [9], while this technology is eventually evolved into virtualization mechanisms like Linux VServer, OpenVZ and Linux Containers (LXC) along with the booming of Linux [10]. Unlike the hardware-level solution of hypervisors, containers realize virtualization at the OS level and utilize isolated slices of the host OS to shield their contained applications [9]. In essence, a container is composed of one or more lightweight images, and each image is a prebaked and replaceable file system that includes necessary binaries, libraries or middlewares for running the application. In the case of multiple images, the read-only supporting file systems are stacked on top of each other to cater for the writable top-layer file system [2]. With this mechanism, as shown in Figure 1b, containers enable applications to share the same OS and even binaries/libraries when appropriate. As such, compared to VMs, containers would be more resource efficient by excluding the execution of hypervisor and guest OS, and more time efficient by avoiding booting (and shutting down) a whole OS [11, 7]. Nevertheless, it has been identified that the cascading layers of container images come with inherent complexity and performance penalty [12]. In other words, the container-based virtualization technology could also negatively impact the corresponding QoS due to its performance overhead.



(a) Hypervisor-based virtual service. (b) Container-based virtual service.

Figure 1: Different architectures of hypervisor-based and container-based virtual services.

3 Fundamental Performance Evaluation of a Single Container

3.1 Performance Evaluation Methodology

Since the comparison between the container's and the VM's performance overheads is essentially based on their performance evaluation, we define our work as a performance evaluation study that belongs to the field of experimental computer science [4, 5]. Considering that "evaluation methodology underpins all innovation in experimental computer science" [13], we employ the methodology DoKnowMe [6] to guide evaluation implementations in this study. DoKnowMe is an abstract evaluation methodology on the analogy of "class" in object-oriented programming. By integrating domain-specific knowledge artefacts, DoKnowMe can be customized into specific methodologies (by analogy

of "object") to facilitate evaluating different concrete computing systems. The skeleton of DoKnowMe is composed of ten generic evaluation steps, as listed below.

(1) Requirement recognition;
(2) Service feature identification;
(3) Metrics and benchmarks listing;
(4) Metrics and benchmarks selection;
(5) Experimental factors listing;
(6) Experimental factors selection;
(7) Experiment design;
(8) Experiment implementation;
(9) Experimental analysis;
(10) Conclusion and documentation.

Each evaluation step further comprises a set of activities together with the corresponding evaluation strategies. The elaboration on these evaluation steps is out of the scope of this paper. To better structure our report, we divide the evaluation implementation into pre-experimental activities and experimental results & analyses.

3.2 Pre-Experimental Activities

3.2.1 Requirement Recognition

Following DoKnowMe, the whole evaluation implementation is essentially driven by the recognized requirements. In general, the requirement recognition is to define a set of specific requirement questions both to facilitate understanding the real-world problem and to help achieve clear statements of the corresponding evaluation purpose. In this case, the basic requirement is to give a fundamental quantitative comparison between the hypervisor-based and container-based virtualization solutions. As mentioned previously, we concretize these two virtualization solutions into VMWare Workstation VMs and Docker containers respectively, in order to facilitate our evaluation implementation (i.e., using a physical machine as a baseline to investigate the performance overhead of a Docker container against a VM). Thus, such a requirement can further be specified into two questions:

RQ1: How much performance overhead does a standalone Docker container introduce over its base physical machine?

RQ2: How much performance overhead does a standalone VM introduce over its base physical machine?

Considering that virtualization technologies could lead to variation in performance of Cloud services [14], we are also concerned with the container's and VM's potential variability overhead besides their average performance overhead:

RQ3: How much performance variability overhead does a standalone Docker container introduce over its base physical machine during a particular period of time?

RQ4: How much performance variability overhead does a standalone VM introduce over its base physical machine during a particular period of time?

3.2.2 Service Feature Identification

Recall that we treat Docker containers as an alternative type of Cloud service to VMs. By using the taxonomy of Cloud services evaluation [15], we directly list the service feature candidates, as shown in Figure 2. Note that a service feature is defined as a combination of a physical property and its capacity, and we individually examine the four physical properties in this study:

- Communication
- Computation
- Memory
- Storage

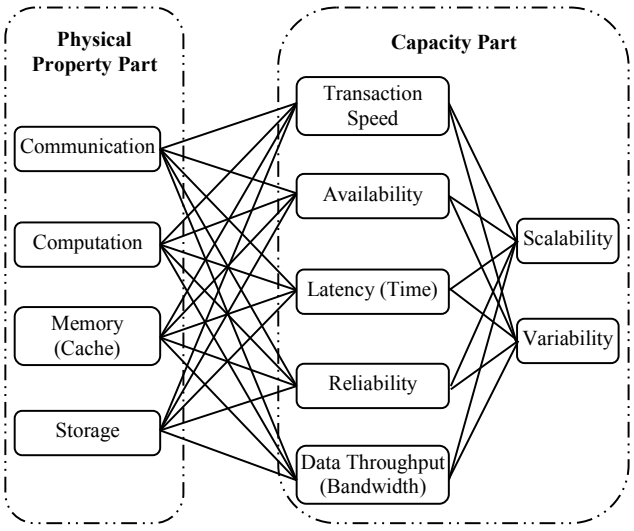


Figure 2: Candidate service features for evaluating Cloud service performance (cf. [15]).

3.2.3 Metrics/Benchmarks Listing and Selection

The selection of evaluation metrics usually depends on the availability of benchmarks. According to our previous experience of Cloud services evaluation, we choose relatively lightweight and popular benchmarks to try to minimize the benchmarking overhead, as listed in Table 1. For example, Iperf has been identified to be able to deliver more precise results by consuming less system resources. In fact, except for STREAM that is the de facto memory evaluation benchmark included in the HPC Challenge Benchmark (HPCC) suite, the other benchmarks are all Ubuntu's built-in utilities.

Physical Property	Capacity Metric	Benchmark	Version
Communication	Data Throughput	Iperf	2.0.5
Computation	(Latency) Score	HardInfo	0.5.1
Memory	Data Throughput	STREAM	5.10
Storage	Transaction Speed	Bonnie++	1.97.1
Storage	Data Throughput	Bonnie++	1.97.1

Table 1: Metrics and benchmarks for this evaluation study.

In particular, although Bonnie++ only measures the amount of data processed per second, the disk I/O transactions are on a byte-by-byte basis when accessing small size of data. Therefore, we consider to measure storage transaction speed when operating byte-size data and measure storage data throughput when operating block-size data. As for the property computation, considering the diversity in CPU jobs (e.g., integer and floating-point calculations), we employ HardInfo that includes six micro-benchmarks to generate performance scores, as briefly explained in Table 2. HardInfo is a tool package that can summarize the information about the host machine's hardware and operating system, as well as benchmarking the CPU. In this study, we employ HardInfo for CPU benchmarking only.

Benchmark	Brief Explanation
CPU Blowfish	Encrypting blocks of random data using the Blowfish algorithm.
CPU CryptoHash*	Checking the ability of the computer to find the hash of a specific test file.
CPU Fibonacci	Calculating the 42*nd* Fibonacci number.
CPU N-Queens	Solving the combinatorially hard chess problem of placing N queens on an $N \times N$ chessboard such that no queen can attack any other.
FPU FFT	Computing a fast Fourier transform.
FPU Raytracing	Generating an image by tracing the path of light through pixels in an image plane and simulating the effects of its encounters with virtual objects.

*The higher the better. (The lower the better for the others.)

Table 2: Micro-benchmarks included in HardInfo.

When it comes to the performance overhead, we use the business domain's *Overhead Ratio*[1] as an analogy to its measurement. In detail, we treat the performance loss compared to a baseline as the expense, while imagining the baseline performance to be the overall income, as defined in Equation (1).

$$O_p = \frac{|P_m - P_b|}{P_b} \times 100\% \quad (1)$$

where O_p refers to the performance overhead; P_m denotes the benchmarking result as a measurement of a service feature; P_b indicates the baseline performance of the service feature; and then $|P_m - P_b|$ represents the corresponding performance loss. Note that the physical machine's performance is used as the baseline in our study. Moreover, considering possible observational errors, we allow a margin of error for the confidence level as high as 99% with regarding to the benchmarking results. In other words, we will ignore the difference between the measured performance and its baseline if the calculated performance overhead is less than 1% (i.e. if $O_p < 1\%$, then $P_m = P_b$).

[1] http://www.investopedia.com/terms/o/overhead-ratio.asp

3.2.4 Experimental Factor Listing and Selection

The identification of experimental factors plays a prerequisite role in the following experimental design. More importantly, specifying the relevant factors would be necessary for improving the repeatability of experimental implementations. By referring to the experimental factor framework of Cloud services evaluation [16], we choose the resource- and workload-related factors as follows.

The resource-related factors:

- *Resource Type:* Given the evaluation requirement, we have essentially considered three types of resources to support the imaginary Cloud service, namely physical machine, container and VM.
- *Communication Scope:* We test the communication between our local machine and an Amazon EC2 t2.micro instance. The local machine is located in our broadband lab at Lund University, and the EC2 instance is from Amazon's available zone ap-southeast-1a within the region Asia Pacific (Singapore).
- *Communication Ethernet Index:* Our local side uses a Gigabit connection to the Internet, while the EC2 instance at remote side has the "Low to Moderate" networking performance defined by Amazon.
- *CPU Index:* The physical machine's CPU model is Intel Core™2 Duo Processor T7500. The processor has two cores with the 64-bit architecture, and its base frequency is 2.2 GHz. We allocate both CPU cores to the standalone VM upon the physical machine.
- *Memory Size:* The physical machine is equipped with a 3GB DDR2 SDRAM. When running the VMWare Workstation Pro without launching any VM, "`watch -n 5 free -m`" shows a memory usage of 817MB while leaving 2183MB free in the physical machine. Therefore, we set the memory size to 2GB for the VM to avoid (at least to minimize) the possible memory swapping.
- *Storage Size:* There are 120GB of hard disk in the physical machine. Considering the space usage by the host operating system, we allocate 100GB to the VM.
- *Operating System:* Since Docker requires a 64-bit installation and Linux kernels older than

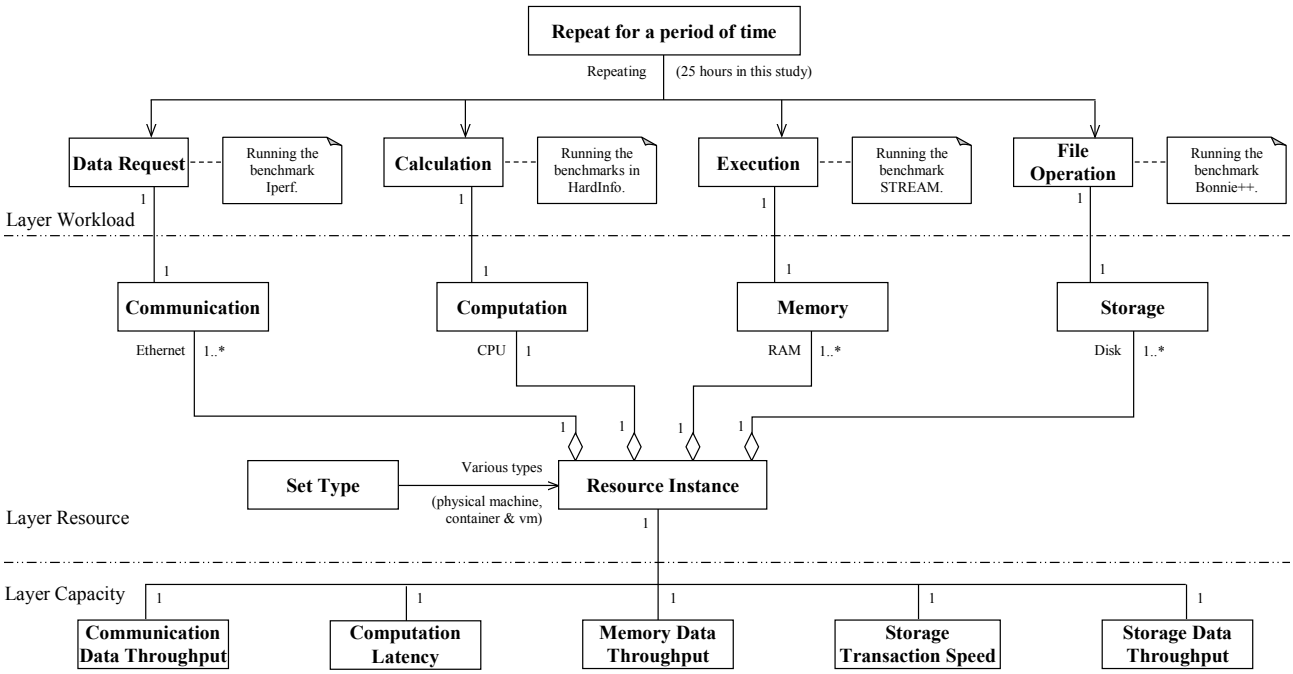


Figure 3: Experimental blueprint for evaluating three types of resources in this study.

3.10 do not support all the features for running Docker containers, we choose the latest 64-bit Ubuntu 15.10 as the operating system for both the physical machine and the VM. In addition, according to the discussions about base images in the Docker community [17, 18], we intentionally set an OS base image (by specifying `FROM ubuntu:15.10` in the Dockerfile) for all the Docker containers in our experiments. Note that a container's OS base image is only a file system representation, while not acting as a guest OS.

The workload-related factors:

- *Duration:* For each evaluation experiment, we decided to take a whole-day observation plus one-hour warming up (i.e. 25 hours).
- *Workload Size:* The experimental workloads are predefined by the selected benchmarks. For example, the micro-benchmark CPU Fibonacci generates workload by calculating the 42*nd* Fibonacci number (cf. Table 2). In particular, the benchmark Bonnie++ distinguishes between reading/writing byte-size and block-size data.

3.2.5 Experimental Design

It is clear that the identified factors are all with single value except for the *Resource Type*. Therefore, a straightforward design is to run the individual benchmarks on each of the three types of resources independently for a whole day plus one hour.

Furthermore, following the conceptual model of IaaS performance evaluation [19], we record the experimental design into a blueprint both to facilitate our experimental implementations and to help other evaluators replicate/repeat our study. In particular, the experimental elements are divided into three layers (namely Layer Workload, Layer Resource, and Layer Capacity), as shown in Figure 3. To avoid duplication, we do not elaborate the detailed elements in this blueprint.

3.3 Experimental Results and Analyses

3.3.1 Communication Evaluation Result and Analysis

Docker creates a virtual bridge `docker0` on the host machine to enable both the host-container and the container-container communications. In particular, it is the Network Address Translation (NAT) that forwards containers' traffic to external networks. For the purpose of "apple-to-apple" comparison, we also configure VM's network type as NAT that uses the `VMnet8` virtual switch created by VMware Workstation.

Using NAT, both Docker containers and VMs can establish outgoing connections by default, while they require port binding/forwarding to accept incoming connections. To reduce the possibility of configurational noise, we only test the outgoing communication performance, by setting the remote EC2 instance to Iperf server and using the local machine, container and VM all as Iperf clients.

The benchmarking results of repeating `iperf -c XXX.XXX.XXX.XXX -t 15` (with a one-minute interval between every two consecutive trials) are listed in Table 3. The `XXX.XXX.XXX.XXX` denotes the external IP address of the EC2 instance used in our experiments. Note that, unlike the other performance features, the communication data throughput delivers periodical and significant fluctuations, which might be a result from the network resource competition at both our local side and the EC2 side during working hours. There-

fore, we particularly focus on the longest period of relatively stable data out of the whole-day observation, and thus the results here are for rough reference only.

Resource Type	Average	Standard Deviation
Physical machine	29.066 Mbits/sec	1.282 Mbits/sec
Container	28.484 Mbits/sec	1.978 Mbits/sec
Virtual machine	12.843 Mbits/sec	2.979 Mbits/sec

Table 3: Communication benchmarking results using Iperf.

Given the extra cost of using the NAT network to send and receive packets, there would be unavoidable performance penalties for both the container and the VM. Using Equation (1), we calculate their communication performance overheads, as illustrated in Figure 4.

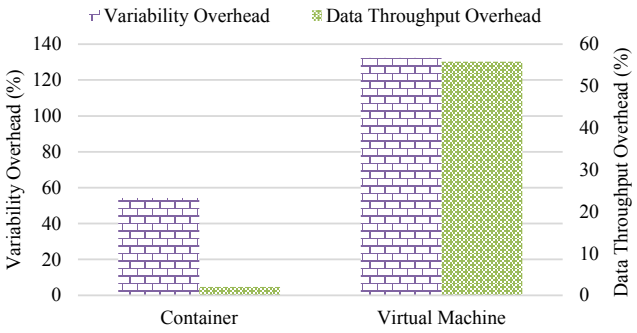


Figure 4: Communication data throughput and its variability overhead of a standalone Docker container vs. VM (using the benchmark Iperf).

A clear trend is that, compared to the VM, the container loses less communication performance, with only 2% data throughput overhead and around 54% variability overhead. However, it is surprising to see a more than 55% data throughput overhead for the VM. Although we have double checked the relevant configuration parameters and redone several rounds of experiments to confirm this phenomenon, we still doubt about the hypervisor-related reason behind such a big performance loss. We particularly highlight this observation to inspire further investigations.

3.3.2 Computation Evaluation Result and Analysis

Recall that HardInfo's six micro benchmarks deliver both "higher=better" and "lower=better" CPU scores (cf. Table 2). To facilitate experimental analysis, we use the two equations below to standardize the "higher=better" and "lower=better" benchmarking results respectively.

$$HB_i = \frac{Benchmarking_i}{\max(Benchmarking_{1,2,\ldots,n})} \quad (2)$$

$$LB_i = \frac{\frac{1}{Benchmarking_i}}{\max\left(\frac{1}{Benchmarking_{1,2,\ldots,n}}\right)} \quad (3)$$

where HB_i further scores the service resource type i by standardizing the "higher=better" benchmarking result $Benchmarking_i$; and similarly, LB_i represents the standardized "lower=better" CPU score of the service resource type i. Note that Equation (3) essentially offers the "lower=better" benchmarking results a "higher=better" representation through reciprocal standardization.

For the purpose of conciseness, here we only specify the standardized experimental results, as shown in Table 4. Exceptionally, the container and VM have slightly higher CPU N-Queens scores than the physical machine. Given the predefined observational margin of error (cf. Section 3.2.3), we are not concerned with this trivial difference, while treating their performance values as equal to each other in this case.

Benchmark	Physical machine	Container	VM
CPU Blowfish	1	0.986	0.942
CPU CryptoHash	1	0.992	0.943
CPU Fibonacci	1	0.999	0.976
CPU N-Queens	0.996	1	0.997
FPU FFT	1	0.966	0.924
FPU Raytracing	1	0.968	0.941

Table 4: Standardized computation benchmarking results using HardInfo.

We can further use a radar plot to help ignore the trivial number differences, and also help intuitively contrast the performance of the three resource types, as demonstrated in Figure 5. For example, the different polygon sizes clearly indicate that the container generally computes faster than the VM, although the performance differences are on a case-by-case basis with respect to different CPU job types.

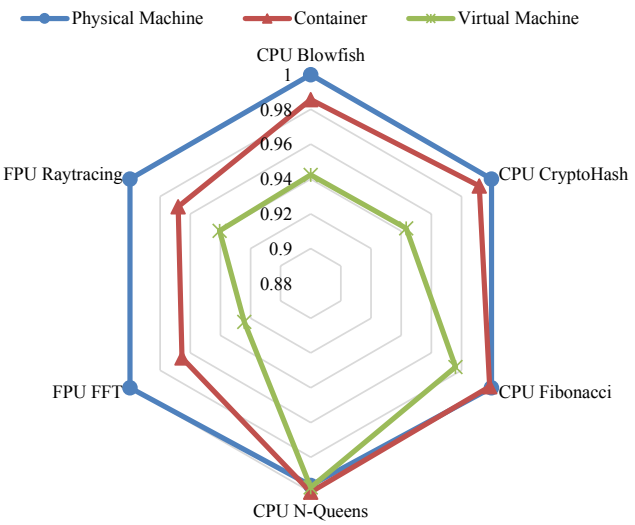


Figure 5: Computation benchmarking results by using HardInfo.

Nevertheless, our experimental results do not display any general trend in variability of those resources' computation scores. As can be seen from the calculated performance overheads (cf. Figure 6), the VM

does not even show worse variability than the physical machine when running CPU CryptoHash, CPU N-Queens and FPU Raytracing. On the contrary, there is an almost 2500% variability overhead for the VM when calculating the 42*nd* Fibonacci number. In particular, the virtualization technologies seem to be sensitive to the Fourier transform jobs (the benchmark FPU FFT), because the computation latency overhead and the variability overhead are relatively high for both the container and the VM.

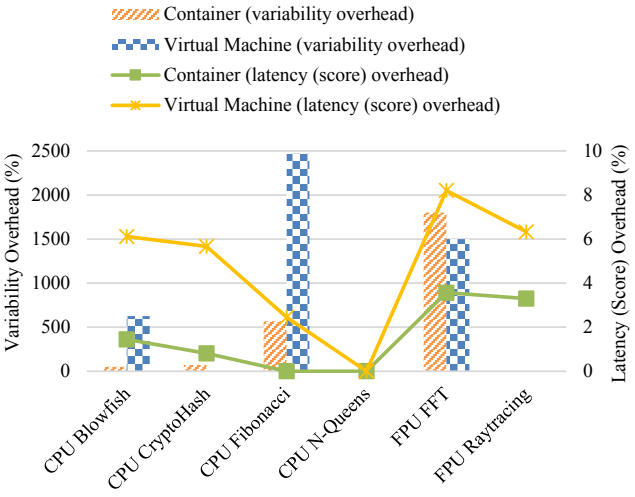


Figure 6: Computation latency (score) and its variability overhead of a standalone Docker container vs. VM (using the tool kit HardInfo).

3.3.3 Memory Evaluation Result and Analysis

STREAM measures sustainable memory data throughput by conducting four typical vector operations, namely Copy, Scale, Add and Triad. The memory benchmarking results are listed in Table 5. We further visualize the results into Figure 7 to facilitate our observation. As the first impression, it seems that the VM has a bit poorer memory data throughput, and there is little difference between the physical machine and the Docker container in the context of running STREAM.

Operation (MB/s)	Physical machine	Container	VM
Copy	2902.685	2914.023	2818.291
(Std. Dev.)	(4.951)	(12.579)	(57.633)
Scale	2916.247	2910.485	2765.737
(Std. Dev.)	(3.783)	(14.488)	(59.193)
Add	3335.634	3332.822	3159.188
(Std. Dev.)	(3.765)	(14.405)	(58.385)
Triad	3341.327	3340.416	3204.811
(Std. Dev.)	(3.976)	(26.361)	(59.004)

Table 5: Memory benchmarking results using STREAM.

By calculating the performance overhead in terms of memory data throughput and its variability, we are able to see the significant difference among these three types of resources, as illustrated in Figure 8. Take the operation Triad as an example, although the container performs as well as the physical machine on average, the variability overhead of the container is more than 500%; similarly, although the VM's Triad data throughput overhead is around 4% only, its variability overhead is almost 1400%. In other words, the memory performance loss incurred by both virtualization techniques is mainly embodied with the increase in the performance variability.

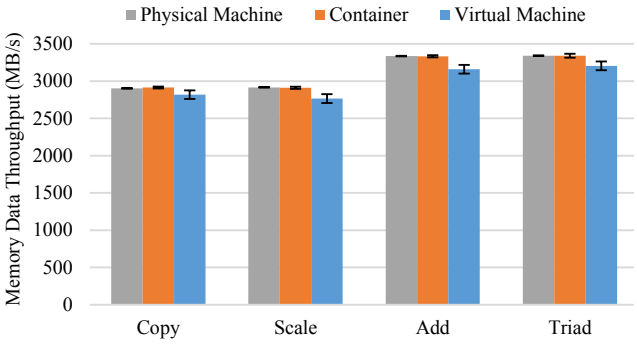


Figure 7: Memory benchmarking results by using STREAM. Error bars indicate the standard deviations of the corresponding memory data throughput.

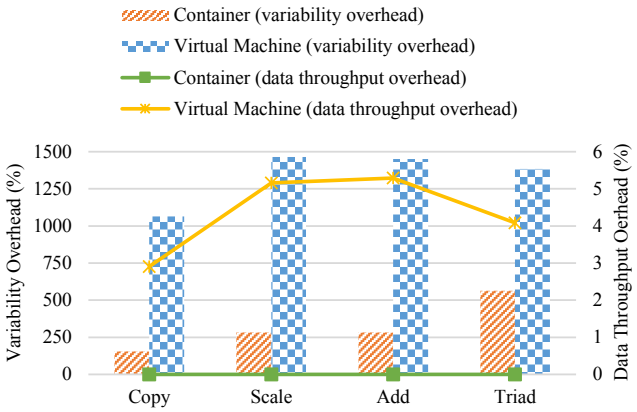


Figure 8: Memory data throughput and its variability overhead of a standalone Docker container vs. VM (using the benchmark STREAM).

In addition, it is also worth notable that the container's average Copy data throughput is even slightly higher than the physical machine (i.e. 2914.023MB/s vs. 2902.685MB/s) in our experiments. Recall that we have considered a 1% margin of error. Since those two values are close to each other within this error margin, here we ignore such an irregular phenomenon as an observational error.

3.3.4 Storage Evaluation Result and Analysis

For the test of disk reading and writing, Bonnie++ creates a dataset twice the size of the involved RAM memory. Since the VM is allocated 2GB of RAM, we also restrict the memory usage to 2GB for Bonnie++

on both the physical machine and the container, by running "`sudo bonnie++ -r 2048 -n 128 -d / -u root`". Correspondingly, the benchmarking trials are conducted with 4GB of random data on the disk. When Bonnie++ is running, it carries out various storage operations ranging from data reading/writing to file creating/deleting. Here we only focus on the performance of reading/writing byte- and block-size data.

To help highlight several different observations, we plot the trajectory of the experimental results along the trial sequence during the whole day, as shown in Figure 9. The first surprising observation is that, all the three resource types have regular patterns of performance jitter in block writing, rewriting and reading. Due to the space limit, we do not report their block rewriting performance in this paper. By exploring the hardware information, we identified the hard disk drive (HDD) model to be ATA Hitachi HTS54161, and its specification describes "It stores 512 bytes per sector and uses four data heads to read the data from two platters, rotating at 5,400 revolutions per minute". As we know, the hard disk surface is divided into a set of concentrically circular tracks. Given the same rotational speed of an HDD, the outer tracks would have higher data throughput than the inner ones. As such, those regular patterns might indicate that the HDD heads sequentially shuttle between outer and inner tracks when consecutively writing/reading block data during the experiments.

The second surprising observation is that, unlike most cases in which the VM has the worst performance, the container seems significantly poor at accessing the byte size of data, although its performance variability is clearly the smallest. We further calculate the storage performance overhead to deliver more specific comparison between the container and the VM, and draw the results into Figure 10. Note that, in the case when the container's/VM's variability is smaller than the physical machine's, we directly set the corresponding variability overhead to zero rather than allowing any performance overhead to be negative. Then, the bars in the chart indicate that the storage variability overheads of both virtualization technologies are nearly negligible except for reading byte-size data on the VM (up to nearly 200%). The lines show that the container brings around 40% to 50% data throughput overhead when performing disk operations on a byte-by-byte basis. On the contrary, there is relatively trivial performance loss in VM's byte data writing. However, the VM has roughly 30% data throughput overhead in other disk I/O scenarios, whereas the container barely incurs overhead when reading/writing large size of data.

Our third observation is that, the storage performance overhead of different virtualization technologies can also be reflected through the total number of the iterative Bonnie++ trials. As pointed by the maximum x-axis scale in Figure 9, the physical machine, the container and the VM can respectively finish 150, 147 and 101 rounds of disk tests during 24 hours. Given this information, we estimate the container's and the VM's storage performance overhead to be 2% (= |147−150|/150) and 32.67% (= |101−150|/150) respectively.

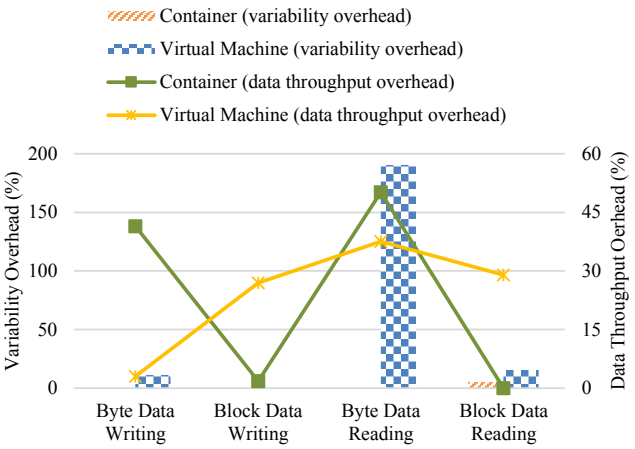
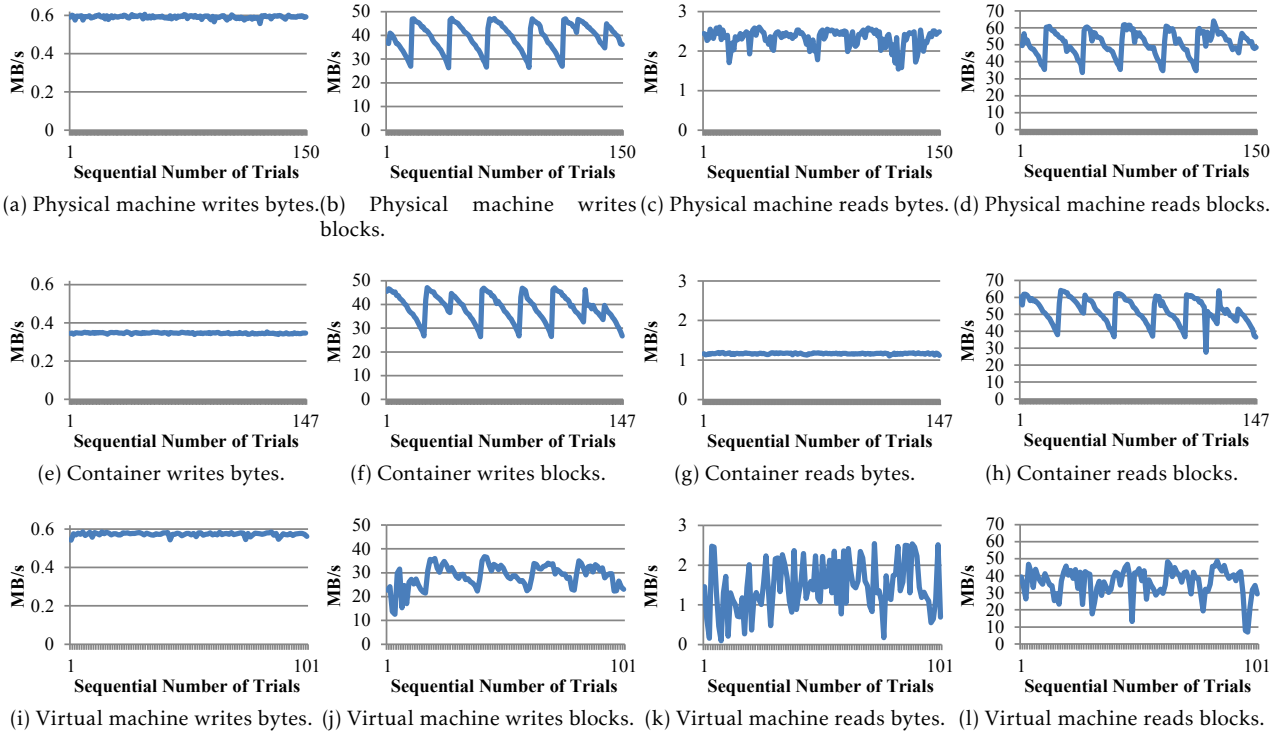


Figure 10: Storage data throughput and its variability overhead of a standalone Docker container vs. VM (using the benchmark Bonnie++).

3.4 Performance Evaluation Conclusion

Following the performance evaluation methodology DoKnowMe, we draw conclusions mainly by answering the predefined requirement questions. Driven by RQ1 and RQ2, our evaluation result largely confirms the aforementioned qualitative discussions: The container's average performance is generally better than the VM's and is even comparable to that of the physical machine with regarding to many features. Specifically, the container has less than 4% performance overhead in terms of communication data throughput, computation latency, memory data throughput and storage data throughput. Nevertheless, the container-based virtualization could hit a bottleneck of storage transaction speed, with the overhead up to 50%. Note that, as mentioned previously, we interpret the byte-size data throughput into storage transaction speed, because each byte essentially calls a disk transaction here. In contrast, although the VM delivers the worst performance in most cases, it could perform as well as the physical machine when solving the N-Queens problem or writing small-size data to the disk. By further comparing the storage filesystems of those two types of virtualization technologies, we believe that it is the copy-on-write mechanism that makes containers poor at storage transaction speed.

Driven by RQ3 and RQ4, we find that the performance loss resulting from virtualizations is more visible in the performance variability. For example, the container's variability overhead could reach as high as over 500% with respect to the Fibonacci calculation and the memory Triad operation. Similarly, although the container generally shows less performance variability than the VM, there are still exceptional cases: The container has the largest performance variation in the job of computing Fourier transform, whereas even the VM's performance variability is not worse than



(a) Physical machine writes bytes. (b) Physical machine writes blocks. (c) Physical machine reads bytes. (d) Physical machine reads blocks.

(e) Container writes bytes. (f) Container writes blocks. (g) Container reads bytes. (h) Container reads blocks.

(i) Virtual machine writes bytes. (j) Virtual machine writes blocks. (k) Virtual machine reads bytes. (l) Virtual machine reads blocks.

Figure 9: Storage benchmarking results by using Bonnie++ during 24 hours. The maximum x-axis scale indicates the iteration number of the Bonnie++ test (i.e. the physical machine, the container and the VM run 150, 147 and 101 tests respectively).

the physical machine's when running CryptoHash, N-Queens, and Raytracing jobs.

4 Container-based Application Case Study and Performance Optimization

4.1 Motive from a Background Project

Adequate pricing techniques play a key role in successful Cloud computing [20]. In the de facto Cloud market, there are generally three typical pricing schemes, namely on-demand pricing scheme, reserved pricing scheme, and spot pricing scheme. Although the fixed pricing schemes are dominant approaches to trading Cloud resources nowadays, spot pricing has been broadly agreed as a significant supplement for building a full-fledged market economy for the Cloud ecosystem [21]. Similar to the dynamic pricing in the electricity distribution industry, the spot pricing scheme here also employs a market-driven mechanism to provide spot service at a reduced and fluctuating price, in order to attract more demands and better utilize idle compute resources [22].

Unfortunately, the backend details behind changing spot prices are invisible for most of the Cloud market participants. In fact, unlike the static and straightforward pricing schemes of on-demand and reserved Cloud services, the market-driven mechanism for pricing Cloud spot service has been identified to be complicated both for providers to implement and for consumers to understand.

Therefore, it has become popular and valuable to take Amazon's spot service as a practical example to investigate Cloud spot pricing, so as to encourage and facilitate more players to enter the Cloud spot market. We are currently involved in a project on Cloud spot pricing analytics by using the whole-year price history of Amazon's 1053 types of spot service instances. Since Amazon only offers the most recent 60-day price trace to the public for review, we downloaded the price traces monthly in the past year to make sure the completeness of the whole-year price history.

In this way of downloading price traces, it is clear that around half of the overall raw data are duplicate. Moreover, the original price trace is sorted by the timestamp only, as shown in Table 6. To analyze spot service pricing on an instance-by-instance basis, however, the collected price data need not only to be sorted by timestamp but also to be distinguished by the other three attributes, i.e. Instance Type, Operating System and Zone. Thus, we decided to implement a preprocessing program to help clean the data, including removing the duplicate price records and categorizing price traces for individual service instances.

Tag	Price ($)	Timestamp	Instance Type	Operating System	Zone
SPOTINSTANCEPRICE	0.072700	2017-10-04T13:39:11+0000	m1.large	Windows	us-east-1a
SPOTINSTANCEPRICE	0.552500	2017-10-04T13:39:11+0000	c3.8xlarge	SUSE Linux	us-east-1b
SPOTINSTANCEPRICE	0.546700	2017-10-04T13:39:11+0000	c3.8xlarge	SUSE Linux	us-east-1e
SPOTINSTANCEPRICE	0.452500	2017-10-04T13:39:11+0000	c3.8xlarge	Linux/UNIX	us-east-1b
SPOTINSTANCEPRICE	0.446700	2017-10-04T13:39:11+0000	c3.8xlarge	Linux/UNIX	us-east-1e
SPOTINSTANCEPRICE	0.400800	2017-10-04T13:39:11+0000	c3.2xlarge	Windows (Amazon VPC)	us-east-1e
SPOTINSTANCEPRICE	0.417800	2017-10-04T13:39:11+0000	c3.4xlarge	SUSE Linux	us-east-1c
SPOTINSTANCEPRICE	0.317800	2017-10-04T13:39:11+0000	c3.4xlarge	Linux/UNIX	us-east-1c
SPOTINSTANCEPRICE	0.039600	2017-10-04T13:39:10+0000	m1.small	SUSE Linux	us-east-1b
SPOTINSTANCEPRICE	0.009600	2017-10-04T13:39:10+0000	m1.small	Linux/UNIX	us-east-1b
SPOTINSTANCEPRICE	0.434100	2017-10-04T13:39:10+0000	i3.2xlarge	SUSE Linux (Amazon VPC)	us-east-1c
SPOTINSTANCEPRICE	0.334100	2017-10-04T13:39:10+0000	i3.2xlarge	Linux/UNIX (Amazon VPC)	us-east-1c
SPOTINSTANCEPRICE	0.200000	2017-10-04T13:39:09+0000	c4.large	SUSE Linux (Amazon VPC)	us-east-1d
SPOTINSTANCEPRICE	0.100000	2017-10-04T13:39:09+0000	c4.large	Linux/UNIX (Amazon VPC)	us-east-1d
SPOTINSTANCEPRICE	0.281700	2017-10-04T13:39:09+0000	m3.2xlarge	Windows (Amazon VPC)	us-east-1c
SPOTINSTANCEPRICE	0.211200	2017-10-04T13:39:09+0000	m2.xlarge	SUSE Linux	us-east-1d
SPOTINSTANCEPRICE	0.111200	2017-10-04T13:39:09+0000	m2.xlarge	Linux/UNIX	us-east-1d

Table 6: A small piece of Amazon's spot price trace.

4.2 WordCount-alike Solution

Given the aforementioned requirement of data cleaning, we propose a WordCount-alike solution by analogy. WordCount is a well-known application that calculates the numbers of occurrences of different words within a document or word set. In the domain of big data analytics, WordCount is now a classic application to demonstrate the MapReduce mechanism that has become a standard programming model since 2004 [23]. As shown in Figure 11, the key steps in a MapReduce workflow are:

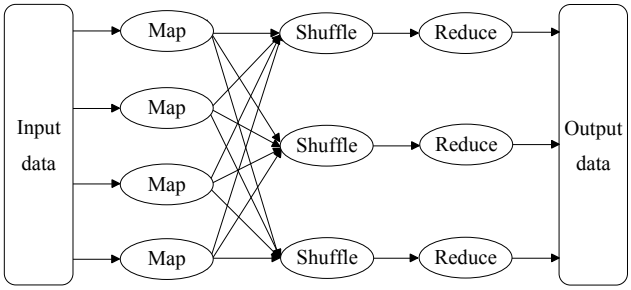


Figure 11: Workflow of the MapReduce process.

(1) The initial input source data are segmented into blocks according to the predefined split function and saved as a list of key-value pairs.

(2) The mapper executes the user-defined map function which generates intermediate key-value pairs.

(3) The intermediate key-value pairs generated by mapper nodes is sent to a specific reducer based on the key.

(4) Each reducer computes and reduces the data to one single key-value pair.

(5) All the reduced data are integrated into the final result of a MapReduce job.

Benefiting from MapReduce, applications like WordCount can deal with large amounts of data parallelly and distributedly. For the purpose of conciseness, we use a three-file scenario to demonstrate the process of MapReduce-based WordCount, as illustrated in Figure 12. In brief, the input files are broken into a set of <key, value> pairs for individual words, then the <key, value> pairs are shuffled alphabetically to facilitate summing up the values (i.e. the occurrence counts) for each unique key, and the reduced results are also a set of <key, value> pairs that directly act as the output in this case.

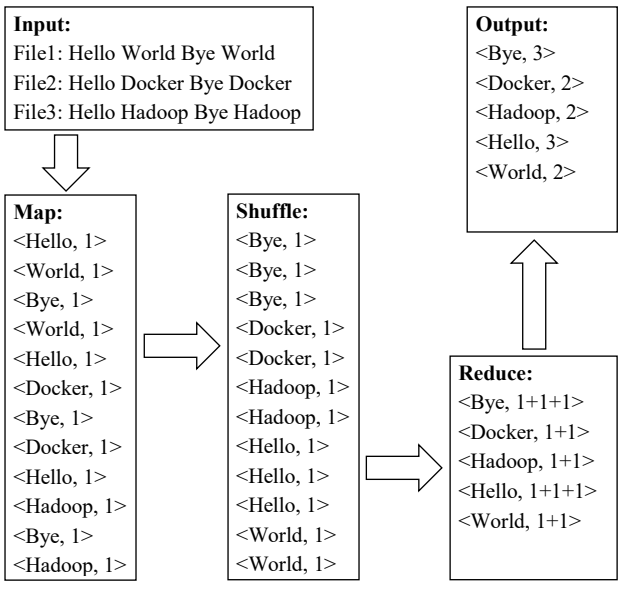


Figure 12: A Three-file Scenario of MapReduce-based WordCount.

Recall that the information of Amazon's spot price trace is composed of Tag, Price, Timestamp, Instance Type and Zone (cf. Table 6). By considering the value in each information field to be a letter, we treat every

single spot price record as an English word. For example, by random analogy, the spot price record "tag price1 time1 OS1 zone1" can be viewed as a six-letter word like "Hadoop", as highlighted in Figure 13.

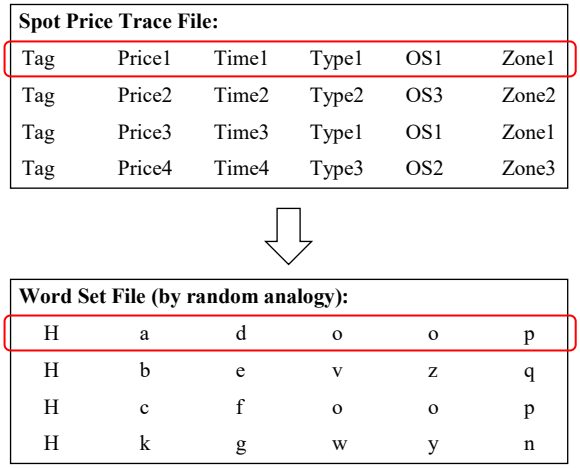


Figure 13: Treating spot price records as six-letter words.

As such, we are able to follow the logic of WordCount to fulfill the needs of cleaning our collected spot price history. In particular, instead of summing up the occurrences of the same price records, the duplicate records are simply ignored (removed). Moreover, the price records are sorted by the order of values of Instance Type, Operating System, Zone and Timestamp sequentially during the shuffling stage, while Tag and Price are not involved in data sorting. To avoid duplication, here we do not further elaborate the MapReduce-based data cleaning process.

4.3 Application Environment and Implementation

Unlike using "just-enough" environment for micro-level performance evaluation, here we employ "fair-enough" hardware resource to implement the MapReduce-based WordCount application, as listed in Table 7. In detail, the physical machine is Dell PowerEdge T110 II with the CPU model Intel Xeon E3-1200 series E3-1220 / 3.1 GHz. When preparing the MapReduce framework, we choose Apache Hadoop 2.7.0 [24] running in the operating system Ubuntu Server 16.04, and using Docker containers to construct a Hadoop cluster. In particular, Docker allows us to create an exclusive bridge network for the Hadoop cluster by using a specific name, e.g., `sudo docker network create --driver=bridge cluster`.

Environmental Item	Specification
Physical Machine	Dell PowerEdge T110 II
Operating System	Ubuntu Server 16.04
Java Environment	JDK 1.8.0
MapReduce Framework	Hadoop 2.7.0

Table 7: Summary of application environment.

In our initial implementation, we realized a three-node Hadoop cluster by packing Hadoop 2.7.0 into a Docker image and starting one master container and two slave ones. As illustrated in Figure 14, such a Hadoop cluster can logically be divided into a distributed processing layer (i.e. MapReduce workflow) and a distributed file-system layer (i.e. Hadoop Distributed File System (HDFS) in this case). When running a MapReduce application, the main interactions are:

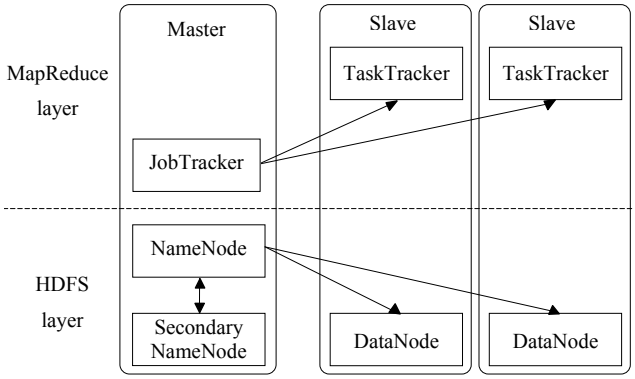


Figure 14: A three-node Hadoop cluster.

(1) A MapReduce job can run in a Hadoop cluster.

(2) The JobTracker in the cluster accepts a job from the MapReduce application, and locates relevant data through the NameNode.

(3) Suitable TaskTrackers are selected and then accept the tasks delivered by the JobTracker.

(4) The JobTracker communicates with the TaskTrackers and manages failures.

(5) When the collaboration among the TaskTrackers finish the job, the JobTracker updates its status and returns the result.

4.4 Performance Optimization Strategies

It is known that the performance of MapReduce applications can be tuned by adjusting the various parameters in the three configuration files of Hadoop. However, it is also clear that there is no one-size-fits-all approach to performance tuning. Therefore, we conducted a set of performance evaluation of our data cleaning application, in order to come up with a set of optimization strategies at least for this case.

- **Setting Timeout for Tasks.** A map or reduce task can be blocked or failed during runtime for various reasons, which would slow down the execution, and even result in the failure, of the whole MapReduce job. Therefore, by using TaskTracker to kill the blocked/failed tasks after a proper time span, those tasks will be able to be relaunched to save some waiting time. Given our relatively small size of data, we reduce the default timeout value from ten minutes to one minute, as shown below.

```
<!--Configuration in mapred-site.xml-->
<property>
  <name>mapred.task.timeout</name>
  <value>60000</value>
</property>
```

- **Turn on Out-of-Band Heartbeat.** Unlike regular heartbeats, the out-of-band heartbeat is triggered when a task is complete or failed. As such, JobTracker will be noticed the first time when there are free resources, so as to assign them to new tasks and eventually to save time. The configuration for turning on out-of-band heartbeat is specified as follows.

```
<!--Configuration in mapred-site.xml-->
<property>
  <name>mapreduce.tasktracker.
          outofband.heartbeat</name>
  <value>true</value>
</property>
```

- **Setting Buffer.** To begin with, we are concerned with a threshold percentage of buffer, and a background thread will be issued to spill buffer contents to hard disk when the threshold is reached. Inspired by the storage micro-benchmarking results (cf. Section 3.3.4), we decided to increase the threshold from 80% to 90% of buffer. As for the amount of memory to be buffer size, we double the default value (i.e. 100MB) for an intuitive test. These two parameters can be adjusted respectively as shown below.

```
<!--Configuration in mapred-site.xml-->
<property>
  <name>io.sort.spill.percent</name>
  <value>0.9</value>
</property>
<property>
  <name>io.sort.mb</name>
  <value>200</value>
</property>
```

- **Merging Spilled Streams.** As a continuation of spilling buffer contents to hard disk, the intermediate streams from multiple spill threads are merged into one single sorted file per partition which is to be fetched by reducers. Thus, we can control how many of spills will be merged into one file at a time. Since the smaller merge factor incurs more parallel merge activities and more disk IO for reducers, we decided to increase the merge value from 10 to 100, as shown below.

```
<!--Configuration in mapred-site.xml-->
<property>
  <name>io.sort.factor</name>
  <value>100</value>
</property>
```

- **LZO Compression.** Recall that the data we are dealing with are plain texts. The text data can generally be compressed significantly to reduce the usage of hard disk space and transmission bandwidth, and correspondingly to save the time taken for data copying/transferring. As a lossless algorithm with high decompression speed, Lempel-Ziv-Oberhumer (LZO) is one of the compression mechanisms supported by the Hadoop framework. In addition to various benefits and characteristics in common, LZO's block structure is particularly split-friendly for parallel processing in MapReduce jobs [25]. Therefore, we install and enable LZO for Hadoop by specifying the configurations below.

```
<!--Configuration in core-site.xml-->
<property>
  <name>io.compression.codecs</name>
  <value>org.apache.hadoop.io.compress.
          GzipCodec,org.apache.hadoop.
          io.compress.DefaultCodec,
       com.hadoop.compression.lzo.
          LzoCodec,com.hadoop.
          compression.lzo.LzopCodec,
       org.apache.hadoop.io.compress.
          BZip2Codec</value>
</property>
<property>
  <name>io.compression.codec.
                  lzo.class</name>
  <value>com.hadoop.compression.
              lzo.LzoCodec</value>
</property>

<!--Configuration in mapred-site.xml-->
<property>
  <name>mapreduce.map.output.
                  compress</name>
  <value>true</value>
</property>
<property>
  <name>mapreduce.map.output.
              compress.codec</name>
  <value>com.hadoop.compression.
               lzo.LzoCodec</value>
</property>
```

- **Doubling Slave Nodes.** In distributed computing, a common scenario is to employ more resources to deal with more workloads. Recall that map and reduce tasks of a MapReduce job are distributed to the slave nodes in a Hadoop cluster, and the physical machine used in this study has a Quad-Core processor. To obtain some quick clues at this current stage, we try to improve our application's performance by doubling the slave nodes, i.e. extending the original three-node cluster (with two slave nodes) into a five-node one (with four slave nodes). Note that, in this optimization strategy, we keep the other configurations settings by default.

4.5 Performance Evaluation Results

Due to the time limit, we follow "one factor at a time" to perform evaluation of the aforementioned optimization strategies. Furthermore, to confirm the effects of

these optimization strategies, we are concerned with 2GB+ and 5GB+ data respectively in the performance evaluation. Correspondingly, we draw the evaluation results in Figure 15 and 16 respectively.

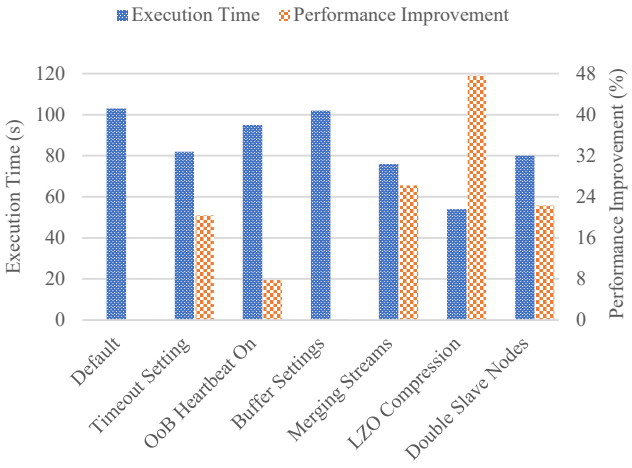


Figure 15: Performance optimization for 2GB+ data cleaning.

The results show three relatively different patterns of optimization effects. First, LZO compression acts as the most effective optimization strategy, and it can improve the performance nearly by 50% when dealing with 2GB+ data. Second, we can expect similar and moderate performance improvement by three individual strategies, such as merging more spilled streams, reducing the timeout value, and doubling the slave nodes. Third, out-of-band heartbeat and buffer settings seem not to be influential optimization strategies in this case.

These different optimization patterns could be closely related to the data characteristics of our application. On one hand, since text data can be compressed significantly [26], our application mostly benefits from the optimization strategy of LZO compression. On the other hand, since the current price traces used in this study are still far from "big data", the buffer setting could not take clear effects until the data size reaches TB levels.

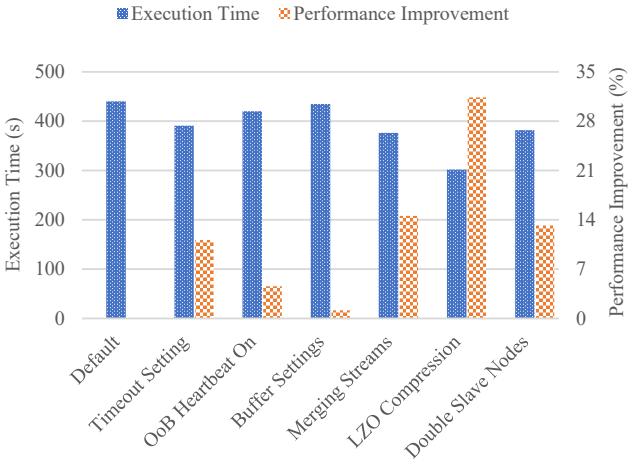


Figure 16: Performance optimization for 5GB+ data cleaning.

5 Related Work

Although the performance advantage of containers were investigated in several pioneer studies [3, 27, 10], the container-based virtualization solution did not gain significant popularity until the recent underlying improvements in the Linux kernel, and especially until the emergence of Docker [28, 29]. Starting from an open-source project in early 2013 [7], Docker quickly becomes the most popular container solution [2] by significantly facilitating the management of containers. Technically, through offering the unified tool set and API, Docker relieves the complexity of utilizing the relevant kernel-level techniques including the LXC, the cgroup and a copy-on-write filesystem. To examine the performance of Docker containers, a molecular modeling simulation software [30] and a postgreSQL database-based Joomla application [31] have been used to benchmark the Docker environment against the VM environment.

Considering the uncertainty of use cases (e.g., different workload densities and QoS requirements), at this current stage, a baseline-level investigation would be more useful and helpful for understanding the fundamental difference in performance overhead between those two virtualization solutions. A preliminary study has particularly focused on the CPU consumption by using the 100000! calculation within a Docker container and a KVM VM respectively [32]. Nevertheless, the concerns about other features/resources like memory and disk are missing. Similarly, the performance analysis between VM and container in study [33] is not feature-specific enough (even including security that is out of the scope of performance). On the contrary, by treating Docker containers as a particular type of Cloud service, our study considers the four physical properties of a Cloud service [15] and essentially gives a fundamental investigation into the Docker container's performance overhead on a feature-by-feature basis.

The closest work to ours is the IBM research report on the performance comparison of VM and Linux containers [34]. In fact, it is this incomplete report (e.g., the container's network evaluation is partially missing) that inspires our study. Surprisingly, our work denies the IBM report's finding "containers and VMs impose almost no overhead on CPU and memory usage" that was also claimed in [35], and we also doubt about "Docker equals or exceeds KVM performance in every case". In particular, we are more concerned with the overhead in performance variability.

Within the context of MapReduce clusters, Xavier et al. [10] conducted experimental comparisons among the three aforementioned types of container-based virtual environments, while a set of other studies particularly contrasted performance of OpenVZ with the hypervisor-based virtualization implementations including VMWare, Xen and KVM [3, 27]. The significant difference between these studies and ours is that we focus more on the performance optimization of a

container-based MapReduce cluster in a specific application scenario.

Note that, although there are also performance studies on deploying containers inside VMs (e.g., [36, 37]), such a redundant structure might not be suitable for an "apple-to-apple" comparison between Docker containers and VMs, and thus we do not include this virtualization scenario in our study.

6 Conclusions and Future Work

It has been identified that virtualization is one of the foundational elements of Cloud computing and helps realize the value of Cloud computing [38]. On the other hand, the technologies for virtualizing Cloud infrastructures are not resource-free, and their performance overheads would incur negative impacts on the QoS of the Cloud. Since hypervisors that currently dominate the Cloud virtualization market are a relatively heavyweight solution, there comes a rising trend of interest in its lightweight alternative [7], namely the container-based virtualization. Their mechanism difference is that, the former manages the host hardware resources, while the latter enables sharing the host OS. Although straightforward comparisons can be done from the existing qualitative discussions, we conducted a fundamental evaluation study to quantitatively understand the performance overheads of these two different virtualization solutions. In particular, we employed a standalone Docker container and a VMWare Workstation VM to represent the container-based and the hypervisor-based virtualization technologies respectively.

Recall that there are generally two stages of performance engineering in ECS, for revealing the primary performance of specific (system) features and investigating the overall performance of real-world applications respectively. In addition to the fundamental performance of a single container, we also studied performance optimization of a container-based MapReduce application in terms of cleaning Amazon's spot price history. At this current stage, we only focused on one factor at a time to evaluate the optimization strategies ranging from setting task timeout to doubling slave nodes.

Overall, our work reveals that the performance overheads of these two virtualization technologies could vary not only on a feature-by-feature basis but also on a job-to-job basis. Although the container-based solution is undoubtedly lightweight, the hypervisor-based technology does not come with higher performance overhead in every case. At the application level, the container technology is clearly more resource-friendly, as we failed in building VM-based MapReduce clusters on the same physical machine. When it comes to container-based MapReduce applications, it seems that the effects of performance optimization strategies are closely related to the data characteristics. For dealing with text data in our case study, LZO compression can bring the most significant performance improvement.

Due to the time and resource limit, our current investigation into the performance of container-based MapReduce applications is still an early study. Thus, our future work will be unfolded along two directions. Firstly, we will adopt sophisticated experimental design techniques (e.g., the full-factorial design) [39] to finalize the same case study on tuning the MapReduce performance of cleaning Amazon's price history. Secondly, we will gradually apply Docker containers to different real-world applications for dealing with different types of data. By employing "more-than-enough" computing resource, the application-oriented practices will also be replicated in the hypervisor-based virtual environment for further comparison case studies.

Acknowledgment This work is supported by the Swedish Research Council (VR) under contract number C0590801 for the project Cloud Control, and through the LCCC Linnaeus and ELLIIT Excellence Centers. This work is also supported by the National Natural Science Foundation of China (Grant No.61572251).

References

[1] Z. Li, M. Kihl, Q. Lu, and J. A. Andersson, "Performance overhead comparison between hypervisor and container based virtualization," in *Proceedings of the 31st IEEE International Conference on Advanced Information Networking and Application (AINA 2017)*. Taipei, Taiwan: IEEE Computer Society, 27-29 March 2017, pp. 955–962, https://doi.org/10.1109/AINA.2017.79.

[2] C. Pahl, "Containerization and the PaaS Cloud," *IEEE Cloud Computing*, vol. 2, no. 3, pp. 24–31, May/June 2014, https://doi.org/10.1109/MCC.2015.51.

[3] J. P. Walters, V. Chaudhary, M. Cha, S. G. Jr., and S. Gallo, "A comparison of virtualization technologies for HPC," in *Proceedings of the 22nd International Conference on Advanced Information Networking and Applications (AINA 2008)*. Okinawa, Japan: IEEE Computer Society, 25-28 March 2008, pp. 861–868, https://doi.org/10.1109/AINA.2008.45.

[4] P. J. Denning, "Performance evaluation: Experimental computer science at its best," *ACM SIGMETRICS Performance Evaluation Review*, vol. 10, no. 3, pp. 106–109, Fall 1981, https://doi.org/10.1145/1010629.805480.

[5] D. G. Feitelson, "Experimental computer science," *Communications of the ACM*, vol. 50, no. 11, pp. 24–26, November 2007, https://doi.org/10.1145/1297797.1297817.

[6] Z. Li, L. O'Brien, and M. Kihl, "DoKnowMe: Towards a domain knowledge-driven methodology for performance evaluation," *ACM SIGMETRICS Performance Evaluation Review*, vol. 43, no. 4, pp. 23–32, 2016, https://doi.org/10.1145/2897356.2897360.

[7] D. Merkel, "Docker: Lightweight Linux containers for consistent development and deployment," *Linux Journal*, vol. 239, pp. 76–91, March 2014.

[8] X. Xu, H. Yu, and X. Pei, "A novel resource scheduling approach in container based clouds," in *Proceedings of the 17th IEEE International Conference on Computational Science and Engineering (CSE 2014)*. Chengdu, China: IEEE Computer Society, 19-21 December 2014, pp. 257–264, https://doi.org/10.1109/CSE.2014.77.

[9] D. Bernstein, "Containers and Cloud: From LXC to Docker to KuBernetes," *IEEE Cloud Computing*, vol. 1, no. 3, pp. 81–84, September 2014, https://doi.org/10.1109/MCC.2014.51.

[10] M. G. Xavier, M. V. Neves, and C. A. F. D. Rose, "A performance comparison of container-based virtualization systems for MapReduce clusters," in *Proceedings of the 22nd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP 2014)*. Turin, Italy: IEEE Press, 12-14 February 2014, pp. 299–306, https://doi.org/10.1109/PDP.2014.78.

[11] C. Anderson, "Docker," *IEEE Software*, vol. 32, no. 3, pp. 102–105, May/June 2015, https://doi.org/10.1109/MS.2015.62.

[12] T. Banerjee, "Understanding the key differences between LXC and Docker," https://www.flockport.com/lxc-vs-docker/, August 2014.

[13] S. M. Blackburn, K. S. McKinley, R. Garner, C. Hoffmann, A. M. Khan, R. Bentzur, A. Diwan, D. Feinberg, D. Frampton, S. Z. G. M. H. A. H. M. J. H. Lee, J. E. B. Moss, A. Phansalkar, D. Stefanovik, T. VanDrunen, D. von Dincklage, and B. Wiedermann, "Wake up and smell the coffee: Evaluation methodology for the 21st century," *Communications of the ACM*, vol. 51, no. 8, pp. 83–89, August 2008, https://doi.org/10.1145/1378704.1378723.

[14] A. Iosup, N. Yigitbasi, and D. Epema, "On the performance variability of production Cloud services," in *Proceedings of the 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2011)*. Newport Beach, CA, USA: IEEE Computer Society, 23-26 May 2011, pp. 104–113, https://doi.org/10.1109/CCGrid.2011.22.

[15] Z. Li, L. O'Brien, R. Cai, and H. Zhang, "Towards a taxonomy of performance evaluation of commercial Cloud services," in *Proceedings of the 5th International Conference on Cloud Computing (IEEE CLOUD 2012)*. Honolulu, Hawaii, USA: IEEE Computer Society, 24-29 June 2012, pp. 344–351, https://doi.org/10.1109/CLOUD.2012.74.

[16] Z. Li, L. O'Brien, H. Zhang, and R. Cai, "A factor framework for experimental design for performance evaluation of commercial Cloud services," in *Proceedings of the 4th International Conference on Cloud Computing Technology and Science (CloudCom 2012)*. Taipei, Taiwan: IEEE Computer Society, 3-6 December 2012, pp. 169–176, https://doi.org/10.1109/CloudCom.2012.6427525.

[17] StackOverflow, "What is the relationship between the docker host OS and the container base image OS?" http://stackoverflow.com/questions/18786209/what-is-the-relationship-between-the-docker-host-os-and-the-container-base-image, September 2013.

[18] Reddit, "Do I need to use an OS base image in my Dockerfile or will it default to the host OS?" https://www.reddit.com/r/docker/comments/2teskf/do_i_need_to_use_an_os_base_image_in_my/, January 2015.

[19] Z. Li, L. O'Brien, H. Zhang, and R. Cai, "On the conceptualization of performance evaluation of IaaS services," *IEEE Transactions on Services Computing*, vol. 7, no. 4, pp. 628–641, October-December 2014, https://doi.org/10.1109/TSC.2013.39.

[20] C. Weinhardt, A. Anandasivam, B. Blau, N. Borissov, T. Meinl, W. Michalk, and J. Stößer, "Cloud computing a classification, business models, and research directions," *Business and Information Systems Engineering*, vol. 1, no. 5, pp. 391–399, October 2009, https://doi.org/10.1007/s12599-009-0071-2.

[21] V. Abhishek, I. A. Kash, and P. Key, "Fixed and market pricing for Cloud services," in *Proceedings of the 7th Workshop on the Economics of Networks, Systems, and Computation (NetEcon 2012)*. Orlando, FL, USA: IEEE Computer Society, 20 March 2012, pp. 157–162, https://doi.org/10.1109/INFCOMW.2012.6193479.

[22] Z. Li, H. Zhang, L. O'Brien, S. Jiang, Y. Zhou, M. Kihl, and R. Ranjan, "Spot pricing in the cloud ecosystem: A comparative investigation," *Journal of Systems and Software*, vol. 114, pp. 1–19, April 2016, https://doi.org/10.1016/j.jss.2015.10.042.

[23] Q. Lu, Z. Li, M. Kihl, L. Zhu, and W. Zhang, "CF4BDA: A conceptual framework for big data analytics applications in the cloud," *IEEE Access*, vol. 3, pp. 1944–1952, October 2015, https://doi.org/10.1109/ACCESS.2015.2490085.

[24] Apache Software Foundation, "Mapreduce tutorial," https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html, October 2017.

[25] A. Viswanathan, "A guide to using LZO compression in Hadoop," *Linux Journal*, vol. 2012, no. 220, August 2012, article no. 1.

[26] WinZip, "Varying file compression explored," http://kb.winzip.com/kb/entry/326/, September 2017.

[27] J. Che, C. Shi, Y. Yu, and W. Lin, "A synthetical performance evaluation of OpenVZ, Xen and KVM," in *Proceedings of the 2010 IEEE Asia-Pacific Services Computing Conference (APSCC 2010)*. Hangzhou, China: IEEE Computer Society, 6-10 December 2010, pp. 587–594, https://doi.org/10.1109/APSCC.2010.83.

[28] D. Strauss, "Containers - not virtual machines - are the future Cloud," *Linux Journal*, vol. 228, pp. 118–123, April 2013.

[29] A. Karle, "Operating system containers vs. application containers," https://blog.risingstack.com/operating-system-containers-vs-application-containers/, May 2015.

[30] T. Adufu, J. Choi, and Y. Kim, "Is container-based technology a winner for high performance scientific applications?" in *Proceedings of the 17th Asia-Pacific Network Operations and Management Symposium (APNOMS 2015)*. Busan, Korea: IEEE Press, 19-21 August 2015, pp. 507–510, https://doi.org/10.1109/APNOMS.2015.7275379.

[31] A. M. Joy, "Performance comparison between Linux containers and virtual machines," in *Proceedings of the 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA 2015)*. Ghaziabad, India: IEEE Press, 14-15 February 2015, pp. 507–510, https://doi.org/10.1109/ICACEA.2015.7164727.

[32] K.-T. Seo, H.-S. Hwang, I.-Y. Moon, O.-Y. Kwon, and B.-J. Kim, "Performance comparison analysis of Linux container and virtual machine for building Cloud," *Advanced Science and Technology Letters*, vol. 66, pp. 105–111, December 2014, https://doi.org/10.14257/astl.2014.66.25.

[33] R. K. Barik, R. K. Lenka, K. R. Rao, and D. Ghose, "Performance analysis of virtual machines and containers in cloud computing," in *Proceedings of the 2016 International Conference on Computing, Communication and Automation (ICCCA 2016)*. Greater Noida, India: IEEE Press, 29-30 April 2016, pp. 1204–1210, https://10.1109/CCAA.2016.7813925.

[34] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio, "An updated performance comparison of virtual machines and Linux containers," in *Proceedings of the 2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2015)*. Philadelphia, PA, USA: IEEE Press, 29-31 March 2015, pp. 171–172, https://doi.org/10.1109/ISPASS.2015.7095802.

[35] Z. Kozhirbayev and R. O. Sinnott, "A performance comparison of container-based technologies for the Cloud," *Future Generation Computer Systems*, no. 68, pp. 175–182, March 2017, https://doi.org/10.1016/j.future.2016.08.025.

[36] R. Dua, A. R. Raja, and D. Kakadia, "Virtualization vs containerization to support PaaS," in *Proceedings of the 2014 IEEE International Conference on Cloud Engineering (IC2E 2015)*. Boston, Massachusetts, USA: IEEE Computer Society, 10-14 March 2014, pp. 610–614, https://doi.org/10.1109/IC2E.2014.41.

[37] S. F. Piraghaj, A. V. Dastjerdi, R. N. Calheiros, and R. Buyya, "Efficient virtual machine sizing for hosting containers as a service," in *Proceedings of the 11th World Congress on Services (SERVICES 2015)*. New York, USA: IEEE Computer Society, 27 June - 2 July 2015, pp. 31–38, https://doi.org/10.1109/SERVICES.2015.14.

[38] S. Angeles, "Virtualization vs. Cloud computing: What's the difference?" http://www.businessnewsdaily.com/5791-virtualization-vs-cloud-computing.html, January 2014.

[39] D. C. Montgomery, *Design and Analysis of Experiments*, 8th ed. Hoboken, NJ: John Wiley & Sons, Inc., April 2012.