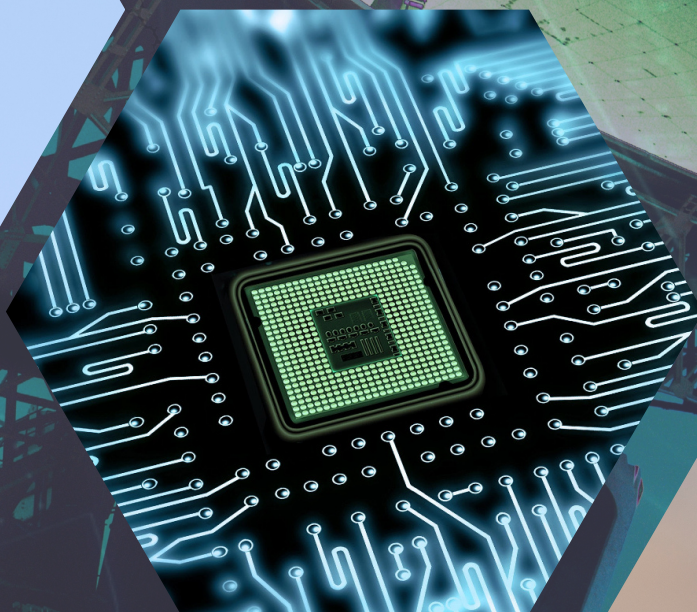


Advances in Science, Technology & Engineering Systems Journal



VOLUME 4-ISSUE 1 | JAN-FEB 2019

EDITORIAL BOARD

Editor-in-Chief

Prof. Passerini Kazmerski
University of Chicago, USA

Editorial Board Members

Prof. Rehan Ullah Khan
Qassim University, Saudi
Arabia

Prof. María Jesús Espinosa
Universidad Tecnológica
Metropolitana, Mexico

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

Dr. Nguyen Tung Linh
Electric Power University,
Vietnam

Dr. Omeje Maxwell
Covenant University, Nigeria

Gussan Maaz Mufti
Bahria University Islamabad, Pakistan

Mohamed Mohamed Abdel-Daim
Suez Canal University, Egypt

Regional Editors

Dr. Hung-Wei Wu
Kun Shan University, Taiwan

Dr. Maryam Asghari
Shahid Ashrafi Esfahani, Iran

Dr. Shakir Ali
Aligarh Muslim University, India

Dr. Ahmet Kayabasi
Karamanoglu Mehmetbey
University, Turkey

Dr. Ebubekir Altuntas
Gaziosmanpasa University,
Turkey

Dr. Sabry Ali Abdallah El-Naggar
Tanta University, Egypt

Mr. Aamir Nawaz
Gomal University, Pakistan

Dr. Gomathi Periasamy
Mekelle University, Ethiopia

Dr. Walid Wafik Mohamed Badawy
National Organization for Drug Control
and Research, Egypt

Aamir Nawaz
Gomal University, Pakistan

Abdullah El-Bayoumi
Cairo University, Egypt

Ayham Hassan Abazid
Jordan university of science and
technology, Jordan

Editorial

Advances in Science, Technology and Engineering Systems Journal (ASTESJ) is an online-only journal dedicated to publishing significant advances covering all aspects of technology relevant to the physical science and engineering communities. The journal regularly publishes articles covering specific topics of interest.

Current Issue features key papers related to multidisciplinary domains involving complex system stemming from numerous disciplines; this is exactly how this journal differs from other interdisciplinary and multidisciplinary engineering journals. This issue contains 32 accepted papers in Computer Science and Telecom domains.

Editor-in-chief
Prof. Passerini Kazmersk

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Volume 4 Issue 1

January-February 2019

CONTENTS

<i>Automation System for Regulation Optimization in Power Transformer Design</i> Tatjana Šimović , Mislav Gazdović	01
<i>Spectral Grid Impedance Identification on the Low-, Medium- and High-Voltage Level – System Design, Realization and Measurement Results of Grid Impedance Measurement Devices</i> Hauke Wilken, Michael Jordan, Detlef Schulz	08
<i>A Wearable Exoskeleton Rehabilitation Device for Paralysis – A Comprehensive Study</i> Ahmed Roshdy, Samer Al Kork, Sherif Said, Taha. Beyrouthy	17
<i>A Practical PIR-based Scheme for Discovering Nearby Places for Smartphone Applications</i> Maryam Hezaveh, Carlisle Adams	27
<i>A Proposed Architecture for Parallel HPC-based Resource Management System for Big Data Applications</i> Waleed Al Shehri, Maher Khemakhem, Abdullah Basuhail, Fathy E. Eassa	40
<i>Morphological and Optoelectrical Characterization of Silicon Nanostructures for Photovoltaic Applications</i> Babacar Dieng, Moussa Toure, Modou Beye, Diouma Kobor, Amadou Seidou Maiga	45
<i>Extending the Life of Legacy Robots: MDS-Ach via x-Ach</i> Daniel M. Lofaro, Magdalena Bugajska, Donald Sofge	50
<i>State Estimation based Echolocation Bionics and Image Processing based Target Pattern Recognition</i> David Kondru, Mehmet Celenk, Xiaoping A. Shen	73
<i>Web Authentication: no Password; Listen and Touch</i> Viorel LUPU	84
<i>Hypervolume-Based Multi-Objective Reinforcement Learning: Interactive Approach</i> Hiroyuki Yamamoto, Tomohiro Hayashida, Ichiro Nishizaki, Shinya Sekizaki	93
<i>Optimal Designs of Constrained Accelerated Life Testing Experiments for Proportional Hazards Models</i> Xiaojian Xu, Wanyi Huang	101

<i>Fuzzy Logic Implementation for Enhanced WCDMA Network Using Selected KPIs</i>	114
Nosiri Onyebuchi Chikezie, Onyenwe Ezinne Maureen, Ekwueme Emmanuel Uchenna	
<i>Conducted and Radiated Interference on Interconnection's Lines</i>	125
Patricio E. Munhoz-Rojas	
<i>Can parallelization save the (computing) world?</i>	141
János Végh, József Vásárhelyi, Dániel Drótos	
<i>Low Contrast Image Enhancement Using Convolutional Neural Network with Simple Reflection Model</i>	159
Bok Gyu Han, Hyeon Seok Yang, Ho Gyeong Lee, Young Shik Moon	
<i>Photodecoloration of Methyl Orange Solution Assisted by ZrS₃ Powders</i>	165
Sofya Artemkina, Anastassiia Poltarak, Pavel Poltarak, Igor Asanov, Vladimir Fedorov	
<i>A Novel Pulse Position Modulator for Compressive Data Acquisition</i>	171
Constantine A. Pappas	
<i>Enhancing and Monitoring Patient Outcomes Through Customized Learning</i>	183
Majed Almotairi, Mohammed Abdulkareem Alyami, Yeong-Tae Song	
<i>Robot-Assisted Posture Emulation for Visually Impaired Children</i>	193
Fang-Lin Chao, Hung-Chi Chu, Liza Lee	
<i>Development of Application Specific Electronic Nose for Monitoring the Atmospheric Hazards in Confined Space</i>	200
Muhammad Aizat Bin Abu Bakar, Abu Hassan Bin Abdullah, Fathinul Syahir Bin Ahmad Sa'ad	
<i>Trajectory Tracking Control Optimization with Neural Network for Autonomous Vehicles</i>	217
Samuel Oludare Bamgbose, Xiangfang Li, Lijun Qian	
<i>PAPR and BER Performances of OFDM System with Novel Tone Reservation Technique Over Frequency Non-Selective Fading Channel</i>	225
Moftah Ali, Raveendra K. Rao, Vijay Parsa	
<i>Critical Embedded Systems Development Using Formal Methods and Statistical Reliability Metrics</i>	231
Jonathan Lockhart, Carla Purdy, Philip Wilsey	
<i>Observing and Forecasting the Trajectory of the Thrown Body with use of Genetic Programming</i>	248
Konstantin Mironov, Ruslan Gayanov, Dmiriy Kurenov	

<i>Performance Investigation of Semiconductor Devices using Commutation-speed based methodology for the application of Boost Power Factor Correction</i>	258
Barkha Parkash, Ajay Poonjal Pai, Wei Tian, Ralph Kennel	
<i>Strategies of the Level-By-Level Approach to the Minimal Route</i>	268
Nikolay Starostin, Konstantin Mironov	
<i>Over-The-Air Testing of Automotive Antennas and Wireless Links in The Installed State on The Basis of LTE Downlink Communication Parameters</i>	282
Philipp Bertl, Lisa Jäger, Andreas Schwind, Frank Wollenschläger, Christian Bornkessel, Matthias A. Hein	
<i>Insight into the IEEE 802.1 Qcr Asynchronous Traffic Shaping in Time Sensitive Network</i>	292
Zifan Zhou, Michael Stübert Berger, Sarah Renée Ruepp, Ying Yan	
<i>A Study on the Efficiency of Hybrid Models in Forecasting Precipitations and Water Inflow Albania Case Study</i>	302
Eralda Gjika, Aurora Ferrja, Arbesa Kamberi	
<i>Talk Show's Business Intelligence on Television by Using Social Media Data in Indonesia</i>	311
Eris Riso, Abba Suganda Girsang	
<i>Application Layer Security Authentication Protocols for the Internet of Things: A Survey</i>	317
Shruthi Narayanaswamy, Anitha Vijaya Kumar	
<i>Managing and Optimizing Quality of Service in 5G Environments Across the Complete SLA Lifecycle</i>	329
Evgenia Kapassa, Marios Touloupou, Panagiotis Stavrianos, Georgios Xylouris, Dimosthenis Kyriazis	

Automation System for Regulation Optimization in Power Transformer Design

Tatjana Šimović^{*1}, Mislav Gazdović²

¹Končar Power Transformers Ltd., A joint venture of Siemens and Končar /Zagreb 10000, Croatia

²Montelektro d.o.o./Kastav 51215, Croatia

ARTICLE INFO

Article history:

Received: 29 November, 2018

Accepted: 17 December, 2018

Online: 11 January, 2019

Keywords:

Power transformer regulation

Tap changer selection

Tap changer database

ABSTRACT

Large power transformers generally include a customer request for a technically appropriate regulation unit. The selection process of the regulation unit consists of defining the required input data, performing mathematical calculations necessary to find the technical limit values that the regulation unit has to satisfy, and finally optimizing and selecting the appropriate regulation unit. The number of possible permutations consists of thousands of different combinations, depending on the type of regulation and other technical limitations. The automation system presented in this paper significantly reduces the time required to obtain the optimal regulation unit solution in both offer and order stages of the project, providing significant overall productivity increase. This paper presents an example of managing a part of a complex system such as power transformer design using a software solution. The process of finding the optimal regulation unit „manually” can take up to several hours. Implementing the developed algorithm and introducing Tap Changer Selection application, the required time is reduced to several minutes. This represents significant time savings and reduced possibility of errors, thus improving the power transformer design process.

1. Introduction

This paper is an extension of work originally presented in the conference Proceedings of the 41st International Convention for Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018) [1].

Large power systems such as electrical grids of entire cities, countries or continents are systems which must be continuously regulated in order to keep the voltage supply constant, thereby ensuring all the electrical devices and equipment to work in the expected voltage range. In such large systems, disturbances in the load and/or supply are inevitable – i.e. under greater load, the current increases, and consequently the voltage drops along the transmission lines which connect the power supply to the power consumption. It is important to emphasize that there are two main types of voltage deviations. The first is caused by a change of load on the lower voltage side of the power transformer. Since load through the power transformer has changed, the change of the position of the regulation unit is required in order to maintain the voltage at a constant level. The second voltage disorder is the case when voltage is changed on the higher voltage side of the power

transformer as a consequence of a certain kind of network disturbances [2]. Handling such disturbances and maintaining a constant voltage supply is (among others) the responsibility of the power transformer as one of the most expensive parts of the power grid. For this purpose, the regulating power transformer is equipped with a regulating winding which is connected to the regulation unit - as shown in Figure 1 - usually On Load Tap Changer (OLTC).

The OLTC principle was patented in 1927 by Dr. Bernhard Jansen. Together with the industrial development and economic growth over the years, the overall power consumption took an upward trend, following the expansion of the electrical grid. The development of OLTCs was accelerated over the years due to the steady increase of the transmission voltage and power. The voltage regulation basic principle is adding or subtracting turns from either the primary or the secondary winding. The OLTC alters the power transformer turns ratio in a number of predefined steps and in that way changes the secondary (or primary) side voltage [3].

As opposed to the OLTC, the DETC (De Energized Tap Changer) must not be operated while the transformer is energized.

*Corresponding Author: Tatjana Šimović, Zagreb 10000, Croatia, +385989889304, Email: tatjana.simovic@siemens.com

This means that for the DETC to operate, the transformer must be disconnected from the network for a short period of time. Due to the mentioned limitations, the DETC is used only when the regulation capabilities are not a primary concern. This is why an OLTC regulation unit is more and more becoming a standard customer request for power transformer regulation.



Figure 1: Power transformer with a regulation unit

Generally, many different regulation units are technically appropriate and may be used for a particular transformer design. In order to minimize the overall costs while still managing to provide the required regulation capabilities, one of the transformer manufacturers' challenges is to optimize the regulation unit selection process. Since the regulation unit (i.e. tap changer) can make around 10 – 20% of the total material cost of the transformer, the overall effect of optimizing the regulation unit becomes significant.

Related work on this area consists of different power transformer design tools which (at most) consist of a regulation unit database, expecting the user to choose the appropriate regulation unit from the list (manually). A tool example presenting automated regulation unit selection functionality in the power transformer industry is ABB Compas tool [4]. Since this tool is limited only to ABB types of tap changers, the main motivation for this work was to provide a faster and more efficient way of selecting the optimal regulation unit also for manufacturers other than ABB, e.g. MR (Germany) and C.A.P.T. (Italy) for OLTC and DETC regulation units since this is a common customer request. The main advantage of this application becomes evident in the offer stage where quick and correct selection of the optimal regulation unit is important for the final transformer design cost estimation.

2. Transformer Regulation

The transformer is an electrical static device without moving parts, used to transform electrical power from one circuit to another without electrical connection and without changing frequency. Transformer switches the alternating current of the predetermined electrical voltage into alternating current of a higher or lower electrical voltage using the effect of mutual induction. The only moving part in a power transformer, the On Load Tap Changer, is one of the main contributors to the failure rates of high voltage power transformers [5]. Taking this into consideration, it

makes sense to standardize and automate the regulation unit selection process, which was the main motivation for this work.

The introduction of OLTCs improved the operating efficiency of electrical systems considerably and this technique found acceptance worldwide. In general the percentage of transformers equipped with OLTCs is increasing with the increase of the load density and interconnection of electrical networks. In addition, OLTCs applied in industrial process transformers as regulating units in the chemical and metallurgical industry is another important field of application [6].

Basic connection diagrams for regulation are typically chosen with regard to system conditions and size/weight limits during transportation.

Linear regulation (Figure 2) is applied for a simple design transformer and regulation unit, with regulating range up to 20 percent of nominal voltage. The taps are added or subtracted in the series with the main winding. This kind of regulation ensures smallest ohmic losses.



Figure 2: Basic connection diagram linear regulation

For linear regulation, voltage across tap winding U_{TV} is equal phase to phase voltage across regulation U_R .

$$U_{TV} = U_R \quad (1)$$

Most common type of regulation is a reversing change-over selector (plus/minus switching), which allows doubled tapping range. Regulated winding is connected with the basic winding in series (Figure 3). The total number of taps decreases or increases, depending on position of tap selector. Boost and buck connection (boost - vectorial addition to main winding and buck - vectorial subtraction to main winding) enables to increase the regulating range or to reduce the number of tapped winding.

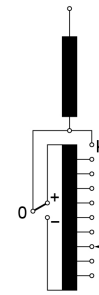


Figure 3: Basic connection diagram with reversing switch

Regulation with reversing change-over selector, for constant induction, depends on regulating voltage and number of step of the tap winding:

$$U_{TV} = \frac{U_R}{2} \cdot \frac{m}{m-1} \quad (2)$$

Coarse change over selector or coarse/fine switching requires more complicated winding layout – it is regulation with two windings, regulated winding and coarse winding insertion (Figure 4). Arrangement of coarse/fine selector has the electrical length of the fine tap winding plus one step. This kind of regulation offers the lower copper losses in the tap position with the minimum number of effective turns. Coarse/fine regulation offers for some industrial applications possibility of large number of operations using special design with up to 5 coarse taps (107 operations i.e. operating positions) [7].

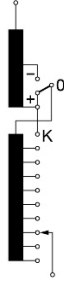


Figure 4: Basic connection diagram with coarse tap selector

Regulation with coarse fine change over selector across tap winding:

$$U_{TV} = \frac{U_R}{2} \quad (3)$$

Voltage across coarse tap winding $U_{CV}=U_{TV}$

where:

U_{TV} is voltage across tap winding in kV;

U_R is regulating range phase to phase in kV;

U_{CV} is voltage across coarse winding in kV;

m is number of steps of tap winding

3. Manual Selection and Calculations

The selection of appropriate OLTC should be made very carefully, because the OLTC is an important factor with respect to the transformer's reliability and cost. Each type of regulation unit is available in a number of variants, with different values of maximum rated through-current, number of phases, highest insulation level, tap selector size and basic connection diagrams. Therefore, the type designation of a regulation unit depends on these features. A guideline for the OLTC selection is given in IEC 60542, but some special applications as HVDC transformers or phase shifters are not described in detail.

When selecting the appropriate type of OLTC or DETC manually, a number of parameters and technical limits have to be considered.

First of all, the rated through current of the regulation unit must not be less than the transformer highest current value of the regulating winding. For regulation at the neutral end with constant induction (constant flux voltage variation, CFVV), the rated through current in case of a three-phase full transformer is derived

using the rated power as the product of current and voltage (one or three-phase).

For a three-phase autotransformer with regulation in neutral end, variable induction (variable flux voltage variation, VFVV), $S_n = \text{const.}$, the rated through current depends on both regulated and non-regulated voltage levels:

$$I_{max} = \frac{S_n \times 1000}{\sqrt{3}} \times \left(\frac{1}{U_{2min}} - \frac{1}{U_1} \right) \quad (4)$$

where:

U_1 is the rated line voltage of the non-regulated side in kV

U_{2min} is the rated line voltage in the minimum tap position in kV

In case of **one-phase transformer**, the rated through current is multiplied by a factor of $\sqrt{3}$ since the rated power is per phase and the rated voltages are phase voltages.

Maximum rated through current is the current which the regulation unit is transferring from one tap to the other at relevant step voltage.

Step voltage U_{step} [V] is the phase voltage between the taps which depends on the required regulated voltage range and the required number of positions. This value is constant in case of CFVV (constant flux voltage variation), unlike in the case of VFVV (variable flux voltage variation) where different regulating positions have different volt/turn ratio and therefore different step voltage values:

$$U_{step} = \begin{cases} \frac{U_{1max} - U_{1min}}{\sqrt{3}(n-1)} \times 1000 & ; \text{CFVV regulation} \\ \max(U_k - U_{k-1}) & ; \text{VFVV regulation} \end{cases} \quad (5)$$

where:

U_{1max} and U_{1min} are the required maximum and minimum regulated line voltage in [kV] (equivalent three-phase bank line voltage in case of a single-phase transformer);

n is the required number of regulating positions (including the rated position);

$U_k - U_{k-1}$ is the (phase) voltage difference in [V] in the regulating position k and $k-1$ (for positions $k = 2$ to n).

The required regulation unit external insulation level depends on the insulation level to ground of the side where the regulation unit is connected. The withstand voltages of the external insulation are standardized by international standards and correspond to the highest voltage for equipment [8]. In case of regulation in neutral end (common for full transformer or VFVV autotransformer), the insulation level is usually the same as the insulation level of the transformer neutral (unless otherwise specified). In case of regulation in line (common for CFVV autotransformers), the insulation level is the same as the lower voltage side to which the regulation unit is connected. The insulation level has to be checked against the applicable technical guide.

The internal insulation level affects the regulation unit tap selector size, whose price can vary significantly, meaning that the proper selection is of the essence in choosing the optimal regulation unit. Highest operating voltage across the regulating winding is the value that has to be compared with the highest

permissible phase service voltage across the regulating winding described in the technical guide, which defines several internal insulation distances to be checked, most important of which are distances 'a' and 'b'. For insulating distance 'a' between start and end of the regulating winding and insulating distance 'b' between the fine tap selector contacts of different phases, a statistics-based estimation is used. The logic behind the estimation is assuming linear distribution, and additionally applying a „non-linearity“ factor depending on the regulation concept (reg.in main / coarse-fine / reversing / linear) and transformer type (full / auto). The mentioned factor is determined using statistical analysis derived from experimental data. The differences in factors arise from the specifics of the distribution of electrical field in various power transformer types.

With some winding arrangements (neutral-end in autotransformers and line-end regulation in delta connected windings), very high voltages may occur. These voltages are significantly influenced by the choice of the regulation concept (linear, coarse fine or reversing regulation). In such cases, additional tools are used to determine the electrical field distribution in the power transformer (FEM analysis). Since the tap selector size price varies significantly, this can also be subject to optimization.

The switching capacity P_{StN} [kVA] determines the capability of the regulation unit to switch between regulation steps. Since the worst case has to be considered, the required switching capacity in the regulation unit is the product of the relevant step voltage and maximum rated through current (both described above):

$$P_{StN} = \frac{I_{max} \times U_{step}}{1000} \quad (6)$$

The required switching capacity for a specific contact in an OLTC is based on the relevant step voltage and current but is also determined by the design and circuit of the OLTC. The switching capacity itself is primarily a function of the contact design, contact speed and arc-quenching agent [9].

The selected regulation unit must be within the switching capacity diagram (example in Figure 5). Limit values are given in the manufacturer's technical guide and may restrict certain types of tap changers in some cases.

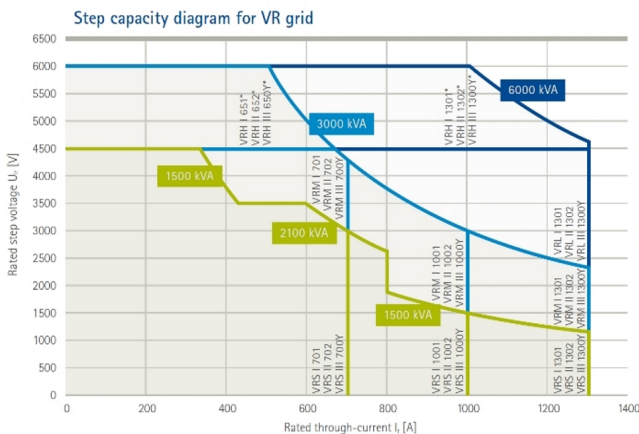


Figure 5: Step capacities diagram

The number of tap positions and the number of contacts have to be considered as well - different types have different maximum regulation range, depending on the regulation concept. A higher

number of positions may also increase the voltage stress of the regulating winding, so this has to be taken into consideration (as described in the internal voltage calculations above). For rectifier or furnace transformers, wide tapping range should be considered.

The predicted contact life of fixed and moving contacts is a function of the rated through current and has to be compared with the diagram for a specific regulation unit which can generally be found in the regulation unit manufacturer's technical guide.

The recovery voltage occurs on the open change-over selector contacts during switching sequence. It depends on the geometric winding arrangement and the capacitances C_1 and C_2 . Recovery voltage can cause switching sparks or low-energy arcs, thereby producing unwanted gas which has a negative effect on the characteristics of the insulation media (usually insulating oil). The recovery voltage and breaking current values have to be checked in + and - selector switch positions:

$$|\vec{U}_{w+}| = \sqrt{\left(\frac{U_1}{2} + \frac{U_f}{2}\right)^2 + \left(\frac{U_1}{2\sqrt{3}} \times \frac{C_2}{C_1 + C_2}\right)^2} \quad (7)$$

$$|\vec{U}_{w-}| = \sqrt{\left(\frac{U_1}{2} - \frac{U_f}{2}\right)^2 + \left(\frac{U_1}{2\sqrt{3}} \times \frac{C_2}{C_1 + C_2}\right)^2} \quad (8)$$

The breaking current I_{S+} and I_{S-} :

$$\vec{I}_{S+} = \frac{\vec{U}_1}{2\sqrt{3}} \times \omega C_2 + j \frac{\vec{U}_1 + \vec{U}_f}{2} \times \omega(C_1 + C_2) \quad (9)$$

$$\vec{I}_{S-} = \frac{\vec{U}_1}{2\sqrt{3}} \times \omega C_2 + j \frac{\vec{U}_1 - \vec{U}_f}{2} \times \omega(C_1 + C_2) \quad (10)$$

U_{w+} and U_{w-} are the recovery voltages in [kV];

I_{S+} and I_{S-} are breaking currents in [mA];

C_1 and C_2 are the capacitance values in [pF] of the regulating winding (as shown in Figure 6) which can be approximated assuming cylindrical winding system [10].

U_l is the rated regulated voltage in [kV];

U_f is the preselector voltage in [kV];

$\omega = 2\pi f$, where f is the power frequency in [Hz].

Example for recovery voltage calculation for single phase autotransformer with regulation of common winding (Figure 7) in wye connection (fine regulation), input data:

- Rated power: 200MVA
- Rated voltages and relevant tapping range: 400 / 230 +10 x 1,5%, -10 x 1,5% / 24 kV
- Step voltage: 1992 V
- Max. tapping current: 1772 A
- Test Voltages:
 - a...340 kV (1,2/50µs); 65 kV (50 Hz)
 - $U_{m...}$ 950 kV (1,2/50µs); 395 kV (50 Hz)

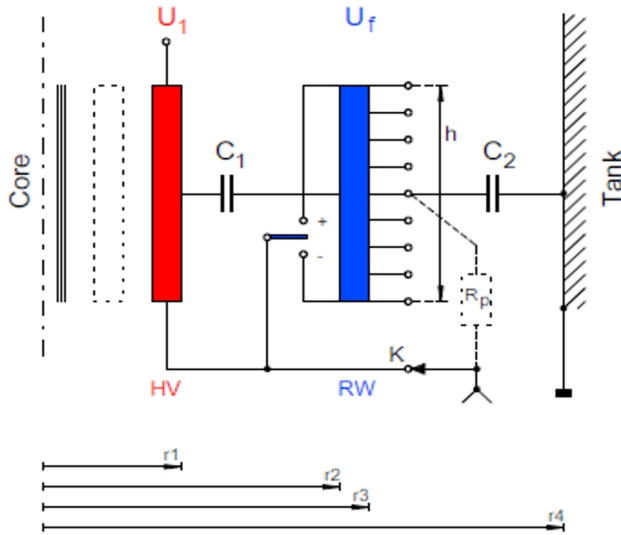


Figure 6: Recovery voltage – capacitances per phase

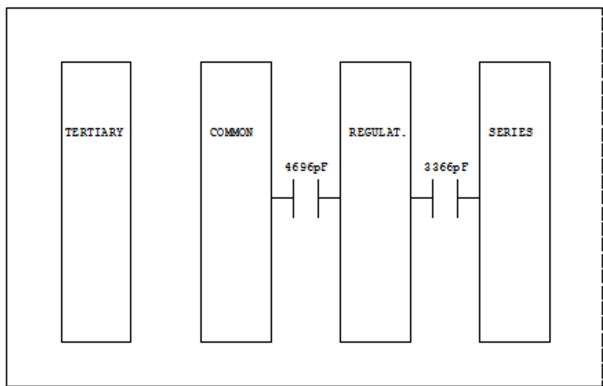


Figure 7: Autotransformer winding arrangement wye connection

According to input data (capacitances $C_1=4696\text{pF}$ and $C_2=3366\text{pF}$), following recovery voltages and breaking current are calculated:

$$U_{W+}=8,3\text{kV}; U_{W-}=28,2\text{kV}$$

$$I_{S+}=21\text{mA}; I_{S-}=71,3\text{mA}$$

According to Figure 8., the potential measure is not needed.

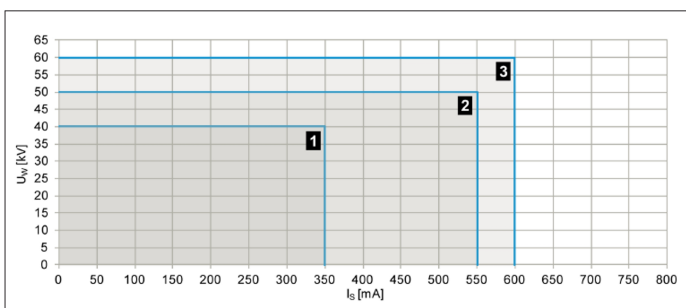


Figure 8: Recovery voltage and breaking current diagram

The recovery voltage permissible value is the function of the calculated breaking current value. The mentioned $U_w(I_s)$ function for each tap selector type and size can be found in the technical data of the regulation unit. If the calculated values of recovery voltages are exceeded, it is necessary to install tie-in resistors or

select a different tap selector. The installation of the tie-in resistors affects the regulation unit price as well as overall transformer dimensions, so a solution with bigger tap selector size and therefore higher permissible values of recovery voltage should be considered in such cases. Furthermore, the tie-in resistor dimensions are defined by the regulation unit manufacturer specifically for each design, which can cause an additional unwanted time delay in the transformer final design process. For phase shifting transformers, using regulation unit with two-way change-over selector, one can overcome problems during the change-over selector operation and change of winding connections between phases of the system voltage [11].

4. Tap Changer Selection Application and Database

Due to the complexity of the regulation unit technical limits and cross-dependencies between different types, a relational database is developed in such an architecture that an algorithm can quickly go through all of the possible permutations and compare the actual calculated values with every combination of regulation unit's limits. Also, the database management user interface (UI) is developed in MS Access environment. Within the database, SQL is used for querying data and sending it to the application for further analysis in the TCS algorithm (Tap Changer Selection). The algorithm then compares the queried data with the calculation results which are calculated in the background as thoroughly described in chapter III. The algorithm also takes into account different possibilities of transformer regulation: for example, a three phase transformer can either be regulated with a single three-phase regulation unit or three single-phase regulation units. Furthermore, autotransformer can be designed with regulation in main winding, meaning that the regulation unit does not have a common neutral end. Therefore, autotransformers with such regulation can only be regulated with three single-phase regulation units (three-phase regulation unit requires a neutral end common point). Also, different regulation concepts have different possibilities for number of positions and number of contacts.

All of the mentioned above is automatically checked for every single regulation unit combination queried from the database. This is achieved using a foreach loop which (in each step) calls for different modules used for this purpose. All of the modules and functions are developed in C#, following object-oriented software development principles (Model-View-Controller, MVC pattern). The input for the algorithm that is expected from the application user is the same data that the electrical designer uses in power transformer design (MVA rating, rated voltages and required regulation range, regulation concept, connection of regulating winding, LI and AC voltage levels i.e. external insulation levels to ground and number of tap positions), making the application very user-friendly for any power transformer designer. This data model also makes automatic data import from transformer design software very simple. The algorithm output is the list of all the possible solutions for regulation units (tap changers) that satisfy the given input.

The results are finally sorted according to the expected total cost, as shown in Figure 9. Therefore, the application user can simply choose the first regulation unit from the generated list knowing that this is the optimal solution. Other regulation units from the list are not optimal but are also technically acceptable. These can be selected in case of a specific request (e.g. higher

external insulation level requested by the customer). Finally, a report containing all the calculations together with the technical limits can be generated automatically.

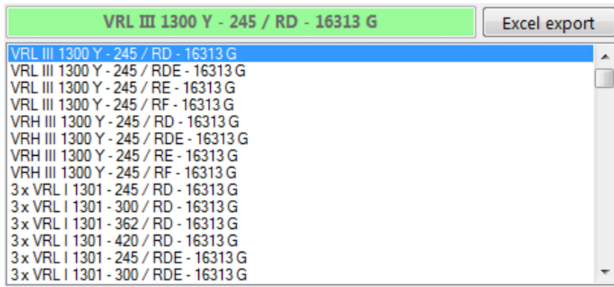


Figure 9: TCS results

4.1. Web application

Together with the Tap Changer Selection application, an additional tool (Figure 10) is developed for accessing the final order data through a Web interface.

The *RSB Base* application allows reading all the necessary technical data from the ordering data sheets, saves this data to the database and gives the possibility to search, analyze and report through user-friendly interface.

Features and functions of the MVC WEB application and the database:

- **Tables**
OLTC data loaded from ordering documentation, gets entered into the database tables. An intuitive UI provides an organized way to represent the data. Tables contain all necessary fields from ordering documentation as well as coded tables for data analyzing and reporting.
- **Forms**
Application provide functionality with controls that allow inserting new ordering data, editing, deleting, page through, sort, filter and search tables and queries, modify ordering data, as well as creating reports.
- **Reports**
The application allows the user to manipulate stored data in various ways. One can use application's built-in function for creating reports to help in analyzing the OLTC data, e.g., calculation of motor drive equipment and electronics in detail.
- **Data functions**
Application software provides features to organize data (CRUD - Create Read Update and Delete) and functions for sorting, filtering, reporting etc.

5. Results and Future Work

Possible future work includes improvements by developing a faster OLTC search algorithm. The database contains all the necessary technical information about the offered and executed projects. The current available size is approximately 400 items with more than 90 properties (with a growth tendency). By increasing database and displaying equipment in the motor drive unit in detail, the possibility to make a mistake in the transformer design offer stage is already significantly reduced. In TCS application, different simplifying estimations were made

(depending on the regulation concept, linear distribution and nonlinear factor were applied) which are still accurate enough to select an OLTC properly. There is room for expansion in the application in order to better cover some special transformer's designs, such as HVDC, limit cases or arrangement with enforced current splitting. However, the results are already visible in the offering phase, by reducing the time of OLTC selection up to 90% compared to the manual selection. This means a lot of working hours in the design office have been saved. Further improvements are reflected in the pooling data from application database itself and from the web application that is related to the OLTC and motor drive unit equipment ordering data in one general database.

Project No	File name	No of col.	Type	No of ph.	Rated curr. (A)	Y /D	Um (kV)	Tap sel. size
WEL.100	B52 64 506.pdf	1	VM	III	350	Y	170	D
Kom 115	B52 54 121.pdf	1	VM	I	1200		300	RDE
TAYAB. 300	B52 64 490.pdf	3	M	I	1200		170	B
GEN. 300	B52 64 499.pdf	1	VM	III	1200	Y	72,5	C
JEB 388	B52 64 488.pdf		VRM	III	1000	Y	245	RDE
Ten. 400	B52 64 496.pdf	1	VRM	III	1300	Y	170	RC

Figure 10: Web interface – homescreen

6. Conclusion

The importance of regulation in power transformers shows that optimizing and choosing the appropriate regulation unit may have a significant impact on the transformer overall cost, making it one of the essential parts of power transformer design. Management of such a complex system by implementing a software solution is presented in this paper. Comparison with Compas tool is not entirely possible, because TCS application works with types of OLTCs and DETCs other than ABB. It could be said that TCS and Compas largely cover all the needs of the transformer manufacturer for the appropriate regulation unit. In TCS application, dozens of tables and diagrams, as well as hundreds of pages of technical data from different manufacturers are incorporated in the database. Nevertheless, the developed algorithm performs all the calculations and searches through all of the possibilities in a matter of seconds, regardless of the regulation unit manufacturer. This represents significant time savings and reduced possibility of errors, providing productivity improvement in power transformer design process. Selection of OLTC is a critical point in the transformer design process due to a large number of correct solutions, only one of them being optimal while also technically acceptable and cost-effective. It can be concluded that the TCS application connects the physical properties of the OLTC and the information technology, which has led to material and time savings in transformer design process

Conflict of Interest

The authors declare no conflict of interest.

References

- [1] M.Gazdović, T.Šimović, “Digitalization of Regulation Unit Selection Process in Power Transformer Design” Proceedings of the 41st International Convention for Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018).
- [2] G.Leci, A.Marušić “Koodinirana regulacija napona energetskih transformatora” 11. savjetovanje HRO CIGRE, 2013.
- [3] IEC 60214-2 Tap-changers – Part 2: Application guide, 2004
- [4] ABB, Tap-changer selection program Compas [Web]
<http://new.abb.com/products/transformers/interactive-tools/tap-changer-selection-program-compas>
- [5] R.Levi, “On Load Tap Changer Condition Assessment Using Dynamic Recording and Measurement (DRM)”, 2017.
- [6] A.Krämer, “On-load tap-changers for power transformers,” Maschinenfabrik Reinhausen, 2000.
- [7] Maschinenfabrik Reinhausen GmbH “TD - General Section”, 2003.
- [8] IEC 60076 – 3 Power transformers – Part 3: “Insulation levels, dielectric tests and external clearances in air”, 2013.
- [9] D.Dohnal, “On-Load Tap-Changers for power transformers,” Maschinenfabrik Reinhausen, 2013.
- [10] M.Gazdović, “TCS help”, Končar Power Transformers, 2017., unpublished.
- [11] T.Šimović, M.Stanić, L.Kirchner, “200 MVA phase shifting autotransformer for HE Senj and the possibility to change-over between active and reactive power flow control in energized state“, 4rd International Colloquium "Transformer Research and Asset Management", May 10-12, 2017.

Spectral Grid Impedance Identification on the Low-, Medium- and High-Voltage Level – System Design, Realization and Measurement Results of Grid Impedance Measurement Devices

Hauke Wilken*, Michael Jordan, Detlef Schulz

Helmut-Schmidt-University, Chair of Electrical Power Systems, Holstenhofweg 85, 22043 Hamburg, Germany

ARTICLE INFO

Article history:

Received: 15 August, 2018

Accepted: 30 December, 2018

Online: 12 January, 2019

Keywords:

Grid impedance

Power quality

Grid integration

ABSTRACT

The frequency dependent grid impedance at the terminals of an electrical supply grid is an essential parameter for power quality and grid feedback analysis. In this contribution the identification of the grid impedance is achieved using spectral excitation currents at the corresponding grid connection point, which generate measurable changes in the grid voltage spectra. Different measurement systems based on this method have been successfully realized for the low- and medium-voltage level. A measurement system for the high-voltage level is currently being realized. The application of the measurement systems are facing quite different external framework conditions leading to specific system designs for each measurement device. The spectral grid impedance identification on these voltage levels is done through fast switching of an ohmic load with power electronic. The requirements and setup of the grid impedance measurement systems are outlined and measurement results on the low- and medium-voltage level are presented and discussed. The systems can be used to evaluate grid connection points and to verify other active or passive grid impedance identification methods on different voltage levels.

1. Introduction

This paper is an extension of work originally presented in 2017 at the IEEE Power & Energy Society General Meeting [1]. The extensions consider essential aspects on the relationships of the grid impedance and the influence of harmonics on the grid voltages, especially in relation with renewable energy systems. Additionally the measurement results and possible benefits are described in more detail. The harmonic requirements are discussed in relation to the experimental results. Further on a new possible measurement procedure for an improved grid integration of renewable energy systems is presented. Finally the connection aspects and structure of the devices are explained in more detail.

The share of decentralized renewable energy systems will continue to grow in the upcoming years. Their share in the total electricity generation in Germany is targeted to rise up to 80 % by 2050. This changes the power flows in the grid, which have an effect on the grid voltage levels, the loads of the grid equipment and the protection technology. Therefore, a strategic grid reinforcement, grid optimization and grid expansion is required to master the future challenges.

The electrical characteristics of a point of common coupling (PCC) are determined by the grid voltage and the internal

impedance at the PCC. At the fundamental frequency these parameters define the short circuit power and hence the capacity of the PCC. To evaluate harmonics above the fundamental frequency of power systems connected to the PCC, the spectral characteristic of both parameters are needed. In most cases the grid voltages can be easily measured and spectrally analyzed. The grid impedance at the fundamental frequency can be approximately obtained from grid simulation programs. It is typically not possible to determine the grid impedance above the fundamental frequency through simulation because this requires detailed spectral models of the power system utilities. The spectral models however are often not available or just valid for low frequencies. Therefore an exact determination of the spectral grid impedance over a wide spectral range can only be accomplished by measurement.

The frequency-dependent grid impedance can be used to evaluate network perturbations (harmonics, flicker) [2]. At the Helmut Schmidt University, measurement devices for the determination of the time- and frequency-dependent grid impedances at the low-, medium- and high-voltage level have been developed [3], [4], [5].

2. Relevance and interrelation of the grid impedance

The grid impedance is a complex variable that describes the resulting voltage uplift or voltage drop when power is fed into the

*Hauke Wilken, +49 40 6541-2329, hauke.wilken@hsu-hh.de

grid or power is taken from the grid. It is composed of the proportions of the individual impedances of the used grid equipment. In the following, the 50 Hz component of the grid impedance of a grid connection point / point of common coupling (PCC) in the medium-voltage grid is described. It essentially consists of the higher-level 110 kV grid, the 110/10 kV transformer and the medium voltage cable, shown in Fig. 1. For this purpose, one ohmic and one inductive component are used.

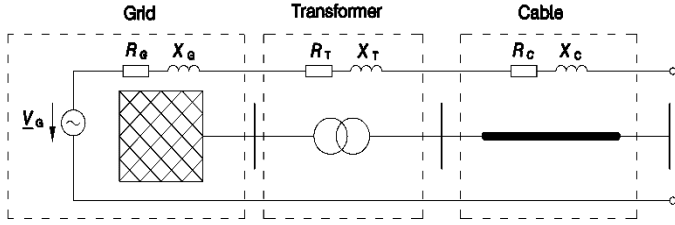


Figure 1. Schematic of grid equipment

2.1. Superordinate 110-kV-Grid

The proportion from the higher-level 110 kV grid can be described with the terminal voltage $V_{nG}/\sqrt{3}$ as:

$$\underline{Z}_G = R_G + jX_G = \frac{1,1 \cdot V_{nG}}{\sqrt{3} \cdot I_k''} \quad (1)$$

According to DIN EN 60909, a safety margin of 10 % is taken into account for the short-circuit current I_k'' . The impedance \underline{Z}_G is related to the short circuit apparent power S_k'' . The amount can be stated, with usual values according to Table 1, to:

$$|\underline{Z}_G| = \frac{1,1 \cdot V_{nG}^2}{S_k''} \quad (2)$$

Table 1: Typical short-circuit currents and short-circuit powers [6]

V_{nG}	I_k''	S_k''
10 kV	29 kA	0,5 GVA
110 kV	42 kA	8 GVA
220 kV	63 kA	24 GVA
380 kV	80 kA	53 GVA

The relevant reactance X_G can be equated approximately with the impedance Z_G for the high voltage level. Since the transformer in most cases has a step control, the regulated voltage value of the low-voltage side of is used for V_{nG} . The reactance / resistance ratio can be estimated with a factor of 6 if unknown [7].

2.2. 110/10-kV-Transformer

The impedance of the transformer can be calculated for the middle position of the step control with the relative short-circuit voltage of v_k and copper losses of P_{Co} to:

$$X_T \approx Z_T = \frac{V_{nG}^2}{S_{kT}} = \frac{V_{nG}^2}{\frac{S_{rT}}{v_k}} \quad (3)$$

$$R_T = \frac{V_{nG}^2 \cdot P_{Co}}{S_{rT}^2} \quad (4)$$

2.3. Medium-voltage cable

The impedance of the medium-voltage cable depends on the corresponding length as well as the resistance coating R'_C and reactance coating X'_C .

2.4. Grid impedance at the point of common-coupling

The impedance Z_{PCC} at the grid connection point / the point of common-coupling (PCC) results from the sum of the individual impedances. It is related to the short-circuit power $S_{k,PCC}$ and the voltage V_{PCC} as follows:

$$S_{k,PCC} = \frac{V_{PCC}^2}{Z_{PCC}} \quad (5)$$

Fig. 2 shows the grid impedance as well as the short circuit apparent power of a PCC as a function of increasing cable length. It shows the indirect proportionality between $S_{k,PCC}$ and Z_{PCC} . This relationship plays the decisive role in the calculation of the permissible voltage change by connection of loads or generation units.

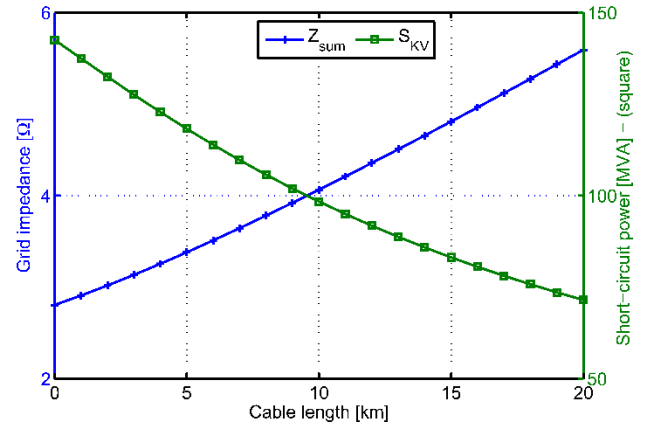


Figure 2. Characteristic of grid impedance (solid) and short-circuit power (dashed)

2.5. Voltage change

According to the technical guideline for the connection of power generation units in Germany, the voltage change caused at the PCC Δv_a in the medium-voltage grid must not exceed 2 % [7]. Using the power ratio of $S_{k,PCC}$ to the apparent power of the power generation unit, the voltage change can be roughly estimated. With a ratio smaller than 50, a closer look is necessary, since the power generation unit has more influence on the grid voltages. For this, the grid impedance angle as well as the injected reactive current must be taken into account, which can be set at the medium voltage level via the power factor between $\cos \varphi_a = 0,95_{ind}$ and $\cos \varphi_a = 0,95_{cap}$ (inductive / capacitive power feed-in and over- / under-excited operation respectively). The voltage change is a function of the grid impedance angle ψ_{PCC} at the PCC, the maximum system apparent power $S_{a,max}$, the grid impedance \underline{Z}_{PCC} and the voltage V_{PCC} .

$$\Delta v_a = \frac{S_{a,max} \cdot \cos(\psi_{PCC} + |\varphi_a|)}{S_{k,PCC}} \quad (6)$$

Solving this relationship to the maximum apparent power of the generation unit that can be operated at the corresponding PCC, and using the ohmic and inductive components of the grid impedance Z_{PCC} , yields:

$$S_{a,max} = \frac{\Delta v_a \cdot V_{PCC}^2}{|R_{PCC} \cdot \cos(\varphi_a) - X_{PCC} \cdot \sin(\varphi_a)|} \quad (7)$$

It can be observed that for a power feed-in with a power factor of $\cos \varphi_a = 1$, only the resistive component is relevant for the voltage uplift.

3. Grid feedback

There are a number of standards for ensuring the proper operation of electrical power generation units and electrical loads connected to the power grid. At the PCC, the grid voltages must meet certain criteria that are understood by the grid quality. Failures and outages of the energy supply grid (long and short-term outages) fall under the security of supply. In reference with the grid impedance, in particular the voltage form and the influence of generating units and consumers on the grid voltages are of interest. The grid voltages should be ideally sinusoidal and have a constant frequency (50 or 60 Hz). Grid perturbations due to harmonics (also intermediate and subharmonic), voltage fluctuations and unbalances as well as transients (also voltage dips and overvoltages) reduce the grid quality. Grid perturbations result from non-linear equipment and switching operations. Especially the increasing proportion of power electronics in many devices and also renewable energy systems can lower the power quality by taking or feeding in non-ideal sinusoidal currents. The permitted limits for the harmonic currents and also for the harmonic voltages must not be exceeded [8]. Harmonics can have several bad effects in the electrical energy supply network:

- Reduction of efficiency
- Destruction of insulation
- Reduction of lifetime (early failure of insulation)
- Thermal overload of equipment
- Malfunction of protection and measuring devices
- Attenuation and distortion of signals (eg frequency remote control, grid communication...)

As the number of power electronics within consumers and power generation units connected to the electrical energy supply grid increases, the effects increase. With measured frequency characteristics of the grid impedances, the grid perturbations can be better and more objectively evaluated. Resonances in the frequency characteristic of the grid impedances can be detected at an early stage.

In general, high short circuit power and hence a small grid impedance value results in lower voltage drops when connecting loads, and also in a low voltage uplift when connecting power generation units as well as in reduced grid perturbations.

Table 4 summarizes limits for the allowable harmonic voltages at the low, medium and high voltage levels. The harmonic limits of standard EN50160 describe the minimum requirements for PCCs. They are therefore higher than the values of the technical guidelines for power generation units, including renewable energy systems ([7], [9]). Deviating from these limitation values can be

tolerated by grid operators with higher values in subareas of their electrical grids in coordination with customers if a negative influence on consumers and power generation plants can be avoided.

Table 2: Selected limitation values for certain harmonics at low-, medium- and high-voltage level

Order v	Vv/Vn at low- and medium-voltage level (acc. to [8])	Vv/Vn at medium-voltage level (acc. to [7])	Vv/Vn at high-voltage level (acc. to [8])	Vv/Vn at high-voltage level (acc. to [9])
5	6.0 %	0.5 %	5.0 %	0.25 %
7	5.0 %	1 %	4.0 %	0.5 %
11	3.5 %	1 %	3.0 %	0.5 %
13	3.0 %	0.85 %	2.5 %	0.4 %
17	2.0 %	0.65 %	tba.	0.3 %
19	1.5 %	0.6 %	tba.	0.25 %
23	1.5 %	0.5 %	tba.	0.2 %
25	1.5 %	0.4 %	tba.	0.15 %

When connecting high power wind farms via HVDC systems to the high-voltage level (especially in the offshore area), the harmonic emissions of the entire system must be considered up to a frequency of 10 kHz [9]. In addition, the THD (Total Harmonic Distortion) is used to assess the power and voltage quality of renewable energy systems:

$$THD_V = \frac{\sqrt{\sum_{i=2}^{40} V_i^2}}{V_1} \quad (8)$$

At the medium voltage level, the THD_V must be below 8 % [7]. At the high voltage level, no limit value is specified for the THD_V in [9] and [8]. Furthermore, limitation values for the short- and long-term flicker must be complied, which were especially relevant for older directly coupled wind energy systems and have higher values than current systems using inverters.

The basis of the limitation values for harmonic voltages and currents is the assumption that the grid impedance consists of an ohmic and an inductive part, as described in section 2.1. However, capacitive elements in the grid equipment as well as non-linear behavior of consumers and power generation units with power electronics lead to a strong deviation from this assumption and to the occurrence of resonance frequencies. At these resonant points the limitation values can be exceeded due to power feed-in of renewable energy systems. Even instabilities can occur. In this case, the power must be reduced and possibly additional filters must be installed.

The design of grid-side and internal filters can be improved with knowledge of the grid impedance [10]. The internal control of the voltage and current regulators used in the inverters can also be adapted to the network impedance to avoid possible current fluctuations and instability problems [11].

4. Basic Principle of applied Grid Impedance Measurement Method

Several methods have been introduced to measure the grid impedance. They can be categorized into invasive and non-

invasive techniques. Non-invasive techniques assess natural power flow variations caused by fluctuating loads or generators with statistical methods. An example of a grid impedance estimation with a statistical methods can be found in [12]. However, in most cases a high precision over a wide frequency range cannot be achieved with these methods.

Commonly used invasive techniques inject harmonic or interharmonic current signals to a PCC. In [13] an inverter system is used to measure the grid impedance on the low-voltage level by sweeping the frequency of the injected current signals in the range from 100 Hz to 10 kHz. Inverter based grid impedance measurement methods are applied almost exclusively on the low-voltage level. To measure the grid impedance at medium- or high-voltage level with this method, the inverter systems need to have very high output power. Further on, the inverter system is coupled with a special transformer to the grid. Transformers with high rated power that offer the necessary broad frequency range are not available on the market and have to be specially designed, which makes this method complex and cost-intensive. On the medium- and high-voltage level active switching of transformers [14] or capacitor banks [15] have been commonly used to identify the grid impedance with good results. The switching has been done by circuit breakers with a number of maximum cycles of operations, which strongly reduces the possible repetition rate. Further on, the duration of the generated excitation pulse signals are very short and the pulse amplitude depends on the switching moment, which cannot be controlled accurately with the circuit breakers. Therefore the spectral excitation of these current pulses is only sufficient at certain frequencies and cannot be varied in practice. These drawbacks can be overcome by generating pulsed current signals with power electronic switches and ohmic loads.

As shown in Fig. 3 a power electronic switch is used to pulse an ohmic load in order to excite the grid. The advantage of this method compared to switching loads with circuit breakers lies in the flexibility of the pulse patterns that can be applied to the power electronic switch. The voltages and currents of the first interval where the switch is open (subscript 1) and the second interval where the load is pulsed by the switch (subscript 2) are transformed into the frequency domain with FFT-algorithms. In the frequency domain the complex spectra of the grid impedance $\underline{Z}_G(\omega)$ can be calculated as follows:

$$\underline{Z}_G(\omega) = \frac{V_1(\omega) - V_2(\omega)}{I_1(\omega) - I_2(\omega)} = \frac{\Delta V(\omega)}{\Delta I(\omega)} \quad (9)$$

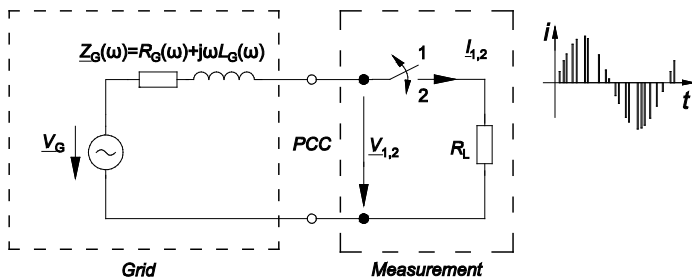


Figure 3. Grid impedance measurement with pulsed ohmic load

Strategies to excite the grid with pulsed current sequences can be found in [16], [17] and [18]. With periodic pulse signals the grid impedance can be precisely measured at specific frequencies.

Spectral ranges of the grid impedance may be measured with a frequency sweep of periodic pulse signals. Depending on the spectral range and number of steps this method can be very time consuming. Another option is to switch the load with random pulse patterns. The resulting current signals have broad frequency spectra but with comparatively low amplitudes. In order to obtain good measurement results several measurement iterations in succession using random signals can be used.

On the low-voltage level with a four-wire system the pulsed current signals can be generated sequentially between two conductor lines or between a conductor line and the neutral line to measure the corresponding loop impedance parameters. Further on, equivalent line impedance parameters can be calculated out of the loop impedance parameters [18]. In three-wire systems as on the medium- and high-voltage level the loop and equivalent line impedance parameters can be measured in the same way. It is also possible to determine coupling impedance parameters between the lines with this method [4].

5. System Design and External Technical Parameters for the Realization of the Measurement Devices

The most important relevant technical parameters for the realization of the grid impedance measurement devices are given by the structure of the grids:

- Nominal voltage
- Grid structure
- Typical short-circuit powers

These parameters have to be compromised with the application and the definition of the subsequent measurement results:

- Frequency range of interest (only 50 Hz or spectral range)
- Repetition rate of measurements (in order to evaluate time changes)
- Measurement accuracy
- Framework conditions:
 - Price
 - Transportation
 - Grid connection

A. Technical Requirements

The power electronic switch is directly connected to the terminals of the different voltage systems. At the low-voltage level (0.4 kV) the effort is relatively low since there are single power electronic components e.g. Mosfets or IGBTs that fully offer the necessary blocking voltage. On the medium-voltage level (20 kV) a series connection of multiple power electronic components is required in order to achieve the high blocking voltages. At the high-voltage level (110 kV) there is an enormous effort. An extremely high demand of power electronic devices is necessary.

Another aspect concerns the spacing between parts within the measurement devices which are subjected to the grid voltages and other parts that are connected to ground, refer to Table III. At the low-voltage level a compact design on a printed circuit board is possible. At the medium- and high-voltage level this is not possible

due to the high grid voltages. The breakdown voltage in air amounts to about 30 kV/cm. A homogenous design however cannot be realized, so that a typical design value of 1 kV/cm can be used for an air-insulated design. The measurement system for the medium-voltage level has been realized with an air-insulated setup. At the high-voltage level distances of over 1 m would have to be held, making a compact air-insulated design impossible. Therefore ester is used as insulation medium.

Table 3: Overview of selected permitted isolation distances in air according to EN 61936-1

Highest voltage of facility	Minimum distances (line to earth and line to line)
12 kV	90 mm (indoor) 120 mm (indoor) 160 mm (indoor)
24 kV	160 mm 220 mm 270 mm
123 kV	900 mm 1100 mm

B. Grid Connection

The grid connections of the measurement systems are realized by circuit breakers so that possible short-circuits within the devices can be disconnected. On the one hand the dimensions and costs of the measurement devices for the different voltage levels strongly deviate. On the other hand the effort and costs for the connection to the grid also deviate severely. At the low-voltage level the power switch is very compact and grid connection can be done with simple cables and connectors. At medium- and high-voltage level the design, weight and laying of the cables as well as the necessary mountings are more complicated. Mounting of the sealing ends has to be done by specialized technical personnel. Especially at the high-voltage level the effort and costs are extremely high.

C. Organization Aspects

Dimensions and weight of the different grid impedance measurement systems vary to each other. At first the transportation of the systems differentiates a lot. The device for the low-voltage level can be easily carried. The measurement system for the medium-voltage level requires a small truck with a crane for transportation. The very large measurement container for the high-voltage level needs a heavy-weight transportation and for its placing a special crane. Additionally the area for assembly has to be considered. The measurement system for medium- and high-voltage level have to be put on solid ground, so that in some cases a fundament is needed. In order to connect these measurement devices to the grid the personnel requires qualification to obtain the authorization to connect and disconnect the devices to the grid.

6. Practical Realizations

Based on the method stated in section 4, measurement systems for the low- and medium-voltage level have been realized to measure the spectral grid impedance. Further on, a measurement system for the high-voltage level is currently being realized.

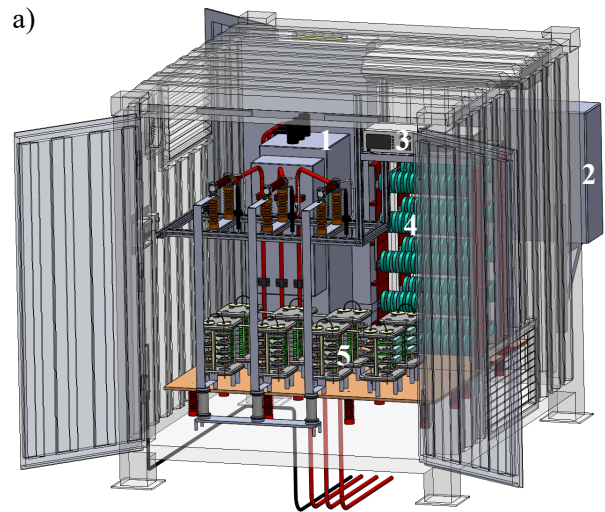
A. Low-Voltage Level (0.4 kV)

Fig. 4 shows the grid impedance measurement system for the low-voltage level. The system can measure the complex loop impedance parameters $Z_{ab}(j\omega)$, $Z_{bc}(j\omega)$ and $Z_{ca}(j\omega)$ as well as

the line impedance parameters $Z_a(j\omega)$, $Z_b(j\omega)$ and $Z_c(j\omega)$ in four-wire systems with the outer conductor lines a, b and c up to 0.4 kV in the frequency range from DC to 150 kHz. The system generates asymmetrical pulsed current signals between two lines for the measurement, whereby every line combination can be pulsed sequentially. The maximum amplitude of the pulsed current signals can be varied between 1 A_{peak} to 20 A_{peak}.



Figure 4. Grid impedance measurement system for the low-voltage level



- 1) SF6- circuit-breaker (20 kV)
- 2) Air conditioning
- 3) Measurement and control units
- 4) High power resistors
- 5) Power electronic

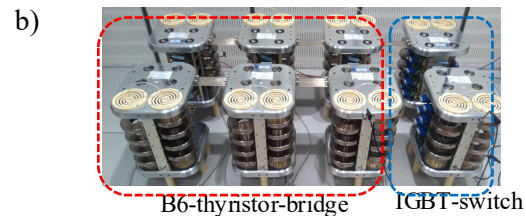


Figure 5. 3D CAD model (a), photo of the power electronic grid excitation circuit (b) and photo of the medium-voltage measurement container

B. Medium-Voltage Level (20 kV)

The setup of the grid impedance measurement system for the medium-voltage level is illustrated as a 3D-CAD model in Fig. 5 a). Fig. 5 c) shows a photo of the system. All components have been installed within a type-tested container. Additionally, to the power electronic circuit (Fig. 5 b)) and the load resistor a circuit-breaker, highly precise voltage and current transducers, high-voltage fuses, overvoltage suppressors, an air-conditioning system and a control and data recording system are installed in the system.

In most cases medium-voltage grids are operated with isolated or compensated starpoints. There is no neutral line/point explicitly available to identify the grid impedances. Therefore the method used on the low-voltage level is adapted, so that only the three conductor line impedance parameters are determined [19]. The system can measure the complex loop- and line impedance parameters in three-wire systems up to 20 kV_{rms} in the frequency range from DC to 20 kHz. The measurement device generates asymmetrical pulsed current signals between two outer conductor lines, whereby every combination can be sequentially pulsed. The maximum pulse amplitude can be varied from 1 A_{peak} up to 100 A_{peak}. The measured loop impedance parameters can then be rearranged to the equivalent impedance parameters of each outer conductor line:

$$\underline{Z}_a(j\omega) = \frac{1}{2} \cdot [\underline{Z}_{ab}(j\omega) - \underline{Z}_{bc}(j\omega) + \underline{Z}_{ca}(j\omega)] \quad (10)$$

$$\underline{Z}_b(j\omega) = \frac{1}{2} \cdot [\underline{Z}_{ab}(j\omega) + \underline{Z}_{bc}(j\omega) - \underline{Z}_{ca}(j\omega)] \quad (11)$$

$$\underline{Z}_c(j\omega) = \frac{1}{2} \cdot [-\underline{Z}_{ab}(j\omega) + \underline{Z}_{bc}(j\omega) + \underline{Z}_{ca}(j\omega)] \quad (12)$$

The power electronic grid excitation circuit is shown in Fig. 6. The B6-thyristor-bridge allows to select two outer conductor lines at a time. Then the IGBT-switch is pulsing the load resistor in order to excite the grid in the frequency region of interest. Each thyristor consists of 8 thyristors in series. The IGBT-switch is build up out of 16 IGBTs connected in series. The high power resistor can be varied in its resistance value in order to adapt the grid excitation current to the short-circuit power at the measurement point.

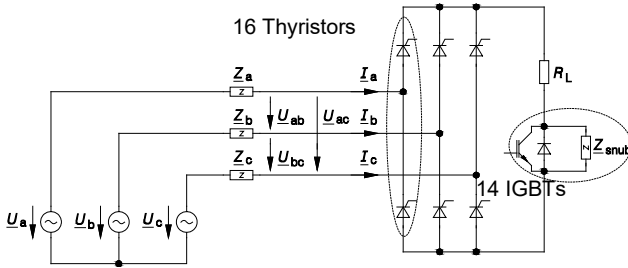


Figure 6. Grid impedance measurement circuit on the 20 kV medium-voltage level

C. High-Voltage Level (110 kV)

The grid excitation circuit for the high-voltage level is shown in Fig. 7. In comparison to the medium-voltage level there is a B6-diode bridge in combination with an IGBT and a resistive load. The B6-diode-bridge offers a simpler design and does not require control signals. Therefore the selection of the outer conductor lines has to be realized by the power switch.

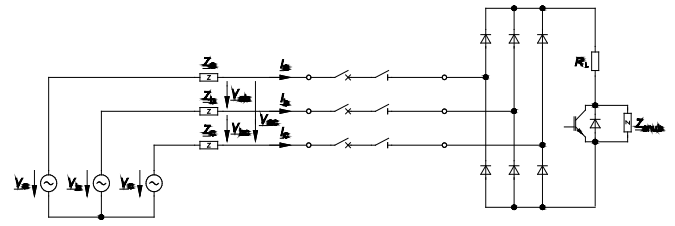
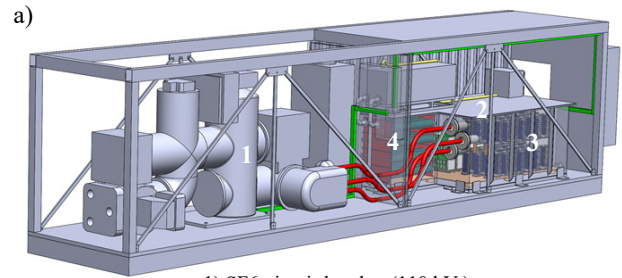


Figure 7. Schematic circuit diagram for the grid impedance measurement on the 110 kV high voltage level

Fig. 6 shows a 3D-CAD model and photos of the current construction status. The power electronic circuit and the load are installed within a special tank filled with Ester as insulation medium. Each diode in the schematic diagram consists of 6 diode-stacks with each having 12 diodes in series. In total there are 432 diodes in the tank. The IGBT is built up of 15 stacks with each stack containing 6 IGBTs (90 in total). The peak current of the power electronic amounts to 200 A resulting in a possible peak grid excitation current of 22 MW. The temperature rising has to be observed and compensated by an air-conditioning system. The number and time length of the grid excitation correlates with the thermal energy and can be controlled by the measurement software system.



- 1) SF6-circuit-breaker (110 kV)
- 2) Tank filled with ester as isolation medium
- 3) Power electronic modules
- 4) High power resistors



Figure 8. 3D CAD model of 110 kV grid impedance measurement system (a) and photos of the current construction status (b)

Basic specifications of the measurement systems for the low-, medium- and high voltage level are stated in Tab. I. It can be seen that dimensions, weight and costs are highly deviating.

7. Measurement Results and Relevance for Grid Integration

On weekdays, there are generally noticeably different frequency profiles of the grid impedances than on weekends. There are also fluctuations between midday, evening and night hours. For the evaluation of grid perturbations, the frequency dependent grid impedance below 2 kHz (40th harmonic) is especially relevant. The

attenuation and frequency of the resonance points can vary greatly over time. At low attenuation, the parallel resonance has a lower frequency width but higher magnitude and at high attenuation vice versa. The damping is mainly determined by the local electrical loads. For Flicker, the 50 Hz grid impedance is crucial, where a single measurement is sufficient, since this usually varies only slightly in time. For the evaluation of harmonics a measurement of the frequency dependent grid impedance over several days or partial weeks is necessary. In particular, if temporally fluctuating feed-in capacities of renewable energies are to be investigated.

Table 3: Specifications of the grid impedance measurement systems for different voltage levels

Parameter	Low-Voltage Level (0.4 kV)	Medium-Voltage Level (20 kV)	High-Voltage Level (110 kV)
Dimensions (L×W×H)	0,5×0,4×0,2 m	2,9×2,5×2,5 m	12,2×2,5×2,7 m
Weight	7 kg	3,65 t	~26 t
Excitation current	1 – 20 A	1 – 100 A	200 A
Device costs	~8.000 €	~200.000 €	~2.500.000 €
Connection costs	0 €	~10.000 €	~500.000 €

A. Low-Voltage Level (0.4 kV)

Fig. 9 shows the frequency characteristics of the loop impedances between the conductor lines (“a”, “b” and “c”) and the neutral line (“N”) measured at a low-voltage PCC. The absolute values are increasing with frequency and show some weak parallel and series resonances. Due to a high number of single phase loads there are strong asymmetries in the characteristics. Up to a frequency of about 30 kHz the frequency characteristics are quite similar, but at higher frequencies especially the phase angles show large differences. Some resonances vary over time caused by load profiles, which especially change the damping of resonances. This effect is illustrated in Fig. 10.

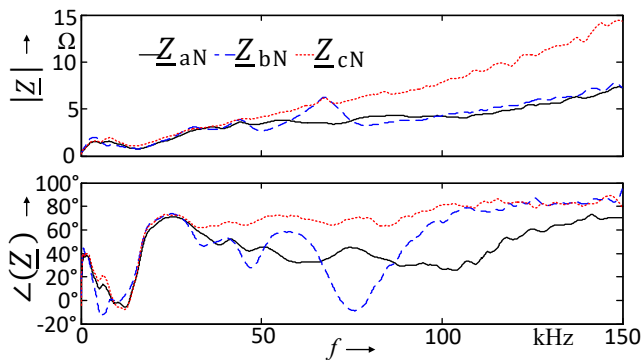


Figure 9. Absolute values and phase angles of the impedances between the conductor lines (“a”, “b” and “c”) and neutral line (“N”) measured at a low-voltage PCC

The first two parallel and series resonance points are below 3 kHz and are more attenuated during the day than at night. Reason could be a higher number of switching power supplies on the grid and thus higher cross-capacity due to computers and office equipment. Figure 10 shows the temporal variation of the first parallel resonance at approximately 1.6 kHz. It results from the interrelation of transformer and line inductances as well as capacities of cables and loads.

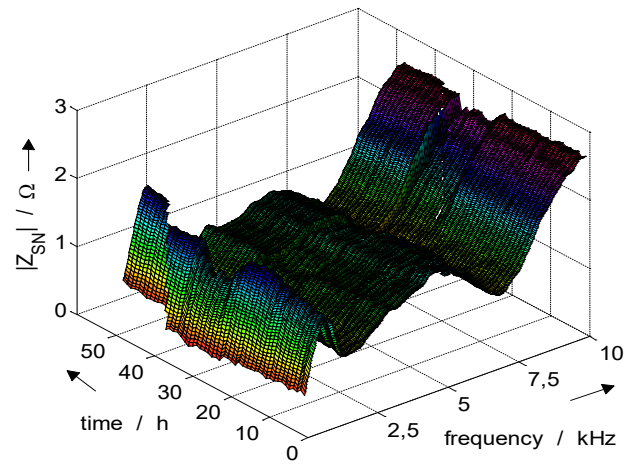


Figure 10. Absolute value of the loop impedance $|Z_{aN}|$ (conductor line “a” to neutral “N”) hourly measured over 3 days at a low-voltage PCC

B. Medium-Voltage Level (20 kV)

Fig. 11 presents the frequency characteristics of the loop impedance parameters measured at a medium-voltage PCC (20 kV). Only small asymmetries appear in the characteristics due to the geometric construction of transformers and power lines. The variation in time and frequency of the positive sequence component is displayed in Fig. 12. Two parallel resonance points can be seen at about 3.8 kHz and 12 kHz with high absolute values of about 400 Ω and 300 Ω, where the phase angle amounts to 0°. Series resonances can be seen at about 7.5 kHz and 15.5 kHz with local minimum impedance values and phase angles of 0°.

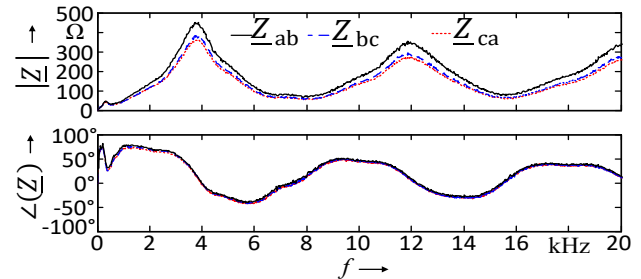


Figure 11. Absolute values and phase angles of the loop impedances between the conductor lines (“a”, “b” and “c”) at a 20 kV medium-voltage PCC

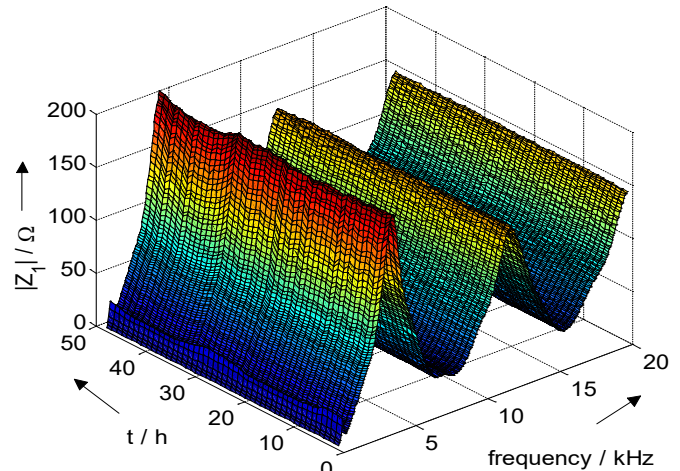


Figure 12. Absolute value of the positive sequence component $|Z_1|$ from DC to 20 kHz hourly measured over two days at a 20 kV medium-voltage PCC

Especially resonances in the frequency range below 1 kHz change over time with a day-night cycle. This effect can be seen in more detail in Fig. 13. The first parallel resonance point at 250 Hz (5th harmonic) shows higher peaks at night and more attenuation with less absolute values in the daytime.

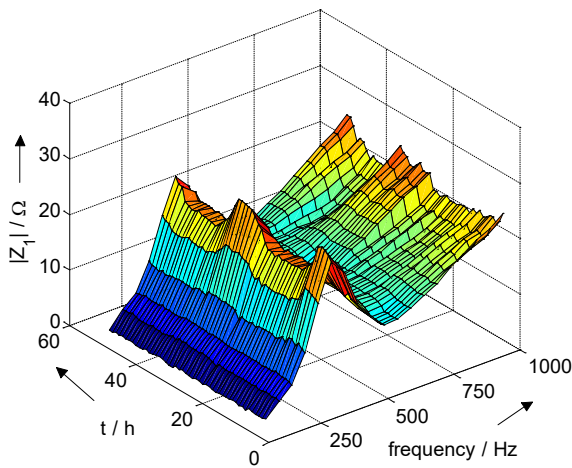


Figure 13. Absolute value of the positive sequence component $|Z_1|$ from DC to 1000 Hz hourly measured over two days at a 20 kV medium-voltage PCC

C. Relevance for Grid Integration

The measurement of the grid impedance can contribute to an improved grid integration of renewable energies. Its knowledge allows a more accurate calculation of the harmonic propagation in the grid. Currently, estimation methods based on measured RMS values of the injected currents are mostly used [2], [7], [9]. But when connecting renewable energy systems to the grid, two sources of harmonics should be considered in general. On the one hand, the renewable energy systems feed harmonic or inter-harmonic currents into the grid. On the other hand, the existing pre-load of the grid with harmonic voltages must be considered. With the frequency-dependent grid impedance and the complex values of the injected currents, the harmonic emission of renewable energies can be determined more accurately. Early detection of resonant frequencies and locations in the frequency characteristics of the grid impedances at the PCC can be used to better select grid equipment when connecting renewable energies. The knowledge of resonance points at the PCC is decisive for the determination of the pulse frequency in the used inverter. It can be varied within certain limits, whereby problems with excessive harmonic voltages at the affected frequencies can be avoided. Operation with variable pulse frequencies can reduce the dependence of the harmonic content on the operating point. Furthermore, the internal power and harmonic controllers can be better adapted to the grid. Power fluctuations and instabilities can be avoided. A lowering of the harmonic voltages can be achieved.

Additionally existing and possibly additionally required filter systems, as well as the renewable energy systems themselves and their inverters, can be designed more efficiently for the PCC. In this way, grid-side and internal filter systems and current controller loops can be better adapted to the PCC in advance as well as in the case of existing renewable energy systems. The dimensioning of a passive filter presupposes knowledge of the grid impedance or its assumed knowledge as well as the harmonic content of the grid. Temporal variations must also be considered in order to avoid

resonances, as this can make the overall system unstable. The grid impedance is also needed for the design of active filters and their stable operation.

Furthermore, the maximum possible feed-in power of a PCC can be determined, as a result of which possibly additional systems can be connected to it. In summary an improved grid integration of renewable energy systems could look like this:

1. Measure the source voltages at the PCC with the identification of the existing harmonic voltage spectra
2. Measurement of the frequency dependent grid impedances with the developed grid impedance measurement systems
3. Identification of resonance points
4. Determination of the maximum grid connection capacity
5. Establishment of a mathematical model of the grid
6. Selection and dimensioning of suitable cables and equipment for connection of renewable energy systems
7. Simulation of the grid connection model and the plant certificate or if available a harmonic model of the renewable energy systems and the selected equipment
8. Detection of possible problems at resonance frequencies and, if necessary, adaptation of the pulse frequency of the inverter
9. Optimized design of filter systems for low harmonics
10. Optimized design of current controllers used within the inverter to avoid stability problems
11. Connection of additional renewable energy systems if limitation values have not yet been reached
12. Summary

This paper presents the challenges and the realization of three measurement devices for the grid impedance identification on the low-, medium- and high-voltage level. Their measurement results allow to determine the grid capacity of points of common coupling (PCC). The grid impedance at the fundamental frequency is directly related to the short-circuit power, so that conclusions about the tolerable connection power for loads or generation units can be stated. Then the measured frequency characteristic can be used to assess the grid feedback and propagation of harmonics especially within the context of an increasing share of renewable energies and their effect on the grid voltages. At first the used measurement method based on the pulsing of a known load using a special power electronic switch concept is described. The grid is excited in the frequency region of interest and the generated current and voltage transients are used to calculate the grid impedance. Then the applications of this method to the low-, medium- and high-voltage level are outlined. Due to the different voltages and short-circuit grid powers as well as grid structures there are quite different concepts necessary. Based on the experience of the small device for the low-voltage level, a 10-foot measurement container for the medium-voltage level has been built. Additionally the development of a 40-foot measurement container for the high-voltage level is introduced. These devices are compared with each other. Finally measurement results on the low- and medium-voltage level conclude this contribution. An improved grid integration of especially renewable energy systems can be enabled. With these novel devices a validation of other grid impedance measurement devices based on different methods can be achieved.

In the future the miniaturization of the presented measurement devices will be further investigated and pursued.

Acknowledgment

The projects “Development of a measurement device for the determination of the time and frequency dependent grid impedance on the medium-voltage level” and “Development of a measurement device for the determination of the frequency dependent grid impedance on the high voltage level up to 110 kV for the assessment of the availability of grid capacities as a system parameter for the dimensioning of energy storages” was/is funded by the German Federal Ministry for Economic Affairs and Energy under the support codes 0325049 (medium-voltage level) and 0325562 (high-voltage level).

References

- [1] H. Langkowski, M. Jordan, T.T. Do, D. Schulz, “Spectral Grid Impedance Identification on Different Voltage Levels – Challenges and Realization,” 2017 IEEE PES General Meeting, July 16-20, 2017, Chicago, 2017.
- [2] D. Schulz, *Power Quality – Theory, Simulation, Measurement and Assessment* (in German), VDE-Verlag, Offenbach (2004), ISBN 3-8007-2757-9.
- [3] H. Langkowski, M. Jordan, T. Do Thanh and D. Schulz, “Grid impedance determination – identification of neutral line impedance,” 21st International Conference on Electricity Distribution CIRED, Frankfurt, 6-9 June 2011.
- [4] H. Langkowski, T. Do Thanh, M. Jordan and D. Schulz, “Grid Impedance Identification considering the Influence of Coupling Impedances,” *IEEE International Symposium on Industrial Electronics (ISIE 2010)*, Bari (Italy), 4-7 July 2010.
- [5] M. Jordan, H. Langkowski, T. Do Thanh and D. Schulz, “Frequency Dependent Grid-Impedance Determination with Pulse-Width-Modulation-Signals,” *IEEE 7th International Conference-Workshop Compatibility and Power Electronics CPE 2011*, Tallinn (Estonia), 1-3 June 2011.
- [6] K. Heuck, K.-D. Dettmann, and D. Schulz, “Elektrische Energieversorgung”, Vieweg, 9th Edition, Wiesbaden, 2013.
- [7] BDEW, “Technical Guideline for Power Generation Plants at Medium-Voltage Grid”, (in German), BDEW, Berlin, 2008.
- [8] DIN EN50160, *Characteristics of the Voltage in Public Electrical Power Supply networks*, (2011).
- [9] VDN, *Technical Guideline for EEG Power Generation Plants at High and Highest Voltage Grid* (in German), VDN, Berlin (2004).
- [10] H. Langkowski, T. T. Do, and D. Schulz, “Grid Impedance Determination – Relevancy for Grid Integration of Renewable Energy Systems”, IEEE IECON-09, Porto (2009).
- [11] M. Liserre, R. Teodorescu and F. Blaabjerg, “Stability of Photovoltaic and Wind Turbine Grid-Connected Inverters for a Large Set of Grid Impedance Values”, *IEEE Transactions on Power Electronics*, Vol. 21, No. 1, Jan. 2006.
- [12] R. Langella and A. Testa, “A New Method for Statistical Assessment of the System Harmonic Impedance and of the Background Voltage Distortion,” *9th International Conference on Probabilistic Methods Applied to Power Systems IEEE*, Stockholm, Sweden, June 2006
- [13] A. Knop and F. W. Fuchs, “High frequency grid impedance analysis with three-phase converter and FPGA based tolerance band Controller,” *Compatibility and Power Electronics (CPE’09)*, Badajoz, May 2009
- [14] C. Xie, S.B. Tennakoon, R. Langella, D. Gallo, A. Testa, and A. Wixon, “Harmonic impedance measurement of a 25 kV single phase AC supply system,” *Proceedings of Ninth International Conference on Harmonics and Quality of Power*, Orlando, vol. 1, pp. 214-219, 1-4 Oct. 2000.
- [15] M. Nagal, W. Xu, and J. Sawada, “Harmonic impedance measurement using three-phase transients,” *IEEE Transactions on Power Delivery*, vol. 13, pp. 272-277, January 1998.
- [16] M. Jordan, T. Do Thanh, H. Langkowski and D. Schulz, “Strategies for Frequency Dependent Grid Impedance Measurement at the Medium- and High-Voltage Level,” *5th International Ege Energy Symposium and Exhibition (IEESE-5)*, Denizli, 27-30 June 2010.
- [17] M. Jordan, F. Grumm, H. Langkowski, T. Do Thanh and D. Schulz, “Online Network Impedance Identification with Wave-Package and Inter-Harmonic Signals,” *XII International School on Nonsinusoidal Currents and Compensation 2015 (IEEE)*, Lagów (Poland), 15-18 June 2015.
- [18] H. Langkowski, M. Jordan, T. Do Thanh and D. Schulz, “Grid impedance determination – identification of neutral line impedance,” *21st International Conference on Electricity Distribution CIRED*, Frankfurt, 6-9 June 2011
- [19] T. Do Thanh, T. Schostan, T. Dettmann and D. Schulz, “Nonsinusoidal Power Caused by Measurements of Grid Impedances at Unbalanced Grid Voltages,” *Proceedings of the IEEE conference of the International School on Nonsinusoidal Currents and Compensation ISNCC*, Lagow, 10-13 June 2008.

A Wearable Exoskeleton Rehabilitation Device for Paralysis – A Comprehensive Study

Ahmed Roshdy, Samer Al Kork*, Sherif Said, Taha. Beyrouthy

College of Engineering and Technology, American University of the Middle East, Kuwait

ARTICLE INFO

Article history:

Received: 20 December, 2018

Accepted: 27 December, 2018

Online : 20 January, 2019

Keywords:

Rehabilitation

EEG

Prosthetic

Emotiv headset

Paralysis

ABSTRACT

As the technology grows scientists and engineers are trying to combine their work to compensate some body parts that is lost. Prosthetic devices grabbed the attention of most of the doctors and engineers working on solution for lost body parts. Generally prosthetic devices are either external wearable devices or internal ones. Such devices may depend on a built on microcontroller or the brain signals from the patient himself. Although it is used for lost body parts it can also be used for rehabilitation, power assistance, diagnostics, monitoring, ergonomics, etc. The currently used devices usually have the disadvantages of big size and high cost. The suggested device in this paper is targeting the design of portable rehabilitation device with light weight and low cost for paralyzed hand people. The suggested device allows the user to train the patient's hand or perform some needed exercises for his impaired hand. This helps the user to restore the normal hand movement and functionality. The device includes two modes of operation to be chosen by the user through the platform built on a microprocessor which can help controlling the exoskeleton to perform the needed exercises or tasks. Collaboration with several healthcare organizations will be considered to verify or test the effectiveness of this exoskeleton.

1. Introduction

This paper is an extension of work originally presented in BioSMART, the 2nd International Conference on Bio-engineering for Smart Technologies” titled ‘A wearable rehabilitation device for paralysis’ [1]. Diseases like stroke may lead to the loss of some parts of the human body and some other people are actually suffering of paralysis, broken bones, spinal cord injury, hemiplegia or traumatic brain injury. People facing such diseases or injuries are actually facing some problems in their day to day activities because of the infected or injured body part. One of the suggested solutions for such cases is the rehabilitation therapy to improve the infected part movement and regain the strength and power of that part. Some of the physiotherapy exercises are needed for the recovering rehabilitation process for the patient to be able to completely or partially restore the normal movement of the paralyzed body part [2]. Although this process is needed for hands or legs rehabilitation, this paper is actually focusing on a device to be used for the hands’ problems and injuries. Humans are using their hands for writing, touching, holding, squeezing, and so many other functionalities. The partial or total loss of this part of the human body means losing so many valuable functions that any human being needs. The full hand with all its many joints act as natural motors to perform routine daily tasks [2]. This paper is meant to target a rehabilitation therapy system for patients with hand disabilities like too weak hands or even paralysis. The

suggested device compared to the existing devices is considered to be low cost, lightweight, customizable, programmable and safe for human practice. All the functions of the device should be running in two different rehabilitation modes of training called active and passive. The design process will move into four phases, which are human interface, processing unit, health care sensors, and display. Each one of them has special task to do. Human interface will be responsible of the interface between the user and the processing unit. Processing unit will transmit commands to the electromechanical part. Electromechanical part will control the movement of the Exoskeleton. Health 10 care sensor will be used to monitor the user’s health to insure the safety. The results will be shown in the display part.

The paper is going through literature review for common wearable devices available in the market showing the mode of control and the controlling method of each one. Also the literature review is covering the different brain waves’ readers as well as the brain waves and the motor controls on the device itself. At the end the proposed wearable device is discussed mainly in hardware with a comparison to the available devices.

2. Literature Review

2.1. Wearable Device

Here in this section is the comparison between the main projects already exist as rehabilitation wearable devices and assistive exoskeletons. In Table 1 the comparison shows the difference in the modes and the methods used in some projects

*Samer Al Kork, American University of the Middle East, +965 2225 1400 Ext. 1732, samer.alkork@aum.edu.kw

www.astesj.com

<https://dx.doi.org/10.25046/aj040103>

that were built for rehabilitation and assistive purposes for either hand or leg usage. Most of the projects are based in using either Virtual Reality (VR) or electromyography (EMG) as an interface for the device. None of the currently used devices has built the system using Electroencephalography (EEG) interface for the wearable device like what is offered in this study. The main problems for the devices already exist in the market are the high cost and the heavy weight because of the heavy materials used in their fabrication [4]. Although there already exist so many devices but only few of them offers the two modes of control passive and active. Herewith the study is based on EEG interface with the two possible modes of operation as well as keeping a light affordable weight and low cost.

Table 1 Some Applications of Wearable Devices

Title	Control Modes	Methods
Design and development of a hand exoskeleton for rehabilitation following stroke [4]	Active	-
Hand Rehabilitation Support system Based on Self-motion control, with Clinical case report Error! Reference source not found.	Active	VR
Current Hand Exoskeleton Technologies for Rehabilitation and Assistive Engineering Error! Reference source not found.	Passive	VR
Design and Development of a Hand Exoskeleton Robot for Active and Passive Rehabilitation [7]	Active/Passive	EMG
An EMG-driven exoskeleton hand robotic training device on chronic stroke [8]	Active	EMG
A review of technological and clinical aspects of robot-aided rehabilitation of upper-extremity after stroke [9]	Active	VR

Hong Kong polytechnic University designed a portable set of robotic hand exoskeleton that can be used or carried anywhere to help stroke patients in opening or closing their hands [8]. It works in an active mode and has 2 degree of freedom for each finger at the MCP and PIP. The robotic hand consists of Velcro straps to hold the hand in place, 5 linear actuators for fingers and a palm support platform. Finger assemblies are used to provide finger's flexion and extension. Also, it is designed in a way that can be used for different finger length. Also, there is an embedded controller that handles the robotic hand various tasks and monitor the EMG signals (sensor) that are used for the closing and opening process of a hand. Moreover, the set contains a wireless remote control system that helps the therapist to configure and select from different training modes (Figure 1).

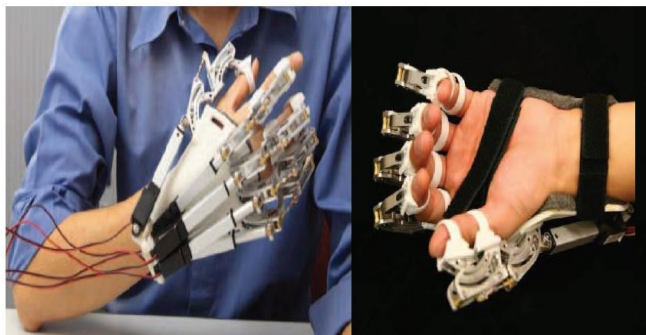


Figure 1. The Set of Hand Exoskeleton and Overview of the Prototype

University of Technology in Sydney Designed and developed a hand exoskeleton for rehabilitation following stroke [4]. The device is achieving full flexion/extension motion of the five fingers of the left hand (impaired hand) based on the motion of the identical digits of the right hand (healthy hand). The device is less than 2 Kg and they chose Aluminum because of its lightweight. The hand exoskeleton has 15 degrees of freedom (DOFs). However, the hand exoskeleton cannot perform abduction/adduction movement; as a consequence, more work needs to be done on the device (Figure 2).



Figure 2. Hand Exoskeleton for Rehabilitation

Kyushu University proposed a hand exoskeleton using three-layered sliding spring mechanism [10]. The idea of this paper is to present a lightweight and compact device to easily use, so under-actuated mechanism is used to reduce the mass and size of the device by limiting the number of actuators into one (Figure 3).

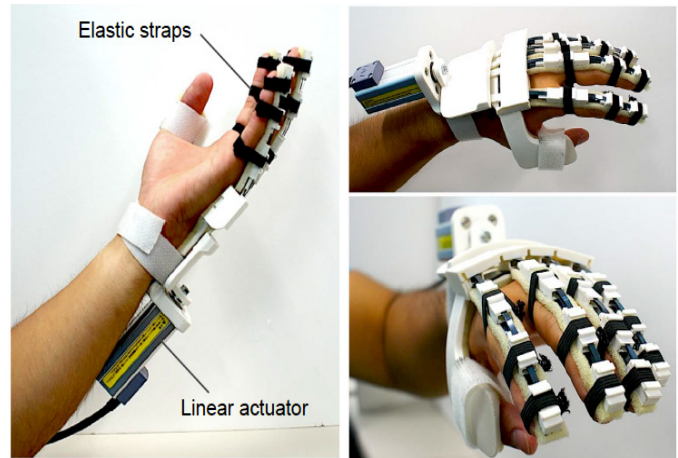


Figure 3. Overview of the Hand Exoskeleton

The weight of the exoskeleton is 320 g. Each finger has three DOF, which is flexion/extension, but the thumb is fixed for the sake of robust grasping. Therefore, the three DOF is actuated through one actuator. as a result; the four fingers will work simultaneously. Three-layered sliding spring consists of 3 springs that is divided into inner (Si), center (Sc) and outer (So) springs and rigid bodies into tip (Rt), inner (ri) and outer (Ro) parts. The mechanism helps to perform flexion motion (Figure 4).

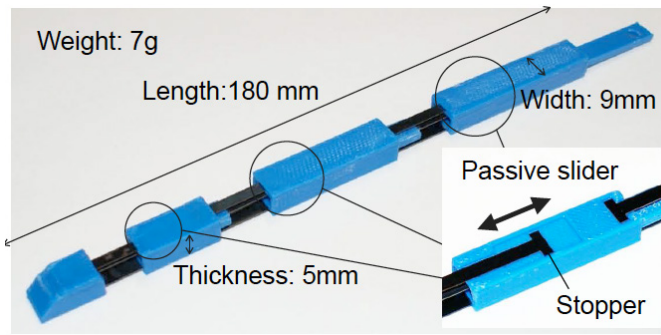


Figure 4. Three-Layered Sliding Spring Mechanism

University of Salford developed a hand assistive exoskeleton that operates in the active mode [11]. It uses virtual reality exerciser to perform the physical therapy exercises. It allows the patient to do the therapy exercises through fun interactive games. This exoskeleton allows hand motion, analysis and recording easily. The exoskeleton has 7 active DOF (Figure 5).

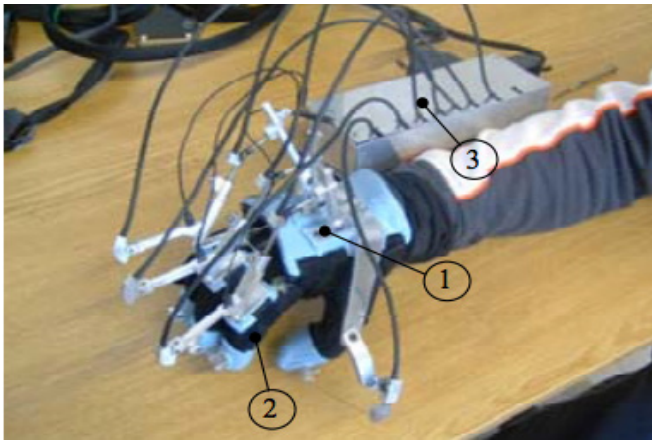


Figure 5. Assistive Exoskeleton

A team at Twenty university develop a low cost portable hand exoskeleton for assistive and rehabilitation in 2016 (Figure 6). They developed it to be for active rehabilitation, cheap, wearable, and portable [12]. This hand controlled by the mussels, they used EMG method. It can be for different sizes. They used 3D printer for some damages exoskeleton's components. The aim of this robotic hand exoskeleton is to assist persons with hand opening disabilities.

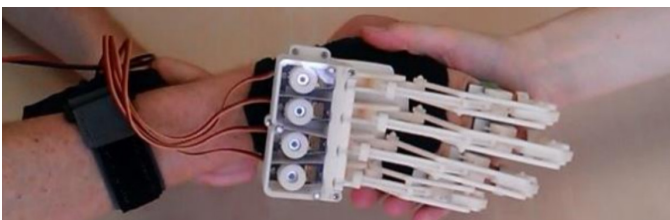


Figure 6. Assistive/Rehabilitation Exoskeleton

2.2. EEG

EEG stands for "Electroencephalography". EEG is a technique used to screen and record electrical signals produced inside the brain (Figure 7). There are two EEG techniques, which are non-invasive and invasive. In the non-invasive technique,

www.astesji.com

electrodes are utilized to record the signals, and they are attached directly to the human head. On the other hand, the invasive technique records the signals by embedding electrodes inside the skull itself till the brain. In addition, there are two different types of communication, the first one is the BCI (brain computer interface), and the second one is BMI (brain machine interface). These two types define the communication method or interface between the brain and the device meant to record the brain signals. Relating the mentioned two communication methods to EEG techniques the BCI communication type follows the non-invasive technique while the BMI type follows the invasive method. Using the EEG will allow the user to control the motors by his own brain signals. The brain signal captured by the EEG is to be sent to set of motors to perform the needed movement or action. These days, this technique has turned out to be affordable and available to general society. Many different products already available in the market are using EEG in order to monitor the brain activity.

EEG Signal Collection

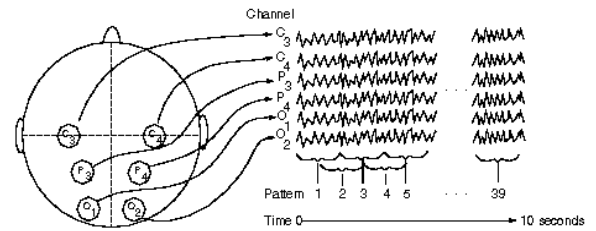


Figure 7. EEG electrodes record signals from brain

The brain is a very complex and essential organ that its main function is to control the whole body. In general, the brain anatomy or structure is consisting of three main sections these are forebrain, midbrain and hindbrain. The forebrain is classified into three parts, which are thalamus, cerebrum, and hypothalamus. The cerebrum represents the largest part of the brain. It is divided into 4 portions called lobes, which are frontal (F), parietal (P), temporal (T), and occipital (O) lobes.

Each area or region has its own function. The frontal lobe is responsible for problem solving, speaking, judgment, emotional expression, thinking, planning and movement. Parietal lobe is more into processing sensory and interpreting visual information (reacting into environment); it allows sensation from muscle and skin, body orientation, and reading. Temporal lobe is associated with behavior, memory, hearing and understanding language. Moreover, the occipital lobe is used for color perception, image recognition and vision or sight (Figure 8).

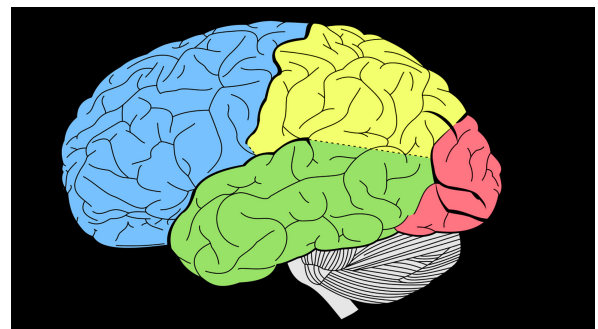


Figure 7. The Main Portions of the Brain

2.3. Emotiv (EPOC+) and the 10-20 System

EPOC follows the 10-20 system electrode placements which is considered to be an international system [13]. This system depends on 10% or 20% separation between the electrodes. It permits 21 electrodes on the surface of the scalp shown in Figure.. Each possible location for the electrodes is described by a letter to identify the lobe and a number to define the location of the hemisphere (left or right side of the brain). It shows the 14 channels of EPOC are distributed as frontal (AF3, AF4, F7, F3, F4, F4), front-central (FC5, FC6), parietal (P8, P7), temporal (T8, T7) and occipital (O1, O2) lobes [14]. The odd numbers refer to the left side of the brain and the even numbers refer to the right side of the brain. The system renamed some locations or points, which are T4, T3, T5 and T6 to be T8, T7, P7 and P8 respectively. Also, the AF and FC are intermediate sites; AF is between Fp and F but FC is between F and C. Moreover, A1 and A2 are the locations of earlobes, DRL and CMS correspond to P3 and P4, which are the reference sensors (Figure 9).

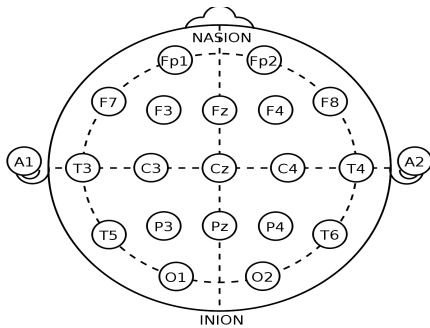


Figure.9. Locations of the reference sensors

2.4. The Motor Area

As the different brain parts are generating different types of waves for each task needs to be performed by the human body, the headset. The primary motor cortex, which is located in the frontal lobe, is responsible for controlling the execution of movement [15]. In fact, it is located in an area called Precentral gyrus. This part of the brain is participating in controlling the movement of different parts of the body like arm, hand, face, foot, etc. (Figure 10). According to the 10-20 system brain map, C3, Cz and C4 are the nearest to the location where motor execution occurs [15], [16]. Unfortunately, the three motor locations aren't available in the EPOC headset. However, there is an easy way to obtain one of the three points; the headset can be tilted a little bit till pointing one of the electrodes to one of the needed three locations.

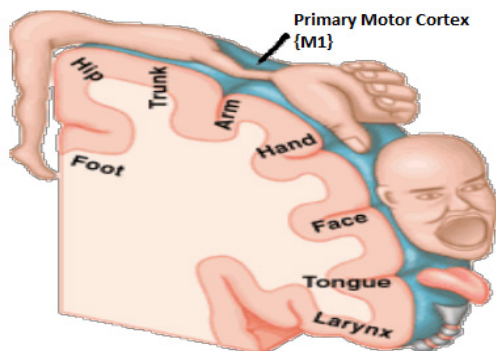


Figure 10. The Location of the Hand Movement in the Motor Cortex

2.5. Brainwaves

The brain signals are different according to the band of frequencies of each one of them. The different frequencies or to be more precise the different band of frequencies are covering all the brain activities whether the human body is moving or not. Each wave generated by the brain is does have particular location to be produced from. The main 5 range of frequencies from the brain are Delta, Beta, Theta, Alpha, and Gamma (Figure 11). Each band has a specific frequency range but a different meaning.

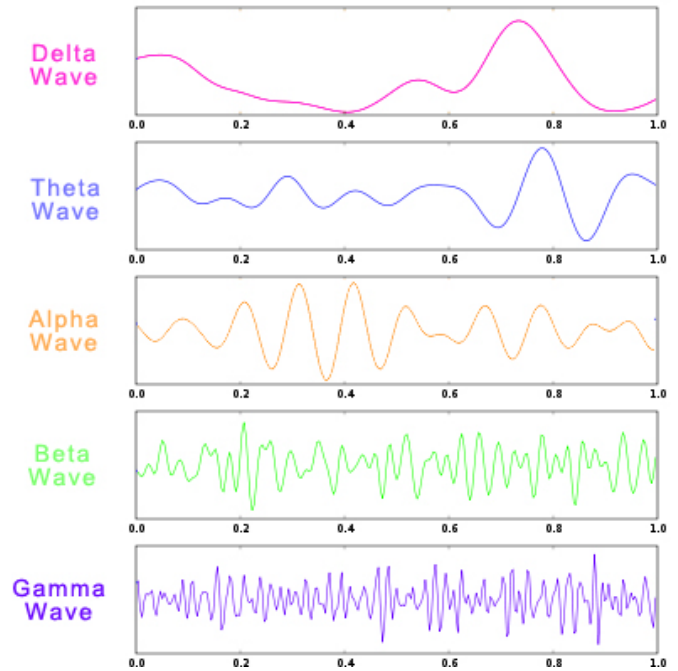


Figure 11. The Main 5 Brainwaves

The first range of frequencies is from 0.5 to 3 Hz and considered to be the Delta range. This range of waves are created while sleeping or relaxing meditation sessions. Also in our scope of work the Delta waves are generated during the normal human body healing during a deep sleep. The next range of waves is the Theta waves and it's in the range of 4 – 8 Hz. Usually these kind of waves generated while sleeping as well or laziness. Third range for the Alpha waves is from 4 to 12 Hz and the got generated by the occipital, parietal and frontal lobes in some cases like closing the eyes and by the frontal lobe during relaxation. Out of the Alpha waves range there is a range from 8-12 Hz is called Mu waves. The Mu waves mainly corresponds to the body movement. In other words, the Mu waves are generated when the human body is relaxed. The next range from 12 to 25 Hz is called Beta waves. Beta waves are generally created in body movements and brain activities like thinking or problem solving. Above 25 Hz signals are called Gamma signals and they are faster than the other waves as noticed from its frequency. Such signals are usually generated with excessive brain work during multi-tasking.

2.6. Virtual Reality

Virtual reality represents an alternative way to be engaged or involved in the rehabilitation therapy routines and not easily get bored by using the rehabilitation gaming system [18],[19]. The rehabilitation gaming system is a virtual reality tool in a three-

dimensional world that let the patient does the therapy exercises through fun interactive games (Figure 12). Also, it can be used as a stimulus to help the patient in the motor imagery process. The VR technology is currently a very up-to-date trend. It is quickly moving towards the mainstream that even big companies are investing in this technology like Samsung, Sony, Apple, etc. They are developing their existing devices to follow this new technology.



Figure 12. Virtual Reality Tool in a Three-Dimensional World

2.7. Electromyography (EMG)

Electromyography EMG measures electrical current inside muscles (Figure 13). EMG is controlled by nervous system. It gains noise while moving through different tissues. Specific Algorithms should be used to reduce the noise in order to obtain accurate EMG signal [20]. Many applications have been implemented using EMG method for exoskeleton hand control; they used muscle signals to control the impaired hand which is driven by an intact organ [8].

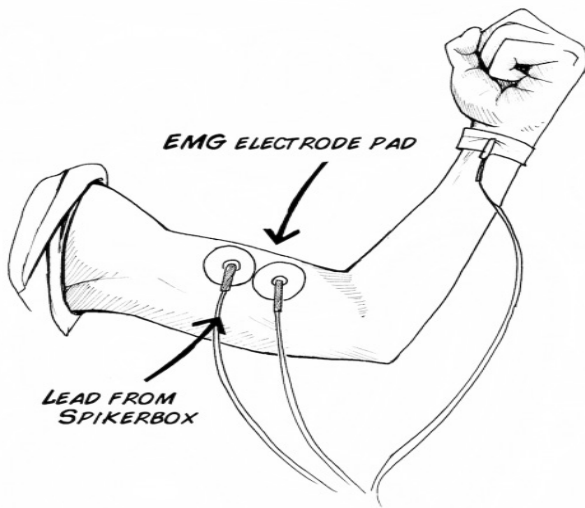


Figure 13. EMG Signals Comes From Muscles

3. Detailed design

3.1 Design Overview

The whole design is described in Figure 18 showing the whole design different stages through the design. In the active mode the EEG headset starts to read the brain signals and detects

the location of the signal and decide the lobe and the side of the brain sending the signal. Next step the headset will wirelessly communicate with the computer through Wi-Fi to start the analysis. After the analysis is done the signal should be compared to previously recorded data to indicate the meaning of the brain signal based on pattern recognition. In the Preprocessing unit, the signal will be filtered. Then the signal goes to the second part where features and specific characteristic will be detected. The final stage the signal will go through is the classification where the imagined movement will be identified and performed. After signal processing, the information taken from the headset will be compared to the previous saved data. Using pattern recognition if the comparison got a match, then it will send the data through Bluetooth ZigBee module to the electromechanical parts to start the movement. This part will then give the instructions to the servo motors to start performing the needed movement. This part includes two big servos that ensure the movement of the arm up and down (Figure 14).

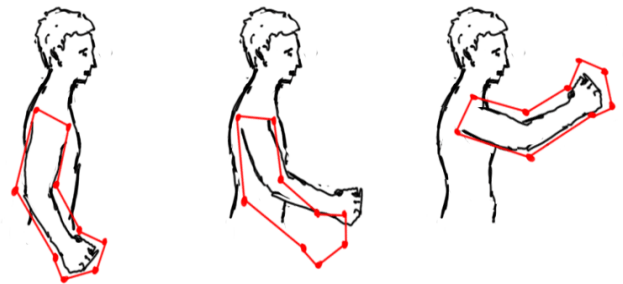


Figure 14. The movement of the arm

The health sensors part includes a body temperature sensor, which allows the user to measure his body temperature [20]. It includes also a pulse oximetry sensor that indicates the arterial oxygen saturation of functional hemoglobin. An LCD screen is present to show the results of some sensors like the body temperature sensor and it is considered to be a major part of the interface between the user and the device. Wi-Fi is used to send all the results to the mobile application. These results will lead to activating or deactivating the mechanical parts (motors). The schematic of the Robotic-Based Rehabilitation system with all its basic parts is shown in Figure 15.

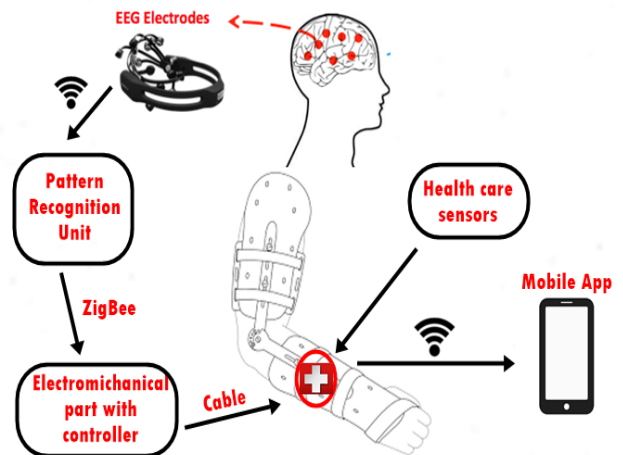


Figure 15. Design overview



Figure 16. Design schematic

3.2 System Design

The overall design and flowchart is represented in Figure 16 and Figure 17. The user enters his/her age. Then, the temperature, ECG, and Pulses of the user are displayed on the LCD. If the values are in the normal ranges, the system continues running, otherwise, it stops. There are specific ranges for each one of them. The normal body temperature tends to be between the range 36.1°C to 37.2°C. On the other hand, the normal human pulses can be classified according the ages to be 40 to 60 pulses for ages between 1 to 10 years and 60 to 100 pulses for ages between 10 to 60 years. Therefore, these things will provide extra security for the user.

Second, the user picks the required mode through the built on switch (left for active, middle for off, right for passive). If the user selected the first mode which is the passive mode, the user will be able to select the needed exercise through the keypad connected to the Arduino microcontroller. After that, the servo motors will start performing the corresponding movements based on the code on the Arduino board. The motors are connected through rods to the wearable device parts like fingers and rest to perform the right movement.

The other mode of operation is the Active mode that can be selected by moving the switch to the left. In the Active mode the movements and therapy will be controlled by the brain signals read by the EEG headset. When the headset reads the brain signal it will pass it to the Arduino for analyzing to recognize the required move Figure 18. In the beginning the signals are filtered whenever read by the EEG to get rid of the noise signals through bandpass filter and notch filter used particularly for 50Hz signals that will be there most of the time from the surroundings and power sources around the patient [14]. After having the clean signal, the next step is getting the information from the signal and translate it into movements or exercise through the motors connected to the wearable device. The signals movement translated orders are saved in a features vector to be performed. The extracted information is classified in order to help in extracting the features into spatial and spectral information. The location of the electrode extracting the signal is considered to be

spectral type of information. On the other hand, spectral information represents the power of the frequency bands. At the end a computer will classify the signals to be able to select the movement. There are many classifiers; however, neural network and SVM are most commonly used [23][24],[14],[26]. The classifier will then save the obtained data in the feature vector with either movement to class A or class B. Comparing these data with the database stored from pre imagined movements will lead in case of matching to moving the signal to Arduino as the brain for the servo motors to perform the exercise and move the device.

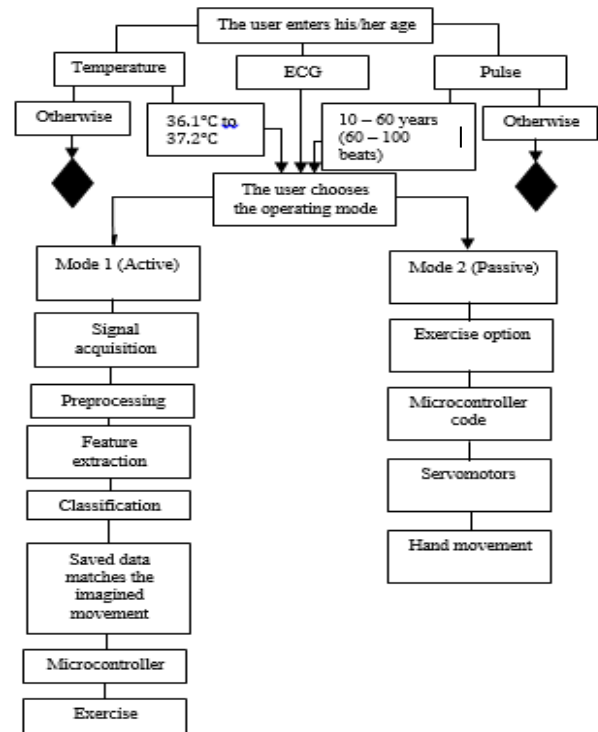


Figure 17. Flowchart of the system design

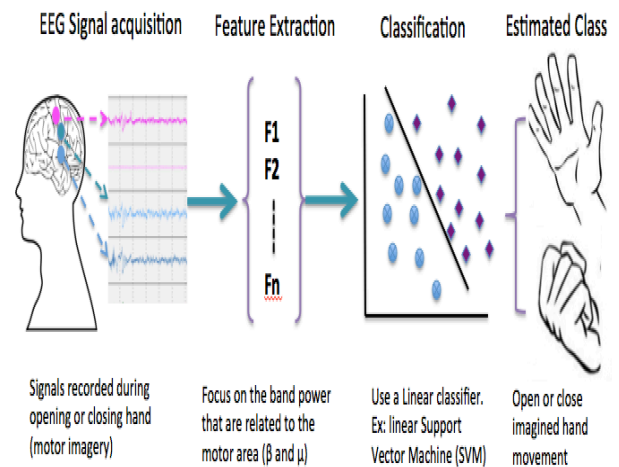


Figure 18. Signal processing steps [21][24]

3.3 Relevant Engineering Application and Calculations

Servomotors was chosen for this project. Four servo motors in two separate joints are needed for each arm. To check the loads

on the motors and make sure about the safety of the human body, it was designed to allow the device to hold weights up to 10 kg. This amount of weight causes the torque on the servo motors while moving. The angle of movement should also be taken into considerations to have the right calculations of the motors mechanical loads. The torque on the servomotors according to the applied force will be calculated using equation (1)

$$\tau = F \times r \times \sin\phi \tag{1}$$

Where τ is the applied torque, F is the force cause by the weight carried ($F = m \times g$), r is the distance, and ϕ is the maximum angle of movement which is 90° . The distance used in our design is 0.74m and the mass of the arm is 3.628kg, and the gravity in Kuwait is 9.793N/A. The resulting torque is 26.29N.m. For such torque and safety, the servomotors will be chosen accordingly. From equation (1) there is a direct proportional relation between the torque and the applied force. Taking into considerations the power consumption, the no load power can be calculated using equation 2 knowing that the maximum current used with no load is 500Ma.

$$P = V \times I = 25 \times 0.5 = 12w \tag{2}$$

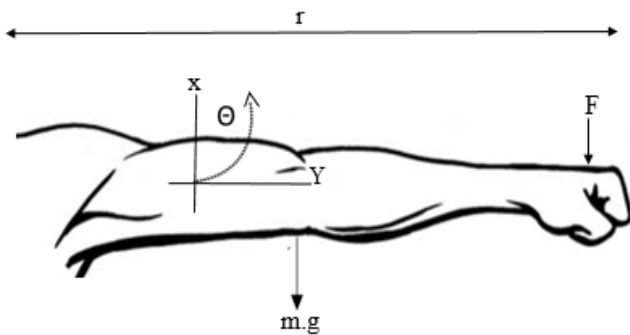


Figure 19. Power Consumption Parameters

3.4 Materials

For the process of building the hand exoskeleton device, so many materials need to be studied in terms of weight, safety, and flexibility in order to choose the suitable one for the project. Iron, aluminum and 3D printed material which is mainly plastic-based are different materials that can be used for exoskeleton. Each one of them has different properties. Iron is heavier than aluminum because of its density, which is 7.1 g/cc. Moreover, Iron is cheaper than aluminum because it does not need many processes to obtain as the aluminum. It is stronger than aluminum. Iron's strength is 169 (MPa) whereas aluminum is 80 (MPa). Aluminum takes time to be manufactured. The features of 3D printer materials are: cheap, flexible, durable, lightweight and available with varies colors. Easy to redesign and model if needs to do any changes. For example; if something is broken we can have reprinted it easily. Furthermore, it is not taking much time to print. The materials that should be used must be environment friendly to provide safety for a long period time. Aluminum is not environment friendly. The releasing of perfluorocarbons during the aluminum smelting process are 9,200 times more harmful than carbon dioxide. Iron is not recommended due to the chemicals which enters into the environment and affect the eco balance, and

cause so many health problems. 3D printer materials like Acrylonitrile Butadiene Styrene (ABS) and Polylatic Acid (PLA) can be considered as environment friendly materials because it is easy to recycle, no waste of materials, and no more unsold product. The comparison is summarized in Table 2 &

Table 3

Table 2. Materials' Specifications

Materials	Price per kg (\$)	Strength (Mpa)	Density	How easy to customize
Iron	0.049 - 0.098	169	7.1 g/cc	Difficult
Aluminum	0.307 - 1.221	80	2.68 g/cc	Difficult
ABS	29.30	25-50	1.01-1.21 Mg/ m ³	Easy
PLA	29.30	36-55	1.25 Mg/ m ³	Easy

Table 3. Advantages & Disadvantages Comparison

Materials	Advantages	Disadvantages
Iron	<ul style="list-style-type: none"> Strong Durable 	<ul style="list-style-type: none"> Heavy Weight Expensive Hard to recycle Not environment friendly Rusts
Aluminum	<ul style="list-style-type: none"> Light Weight Durable Won't Rust Doesn't deteriorate 	<ul style="list-style-type: none"> Not very strong Expensive Hard to recycle Not environment friendly Can easily be water stained
3D Printer Materials	<ul style="list-style-type: none"> Light Weight Cheap Easy to recycle Environment Friendly Flexible Durable Won't Rust Doesn't deteriorate 	<ul style="list-style-type: none"> Not very strong

3.5 Electromechanical System

1) Battery:

For a battery, there are different kinds of batteries and each of them has its own features. First of them is Lithium ion Battery which has very high capacity, normal size, long life, high cost and it is rechargeable. Second one is NiMH which has a small size, low capacity, short life, low cost and non-rechargeable.

Third one is NiCd has big size, low capacity, short life, low cost, and rechargeable. Final one is SLA-6v20 has very big size, very high capacity, long life, high cost, and rechargeable. Energizer NH15BP and Duracell have short life and low capacity so they will not benefit this project well. Lithium ion and SLA-6v20 both are good, but we will choose Lithium ion because its size is smaller. Batteries comparison shown in Table 4.

2) Actuators:

Actuators are mechanical devices that move or turn energy into motion. There are different types of actuators, which are electric

actuators, hydraulic and pneumatic actuators. Electric actuators transfer electrical energy into mechanical energy. There are two types of electric actuators which are direct current motor (DC motor) and servomotors. Brushless DC motor is a type of DC motor that has different properties. It is small in size and weight not to mention its high speed and torque. On the other hand, a servomotor is a type of actuator that its functionality is to control motion. It operates within the limit of the specified angle (angular precision) by receiving a control signal to take further action. Moreover, it uses closed-loop feedback to control speed, torque or position. Pneumatic actuator is a hollow cylinder inside it a piston. The piston moves by applying a pressure from a pneumatic pump (compresses air) to create a force. Hydraulic actuators work similarly to Pneumatic actuators; except for the fact they are driven by liquid (fluid) pressure instead of air pressure. Pneumatic actuators are more expensive than hydraulic actuators. Moreover, hydraulic actuators are stronger or have greater force to move heavy loads.

Table 4. Batteries Types

Specifications	Li-Ion			NiMH	NiCd
	Cobalt	Phosphate	Manganese		
Energy (Wh/Kg)	150-250	100-150	90-120	60-120	45-80
Life Cycle (80% DoD)	500-1000	500-1000	1000-2000	300-500 ³	1000 ³
Charge Time (H)	2-4	1-2	1-2	2-4	1-2
Cell Voltage(v)	14.8	3.7	3.2-3.3	1.2	1.2
Safety Requirements	Protection circuit mandatory			Fuse protection, thermally stable	
Cost	High			Moderate	

3.6 Programming

3) Processor

Nowadays, the advancement in the processors and controller’s technology has made it affordable and possible to achieve most of project requirements. In today’s market, the top leading and the most common three processors/controllers are Raspberry pi, Arduino and Intel Edison. Raspberry pi considers as a tiny computer that has its own operating system. It is Linux based operating system and it can multitask. The board includes a processor, Ram, USB ports and everything a normal computer has. There are four models or generations of raspberry pi, which are Raspberry Pi 1, Raspberry Pi 2, Raspberry Pi 3, and Raspberry zero. Each one of them has different properties and using them depend on the needs. In general, a raspberry pi allows the user to not be restricted or limited to a certain programming language. Also, it is very fast in processing and easy to connect to Internet

On the other hand, Arduino is a microcontroller that consists of software and hardware. It is a board that contains a chip to be programmed and use it to perform several functions. It is flexible that can interact with different devices like LEDs, buttons, speakers, motors, and cameras. For instance, it can read data from sensors, get to blink an LED, etc. There are several types of Arduino but the most popular one is the Arduino UNO. Most Arduinos have the same components, which are a power source, a processor, digital and analog pins, a reset button and a USB port. Arduino is easier to work with and it is better in controlling than Raspberry pi. It is capable of controlling complex external

hardware. Also, it is cheaper than Raspberry pi. There is also another type of microprocessor that considers as a competitor to raspberry pi. It is offered by Intel Company, which is called Intel Edison. It functions as a computer. The board has a processor, RAM, USB port, Bluetooth and WiFi. It is a little bit similar to Raspberry pi but each one of them has its own features. For example, Intel Edison is more expensive than raspberry pi. Also, it has less numbers of USB ports and the processor speed is lower than the one in raspberry pi.

On a separate note, microcontrollers and microprocessors show different weaknesses and strengths depending on the application. There is a difference between microcontroller and microprocessor and each has advantages and disadvantages. In general, microcontrollers are more suitable for controlling devices (such as servos, etc.) whereas the microprocessors are better in processing

4) Interface

There are so many user interfaces to help the user interact with the computer to detect and analyze specific information, or to get some feedback from it. EEG is one of the user interfaces that measured by using electrodes that is placed on the scalp. It can be recorded by the electric fields that are generated by the nerve cells in the brain. EEG has so many advantages such as the characteristic of the electrical recording system because it has high precision and time measurements. Another advantage is that EEG is a very inexpensive device, and can easily be operated with. Poor recording spatial resolution is one of the disadvantages of the EEG. EMOTIV provides two different headsets, which are EPOC+ and Insight. Each one has different properties or features. The main differences that will serve our needs are related to the number of sensors and signal resolution. The EPOC features 16 sensors plus 2 reference sensors. Also, it provides high resolution. On the other hand, the insight has 5 sensors plus 2 reference sensors and it is less accurate than EPOC when detecting the signals (low resolution).

The second user interface is virtual reality. It is an artificial environment that is made by software and presented by the user in such a way that makes it real. Virtual reality also has so many advantages and disadvantages. One of the most important advantages is that the disabled people that are not able to experience reality can explore the virtual world, and experience the full life there. It also allows the user to experience impossible things in real life. One of the biggest drawback of the virtual reality is that people might get addicted to that virtual world which will lead them to forget their responsibilities in the real life.

The third user interface is the electromyography (EMG). It is way to access the health muscles and nerves to control and interact with it. EMG signals can be detected or transferred through a small device called electrodes. It has some advantages and disadvantages. EMG is more accurate, safe, and easy to get the wanted results. Also, EMG signals are not as complicated as the EEG when it comes to finding locations for measuring. However,

noises may occur due to the distance between the user and the signals of muscles.

3.7 Communication

The first category to be chosen is to have wireless communication to give the user the freedom to move while using the device. The Wifi, Bluetooth, and XBEE (pronounced ZigBee) are three different types of wireless communications which varies in the Safety, range of accessibility, reliability, power consumption, and bandwidth.

Starting with WiFi, its security is lower than the other two types with 20-150 meters range of communication. Although the reliability is low it still consumes high power with its wide bandwidth. Second choice is the Bluetooth which has a bit better safety communication than the Wifi. But the main problem is the short range which is 8-30 meters. Still it has low reliability but low power consumption with a narrow bandwidth. The third option is the ZigBee is offering a good safety communication with a good range of communication coverage 20-150 meters. It does have a high reliability compared to the other two options as well as low power consumption with a narrow bandwidth

3.8 Decision Making and Selections

Based on the project needs and requirements, a 3D printed design was used. The pressure sensor chosen to be LPS25MB because it is cheap and light in weight. Also MS5637-02BA03 was chosen to be the selected pressure sensor for its low cost and high accuracy. For the health sensor Pulse and Oxygen in Blood Sensor (SPO2) was chosen due to its accuracy and fast response. And the servo motor ASMC-03B was selected due to its high torque even it's a bit more expensive than other servo motors. Cobalt Li-ion battery got the highest score because of its suitable cell voltage, long life time, good energy, suitable charge time, low cost and it is safe. For the interface, EEG and EPOC allows obtaining more or variety of locations than insight. Based on the comparison, EPOC is more suitable than insight in terms of accuracy and sensors location. Finally, Arduino was chosen because it has the least price above them all. In case something happened during the coding process and the circuit got damaged we can easily buy another one because of its low price. Also, it can help us easily in our project since it has so many libraries and control system to control the exoskeleton.

4. Results and Analysis

The servo motors are responsible for the fingers movement according to the selected exercise. Based on the user selection the servo motors will perform the selected task. In other words, if the user chooses one the Arduino will give the corresponding signals for the motors to perform exercise one which is opening the hand five fingers then close them twice with a delay five seconds after each open or close process. The second exercise that might be selected by the user also is to open the fingers and closing them five times with the same delay time in between. The third and last exercise is opening and closing the fingers nine times. Beside this the medical sensors such as the temperature sensor is helping the assisting the user continuously check his body temperature. Another medical sensor mounted on the device is the pulse oximetry sensor to monitor the arterial oxygen saturation in the hemoglobin. Also a regular sixteen characters two lines LCD is

used to show the measured and monitored data for the user with a Wi-Fi connected between the wearable device and the mobile phone application to send the collected data. Innovator X Post-Op Elbow was used as an angular controller. It provides safety to the user due to the adjustments of the angle position that we can make to prevent hand broken.

The biggest challenge of the project was to build a pattern recognition system for the brain signals to be able to detect the required movement from the brain itself with any other external method. The problem is that making a wrong pattern recognition may hurt the patient if a wrong move was done if the brain signal was not read correctly. A raw brain signal data coming from the EEG was read and analyzed by MATLAB software. MATLAB included a digital band pass and band stop filters to eliminate the noise from the brain signal.

5. Conclusion

Patients suffering from some kind of hand disabilities are facing a problem in their day to day activities that's why in this paper we've suggested a light weight low-cost wearable device compared to the devices used nowadays. The device designed to assist the user through two different modes of operation. The two modes are active and passive and both are helping the user to perform some hand exercises as well as monitoring his body temperature, pulses, ECG, and oxygen percentage.

Exoskeleton is a wearable device that improves the user's performance. There are two kinds of exoskeleton, which are assistive and rehabilitation. Assistive exoskeleton gives a user an extra strength to do different tasks. Rehabilitation exoskeleton reduces the recovery time and has long-term effect. This project will help many people that suffer from diseases like stroke, broken bones, spinal cord injury, hemiplegia and traumatic brain injury. Hands are so important part of the human body, so we will design a hand rehabilitation exoskeleton for stroke patients. This hand works in two modes: passive and active. It will be free size. This exoskeleton expected to increase the strength of the hand, the chance of healing, decrease the pain, the cost, and the recovery time.

Our design contains mainly four parts, first the EEG part, which will read the signals from the brain and send it to the PRU part through the Wi-Fi. PRU stands for pattern recognition unit that its function is to analyze signals. Signals were analyzed using Matlab. The third part, which is electromechanical part, will take the signals from PRU through ZigBee and control the arm servomotor using Arduino. Electromechanical part also contains five small servomotors to control the fingers to do specific exercises. Health care part contains three sensors, which are temperature sensor, pulse sensor, and ECG sensor and the results will be shown in the LCD. The user also can enter his information and choose any exercise using keypad and LCD. The last part is mobile application, which will take the results from the health care part through the Wi-Fi.

As a suggested future work and improvements the device can be improved by selecting the control mode through the brain signals with the keypad used as an extra input method. Also the keypad with the screen can be replaced by one big touch screen.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to thank the American University of the Middle East that gave us the opportunity to carry out our research. We would also like to thank the Robotics Research center for funding the project providing us such great facilities and assistance throughout our research.

References

- [1] F Sayegh, F Fadhli, F Karam, M BoAbbas, F Mahmeed, JA Korbane, S AlKork, T Beyrouthy "A wearable rehabilitation device for paralysis," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp.1-4.
- [2] A. P. Ong and N. T. Bugtai, "Recent Developments of Robotic Exoskeletons for Hand Rehabilitation," Presented at The DLSU Research Congress 2016, pp. 1-2.
- [3] C. A. Lim 2013, "Ironman Project-Design Of Electro-Mechanical Muscle For Elbow Exoskeleton Robot," 2013.
- [4] Rahman, M. A., & Al-Jumaily, A. Design and development of a hand exoskeleton for rehabilitation following stroke. *Procedia Engineering*, 41, 1028-1034, 2012.
- [5] Kawasaki, H., Kimura, H., Ito, S., Nishimoto, Y., & Hayashi, H. Hand rehabilitation support system based on self-motion control, with a clinical case report. In *2006 World Automation Congress* (pp. 1-6). IEEE. (2006, July).
- [6] Heo, P., Gu, G. M., Lee, S. J., Rhee, K., & Kim, J. Current hand exoskeleton technologies for rehabilitation and assistive engineering. *International Journal of Precision Engineering and Manufacturing*, 13(5), 807-824. (2012).
- [7] Sandoval-Gonzalez, O., Jacinto-Villegas, J., Herrera-Aguilar, I., Portillo-Rodriguez, O., Tripicchio, P., Hernandez-Ramos, M., ... & Avizzano, C. Design and Development of a Hand Exoskeleton Robot for Active and Passive Rehabilitation (2016).
- [8] Ho, N. S. K., Tong, K. Y., Hu, X. L., Fung, K. L., Wei, X. J., Rong, W., & Susanto, E. A. An EMG-driven exoskeleton hand robotic training device on chronic stroke subjects: task training system for stroke rehabilitation. In *2011 IEEE international conference on rehabilitation robotics*(pp. 1-5). IEEE (2011, June).
- [9] Babaiasl, M., Mahdioun, S. H., Jaryani, P., & Yazdani, M. A review of technological and clinical aspects of robot-aided rehabilitation of upper-extremity after stroke. *Disability and Rehabilitation: Assistive Technology*, 11(4), 263-280 (2016).
- [10] Arata, J., Ohmoto, K., Gassert, R., Lambercy, O., Fujimoto, H., & Wada, I. A new hand exoskeleton device for rehabilitation using a three-layered sliding spring mechanism. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on* (pp. 3902-3907). IEEE (2013,May).
- [11] Sarakoglou, I., Tsagarakis, N. G., & Caldwell, D. G. (2004, September). Occupational and physical therapy using a hand exoskeleton based exerciser. In *Intelligent Robots and Systems. (IROS 2004). Proceedings. IEEE/RSJ International Conference on* (Vol. 3, pp. 2973-2978). IEEE (2004).
- [12] Capitani, S., Cremoni, A., Lindenroth, L., Secciani, N., Shafr, A., Stilli, A., & Venture, M. Development of low-cost portable hand exoskeleton for assistive and rehabilitation purposes. For The ENTERFACE International Workshop (pp. 1-3) (2016).
- [13] W. D. Hairston, K. W. Whitaker, A. J. Ries, J. M. Vettel, J. C. Bradford, S. E. Kerick, and K. McDowell, "Usability of four commercially-oriented EEG systems," in *Journal of neural engineering*, 11(4), 046018, 2014.
- [14] abilah Hamzah, Haryanti Norhazman, Norliza Zaini and Maizura Sani. Classification of Eeg Signals Based on Different Motor Movement Using Multi-layer Perceptron Artificial Neural Network. *Journal of Biological Sciences*, 16: 265-271 (2016).
- [15] F. Lotte, L. Bougrain, and M. Clerc, "Electroencephalography (EEG)-based Brain Computer Interfaces," Wiley Encyclopedia of Electrical and Electronics Engineering, Wiley, pp. 44, 2015.
- [16] F. Lotte, "A Tutorial on EEG Signal Processing Techniques for Mental State Recognition in Brain-Computer Interfaces," Eduardo Reck Miranda; Julien Castet. *Guide to Brain-Computer Music Interfacing*, Springer, 2014.
- [17] M. M. Moazzami, "EEG signal Processing in Brain-Computer Interface" (Doctoral dissertation, Michigan State University), 2012.
- [18] S AlAwadhi, N AlHabib, D Murad, F AlDeei, M AlHouti, T Beyrouthy, S Al-Kork "Virtual reality application for interactive and informative learning," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp. 1-4.
- [19] S.A. Awadhi, N.A. Habib, D. Al-Murad, F.A. deei, M.A. Houti, T. Beyrouthy, S. Al-Kork "Interactive Virtual Reality Educational Application", *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 4, pp. 72-82 (2018).
- [20] Reaz, M. B. I., Hussain, M. S., & Mohd-Yasin, F. Techniques of EMG signal analysis: detection, processing, classification and applications. *Biological procedures online*, 8(1), 11-35 (2006).
- [21] S. Said, S. AlKork, T. Beyrouthy and M. F. Abdrabbo, "Wearable bio-sensors bracelet for driver as health emergency detection," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSmart), Paris, 2017, pp. 1-4.
- [22] Hasan S., Al-Kandari K., Al-Awadhi E., Jaafar A., Al-Farhan B., Hassan M., Said S. and AlKork S. (2018). Wearable Mind Thoughts Controlled Open Source 3D Printed Arm with Embedded Sensor Feedback System. In *Proceedings of the 2nd International Conference on Computer-Human Interaction Research and Applications - Volume 1: CHIRA*, ISBN 978-989-758-328-5, pages 141-149.
- [23] S. K. A. Kork, I. Gowthami, X. Savatier, T. Beyrouthy, J. A. Korbane and S. Roshdi, "Biometric database for human gait recognition using wearable sensors and a smartphone," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp. 1-4.
- [24] D. Planelles, E. Hortal, A. Costa, A. Ubeda, E. Iaez, and J. M. Azorin, "Evaluating classifiers to detect arm movement intention from eeg signals," *Sensors*, 14(10), pp. 18172-18186, 2014.
- [25] Taha Beyrouthy, Samer Al. Kork, J. Korbane, M. Abouelela, "EEG Mind Controlled Smart Prosthetic Arm – A Comprehensive Study", *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, no. 3, pp. 891-899 (2017).
- [26] T. Beyrouthy, S. K. Al Kork, J. A. Korbane and A. Abdulmonem, "EEG Mind controlled Smart Prosthetic Arm," 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), Balaclava, 2016, pp. 404-409.

A Practical PIR-based Scheme for Discovering Nearby Places for Smartphone Applications

Maryam Hezaveh*, Carlisle Adams

School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada

ARTICLE INFO

Article history:

Received: 10 December, 2018

Accepted: 14 January, 2019

Online : 20 January, 2019

Keywords:

Private Information Retrieval

Privacy

Smartphone applications

Location-Based Services

ABSTRACT

We present a privacy-preserving approach for discovering nearby places of interest to Alice. In this approach, the proposed protocol allows Alice to learn whether there is any place that she is looking for near her. However, the location-based service (LBS) that tries to help Alice to find nearby places does not learn Alice's location. Alice can send a request to the LBS database to retrieve nearby places of interest (POIs) without the database becoming aware of what Alice fetched by using private information retrieval (PIR). The common criticism of previous PIR approaches is that they are not practical for smartphones with limited processing power, memory, and wireless bandwidth due to the computational overhead. Therefore, the main focus of this work is to propose a scheme to reduce the computation cost on the client-side to make PIR appropriate and practical for the smartphone environments, and then apply the proposed PIR to LBS applications. We have implemented our protocol in Percy++ to evaluate its performance over a commercial-grade database of POIs. Our implementation results demonstrate that our approach has faster decode and retrieve time for the nearby POIs on smartphones compared with current similar work.

1. Introduction

Location-based services (LBS) are information services that offer various types of applications based on the location of the user, such as identifying a location of a person or object or place, weather service, parcel and vehicle tracking, etc. LBS retrieves the location of the user from the user's mobile phone via global positioning system (GPS), cell tower triangulation, or wireless local area network (WLAN). LBSs might be helpful to mobile users for safety services. For example, it is beneficial if emergency services can find information about the location details of a user who is in danger. However, users may not be aware that their location information might be shared with other third parties and could be misused, and it could be diverted as a tracking tool. Therefore, the main goal for research communities in this field is to protect the user's location while they are using LBSs.

During the last two decades, privacy-preserving protocols for location-based services have been introduced based on non-cryptographic and cryptographic approaches. Non-cryptographic approaches use trusted third parties to maintain the user's privacy, such as "dummy locations," "K-anonymity" and "cloaking" approaches. Reviews of these approaches and their drawbacks can be found in [1, 2, 3, 4, 5]. The main disadvantage of these

approaches is that the user has to trust and send her information to the trusted third party, which as we mentioned is unacceptable in LBS applications.

Our main goal in this paper is to help LBS users search on their smartphone for particular places of interest (POIs) while keeping their location private from the location-based service provider (LBSP). For example, the user sends her request to the LBS application to find a nearby restaurant, gas station, or ATM. This location could be the exact location of the user or a location to which she wants to travel in the future. These types of applications are especially useful for people who travel or have moved to a new city. A wide range of LBS nearby places applications have been released recently, such as Facebook Nearby Places, AroundMe, NearBy Places, Yelp, FourSquare and Places to help users to identify their nearby locations quickly.

Modern multi-core mobile devices have high-performing processors that are appropriate for cryptographic tasks which can enable location privacy to the LBS applications. Unfortunately, the processing units frequently consume a significant amount of energy, which causes a reduction in the battery life of smartphones. In addition, smartphones devices often have limited bandwidth and memory [6]. Therefore, downloading an entire database for finding POIs around Canada and the U.S., which easily can include more

*Maryam Hezaveh, University of Ottawa, mheza028@uottawa.ca

www.astesj.com

<https://dx.doi.org/10.25046/aj040104>

than ten million entries with respect to a typical commercial POI database (see Appendix A) [7] and require 3 to 4 GB of data storage, is obviously not practical. Furthermore, updating results periodically to make sure they are accurate enough, bandwidth limitation and data usage limitation of smartphones are other important factors which we should keep in mind when we are offering a cryptographic solution for LBS applications with respect to the privacy of the mobile user's location.

1.1. Motivation and Threat Model

Enormous enthusiasm for geographical referencing of individual information is apparent on the web these days. The majority of people currently use smartphones with many complex sensors closely connected to their daily activities. Most of these smartphones have a high-precision localization sensor such as a GPS receiver. GPS devices allow people to tag photographs and occasions and to track their mobility. Moreover, the number of sensors in our environment which interact with smartphones has increased. Although most people like the convenience of using these personal communication devices, there is an inherent trade-off between convenience and privacy. Clients might not be completely aware of exactly how the data about their location is utilized and by whom and what information about their location is being gathered, and subsequently, clients can disregard the potential risk that can happen by using their location information.

Location-aware capabilities allow the service providers to offer different types of application to their users such as the ability to share their location with their friends and other users and to geo-reference their posts. In this way, clients can utilize the location identifier to search and browse for different resources. An essential key for providing these services is to gather real-time location information on clients and additionally, other logical data including client relationships, activities and client provided content perhaps during long intervals of time. Specifically, in nearby location applications, service providers are not only able to collect clients' location information, but are also able to gather personal information by offering clients to write their experience and opinions with regard to reviews and tips on the visited place. Subsequently, clients' historical location data can be identified with relevant and semantic data freely accessible online and can be utilized to discover individual and sensitive data about clients and to develop comprehensive client profiles. The user activities, relationships, interests and mobility patterns could be extracted from these profiles. Although these location-based profiles may be considered helpful to improve and personalize the quality of applications for the clients, they can potentially be utilized for unwanted purposes and can cause different levels of privacy threats. Users' mobility tracks are not only a collection of locations on a map. The content of these tracks includes the users' interests, activities, habits, and relationships. It may also disclose users' private information and secrets. It can expose the users to undesirable commercial and spams, or even threat of physical harm. All of these imply that the negative side-effects of lacking location privacy have increased.

The main aim of this paper is to protect the users' location privacy against a passive adversary, active adversary and malicious service providers while they are using LBSs. We consider the following threats in our architecture:

Passive adversary. A malicious external observer or a malicious LBSP who has access to the data that passes between the user and the database on the communication channel but cannot change the data.

Active adversary. A malicious external observer who has access to data that passes between the user and the database on the communication channel and can insert, modify or delete data.

Malicious Service provider. A malicious server refers to an LBSP that tries to modify, delete or insert new messages in response to the user.

Users should have the privilege of controlling the amount of information (about their location) that is revealed and shared with others. This can be achieved in different ways such as users have a right to choose not to share their location information to untrusted applications, legislating privacy policies to force organizations and service providers to protect their users' location privacy, and designing a system in a privacy-preserving manner so it does not disclose users' location information to others.

1.2. Our Contributions and Assumptions

This current paper is an extension of a paper originally presented in [8]. We first explain our proposed block-based PIR scheme for smartphone applications [8] and then as an extension we apply our proposed PIR scheme to the LBS application. Our privacy-preserving protocol for LBS helps Alice to search for specific nearby POIs on her smartphone by sending a query to the location based service provider (LBSP) over a wireless network. In this scenario, the proposed protocol allows Alice to learn whether there is any place that she is looking for near her. However, the location-based service (LBS) that tries to help Alice to find nearby places does not learn Alice's location. Alice can send a request to the LBS database to retrieve nearby places of interest (POIs) on her smartphone without the database becoming aware of what Alice fetched by using our practical PIR scheme. The LBS server retrieves the query from the database, and returns the results to Alice containing the specific requested POIs type found in the requested location. In order to achieve this, our protocol must fulfill all of the following requirements, as also required in [9]:

1. The LBS server must not learn the exact location of the user. It might only identify a area that is large enough to satisfy the user's privacy in terms of area and the number of POIs it contains.
2. The proposed protocol must have no third parties between the user and the server.
3. The implementation must be computationally practical for resource-constrained hardware such as a smartphone.

4. The proposed approach cannot depend on trusted hardware that does not generally exist on a commercial smartphone.

Our cryptographic approach is based on private information retrieval (PIR) for secure LBS applications that identify nearby places. PIR allows the user to fetch her required information from the database without leaking which information is fetched [10]. The POI database is labeled by the location of POIs; therefore, the LBS server is able to retrieve the POIs depending on the user's location of interest in the requested query. PIR solves most of the previous problems associated with non-cryptographic approaches in LBS. PIR approaches do not have the privacy vulnerabilities of k-anonymity or cloaking, such as single point of attack of their anonymizer or server which tries to help them to apply k-anonymity or generate an obfuscation area. As a result, the information of the user location remains private and secure from all kind of the passive adversary, active adversary and service provider by using PIR approaches.

During the past two decades, various types of PIR-based approaches have been introduced. The common criticism of previous PIR approaches is that they are not practical for smartphones with limited processing power, memory, and wireless bandwidth due to the computational overhead [11, 12]. We ensure that the proposed cryptographic PIR approach is practical for smartphone applications. Based on [13], there are five main time elements that influence the speed of the PIR query:

1. the amount of time that it takes for the client to create a query which has to be private.
2. the amount of communication time that it takes to send the query to the server(s).
3. the amount of time for the server(s) to apply the query to the database.
4. the amount of communication time that it takes to send the response from the server(s) to the client.
5. the amount of time that it takes for the client to decode the response(s) and retrieve the results.

Our approach expands [9] idea of applying a cloaking area to reduce these five factors. Moreover, our approach reduces the amount of time required for the client to process the response and retrieve the results of decoding on the smartphone by approximately 50% compared to [9] by applying the POI types idea to block-based PIR. Reducing the decode time is valuable in our application to satisfy the fifth requirement, so that it can be used on modern smartphones' hardware. The processing cost on the server side is similar to [9], to preserve the privacy of the user's location. Our proposed protocol can be made to support all types of block-based PIR schemes.

In our proposed approach, the identity of the user is not hidden from the service provider, as the results have to be returned to the user. However, if the user wants to keep her identity hidden from LBS, she can use an onion routing technique, such as Tor [14]. Note that keeping the user's location private has priority in an LBS application over keeping the user's identity hidden from LBS,

because if LBS knows the user's location, it is quite easy to identify the user. We should mention that a mobile communications operator is constantly aware of the location of the user based on the cell tower. Therefore, we assume that this operator does not collude with the LBSP.

1.3. Organization of This Work

The rest of this paper is structured as follows. Section 2 presents an overview of previous work regarding PIR schemes and LBS schemes. Section 3 describes the details of our PIR scheme. Section 4 explains the details of our privacy-preserving protocol for LBS. The threat model and the security analysis of our proposed protocol are discussed in Section 5. Section 6 gives an overview of our implementation and compares it with previous work. The limitations of our proposed protocol are discussed in Section 7, and finally Section 8 concludes our paper.

2. Related Work

For greater understanding, we first review the definition of PIR and give a brief overview of different types of PIR schemes. Then, we provide a review of PIR-based approaches for the users' location privacy in LBS applications.

2.1 Review of Private Information Retrieval (PIR)

These days, users are increasingly aware of the privacy requirements of their data in their online activities. But is it actually possible to keep the user's query contents private while she issues a request to online applications? The first answer that comes to your mind when you think about this problem is that the user can send her request to the online application via Tor and communicate over the Tor network [14]. Here, the server has no clue who sent the request for the data; however, in order to fetch the requested data from the database, the server has to be able to access the content of query. Therefore, Tor is not a good option to solve our problem. The main problem that we need to solve is to let a user to send her query to the database without sharing what she searched for. In this scenario, we are trying to protect the content of the query, rather than the identity of the user. Private information retrieval (PIR) is a cryptographic technique that solves the matter of permitting the user to query a database while the content of the user's query is hidden from the database. The need for PIR schemes has been demonstrated in real online activities, such as location-based services, social networks, online research, etc. [9].

In 1995, [10] first introduced the problem of Private Information Retrieval (PIR). Looking at the trivial solution [10] of transferring the entire database to the user to be locally queried, highlights interesting properties. First, it delivers perfect privacy. Second, no information about query or response is leaked, since neither of these are sent across the wire. On the other hand, this approach yields high communication overhead: the size of the whole database. Goldberg [15] presented three important requirements for PIR: privacy, non-triviality, and correctness. For privacy, the database should learn neither the query input nor the

database block retrieved. For non-triviality, communication cost between the client and the server should be less the trivial limit of $O(n)$, where n is the number of bits in the database as seen above. For correctness, the received data from the database must satisfy the user's query. Another requirement which is not considered in most of the previous work for PIR is implementation efficiency. Most of the previous work tried to reduce the communication overhead rather than the computational overhead [16, 17]. This inattention to the computational complexity has caused the introduction of PIR schemes that are not practical for resource-constrained hardware, such as smartphones.

In [10], the author defined the first non-trivial PIR scheme. In 2004, Gasarch [18] described it simply as follows.

Definition 2.1. A one-round k -databases Private Information Retrieval (PIR) scheme with $x \in \{0,1\}^n$ is defined as follows [10, 18].

1. A user wants to find x_i . There exists k databases which all have the same copy of $x = x_1 \dots x_n$. The DBs do not collude with each other.
2. The user flips coins and the combination of the coin flips and i , produces query strings $q_1 \dots q_k$. She sends the query, q_j , to database DB_j .
3. For all j queries, where $1 \leq j \leq k$, DB_j returns an answer string $ANS_j(q_j)$.
4. The user computes x_i using the value of the $ANS_j(q_j)$, the coin flips, and i .

The cost of the defined PIR scheme is $\sum_{j=1}^k |q_j| + |ANS_j(q_j)|$.

Computational PIR (CPIR): The first type of PIR protocols assumes that the adversary and the server(s) have access to limited computational capability to guaranty the user's privacy. Therefore, to breach the security of these protocols, the adversary has to solve a problem which is hard to solve with its limited computational capability. This kind of assumption is usual for cryptography, security, and privacy schemes.

In 1995, [10] proved that it is impossible to have a single-database PIR in the information theoretic security sense. In 1997, [19, 20] proposed the first CPIR to prove that the communication complexity of PIR can be reduced if we want to achieve computational privacy, and we are not willing to achieve information theoretic privacy. In the same year, Kushilevitz and Ostrovsky [21] presented a CPIR protocol which has the same assumption for the computational capability for the adversary, but it uses a single server. Their protocol was the first single-server CPIR. It is based on the Quadratic Residuosity problem that is considered to be difficult to solve. The main advantage of single-server CPIR protocols is that by using the CPIR recursively, the communication complexity of PIR can be improved. Later, different types of single-server CPIR were proposed which tried to reduce the communication cost of PIR, for example, ϕ -hiding problem [22, 23], the presence of one-way trapdoor permutations [24], Pailler homomorphic encryption [25], and the Hidden Lattice problem [16].

In [17], the author proved that none of the previous CPIR schemes were practical, given certain realistic assumptions at the time. However, in 2016, [26] introduced XPIR. They showed that by using lattice-based cryptography, CPIR is of practical value and the conclusion of [17] is no longer valid.

Information Theoretic PIR (IT-PIR): In Information theoretic privacy even if an adversary has unlimited computational capability, he cannot compromise the privacy of the user. In 1995, [10] showed that any single-server IT-PIR scheme must have communication cost at least that of the trivial protocol. Therefore, IT-PIR protocols assume that if you have $k \geq 2$ non-cooperating servers, and each of these servers has a copy of the database, then there exist PIR schemes which achieve complete information theoretic security. Following [10], different types of IT-PIR were proposed which tried to improve [10], such as [9, 15, 27, 28].

By using the idea of multiple servers, we improved the robustness of the PIR, but this can affect privacy if there exists non-responsive servers or/and malicious servers [15, 27]. To handle this issue, Goldberg [15] proposed the privacy threshold in which the total number of the servers must be greater than the privacy threshold. As a result, in order to set a privacy threshold, we need to provide extra responding servers.

Trusted Hardware PIR: The trusted hardware-based PIR is first introduced by [29] in 2006. The trusted hardware-based PIR uses the idea of a tamper-resistant CPU, which is connected to the server and is trusted by the user. The user sends her query to this CPU, where her query is hidden from the server. In this scenario, the CPU is the one who is responsible to fetch the requested information from the database and sends back the results to the user. These types of PIR achieve the low computation and communication costs, but the trusted hardware PIR architecture is secure only if the user can trust the hardware.

Hybrid PIR: In [13], the researcher proposed a hybrid PIR that was a combination of CPIR and IT-PIR to reduce communication costs. Their goal was to combine the positive features of CPIR and IT-PIR to reduce the negative features of each. To achieve a lower bound for both computation and communication costs, they merged the recursion property of CPIR (single-server) approaches and the low computation and communication complexity property of IT-PIR (multiple-server) approaches.

2.2 Review of the PIR-based scheme for Nearby Places

One of the motivations for developing useful and practical PIR schemes is to protect the users' private information while they are using mobile devices with positioning capabilities. In a stationary desktop scenario, when a user tries to query the database or the remote server, the primary concern is leaking information about the query's content. However, in an LBS scenario, when a user queries the LBS server, her location is also revealed to the LBS server. Here, the problem with location privacy is preserving the privacy of the user's real location when she is using the LBS while providing the most precise and acceptable response.

Many of the previous problems of privacy preserving protocols for LBS that we encountered were solved by introducing PIR-based LBS protocols. The idea is to let the user send a query to the LBS server without disclosing her actual location by the PIR scheme. This query typically consists of POIs, which includes a description of the POI and its geographic location.

Most of the existing works which tried to apply PIR to location-based services were based on secure hardware, with a secure coprocessor at the LBS server [3, 5, 30, 31]. The idea of using the secure hardware-based PIR in LBS was first proposed by Hengartner [3]. This hardware performs the trusted computing to hide the user's location from the LBSP. Recent work regarding secure hardware PIR was proposed by [30]. Their PIR technique was similar to [31], however it offered better efficiency, and it was more practical for large datasets. All proposed solutions for secure-hardware PIR claim that the trusted hardware-based PIR method is the only practical PIR scheme [30, 31]. The main disadvantage of all secure hardware PIR schemes is that the proposed architectures are secure only if the user can trust the hardware.

The common criticism of other PIR approaches for location privacy is that the computational overhead is not acceptable and practical for resource-constrained hardware such as a smartphone [11, 12]. In 2008, In [1], the author proposed the first PIR-based approach for location privacy, without using a third party. Their proposed protocol used the idea of the trade-off between efficiency and privacy as defined in [32]. In [1], the researcher proposed a single PIR request for each query approach. In their approach, all queries were indistinguishable, and it was able to achieve strong location privacy. Their proposed protocol included two steps to protect the user's query and information about her location. In the first step, the server and the user engaged in a protocol, which is based on Paillier encryption [33], to determine the index of the user's location cell, without releasing the location to the server. The user uses PIR to retrieve the query results for the target cell in the second step. The advantages of the Ghinita protocol are the nondisclosure of location information and its security for both mobile and stationary users against correlation attacks.

In [5], the author described three drawbacks to [1] protocol. First, it focuses on the nearest neighbor queries. Second, it scans the entire database linearly for each query. Third, it has a high communication complexity. Additionally, the protocol is secure if the privacy of the user is a concern and LBS is not able to learn the user's query, but it is not symmetric for LBS's database privacy since the user can infer the data that are in the same column as her query.

Later, in [9], the author proposed a hybrid solution combining PIR and cloaking to protect the user's privacy without using trusted computing. Their idea of using cloaking reduces the computational cost of PIR and makes it more practical. The user's location privacy relies on the size of the cloaking area. Their PIR approach supports all types of PIR schemes (block-based). Our proposed PIR protocol expands on [9] idea. However, we focus on reducing

the computational complexity on the client side. We explain our proposed protocol in detail in the next section.

3. The Proposed PIR Scheme

Here we present our block-based PIR for location privacy in mobile phone applications. Our solution uses partial queries [10] to reduce communication and computation complexity. Moreover, we structure the database to optimize client computations. This has benefit in our mobile scenario in which the clients (possibly smartphones or IoT sensors) have constrained computational power. In our approach, the user retrieves the exact category of the data, which saves on data processing on the resulting sets. These savings on result set size in turn impact any decode, decrypt, or homomorphic operations which must occur to obtain a result. As these are cryptographic operations, the benefit in result set size reduction is material. Note that our approach is suitable for all applications that need to protect users' privacy while they are searching for data in a database (it is not restricted to just LBS applications).

3.1. Preliminaries

Our proposed protocol can be made to support all types of block-based PIR schemes. We illustrate its usage using multiple server IT-PIR [15] and Shamir secret sharing [34]. As such, the user's query is split into l shares which are then transferred to k servers. This results in communications and computation benefit which we analyze in Section 6. The protocol is robust to byzantine situations in which servers (either malicious or in a service degradation scenario) may fail to respond or may respond with information containing errors.

Our approach to reduction of client computation cost uses the idea of trading off privacy for better performance [9]. In [9], the level of desired-privacy is adjustable and is proportionally related to the number of data items that the database PIR server must process to respond to the client. We extend and improve on this approach in three ways.

First, we divide the database into classes, with each class categorized based on a sub-type of data to be queried. The server returns exactly the subset of the database which pertains to the queried category. By reducing result set size, the client benefits in a number of ways. It is no longer necessary to filter the response data. In addition, the aggregate cost of cryptographic operations, such as decryption or homomorphic computation, is reduced.

Second, If a sub-type has a higher amount of data, a data traffic cost will be higher because of the result size and it will cause a slower response time. On the other hand, if in another sub-type the amount of data is extremely low, it will minimize the result size and lower data traffic. These could help the server to guess the user's sub-type of interest with a high degree of confidence and lead to a loss of privacy. To tackle this problem, our approach equalizes the amount of data in each row of the sub-type in a specific class by adding "null" to all the sub-types which are not equal to the maximum sub-type size in that class. The main advantage of our "null" solution is that if the user is looking for a sub-type with less data, the PIR computation overhead on the

client-side is reduced (when receiving the first null in the decoding process, the computation process stops), and if the user is looking for a sub-type with more data, the PIR computation cost on the client-side increases, without losing privacy.

Finally, in our approach the amount of data is different in different classes (see Figure 1), unlike [9]. As a result, our PIR computation cost depends on the specific class that the user requests. If the user searches for a class with more data, the PIR computation cost increases. If the user searches for a class with less data, the PIR computation cost is reduced.

By considering these three improvements, if the user sends a query to the database for the sub-type of data in each class, the response which is returned to the user not only needs less computation time for decoding, but also does not need to be filtered on the client side to remove non-requested data.

Class	Category Types					
Class-0 00	Sub-type 0	0000	data1	data2		
	Sub-type 1	0001	data 1	-		
	Sub-type 2	0010	data1	data2		
	Sub-type 3	0011	-	-		
	Sub-type 4	0100	-	-		
	Sub-type 5	0101	data1	data2		
	Sub-type 6	0110	data1	data2		
	Sub-type 7	0111	-	-		
	Sub-type 8	1000	-	-		
	Sub-type 9	1001	data1	-		
Class-1 01	Sub-type 0	0000	-			
	Sub-type 1	0001	-			
	Sub-type 2	0010	-			
	Sub-type 3	0011	-			
	Sub-type 4	0100	-			
	Sub-type 5	0101	-			
	Sub-type 6	0110	-			
	Sub-type 7	0111	-			
	Sub-type 8	1000	-			
	Sub-type 9	1001	-			
Class-2 10	Sub-type 0	0000	-	-	-	-
	Sub-type 1	0001	data1	data2	data3	data4
	Sub-type 2	0010	-	-	-	-
	Sub-type 3	0011	data1	data2	-	-
	Sub-type 4	0100	data1	-	-	-
	Sub-type 5	0101	data1	data2	data3	data4
	Sub-type 6	0110	data1	-	-	-
	Sub-type 7	0111	data1	data2	data3	data4
	Sub-type 8	1000	-	-	-	-
	Sub-type 9	1001	-	-	-	-
...						

Figure 1 Sample of the relationship between data, sub-type and classes, as saved in the database.

3.2. The Proposed PIR Scheme

Our PIR protocol has two phases. The first phase is the pre-processing phase in which the whole protocol becomes ready to use, on the server side and also on the client side. In the future, if the client decides to change the level of her privacy, or any changes occur on the server side, this phase can be repeated. The second phase is the execution phase in which the user sends her request to the server. Her request contains the class of data which she searches concatenated with the sub-type category.

Pre-processing Phase contains the following steps:

1. Given a chosen level of the user’s privacy, “Class”, “sub-type” and “data” category are applied to the database.
2. The “class” and “sub-types” are defined to have a number based on their specific categories. As shown in Figure 1, for

example, Class-1 is considered 01 and Sub-type-3 is considered 0011. Note that in this Figure we just showed “10” different “sub-types” for each “class”. This depends on the number of different sub-types of data in the database and also on the level of the user’s privacy which is applied in step 1.

3. Each database index will be the “class” concatenated with the “sub-type”. For example, in Figure 1, the database index for the Class-1||sub-type-3 is considered as 010011.

Execution Phase contains the following steps:

1. The user chooses the sub-type of her interest from the list suggested by the application based on her privacy level. For example, she is looking for Class-1||sub-type-3.
2. The proposed application provides the user’s request, which is an index of the database, and sends it to the server in a way that is hidden from the server. In this example, the request is 010011 which refers to Class-1||sub-type-3.
3. The specific row of database is retrieved from the database and the data present in this row are transmitted back to the user.
4. The user decodes the results and the results are shown on her smartphone.

4. The Proposed Privacy-Preserving Protocol for LBS

The main goal of cryptographic protocols in nearby places is to be able to detect nearby places automatically while the user’s location privacy is considered in the location-based service (LBS) application. Our proposed protocol uses private information retrieval (PIR) to achieve this purpose.

4.1. Problem Statement

Alice has her location as her secret. Alice wants to use a LBS application to search and find nearby places of interest. We propose a protocol that allows Alice to find nearby places for which she is looking. However, the LBS that helps Alice to find her nearby place does not learn Alice’s location. Alice can send a request to the LBS’s database to fetch her nearby places of interest without the LBS being aware of what Alice fetched by using private information retrieval (PIR). Most of the previous PIR schemes are not acceptable in LBS applications because of their use of secure hardware. The focus of this section is to solve the PIR-based LBS issues by offering a practical PIR without using secure hardware or a trusted third party and lower the computational cost on the client side in the smartphone’s application. At the end of this protocol, the proposed application should list the POIs that meet Alice’s search criteria or show her that there is no POI in the selected area.

4.2. The Proposed Privacy-Preserving Protocol for LBS

We first informally describe our proposed protocol via an example. Suppose Alice is located in Ottawa and she wants to look for a specific type of POI, for example a restaurant, near Bank Street. Since she is privacy conscious, she sets her cloaking area to be a 10 km MGRS grid square (see section 4.3). The client

application sends the requested cloaking area to the server. At the same time, the PIR allows the client application to identify which part of the cloaking area has the restaurant, without the server being informed which part is retrieved. All entries in the POI database are indexed by their MGRS block concatenated with the POI type. The row that contains the restaurant(s) is retrieved from the selected MGRS grid square on the database and the results are sent back to Alice. The client application decodes the results and sorts the results, and the nearest restaurants are shown on her phone's local map.

Our protocol follows [9] hybrid solution that uses PIR to preserve the privacy of the user's query and a cloaking scheme in order to make the PIR scheme practical and reduce the computational cost of PIR. The benefits of the hybrid solution are as follows: the location of the user remains secret from the LBSP to a reasonable privacy level chosen by the user without depending on the other users in the selected area; to calculate the cloaking area or cryptographic algorithms we do not need to have a trusted third party; and the computational overhead of the PIR scheme is practical.

Our proposed protocol has two improvements compared to [9]. First, due to the user's request for a specific POI, our proposal categorizes the cloaking area in the database into the POI types. Thus, when the user asks for her POI in her selected cloaking area, the results that are returned to her are of the type that she is looking for. Therefore, our protocol on the client is not required to filter the block of different types of POIs to identify the POI that the user requested. This reduces the computational costs, and saves the battery and data usage on the smartphone. For example, if there are no restaurants near the user, she does not need to wait to decode all POIs in that cloaking area and then filter the restaurant to find that actually the answer is "null".

Second, we propose a new technique based on a static grid-based approach for defining our cloaking area and mapping our POIs to a cloaking area, unlike the approach proposed in [9] which uses the Various-size-grid Hilbert Curve (VHC) technique [35] (see section 4.4).

Our proposed protocol describes how the POI database is initialized and how the protocol generates a cloaking area around the user's exact location, and executes a PIR query on the contents of the requested cloaking area. We name our phases similarly to [9] to highlight the similarities and differences between our phases. Note that each POI consists of 300 bytes that includes longitude and latitude coordinates, name, exact address, the phone number, website address, etc.

The pre-processing phase contains the following steps:

1. An appropriate static grid system is applied on the geolocation plane.
2. POIs are categorized based on their type and saved in the LBS's database.
3. A row of database refers to a cloaking area concatenated with the POI type.

The execution phase contains the following steps:

1. The user selects the area of her interest; it could be her current location as determined through GPS, or some other location that the user may be traveling to in the future.
2. The user selects a preferred level of privacy.
3. The user's corresponding cloaking area is calculated based on the level of her chosen privacy.
4. The user chooses the POI type(s) from the suggested list provided by the client application.
5. The client application sends the cloaking area to the server. Also, the client application identifies which portion of the cloaking area contains the POI type(s), in a way that is hidden from the server.
6. The server receives the request, and finds the database portion corresponding to the cloaking area. A block of rows is retrieved from this portion based on the user's specified POI type. The POIs present in these rows are transmitted back to the client application.
7. The client application decodes and sorts the results, and the nearest POIs are shown on her phone's local map.

4.3. Grid-based Cloaking

In our proposed protocol, the client application extracts the user's location via cell towers, Wi-Fi, or GPS, and it calculates the user's cloaking area by using the military grid reference system (MGRS) technique. MGRS is a geo-coordinate standard for locating points on the Earth [36]. The Earth is divided into grid squares with sizes of 0.1 km, 1 km, 10 km, 100 km, etc., based on the level of accuracy and degree of precision. Our proposed protocol uses MGRS to help ensure the user's location privacy. Each MGRS block is considered as a block in the database that is categorized based on POI type.

considered to be nearby if they are located in the same MGRS block as the user. The user's location privacy level increases if she chooses a larger MGRS block. However, a larger MGRS block includes more POIs, and it affects the computational cost of our proposed protocol. We will discuss this issue in more detail in the following section. Figure 2 shows the different levels of MGRS blocks for the Ottawa, Ontario, area [37].

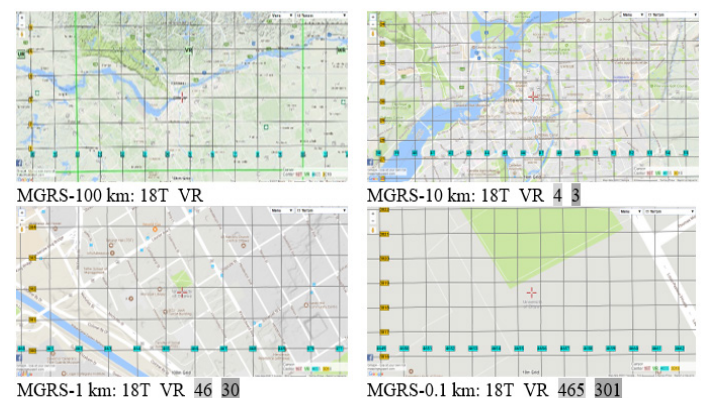


Figure 2 Different MGRS Levels [37]

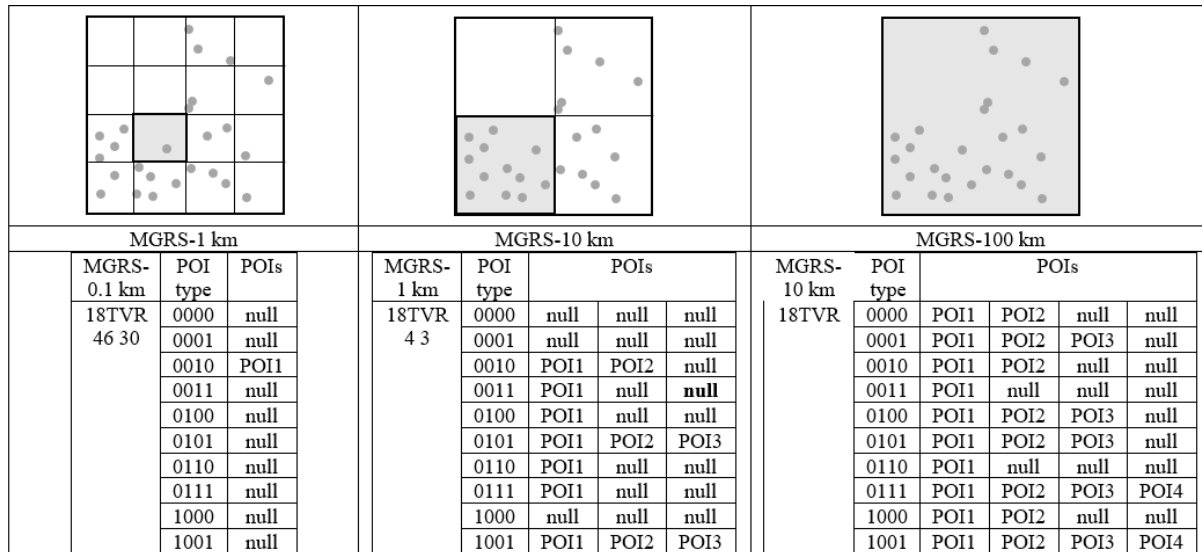


Figure 3 Illustration of the relationship between the MGRS block, POI types, and POIs as saved in the LBS database.

4.4. Location Cloaking

The first step in the location cloaking phase is to apply different MGRS levels on the geographic region, such as Canada and the U.S.. Then, the user’s cloaking area is calculated based on the user’s selected MGRS level and her current location or the location of interest. POIs are considered to be nearby if they are located in the same cloaking area.

The selected MGRS level must be large enough to achieve the privacy of the user’s location within the requested cloaking area, but simultaneously it must be small enough to reduce the computational overhead on the smartphone application to process the results, and also to reduce the communication overhead to transfer the result via the wireless data traffic.

In order to map POIs to a cloaking area, [9] used the Various-size-grid Hilbert Curve (VHC) technique. [9] chose VHC because it could solve the issue of density of POIs based on the geographic region. For example, the data traffic cost increases if the selected area has a high density of POIs (within a city). On the other hand, if the selected area has a lower density of POIs (within a countryside), then the result size decreases and the server is able to estimate the location of the user, which leads to a loss of privacy. VCH can solve this problem by creating a different-sized cloaking area based on the density of POIs. However, this solution has the disadvantage of receiving a list of POIs which may or may not be useful for the user. If the selected area has no POI that the user is looking for, she still has to wait for the client application to calculate the result, which is based on all POIs in a selected VCH region, and then show the result which is actually “null”. This can cause a high computational cost on the client side application.

To manage the computational cost on the smartphone application based on the density of the POIs, and prevent high computational cost in the lower density POIs area, we propose a new technique to map POIs to a cloaking area based on the MGRS fix-sized blocks. First of all, we categorize POIs based on their

types in each MGRS block. In Figure 3 we consider ten POI types per MGRS block (see Appendix A) [7]. Then, each row of the database refers to the MGRS block concatenated with the POI type. As [9] mentioned, the density of POIs varies by geographic area. Therefore, each row of the defined database has a variable size.

To protect the privacy of the user’s location and prevent the server from guessing which POI type is fetched by the user, we need to equalize the number of POIs in each selected cloaking area. Therefore, if the number of POIs in one POI is not the same as the maximum POI size in the selected cloaking area, the rest of the row must be set to “null”. By this technique, our PIR client side computational cost relies on the location of the user and the level of selected MGRS. If the user’s location has low POI density, the PIR client side computation time will decrease. If the user’s location has high POI density, the PIR computational cost on the client side will increase. Note that the server cannot observe the differences between computational costs for queries in different locations, because we equalized the number of POIs in the selected cloaking area. Otherwise, the server which is able to observe the difference between computational cost based on different user’s queries, could guess the user’s location. Figure 3 shows an example of the POI density in the selected area based on different levels of MGRS and illustrates the relationships among the MGRS block, POI types, and POIs as saved in the LBS database.

5. Security Analysis of the Proposed Protocol

The user can use her current location or a location that she wishes to visit in the future. This feature adds one more level of privacy in our protocol because the observer or the location-based service provider (LBSP) is not aware of whether the requested MGRS block corresponds to the user’s current location. Therefore, our proposed protocol has two kinds of privacy: first, protection of the user’s location privacy within the requested MGRS block, and second, the LBSP or an observer does not know whether the

request is the user's current location at the request's time. In both cases, our main goal is to protect the user's location privacy against the LBSP and any other observer.

5.1. Threat model

We can apply our proposed PIR protocol to all existing block-based PIR schemes (CPIR and IT-PIR). In this section, we use the IT-PIR (multi-server) to describe our threat model (as was done by [9]). The primary assumption in IT-PIR schemes is that servers must not communicate with each other to breach the privacy of users' queries. Under this assumption, the IT-PIR protocol itself has been proven secure in [10, 15, 27]. Given the cryptographic security of the IT-PIR scheme, we review the security of our PIR protocol, as well as its security against passive and active adversaries in the following sections.

5.2. Security Analysis

Claim 5.2.1. If B is an MGRS block of level L , chosen by Alice, and T is the type of POI that Alice is searching for, our proposed privacy-preserving protocol for LBS is secure against a *passive adversary* in the block B .

Passive adversary. An external observer or a malicious LBSP who has access to the data exchanged between the user and the database along the communication channel but cannot change the data.

Justification. Our proposed protocol calculates the portion of the database in which Alice wants to find a nearby place of interest. It depends on the level of privacy she selected, such as MGRS-100 m. Therefore, Alice's location privacy and query are limited to the requested portion of the database. The number of POIs in each type of an MGRS block is set to be equal to the maximum number of POIs in that MGRS block by adding "null" to the ends of the other types in our proposed database. Thus, the passive adversary cannot guess which type of POI was fetched by Alice. In other words, if Alice's type of POI changes while she is still in the same MGRS block, then for a new request, Alice will send the same query for the same MGRS block of the database.

Because Alice selected her level of privacy to be, for example, a block of MGRS-100 m, it is impossible for the passive adversary to detect her movement as long as she is in the same MGRS block. Therefore, a correlation attack is actually unachievable since Alice will always send the same query for a given privacy level. Additionally, if Alice knows that she is going to move, she should choose a larger MGRS block that includes both her current location and her next location. Thus, her movement will not be detectable.

The PIR scheme provides the security for our proposed privacy-preserving protocol for LBS against a passive adversary. The only information which the passive adversary can access is the identity of the user. If the user's identity protection is required, we can use TOR. Additionally, the user's query contents and the response of the database can be protected against a malicious

observer by applying end-to-end encryption techniques, for example, TLS (transport layer security) through the wireless communication channel. Both schemes (TOR and TLS) are optional for the user, as they slow down the protocol and cause an additional computational cost.

Claim 5.2.2. If B is an MGRS block of level L chosen by Alice, and T is the type of POI that Alice is searching for, our proposed privacy-preserving protocol for LBS is secure against an *active adversary* in the block B .

Active adversary. A malicious external observer who has access to data exchanged between the user and the database along the communication channel and can delete, insert or modify data. The LBSP is considered to be trusted in Claim 5.2.2.

Justification. Message reordering attack. During this attack, the active adversary attempts to delay and/or reorder requests or responses to mix up results and the communication. Note that in our proposed protocol, if the results of multiple queries are not received in proper sequence, it has no effect on the server or Alice since each result holds the requested information. The adversary also gains no information about Alice's query or location.

Justification. Message tampering attack. Alice will not be able to verify false responses if the adversary starts to send them. Therefore, a DoS (denial-of-service) attack is possible. However, in using this attack, the active adversary will not learn any more information about Alice's query or location, which is the main focus here. Note that this attack can be prevented by using TLS over the communication channel.

Justification. Message insertion or deletion attack. If an active adversary tries to delete or insert data from the server's response or Alice's request, it can cause a DoS attack. The adversary does not receive any information about Alice's request or location. Again, TLS can be used on the communication channel to prevent this attack, if desired.

Justification. Message replay attack. In this attack, if an active adversary starts a replay attack against Alice or the server, it does not affect either of them. The server responds to the requests, and Alice can easily drop multiple responses with the same information. The adversary will not receive any information about Alice's query or location. As with the previous attacks, we can use TLS to prevent this attack as well.

Claim 5.2.3. If B is an MGRS block of level L chosen by Alice, and T is the type of POI that Alice is searching for, our proposed privacy-preserving protocol for LBS is secure against a *malicious server* in the block B .

Malicious Server. A malicious server refers to an LBSP that attempts to insert new messages, modify, or delete messages in response to the user.

Justification. If the malicious server sends a false response, does not return a response, or sends additional messages with the

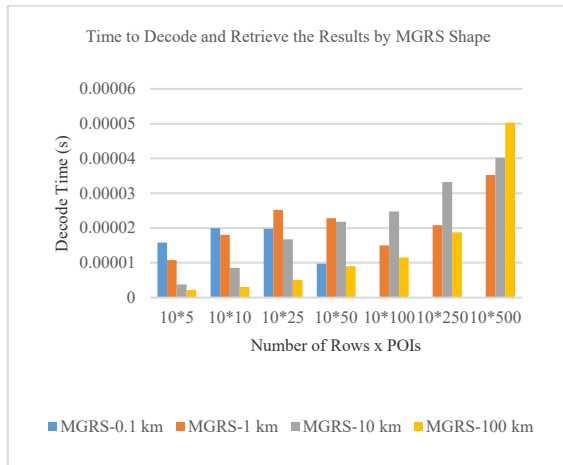


Figure 4 Comparison of time to decode and retrieve the results, by MGRS shape at the client side. It shows the computation time for queries on one MGRS block (ten POI types) for different number of POIs (each POI consists of 300 bytes).

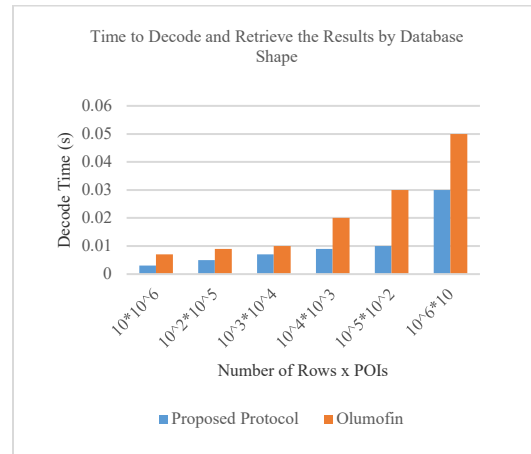


Figure 5 Comparison of decode and result retrieval time by database shape at client side in our proposed protocol and in Olumofin. The results show the computation cost for different-shaped database for queries on a 3 GB database (each POI consists of 300 bytes).

response, it can cause a DoS attack against the user. However, this attack does not enable the malicious server to learn any information about the location of the user or her query. The only thing that could help the malicious server find information about the location of the user is the content of the query, which is protected by using the PIR scheme within the MGRS block.

6. Experimental evaluation and results

We implemented a prototype of our proposed protocol based on an open source PIR protocol called Percy++ [15, 16, 38, 39, 40, 41]. We ran our C++ prototype on a virtual machine with Ubuntu Linux operating system and an Intel® Core™ i7-8550U CPU @ 1.80 GHz, 4GB RAM. We followed all assumptions of [9] in our implementation to compare our results with their approach. We randomly generated and distributed ten million POIs within Canada and the U.S. [7]. Each POI consisted of 300 bytes that included the longitude and latitude coordinates, name, exact address, phone number, website address, etc., of the POI [9]. We set the number of the databases to two to use the Percy++ PIR [9]. For each of these, we applied MGRS to our generated map to create four databases for four levels of user privacy: MGRS-0.1 km, MGRS-1 km, MGRS-10 km, MGRS-100 km.

Recall that the main goal of this section is to decrease the decode time on a smartphone to make the PIR scheme practical for resource-constrained hardware. The decoding time has a direct correlation with the number of POIs in each MGRS block. As we defined the fixed number of POI types for different levels of MGRS blocks (ten types), the PIR decoding time on the smartphone depends on the number of POIs in each type. Therefore, if the number of POIs in a type increases, the decoding time increases. We equalized the number of POIs in all types of MGRS blocks by adding “null” entries, as explained above. Therefore, there can be types that have no data or less than the maximum POI type. This improves the decoding time compared

to a type that has the maximum number of POI. As observed in Figure 4, when the user’s level of privacy is increased (larger MGRS block), the probability of obtaining more POIs per type increases. Therefore, an MGRS block with a maximum of five POIs per type has faster data retrieval compared to 500 POIs per type. However, there are some exceptions. For example, MGRS-0.1 m required less computational time with 50 POIs per type than with five POIs per type. These exceptions can happen if the type that was requested has fewer POIs than 50, and the rest of the data in that type is “null” (to make the total number of POIs equal to the maximum number of POIs in the requested MGRS block). During the decode time, whenever the first “null” block is detected, the decode operation stops and results are returned to the user.

We ran our prototype using different levels of MGRS blocks, and we generated 100 random requests to calculate the average time to decode and retrieve the results. As observed in Figure 4, the probability that we had 500 POIs from one type in an MGRS-0.1 km was equal to zero. This means, for example, that in an area of 0.1 km², there cannot be 500 banks. The opposite scenario may also occur. The probability that we have five POIs of one type in an MGRS-100 km block is rare. Considering this, the MGRS-10 km is the best choice if we want to show the results for an MGRS block with various numbers of POIs/types.

In our proposed protocol, the number of rows in each MGRS block is the same as the number of POI types, which was set to ten types per MGRS block, regardless of the size of the MGRS block. Thus, if our MGRS block becomes larger, the number of types (rows) does not change, but the number of POIs per type increases. We also have different numbers of POIs per type. Thus, in an MGRS block, we have some types that have no data or less than the maximum number of POIs in that MGRS block. This is important because we want to compare our results with [9], and due to the different definition of privacy and the reasons we just mentioned, it is difficult to give an exact comparison.

However, to show the performance of our proposed protocol compared to [9], we followed their implementation by setting the privacy level equal to one. Thus, we considered one large MGRS block that covered Canada and the U.S., and the number of POI types was set equal to the number of rows in [9]. Our implementation results are based on a database of ten million POIs. Figure 5 shows the time for decoding and retrieving the results for various numbers of rows and POIs.

As observed in Figure 5, our performance is approximately 50% better than that of [9] because our method considers the POI type and uses the MGRS which applies a fix-sized cloaking area and a variable-sized block to the database. In our proposed protocol, the user receives exactly the type of POI that she was looking for. By increasing the number of POIs per row, the decode and retrieve time increases in [9] protocol because after decoding the rows of the database, the results must be filtered to show the POI that the user was looking for.

As stated in section 4.3, due to the four levels of MGRS (0.1 km, 1 km, 10 km and 100 km), our database required four different configurations. This was the only disadvantage of our proposed protocol compared to [9]. This could increase processing on the server when adding or removing POIs from the databases (for example, when a restaurant closes or a new one opens in a specific MGRS block).

7. Limitations of our Proposed Protocol

In general, there may be a case in which the user will not find a reasonable POI in the requested cloaking area. Therefore, she may wish to search further in a larger MGRS block (i.e., in a broader geographical area). When this happens, the user's privacy does not decrease in our proposed protocol; it is still guaranteed to the level of the original cloaking area.

As mentioned in section 4.3.3, due to the four levels of MGRS (100 km, 10 km, 1 km and 0.1 km), we required four different configurations for our database. This is a disadvantage of our proposed protocol compared to [9]. This could increase processing on the server when adding or removing POIs from the databases (for example, when a restaurant closes or a new one opens in a specific MGRS block).

Modern smartphones with multi-core processors may be able to handle the 1.5 GB database for Canada and the U.S. that is used in [9] evaluation section, as well as the most recent 3 GB location database [7] that we used in our implementation. However, we should mention that not all people have the most recent smartphones and so our proposal, which reduces computational cost on the client by almost 50%, may be of particular interest for such environments.

8. Conclusion and Future Work

In this paper, we presented a privacy-preserving protocol to help the user search for nearby places of interest while protecting her location's privacy by using PIR. For this purpose, we first

proposed a block-based PIR scheme to decrease the computational overhead on smartphone applications [8]. We demonstrated that by applying our PIR scheme to the LBS, the computational overhead on the client side was reduced by approximately 50% compared to that reported in a previous work [9]. This reduction is valuable for the implementation of PIR in smartphone applications with limited resources. We demonstrated that our proposed LBS protocol is secure against active and passive attacks, as well as against a malicious server that tries to identify information about the user's query and location.

Our approach of retrieving the specific POIs in a cloaking area consumes less computational cost compared with the naive approach of asking the user to download the entire contents of the cloaking area and extract POIs locally. Simultaneously, the user's location privacy is not compromised since the user requests the same cloaking area as if she was requesting the entire contents of the cloaking area. This is a great benefit that reduces the cost of the wireless communication and also the memory usage on the smartphone.

There exist a number of interesting directions for our LBS privacy future work. First, our implementation results are based on [15]; in order to improve the computation, cost of our proposed protocol, we could develop it based on the higher performance block-based PIR such as the hybrid PIR which is proposed in [13]. Second, our proposed protocol could be extended by supporting more complex types of queries. Third, our proposed protocol could be combined with the Vehicle-to-Infrastructure (V2I) and the Vehicle-to-Vehicle (V2V) communication to help the user find her nearby places while she is driving her car. In this scenario, the user receives the latest update about the nearby places for her response from other vehicles or street infrastructure such as traffic lights and signs instead of a solid database. To update information about places and their availability constantly, we could use a Blockchain infrastructure in which other vehicles or street infrastructure are able to update their recent observations about the places. For example, all the infrastructure components on the street are connected to the Blockchain and every change that occurs appears in the block. Now if the user is looking for nearby parking, the traffic light could let her know the nearest parking lot and if there is any spot available by checking the updated list on the Blockchain.

References

- [1] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Private queries in location based services: Anonymizers are not necessary" In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 121–132, 2008. <https://doi.org/10.1145/1376616.1376631>
- [2] W. Hao, and Y-C. Hu, "Location Privacy with Randomness Consistency" In Proceedings on Privacy Enhancing Technologies, 62–82, 2016. <https://doi.org/10.1515/popets-2016-0029>
- [3] U. Hengartner, "Hiding location information from location-based services" In International Conference on Mobile Data Management, 268–272, 2007. <https://doi.org/10.1109/MDM.2007.56>
- [4] M. Hezaveh, and C. Adams, "Privacy Preserving Discovery of Nearby-Friends" In E-Technologies: Embracing the Internet of Things, Lecture

- Notes in Business Information Processing, (289), 41–55, 2017. https://doi.org/10.1007/978-3-319-59041-7_3
- [5] A. Khoshgozaran, C. Shahabi, and H. Shirani-Mehr, “Location privacy: going beyond K-anonymity, cloaking and anonymizers” *Knowledge and Information Systems*, 26(3), 435–465, 2011. <https://doi.org/10.1007/s10115-010-0286-z>
- [6] Y. G. Kim, J. Kong, and S. W. Chung, “A Survey on Recent OS-level Energy Management Techniques for Mobile Processing Units” *IEEE Transactions on Parallel and Distributed Systems*, 2388–2401, 2018. <https://doi.org/10.1109/TPDS.2018.2822683>
- [7] Factual, Global Places-Schema, <https://my.factual.com/data/t/places>, last accessed 2018/11/05.
- [8] M. Hezaveh and C. Adams, “A PIR scheme to improve the computation cost on the client-side of smartphone application” In *IEEE 31th Canadian Conference on Electrical and Computer Engineering*, 1–4, 2018. <https://doi.org/10.1109/CCECE.2018.8447708>
- [9] F. Olumofin, P. K. Tysowski, I. Goldberg, U. Hangartner, “Achieving efficient query privacy for location-based services” *International Symposium on Privacy Enhancing Technologies Symposium*, 93–110 2010. https://doi.org/10.1007/978-3-642-14527-8_6
- [10] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, “Private information retrieval” In *Proceedings of the 36th Annual Symposium on the Foundations of Computer Science*, 41–50, 1995. <https://doi.org/10.1145/293347.293350>
- [11] D. Lin, E. Bertino, R. Cheng, and S. Prabhakar, “Position transformation: a location privacy protection method for moving objects” In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*, 62–71, 2008. <https://doi.org/10.1145/1503402.1503414>
- [12] D. Riboni, L. Pareschi, and C. Bettini, “Privacy in georeferenced context-aware services: A survey” *Privacy in Location-Based Applications*, 151–172, 2009. https://doi.org/10.1007/978-3-642-03511-1_7
- [13] C. Devet and I. Goldberg, “The best of both worlds: Combining information-theoretic and computational PIR for communication efficiency” In *International Symposium on Privacy Enhancing Technologies Symposium*, 63–82, 2014. https://doi.org/10.1007/978-3-319-08506-7_4
- [14] R. Dingledine, N. Mathewson, and P. Syverson, “Tor: the second-generation onion router” In *Proceedings of the 13th conference on USENIX Security Symposium*, 2004. <https://svn.torproject.org/svn/projects/design-paper/tor-design.pdf>
- [15] I. Goldberg, “Improving the robustness of private information retrieval” In *Proceedings of the IEEE Symposium on Security and Privacy*, 131–148, 2007. <https://doi.org/10.1109/SP.2007.23>
- [16] C. Aguilar-Melchor and P. Gaborit, “A lattice-based computationally-efficient private information retrieval protocol” In *Western European Workshop on Research in Cryptology*, 2007. <https://eprint.iacr.org/2007/446.pdf>
- [17] R. Sion, and B. Carbutar, “On the computational practicality of private information retrieval” In *Proceedings of the Network and Distributed Systems Security Symposium*, 2007. <https://zxr.io/research/sion2007pir.pdf>
- [18] W. Gasarch, “A survey on private information retrieval” *The Bulletin of the EATCS*, 82, 72–107, 2004. https://www.researchgate.net/profile/William_Gasarch/publication/266280304_A_survey_on_private_information_retrieval/links/5705098d08ae74a08e270e57.pdf
- [19] B. Chor and N. Gilboa, “Computationally private information retrieval” In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, 304–313, 1997. <https://doi.org/10.1145/293347.293350>
- [20] B. Chor, N. Gilboa, M. Naor, “Private Information Retrieval by Keywords” *Technion-IT, Department of Computer Science*, 1997.
- [21] E. Kushilevitz, and R. Ostrovsky, “Replication is not needed: single database, computationally-private information retrieval” In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*, 364–373, 1997. <https://doi.org/10.1109/SFCS.1997.646125>
- [22] C. Cachin, S. Micali, and M. Stadler, “Computationally private information retrieval with polylog communication” In *International Conference on the Theory and Applications of Cryptographic Techniques*, 402–414, 1999. <https://doi.org/10.1145/258533.258609>
- [23] C. Gentry, and Z. Ramzan, “Single-database private information retrieval with constant communication rate” *International Colloquium on Automata, Languages, and Programming*, 803–815, 2005. https://doi.org/10.1007/11523468_65
- [24] E. Kushilevitz, and R. Ostrovsky, “One-way trapdoor permutations are sufficient for non-trivial single-server private information retrieval” In *International Conference on the Theory and Applications of Cryptographic Techniques*, 104–121, 2000. https://doi.org/10.1007/3-540-45539-6_9
- [25] H. Lipmaa, “An Oblivious Transfer Protocol with Log-Squared Communication” In *International Conference on Information Security*, 314–328, 2005. https://doi.org/10.1007/11556992_23
- [26] C. Aguilar-Melchor, J. Barrier, L. Fousse, and M.O. Killijian, “XPIR: Private information retrieval for everyone” In *Proceedings on Privacy Enhancing Technologies*, 155–174, 2016. <https://doi.org/10.1515/popets-2016-0010>
- [27] A. Beimel and Y. Stahl, “Robust information-theoretic private information retrieval” *Journal of Cryptology*, 20(3), 295–321, 2007. <https://doi.org/10.1007/s00145-007-0424-2>
- [28] Y. Gertner, S. Goldwasser, T. Malkin, “A Random Server Model for Private Information Retrieval” In *2nd International Workshop on Randomization and Approximation Techniques in Computer Science*, 200–217, 1998. https://doi.org/10.1007/3-540-49543-6_17
- [29] S. Wang, X. Ding, R. H. Deng, and F. Bao, F, “Private information retrieval using trusted hardware” *European Symposium on Research in Computer Security*, 49–64, 2006. https://doi.org/10.1007/11863908_4
- [30] E. Fung, G. Kellaris, and D. Papadias, “Combining Differential Privacy and PIR for Efficient Strong Location Privacy” *International Symposium on Spatial and Temporal Databases*, 295–312, 2015. https://doi.org/10.1007/978-3-319-22363-6_16
- [31] S. Papadopoulos, S. Bakiras, and D. Papadias, “Nearest neighbor search with strong location privacy” In *Proceedings of the VLDB Endowment*, 3(1-2), 619–629, 2010. <https://doi.org/10.14778/1920841.1920920>
- [32] G. Ghinita, “Understanding the privacy-efficiency trade-off in location based queries” In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*, 1–5, 2008. <https://doi.org/10.1145/1503402.1503404>
- [33] P. Paillier, “Public key cryptosystems based on composite degree residue classes” In *International Conference on the Theory and Applications of Cryptographic Techniques*, 223–238, 1999. https://doi.org/10.1007/3-540-48910-X_16
- [34] A. Shamir, “How to Share a Secret” *Communications of the ACM*, 22(11), 612–613 1979. <https://doi.org/10.1145/359168.359176>
- [35] A. Pingley, W. Yu, N. Zhang, X. Fu, and W. Zhao, “CAP: A Context-Aware Privacy Protection System For Location-Based Services” In *29th IEEE International Conference on Distributed Computing Systems*, 49–57, 2009. <https://doi.org/10.1109/ICDCS.2009.62>
- [36] S. Hemisphere, “Northern Hemisphere” *Ann Arbor 1001*, 2006. (see also: https://en.wikipedia.org/wiki/Military_Grid_Reference_System)
- [37] Mapping support, <https://mappingsupport.com/p/gmap4.php?tilt=off&mgrs=14SPG34308382&z=5&t=t1>, last accessed 2018/11/05.
- [38] C. Devet, I. Goldberg, and N. Heninger, “Optimally Robust Private Information Retrieval” In *USENIX Security Symposium*, 269–283, 2012. <https://eprint.iacr.org/2012/083.pdf>
- [39] I. Goldberg. Percy++ project on SourceForge, <http://percy.sourceforge.net/>. last accessed 2018/11/05.
- [40] R. Henry, F. Olumofin, I. Goldberg, “Practical PIR for Electronic Commerce” In *Proceedings of the 18th ACM conference on Computer and communications security*, 677–690, 2011. <https://doi.org/10.1145/2046707.2046784>
- [41] W. Lueks, I. Goldberg, “Sublinear Scaling for Multi-Client Private Information Retrieval” In *International Conference on Financial Cryptography and Data Security*, 168–186, 2015. https://doi.org/10.1007/978-3-662-47854-7_10

Appendix A

A global, open, collaborative, standardized points of interest database provided by Factual [7].

	POI Type	Canada	US
1	Automotive	41913	617664
	Automotive, maintenance and repair	27267	285077
	Automotive, maintenance and repair, tires	0	73587
	Automotive, car parts and accessories	6917	79292
	Automotive, car dealers and leasing, car dealers	7729	105690
	Automotive, car dealers and leasing, used car	0	74018
2	Businesses and services, financial	35914	348814
	Businesses and services, financial, bank and finance, bank and credit union	14886	128993
	Businesses and services, financial, financial planning and investments	7919	91282
	Businesses and services, financial, access and bookkeeping	13109	128539
3	Businesses and services	93421	1181684
	Businesses and services, personal care, beauty salons and barbers	26148	325314
	Businesses and services, shipping freight, and material transportation	6232	51966
	Businesses and services, insurance	13687	224441
	Businesses and services, legal, attorney and law offices	12438	226907
	Businesses and services, real estate, real estate agents	9869	140201
	Businesses and services, Telecommunication services	6667	47031
	Businesses and services, computers	9684	92043
	Businesses and services, printing, copying and signage	8696	73781
4	Businesses and services, home improvement	138297	1403076
	Businesses and services, home improvement	93762	962116
	Businesses and services, home improvement, contractors	25612	260638
	Businesses and services, home improvement, ventilating and air conditioning, heating	5970	74885
	Businesses and services, home improvement, plumbing	6279	58550
	Businesses and services, home improvement, electrician	6674	46887
5	Community and government	57185	839554
	Community and government, organization and associations	14688	136389
	Community and government, education and secondary schools		
	Primary and secondary school	14907	192245
	Community and government, day care and preschools	6859	80547
	Community and government, religious, churches	14531	362889
	Community and government, public and social services	6200	67484
6	Healthcare	51359	969517
	Healthcare, dentists	18683	230012
	Healthcare, pharmacies	11840	62652
	Healthcare, hospitals, clinics and medical centers	8221	155416
	Healthcare, physicians	12615	521437
7	Retail	81534	758818
	Retail, furniture and décor	8656	96082
	Retail, fashion, shoes	5946	61335
	Retail, fashion, clothing and accessories	22886	193354
	Retail, fashion, jewelry and watches	6115	67001
	Retail, construction supplies	5776	42307
	Retail, supermarkets and groceries	10862	98231
	Retail, food and beverage, beer, wine and spirits	6014	54942
	Retail, convenience stores	9610	100149
	Retail, glasses	5669	45417
8	Social	179478	1911585
	Social, food and dining, restaurants	101714	1017499
	Social, food and dining, restaurants, fast food	17207	236356
	Social, food and dining, restaurants, dining	16033	0
	Social, food and dining, cafes, coffee and tea houses	15083	117367
	Social, food and dining, restaurants, pizza	12888	136018
	Social, food and dining, restaurants, American	0	225245
	Social, food and dining, restaurants, Chinese	5724	0
	Social, Bars	10829	179100
9	Transportation	17906	206107
	Transportation, gas stations	11783	161466
	Transportation, taxi and car services, car and truck rentals	6123	44641
10	Travel	27937	177682
	Travel, travel agents and tour operators	6505	35613
	Travel, lodging, hotel and motels	21432	142069

A Proposed Architecture for Parallel HPC-based Resource Management System for Big Data Applications

Waleed Al Shehri*, Maher Khemakhem, Abdullah Basuhail, Fathy E. Eassa

Department of Computer Science, King Abdul-Aziz University, Jeddah, KSA

ARTICLE INFO

Article history:

Received: 02 October, 2018

Accepted: 11 January, 2019

Online : 20 January, 2019

Keywords:

High-Performance Computing

Big Data

Load-Balancing

Data-Locality

Resource Management

Parallel Programming models

Power Consumption

ABSTRACT

Big data can be considered to be at the forefront of the present and future research activities. The volume of data needing to be processed is growing dramatically in both velocity and variety. In response, many big data technologies have emerged to tackle the challenges of collecting, processing and storing such large-scale datasets. High-performance computing (HPC) is a technology that is used to perform computations as fast as possible. This is achieved by integrating heterogeneous hardware and crafting software and algorithms to exploit the parallelism provided by HPC. The performance capabilities afforded by HPC have made it an attractive environment for supporting scientific workflows and big data computing. This has led to a convergence of the HPC and big data fields.

However, big data applications usually do not fully exploit the performance available in HPC clusters. This is so due to such applications being written in high-level programming languages and do not provide support for exploiting parallelism as do other parallel programming models.

The objective of this research paper is to enhance the performance of big data applications on HPC clusters without sacrificing the power consumption of HPC. This can be achieved by building a parallel HPC-based Resource Management System to exploit the capabilities of HPC resources efficiently.

1. Introduction

The amount of data produced in the scientific and commercial fields is growing dramatically. Correspondingly, big data technologies, such as Hadoop and Spark, have emerged to tackle the challenges of collecting, processing, and storing such large-scale data.

There are different opinions on the definition of big data resulting from different concerns and technologies. One definition applies to datasets that cannot be realized, managed and analyzed with traditional IT software. This definition reflects two connotations: data volume that is growing and changing continuously; and, this growing volume is different from one big data application to another [1]. A more specific definition based on the multi-V model by Gartner in 2012: “Big Data are high-volume, high-velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization” [2].

While the focus of big data applications is on handling enormous datasets, high-performance computing (HPC) focuses on performing computations as fast as possible. This is achieved by integrating heterogeneous hardware and crafting software and algorithms to exploit the parallelism provided by HPC [3]. The performance capabilities afforded by HPC have made it an attractive environment for supporting scientific workflows and big data computing. This has led to a convergence of the HPC and big data fields.

Unfortunately, there is usually a performance issue when running big data applications on HPC clusters because such applications are written in high-level programming languages. Such languages may be lacking in terms of performance and may not encourage or support writing highly parallel programs in contrast to some parallel programming models like Message Passing Interface (MPI) [4]. Furthermore, these platforms are designed as a distributed architecture, which differs from the architecture of HPC clusters [5].

* Waleed Al Shehri, : Email: waleed.ab2@gmail.com

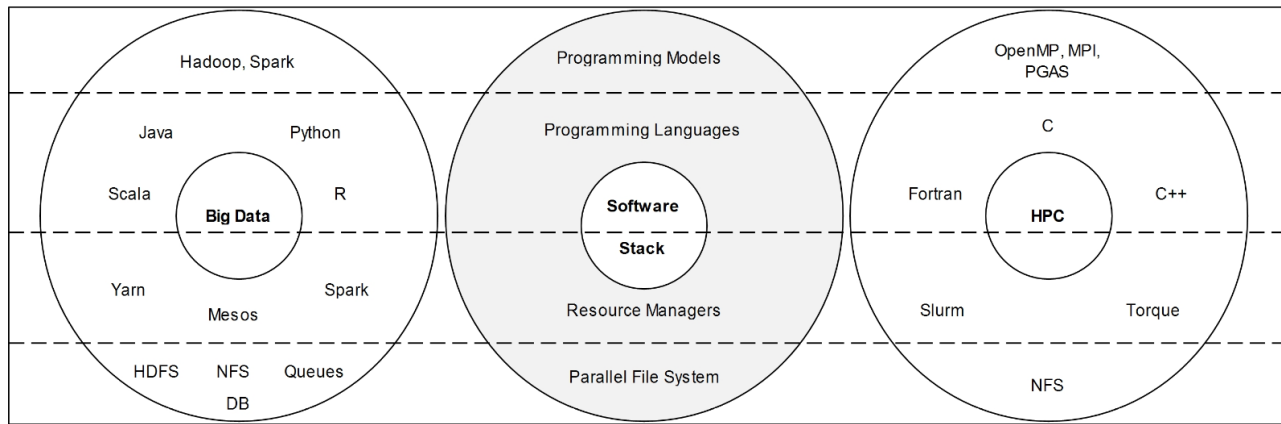


Figure 1. HPC and Big Data Software Stacks.

Additionally, the large volume of big data may hinder parallel programming models such as Message Passing Interface (MPI), Open Multi-Processing (OpenMP) and accelerator models (CUDA, OpenACC, OpenCL) from supporting high levels of parallelism [1].

Furthermore, resources allocation in HPC is one of the prime challenges, especially since HPC and big data paradigms has a different software stack [6] as shown in (Figure 2):

2. Related Work

The related work can be organized based on the different aspects required to fulfill the architecture requirements of this research. The job scheduler concept will be highlighted by considering its features and functionality. Moreover, different comparative studies will be introduced covering big data programming models and parallel programming models to establish the performance gap between them. Other works are presented to show how the performance of these programming models can be enhanced. Finally, research involving data locality approaches and decomposition mechanisms covering the same context of this research is reviewed.

A job scheduler can play an essential role in modern big data platforms and HPC systems. It manages different compute jobs related to different users on homogenous or heterogeneous computational resources. It can have different names that reflect the same mechanism such as scheduler, resource manager, resource management system (RMS), and a distributed resource management system (D-RMS) [7]. Despite significant growth in terms of heterogeneity of resources and job complexity and diversity, job schedulers still have the main core function of job queuing, scheduling and resource allocation, and resource management [8][9]. In [7], many features are analyzed of the most popular HPC and big data schedulers including Slurm, Son of Grid Engine, Mesos, and Hadoop YARN.

Additionally, there are two primary job types: job arrays and parallel jobs. In job arrays, multiple independent processes for a single job identifier can be run with different parameters for each process. In parallel jobs, it is possible to launch each of the processes simultaneously, allowing communication between them during the computation. While HPC schedulers support both types, big data schedulers can support only job arrays. Furthermore, there

are many important features of HPC schedulers, generally not available with big data schedulers, such as job chunking, gang scheduling, network aware scheduling and power-aware scheduling.

Big data jobs are usually considered to be network-bound regarding a large amount of data movement between different nodes among clusters. In [10], traffic forecasting and job-aware priority scheduling for big data processing is proposed by considering the dependencies of the flows. The network traffic for flows of the same job is forecasted via run-time monitoring, then a unique priority for each job is assigned by tagging every packet in the job. Finally, it uses a FIFO order for scheduling flows of the same priority.

In [11], a new backfilling algorithm, known as fattened backfilling, is proposed to provide more efficient backfilling scheduling. In this algorithm, short jobs can be moved forward if they do not delay the first job in the queue. A Resource and Job Management System (RJMS) based on a prolog/epilog mechanism has been proposed in [12]. It allows communication between HPC and Big Data systems by reducing the disturbance on HPC workloads while leveraging the built-in resilience of Big Data frameworks.

Processing tremendous volumes of data on dedicated big data technology is not as fast as processing the data on HPC infrastructure. This fact is recognized when comparing the efficiency of low-level programming models in HPC, which supports more parallelism, with big data technologies that are written with high-level programming languages. Many practical case studies and research have confirmed this fact. In [13], sentiment analysis on Twitter data was conducted for different dataset sizes using an MPI environment that showed better performance than using Apache Spark.

The enhancement of big data programming models can be achieved by integrating them with parallel programming models such as MPI. This approach can be seen in [4] that showed how to enable the Spark environment using the MPI libraries. Although this technique indicates remarkable speedups, it must use shared memory, and there are other overheads as a potential drawback. In [14], a scalable MapReduce framework, named Glasswing, is introduced. It is configured to use a mixture of coarse- and fine-grained parallelism to obtain high performance on multi-core

CPUs and GPUs. The performance of this framework is evaluated using five MapReduce applications with the indication that Glasswing outperforms Hadoop in terms of performance and resource utilization.

Data locality is a critical factor that affects both performance and energy consumption in HPC systems [15]. Many big data frameworks such as MapReduce and Spark can support this concept by sending the computation to where the data resides. In contrast, parallel programming models such as MPI lack this advantage. A novel approach by Yin et al. in [16], named DL-MPI, is proposed for MPI-based data-intensive applications to support data locality computation. It uses a data locality API that allows MPI-based programs to obtain data distribution information for compute nodes. Moreover, it proposes a probability scheduling algorithm for heterogeneous runtime environments that evaluate the unprocessed local data and the computing ability of each compute node.

In [17], a data distribution scheme is used by abstracting NUMA hardware peculiarities away from the programmer and delegating data distribution to a runtime system. Moreover, it uses task data dependence information, which is available with OpenMP 4.0RC2, as a guideline for scheduling OpenMP tasks to reduce data stall times.

Partitioning or decomposition is the first step for designing a parallel program by breaking down problems into small tasks. It includes two main types: domain or data decomposition and function decomposition [18]. These two types can be combined as mixed parallelism that employs an M-SPMD (multiple-single

program multiple data) architecture, which includes both task parallelism (MPMD) and data parallelism (SPMD) [19].

The choice of decomposition type and parallelism paradigm is determined by resource availability. Furthermore, these resources may define the granularity level that the system can support [20]. There have been few empirical studies for performing data decomposition in the HPC field, [21] investigated this approach when designing parallel applications. Additionally, the state-of-practice was studied with probing tools used to perform this function. Moreover, a set of key requirements was derived for tools that support data decomposition and communication when parallelizing applications.

Based on this previous related work and to the best of our knowledge, there is no contribution yet that employs all the previous factors in terms of using a hybrid parallel programming model, decomposition technique and granularity approach to build a HPC-based Resource Management System for enhancing the performance of big data applications and optimizing HPC resource utilization. The contribution of this paper is to address this gap.

3. The Proposed Architecture

The proposed architecture will be built based on different techniques that will be integrated together to constitute a parallel HPC-based Resource Management System that enhances the performance of big data applications on HPC clusters without sacrificing the power consumption of HPC. In more details, the system will have the following techniques (Figure 2):

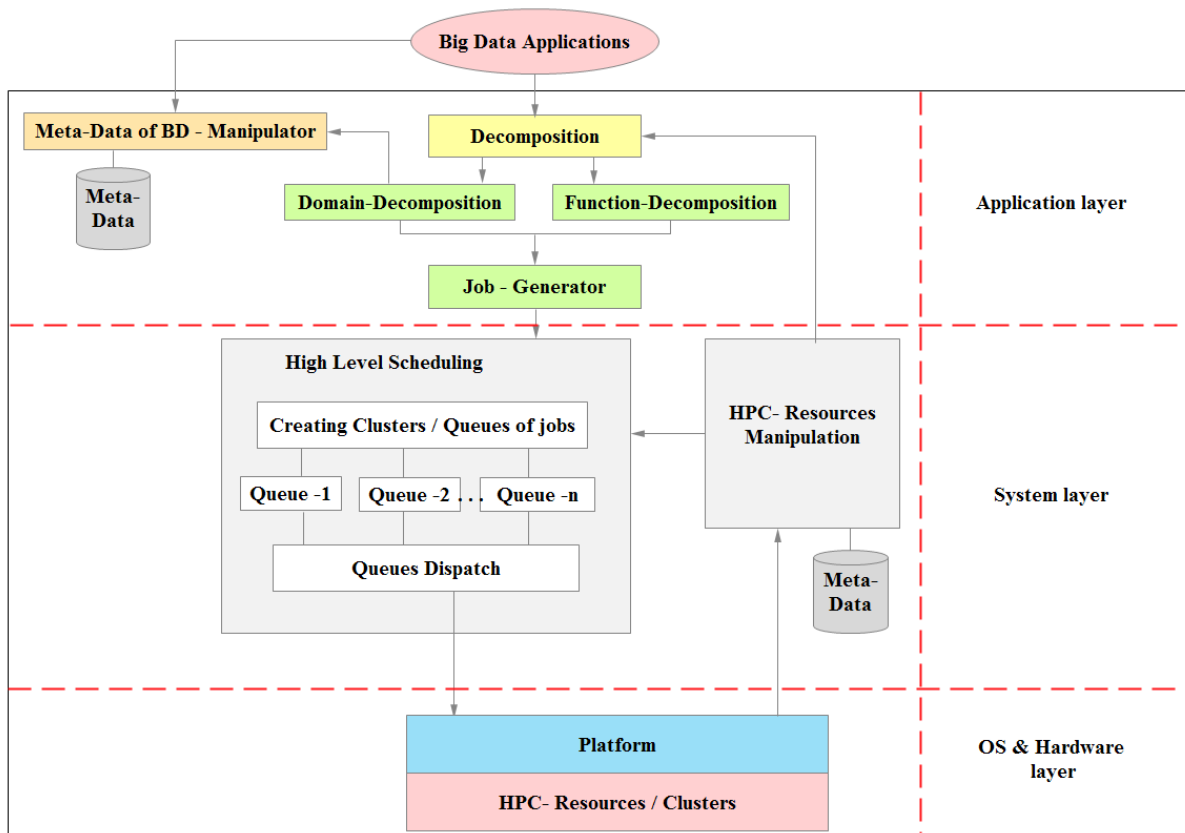


Figure 2: High-level architecture for the proposed system.

• A new technique for HPC resources manipulation

This dynamic technique will have some functionality related to HPC resources including:

- Collecting a resource metadata to constitute a repository containing HPC resources with their capabilities and availabilities.
- Tracking status of HPC resources.
- Updating the metadata repository from time to time.

• New decomposition techniques

In this technique, we will provide a domain decomposition technique and/or a function decomposition technique for the given big data applications targeting both parallel computing architectures: Single Program Multiple Data (SPMD) and/or Multiple Instruction Multiple Data (MIMD). The availability of resources and system architecture can determine the decomposition paradigm and granularity options that efficiently support the resource management system.

• A new high-level scheduling technique

This technique will receive the decomposition results and creates clusters/queues of jobs during scheduling time based on the metadata about HPC resources/ clusters. It will consider data locality and load balancing while performing this task. Once clusters/queues of jobs are created, they will be dispatched to being executed by OS-platform on HPC resources/clusters.

4. Evaluation and comparative study

By comparing our architecture to other techniques, it is noticeable that the proposed architecture considers all the critical factors to achieve the performance and scalability attributes. Primarily, focusing on the topology awareness and building a metadata repository about the availabilities and capabilities of HPC resources can play a critical role to support the decomposition and high-level scheduling techniques. Such metadata can enhance the decision-making about choosing suitable granularity options and parallelism paradigm. Furthermore, the high-level scheduling technique can exploit the HPC resources positively by taking in account data locality and load balancing.

Instead of Integrating some big data and parallel programming models, this architecture constitutes an independent big data platform that employ hybrid parallel programming to support high parallelism for CPUs and GPUs accelerators.

The scalability can be seen from adding more dedicated clusters as needed. Adding more clusters will not affect the essence of each technique in particular, and resource management as a whole system integrating these techniques.

Different performance metrics have to be considered to implement the proposed architecture efficiently. Big data is the primary stream of this architecture, thus data building time is a significant metrics that may affect the performance. This time is required to construct a data structure used for computation and to perform the decomposition technique. Furthermore, employing parallel programming models such as MPI and OpenAcc can affect the computation time positively. From the part of HPC, hardware

utilization metrics is also a cornerstone of the proposed architecture particularly for improving high-level scheduling technique

The novelty of this architecture can be arisen from having metadata about both big data applications and HPC resources, which leads to scheduling current jobs to the most suitable and available resources or cluster.

5. Conclusion

HPC has become an attractive environment for supporting scientific workflows and big data computing due to its performance capabilities. Unfortunately, big data applications usually do not fully exploit these capabilities afforded by HPC clusters, because such applications were written in high-level programming languages that do not encourage parallelism as parallel programming models. Another reason is that the architecture of big data platforms defers from the HPC architecture. A parallel HPC-based Resource Management System is proposed in this paper to enhance the performance of big data applications on HPC clusters without sacrificing the power consumption of HPC. For the future work, the High-level architecture for the proposed system will be developed and evaluated by running some big data applications. Moreover, some performance benchmarks will be provided to reflect the efficiency of our system.

References

- [1] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, 2014.
- [2] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci. (Ny)*, vol. 275, pp. 314–347, 2014.
- [3] D. A. Reed and J. Dongarra, "Exascale computing and big data," *Commun. ACM*, vol. 58, no. 7, pp. 56–68, 2015.
- [4] M. Anderson et al., "Bridging the gap between HPC and big data frameworks," *Proc. VLDB Endow.*, vol. 10, no. 8, pp. 901–912, 2017.
- [5] P. Xuan, J. Denton, P. K. Srimani, R. Ge, and F. Luo, "Big data analytics on traditional HPC infrastructure using two-level storage," *Proc. 2015 Int. Work. Data-Intensive Scalable Comput. Syst. - DISCS '15*, pp. 1–8, 2015.
- [6] H. R. Asaadi, D. Khaldi, and B. Chapman, "A comparative survey of the HPC and big data paradigms: Analysis and experiments," *Proc. - IEEE Int. Conf. Clust. Comput. ICC3*, pp. 423–432, 2016.
- [7] A. Reuther et al., "Scalable system scheduling for HPC and big data," *J. Parallel Distrib. Comput.*, vol. 111, pp. 76–92, 2018.
- [8] P. Jones, "NAS for Job Requirements Queuing / Scheduling Checklist Software," 2018.
- [9] W. Saphir, L. A. Tanner, B. Traversat, W.~Saphier, L.A.~Tanner, and B.~Traversat, "Job Management Requirements for {NAS} Parallel Systems and Clusters," *IPPS Work. Job Sched. Strateg. Parallel Process.*, no. 949, pp. 319–336, 1995.
- [10] Z. Wang and Y. Shen, "Job-Aware Scheduling for Big Data Processing," *Proc. - 2015 Int. Conf. Cloud Comput. Big Data, CCBD 2015*, pp. 177–180, 2016.
- [11] C. Gómez-Martín, M. A. Vega-Rodríguez, and J. L. González-Sánchez, "Fattened backfilling: An improved strategy for job scheduling in parallel systems," *J. Parallel Distrib. Comput.*, vol. 97, pp. 69–77, 2016.
- [12] M. Mercier, D. Glesser, Y. Georgiou, and O. Richard, "Big Data and HPC collocation : Using HPC idle resources for Big Data Analytics," pp. 347–352, 2017.
- [13] D. S. Kumar and M. A. Rahman, "Performance evaluation of Apache Spark Vs MPI: A practical case study on twitter sentiment analysis," *J. Comput. Sci.*, vol. 13, no. 12, pp. 781–794, 2017.
- [14] I. El-Helw, R. Hofman, and H. E. Bal, "Glasswing," *Int. Symp. High-Performance Parallel Distrib. Comput.*, pp. 295–298, 2014.
- [15] D. Unat et al., "Trends in Data Locality Abstractions for HPC Systems," *IEEE Trans. Parallel Distrib. Syst.*, pp. 1–1, 2017.
- [16] J. Yin, A. Foran, and J. Wang, "DL-MPI: Enabling data locality computation

for MPI-based data-intensive applications,” *Proc. - 2013 IEEE Int. Conf. Big Data, Big Data 2013*, pp. 506–511, 2013.

- [17] A. Muddukrishna, P. A. Jonsson, and M. Brorsson, “Locality-aware task scheduling and data distribution for OpenMP programs on NUMA systems and manycore processors,” *Sci. Program.*, vol. 2015, pp. 156–170, 2015.
- [18] B. Ren, S. Krishnamoorthy, K. Agrawal, and M. Kulkarni, “Exploiting Vector and Multicore Parallelism for Recursive, Data- and Task-Parallel Programs,” *Proc. 22nd ACM SIGPLAN Symp. Princ. Pract. Parallel Program. - PPOPP '17*, pp. 117–130, 2017.
- [19] V. Boudet, F. Desprez, and F. Suter, “One-step algorithm for mixed data and task parallel scheduling without data replication,” *Proc. - Int. Parallel Distrib. Process. Symp. IPDPS 2003*, no. October 2015, 2003.
- [20] L. Silva and R. Buyya, “Parallel programming models and paradigms,” *High Perform. Clust. Comput. Archit. ...*, pp. 4–27, 1999.
- [21] A. Meade, D. K. Deeptimahanti, J. Buckley, and J. J. Collins, “An empirical study of data decomposition for software parallelization,” *J. Syst. Softw.*, vol. 125, pp. 401–416, Mar. 2017.

Morphological and Optoelectrical Characterization of Silicon Nanostructures for Photovoltaic Applications

Babacar Dieng^{*1}, Moussa Toure², Modou Beye¹, Diouma Kobor², Amadou Seidou Maiga¹

¹Laboratory of Electronic, Informatics, Telecommunication and Renewable Energy (LEITER), Gaston Berger University of Saint-Louis, Senegal

²Laboratory of Chemistry and Physics of Materials (LCPM), Assane Seck University of Ziguinchor, Senegal

ARTICLE INFO

Article history:

Received: 14 August, 2018

Accepted: 23 December, 2018

Online: 21 January, 2019

Keywords:

Silicon nanostructures

Metal assisted chemical etching

Optoelectrical characterization

ABSTRACT

Metal (silver)-Assisted Chemical Etching (MACE) method is used to fabricate silicon nanostructures like silicon nanowires (SiNWs) and silicon nanocones (SiNCs). The morphological characterization of fabricated SiNWs has shown that 5 minutes is the optimal time of silver deposition on silicon substrate. Silicon nanocones (SiNCs) were also fabricated by etching vertical SiNWs with a $\text{AgNO}_3/\text{HF}/\text{H}_2\text{O}_2$ solution. The optical and electrical properties of SiNWs and SiNCs are analyzed and compared with those of the bulk silicon. The fabricated SiNWs (SiNCs) reduce the surface reflectance and the sheet resistance down to 6% (3%) and 9.143 Ω/sq (6.997 Ω/sq) respectively.

1 Introduction

The energy demand worldwide is largely greater than the energy supply. Thus, the conventional energy sources (oil, coal and nuclear) become expensive. In the other way, these conventional energies have negative impact on the environment and the human health. The explosions of the Chernobyl and Fukushima nuclear power stations cause damage on the human health in many countries [1][2]. Consequently, the development of alternative energy sources is urgently needed. Solar energy, as a clean, abundant and renewable source, is one of the most promising alternatives to conventional energies.

The incident light is converted into electric current by photovoltaic (PV) cells. The PV cells are fabricated by using semiconductor materials. Among the semiconductor used in the PV cell fabrication, the bulk silicon (mono or multi-crystalline) is mostly dominated. The best conversion efficiencies obtained with monocrystalline (multi-crystalline) silicon solar cells and modules are around 26.7% (21.9%) and 24.4% (19.9%) respectively [3]. However, the crystalline silicon represent about 40% of the production cost of bulk silicon modules [4].

Further cost reduction and efficiency enhancement, through reduction of material usage, simplification

of device fabrication as well as optimization of device structure, are required. A way to take up this challenge is to use silicon nanostructures, which theoretically should be both more efficient and less expensive [5].

The common methods to fabricate silicon nanostructures (SiNWs, SiNCs) are dry etching and lithography methods [3][6]. However, these methods are complex, expensive, and unsuitable for mass production. One of the alternative methods is to use Metal Assisted Chemical Etching (MACE) which is based on the strong catalytic activity of metals in $\text{HF}/\text{H}_2\text{O}_2$ aqueous solution [7]. This low cost method is very easy to implement and it allows to realize various silicon nanostructures with excellent optoelectronic properties without changing the purity of silicon.

Dimova Malinovska et al. [8] are the first to etch the silicon by using MACE method. They demonstrated that porous silicon can be obtained by using aluminum metal and $\text{HF}/\text{HNO}_3/\text{H}_2\text{O}$ solution. Since 2002, the MACE method is being improved and a variety of nanostructures are fabricated. SiNWs were firstly fabricated by Peng et al. They used the mixture of HF/AgNO_3 [9][10]. Vertical SiNWs were fabricated by combining MACE and lithography methods [11]. However, the cost of these nanowires is relatively high because of the lithography technique used to deposit the PS sphere masks. Moreover, Au metal was used as catalyst. To realize best silicon nanostructures with-

*Babacar Dieng, Gaston Berger University of Saint-Louis, Senegal, Email: dieng.babacar@ugb.edu.sn

out using any mask patterns and expensive metals is required in the solar cell manufacturing.

The objective of this work is to fabricate silicon nanostructures (nanowires and nanocones) by MACE method in ambient temperature, without using any mask patterns, and to study their structural and optoelectronic properties.

2 Materials and method

The silicon nanostructures were fabricated by using (100) p-type monocrystalline silicon substrate. These samples have a resistivity and a thickness in the ranges of 1-5Ω/cm and 600 - 650 μm, respectively.

The different steps of the SiNWs fabrication by MACE are summarized in the figure 1. First, the samples were cleaned by using, respectively, acetone, ethanol and deionized water in the ultrasonic bath for 15 minutes. After this cleaning, the samples were immersed in a piranha solution (H_2SO_4/H_2O_2) for 10 minutes to eliminate any organic trace. The last step of the cleaning process was to use a diluted HF solution to remove the native silicon oxide SiO_2 and to rinse the samples with deionized water. After the cleaning steps, the samples were immersed in the $HF/AgNO_3$ (4.8M/0.02M) solution for a chosen time in [0.5, 2.5, 5, 10, 15, 30, 60, 120 minutes]. This solution allows to deposit silver (Ag) nanoparticles on the silicon surface. After depositing Ag nanoparticles, samples were directly immersed in HF/H_2O_2 (4.8M/1.176.10⁻³M) mixture. This solution is responsible for the etching silicon. After the etching process, all Ag waste integrated into the silicon substrate must be removed. For this, a diluted HNO_3 solution was used.

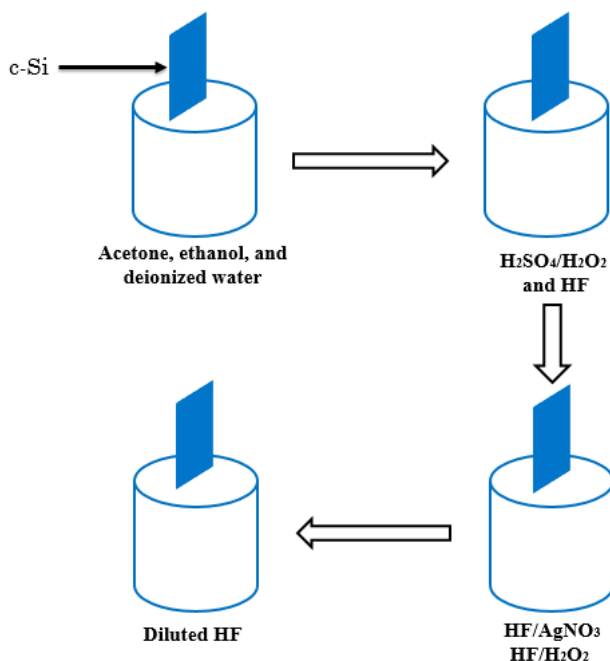


Figure 1: Different steps in the fabrication of the silicon nanowires by MACE

The conical form of silicon nanostructures can be fabricated by MACE via obtained silicon nanowires (SiNWs). For this, SiNWs are introduced into $AgNO_3/HF/H_2O_2$ solution. This solution have double functions. First, it deposit Ag nanoparticles on the top of SiNWs. And second, it etch these SiNWs at the top to form truncated SiNWs like silicon nanocones (SiNCs).

It is interesting then to understand how these nanostructures occur on the surface of silicon substrate.

3 Mechanism of the formation of silicon nanostructures

The growth mechanism of SiNWs is described by Peng et al.[12]. The Ag^+ ions given by $HF/AgNO_3$ seize the valence band electrons of silicon. Thus, these ions become now Ag nucleation. Ag nuclei gradually grow into Ag nanoparticles. The reduction of H_2O_2 oxydant is done at Ag nanoparticles; and the holes are consumed by the silicon oxidation. The produced silicon oxyde is quickly dissolved by HF. Ag nanoparticles travel into the silicon wafer as the silica layer is dissolved, thus the depth of Ag nanoparticles in the pit increase gradually with the increase of the reaction time.

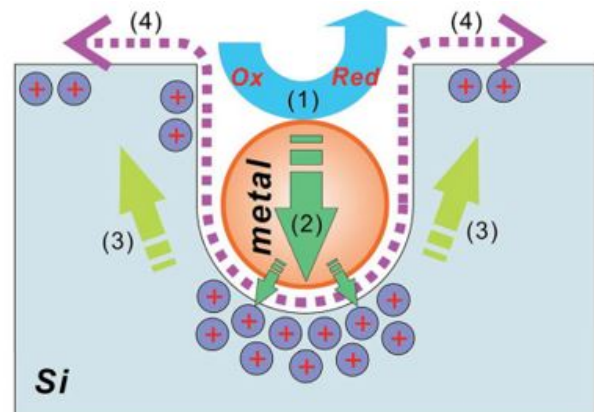
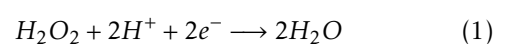
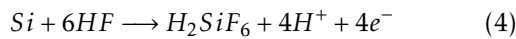
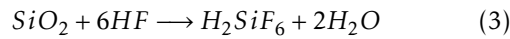
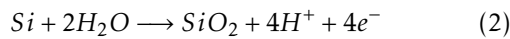


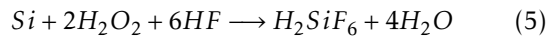
Figure 2: Formation mechanism of silicon nanowires fabrication: (1) reduction of oxidant H_2O_2 , (2) injection of generated holes onto the silicon substrate, (3) migration of holes inside the silicon surfaces, and (4) dissolution of silicon oxide by HF [13]

The Ag nanoparticle side in contact with the etching solution acts as a cathode and serves to reduce H_2O_2 (see the equation 1). The other side of Ag particles which is in contact with the silicon substrate functions as an anode that serves to oxidize silicon which generates H^+ and electrons (see equations 2 to 4). A considerable difference of potentials between the cathode and anode sites is then created and therefore a local oxidation current flows from the cathode site to the anode one.





The overall equation is given by:



The same reactions can explain the formation of silicon nanocones by etching the SiNW tops. Indeed, when the sample is put in $AgNO_3/HF/H_2O_2$ solution, Ag nanoparticles are deposited on the SiNWs. Thus, H_2O_2 oxydises the SiNW tops and silicon oxyde is reduced by HF. These oxydation and dissolution will occur at the same time during the etching process. Then, silicon nanocones will be formed on silicon substrate by truncating the edge top of SiNWs, according to the study of Shimizu et al.[12].

A MERLIN Scanning Electron Microscope (SEM) with Zeiss FEG was used for structural characterization of the elaborated silicon nanowires. The spectral reflectance of fabricated silicon nanostructures is measured by Hitachi UV-VIS-NIR 4001 spectrophotometer equipped with an integrating sphere. Their electrical properties are studied by measuring the sheet resistance with 4-point probes method.

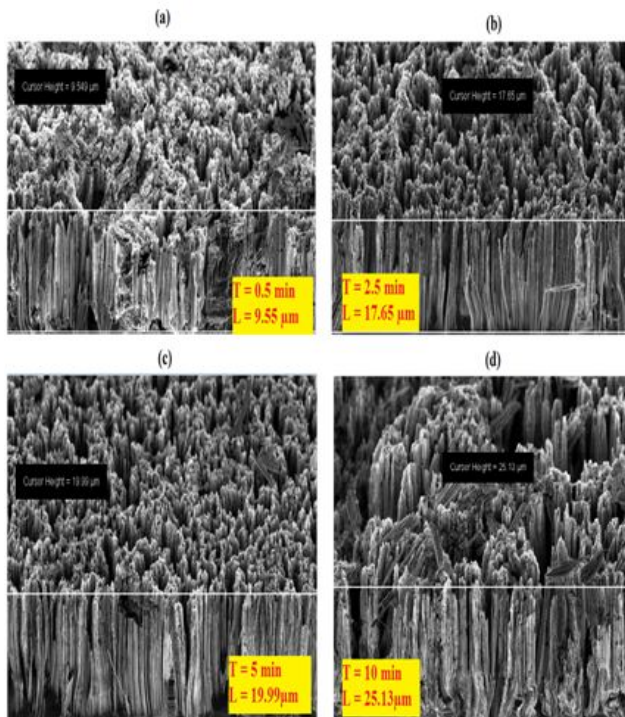


Figure 3: SEM images of silicon nanowires fabricated by the MACE method

4 Results and discussions

Figures 3 and 4 show the SEM characterization results of the fabricated SiNWs. The SiNW height (L) depends on the deposition time of silver nanoparticles (T). The SEM characterization results show that L increases

with T. Therefore continuous film of Ag nanoparticles can be obtained by increasing T. This continuous film with a few holes permits to obtain well-defined SiNWs [13]. However, a spongy structure is shown when T is very high (more than 30 minutes). It can be explained by the break of long SiNWs. Moreover, a structure with too much spaced SiNWs, is obtained for too long deposition time of silver nanoparticles. This spacing of nanowires as function of the etching time could be due to the phenomenon of the coalescence of silver nanoparticles which leads to the etching of the silicon under the silver macroparticles.

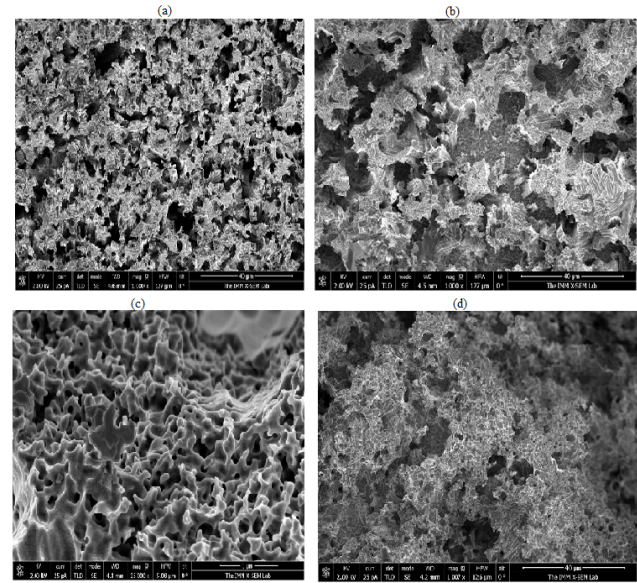


Figure 4: SEM images of silicon nanowires fabricated by MACE method for (a) 15 minutes, (b) 30 minutes, (c) 60 minutes and (d) 120 minutes.

The reflectance of the fabricated SiNWs was measured and the results are shown in figure 5. In this figure, the reflectance of bulk silicon is used as the reference. All SiNWs give very low reflectance in the useful wavelength range. The decrease of reflectance is possible by increasing the SiNWs height. This occurs because very long SiNWs lead to multiple rebounds of the light inside the structure and to the increase of the probability of light absorption. However, the SiNWs must be vertical to the substrate as well found in Figure 3c. Thus, the 5-minutes processed SiNWs give the best antireflective properties. They allow to reduce the surface reflectance of silicon below 6% in the 400-1100 nm wavelength range.

In order to improve the silicon antireflective properties, nanocone arrays were formed on silicon substrate. The 5-minutes processed SiNWs were used to fabricate silicon nanocone arrays because they give the best optical and structural results (figures 3c and 5). The optical characterization results of the fabricated nanocone arrays are shown in the figure 6. According to these results, all the nanocones arrays give the lowest reflectance compared to the SiNWs one. The reflectance around 3% is obtained with the 5-minutes fabricated nanocone arrays.

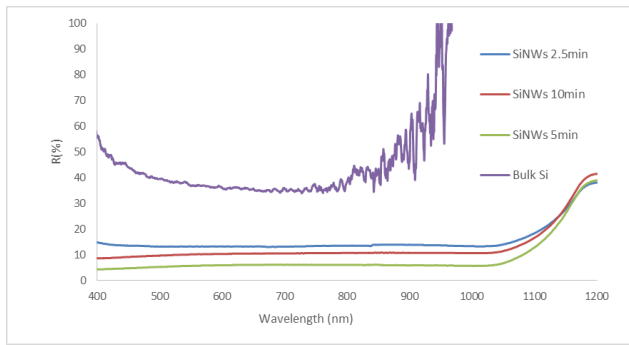


Figure 5: Reflectance of bulk silicon and silicon nanowires (SiNWs).

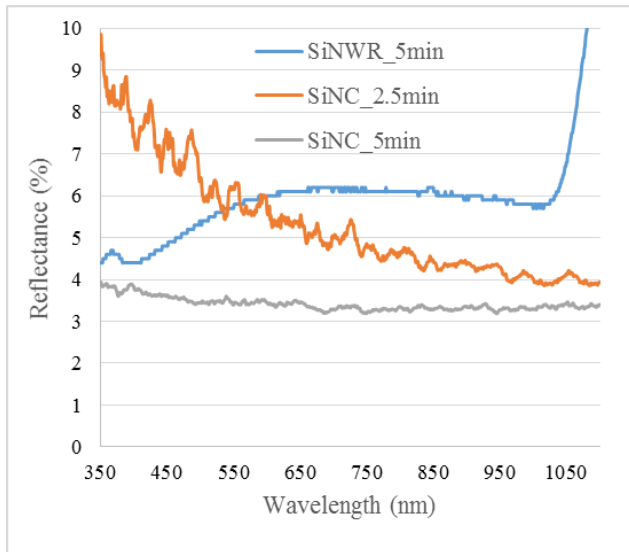


Figure 6: Reflectance of fabricated silicon nanocones (SiNC) compared with SiNW one.

The large index mismatch between air and silicon substrate can be reduced by using ergonomic tapered nanostructures which grade the refractive index from silicon to the air. This grading refractive index is observed for all nanostructure arrays, specially nanocone arrays. They allow to obtain a volume fraction variation between the top and the bottom of nanocones. Furthermore, the nanostructures diffract incoming light and increase the photon path length within the absorber layer.

A comparative study of the electrical properties was also carried out between the bulk monocrystalline silicon, and silicon nanostructures (SiNWs and SiNCs). The surface resistivity called also sheet resistance (R_{sheet}) permits to evaluate the electrical conductivity of a nanostructured surface. It can be defined as the ratio of the voltage drop per unit length to the surface current per width. The common instrument to measure the sheet resistance is the four-point probes method. It consists to apply a current (I) on the outer two probes and to measure the resultant voltage drop (ΔV) between the inner two drop. Thus, the sheet resistance is calculated by the following equation [14]:

$$R_{sheet} = \frac{\pi}{\ln(2)} * \frac{\Delta V}{I} \quad (6)$$

Ohm-per-square (Ω/sq) is the measurement unit of the sheet resistance. The resistivity of sample can be defined as the product of their sheet resistance and thickness. According to the results shown in the figures 7 to 9, the sheet resistance of silicon is reduced by nanostructuring the silicon substrate surface. The sheet resistance of 9.143 and 6.997 Ω/sq have been obtained for silicon nanowires and silicon nanocones respectively.

A material with a low sheet resistance is better to transfer the electrical charge. Furthermore, the resistivity and conductivity can be calculated if the sheet resistance and material thickness are known. This allows for the materials to be electrically characterized, by measuring their sheet resistance. Therefore, the electric conductivity of SiNWs and SiNCs (with low sheet resistances) are far better than that of bulk silicon.

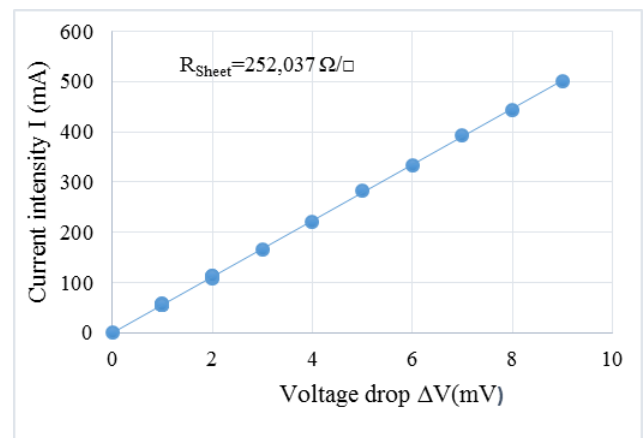


Figure 7: Variation of I(mA) as function of ΔV (mV) of bulk monocrystalline silicon.

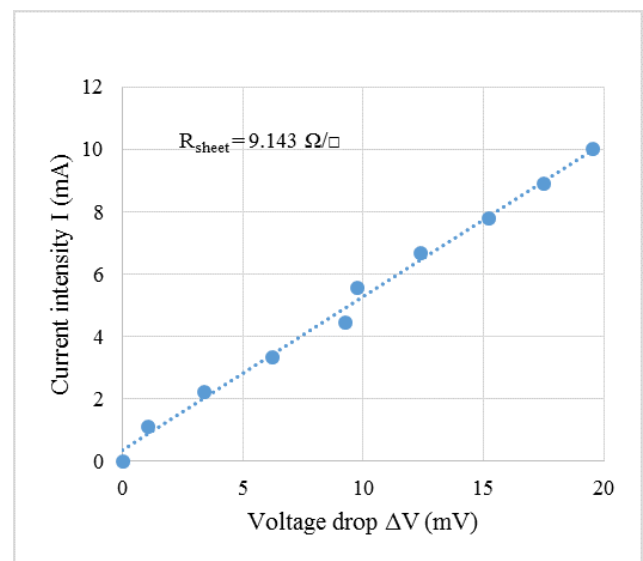


Figure 8: Variation of I(mA) as function of ΔV (mV) of 5-minutes fabricated silicon nanowires.

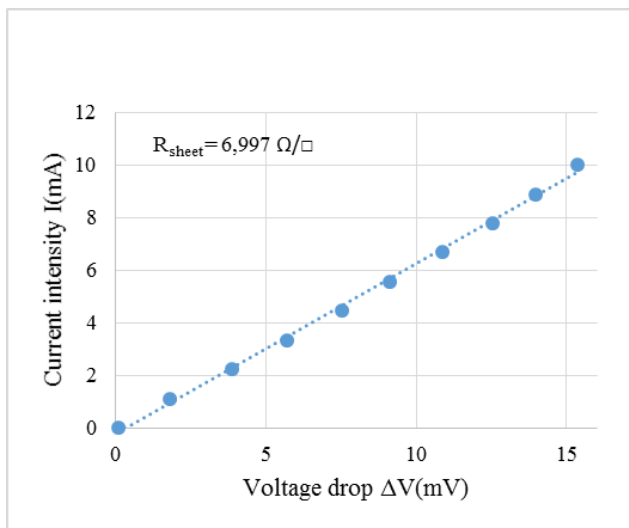


Figure 9: Variation of I (mA) as function of ΔV (mV) of silicon nanocones fabricated by etching 5-minutes processed SiNWs.

5 Conclusion

Silicon nanostructures (SiNWs and SiNCs) have been fabricated by metal-assisted chemical etching. The SEM characterizations have shown that SiNWs fabricated for 5 minutes are vertically straighter than the others. As a result, they exhibit lower reflectance (around 6%) in the 400-1100 nm wavelength range. The SiNCs were also fabricated by etching 5-minutes processed SiNWs. Compared with SiNWs, SiNCs give the best antireflective properties. They allow to reduce the reflection down to 3% in the 350-1050 nm wavelength range. Measurements and comparison of sheet resistances from bulk silicon and silicon nanostructures, have demonstrated an improvement in the property of surface electrical conductivity.

We expect that with high quality surface passivation, performance of solar cells based on these fabricated silicon nanostructures can be greatly enhanced.

Acknowledgment This work was carried out with the support of the CEA-MITIC (Centre d'Excellence Africain en Mathématique, Informatique et TIC).

References

- [1] Bilan du Forum Tchernobyl, organis par l'AIEA, l'OMS et le PNUD, septembre 2005
- [2] S. Alexander et Reuters, Fukushima evacuation has killed more than earthquake and tsunami, survey says, NBC News.com, 10 septembre 2013.
- [3] M.A. Green, Y. Hishikawa, W. Warta, et al, Solar cell efficiency tables (version 50), Progress in Photovoltaics 25, 2017, 668-676
- [4] V. K. Sethi, M. Pandey and P. Shukla, Cost Boundary in Silicon Solar Panel, International Journal of Chemical Engineering and Applications, vol. 2, pp. 372-375, 2011.
- [5] Bo Hua, Qingfeng Lin, Qianpeng Zhang and Zhiyong Fan, Efficient photon management with nanostructures for photovoltaics, Nanoscale, 5 2013, 66276640
- [6] H. Wang, K. Lai, Y. Lin, C. Lin, and J. He, Periodic Si nanopillars arrays fabricated by colloidal lithography and catalytic etching for broadband and omnidirectional elimination of Fresnel reflection, Langmuir 26 (15), 2010, 12855-12858
- [7] J. Lui, M. Ashmkan, B. Wang, and F. Yi, Fabrication and reflection properties of silicon nanopillars by cesium chloride self-assembly and dry etching, Applied Surface Science 258, 2012, 8825-8830
- [8] D. Dimova Malinowska, M. Sendova Vassileva, N. Tzenov, M. Kamenova, Preparation of thin porous silicon layers by stain etching, Thin solid films, vol. 297, pp. 9-12, 1997.
- [9] X. Li, Metal assisted chemical etching for high aspect ratio nanostructures: A review of characteristics and applications in photovoltaics, Current Opinion in Solid State and Materials Science 16, 2012, 7181
- [10] K. Peng, Y. Yan, S. Gao, and J. Zhu, Synthesis of Large-Area Silicon Nanowire Arrays via Self-Assembling Nanoelectrochemistry, Adv. Mater. No. 16, 2002, 1164-1167
- [11] K Peng, Y Yan, S Gao, and J Zhu, Dendrite-assisted grow of silicon nanowires in electroless metal deposition, Advanced functional materials, Vol. 13, No. 2, 2003, 127-132
- [12] T. Shimizu, N. Tanaka, Y/ Tada, Y. Hara, N. Nakamura, J. Taniuchi, K. Takase, T. Ito, and S. Shingubara, Fabrication of nanocone arrays by two step metal assisted chemical etching method, Microelectronic Engineering 153, 2016, 5559
- [13] Z. Huang, N. Geyer, P. Werner, J. De Boor, and U. Gsele, Metal-Assisted Chemical Etching of Silicon: A Review, Adv. Mater., 23, pp. 285-308, 2011
- [14] C. A. Bishop, Process Diagnostics and Coating Characteristics, Vacuum Deposition onto Webs, Films and Foils, 2015

Extending the Life of Legacy Robots: MDS-Ach via x-Ach

Daniel M. Lofaro^{*1}, Magdalena Bugajska², Donald Sofge³

¹Navy Center for Applied Research in Artificial Intelligence, U.S. Naval Research Laboratory, Washington D.C.

²Intelligent Systems Section, U.S. Naval Research Laboratory, Washington D.C.

³Distributed Autonomous Systems Group, U.S. Naval Research Laboratory, Washington D.C.

ARTICLE INFO

Article history:

Received: 25 August, 2018

Accepted: 22 December, 2018

Online: 21 January, 2019

Keywords:

Legacy Robots

Middleware

Control Framework

Real-Time

ABSTRACT

Our work demonstrates how to use contemporary software tools on older or “legacy” robots while keeping compatibility with the original control, tools, and calibration procedures. This is done by implementing a lightweight middle-ware called MDS-Ach connected directly to the hardware communications layer of the robot’s control system. The MDS-Ach middle-ware, which relies on the x-Ach methodology, was specifically designed for Xitome Mobile Dexterous Social (MDS) Robot which was released in 2008. The MDS Robot is actively used in multiple research facilities including the United States Naval Research Laboratory. This middle-ware gives the MDS Robot the bleeding edge software capabilities of today’s robot by implementing the x-Ach real-time processes based computer control architecture. MDS-Ach controls the robot over its low level hardware communications interface (CAN-Bus). This communication controlled and implemented by a real-time daemon process. Controllers communicate with the real-time daemon via a ring buffer shared memory with network capabilities. The ring buffer shared memory is a “first-in-last-out” and is non-head-of-line blocking. All of the latter ensures non-blocking reading and writing of the latest data even while newer data is being added to the buffer. The UDP and TCP protocols can be implemented depending on reliability and timing requirements. Secure communication between networked controllers is implemented via tunneling over SSH if needed. The MDS-Ach middle-ware is designed to allow for simple and easy development with modern robotic tools while adding accessibility and usability to our non-hardware-focused partners. We present an implementation of real-time collision avoidance and a robust inverse kinematics solutions within the MDS-Ach system. We include detailed examples of collision avoidance, inverse kinematics implementation, and the software architecture and tools.

1 Introduction

The goal of this work is to extend the life of legacy robots that have high quality electro-mechanical hardware by allowing them to make use of modern day robotics software frameworks such as the Robot Operating System (ROS) [1], while keeping compatibility with their existing calibration and monitoring tools. The significance and the novelty of this work is in its focus on the compatibility with *legacy robotic systems* via the use of a lightweight middleware, a non-head-of-line blocking buffer design, and real-time network daemon. This document details our efforts in creat-

ing the latter for the MDS Robot. The MDS Robot by Xitome is a high degree of freedom (DOF) Mobile Dexterous Social (MDS) robot [2]. This research-grade robot originally debuted in 2008. The physical robot hardware was custom built to support human-robot interaction research. This includes the two 6-DOF arms each with a 7-DOF hand. Each hand has four under-actuated fingers. The robot also includes a 4-DOF neck and a 17-DOF face. All of the latter items enable different facial expressions, physical gestures, and grasping within a single robotic platform.

Given the goals of this work, our system must meet

^{*}Corresponding Author: Daniel M. Lofaro, U.S. Naval Research Laboratory, Washington D.C., USA - daniel.lofaro@nrl.navy.mil

Table 1: Summary of existing software robotic frameworks [4]

Framework	Concurrency Model	Data Sharing	Focus
ROS [1]	Processes	Proprietary TCP	Mobile Robots, Vision, and AI
ROS 2.0 [3]	Processes	Selectable DDS	ROS 1.0 + Real-Time Control
OpenRDK [4]	Threads	Shared memory, Proprietary TCP/UDP	Mobile Robots
Player [5]	Threads	Client/Server over TCP	Mobile Robots
YARP [6]	Processes	shared memory, proprietary TCP	Mobile & Serial Kinematic Chain Robots
MARIE [7]	Processes	Many (3rd Party)	Connecting Different Frameworks
OpenRTM-aist [8]	Threads	CORBA	General Robotics
Orca [9]	Processes	ICE[10]	Mobile Robots
JAUS [11]	Processes	Proprietary TCP/UDP	Standardization, Unmanned Systems
x-Ach [12]	Processes	Shared Memory, Proprietary TCP/UDP	Real-time Robotics

specific requirements. Firstly, the well-tuned and refined calibration procedures that are extensively documented for the robot must remain unchanged while extending the robot's capabilities. Similarly, existing monitoring tools must remain functional, but often cannot be modified. Furthermore, with robotics and Artificial Intelligence (AI) research back in the spotlight and becoming more mainstream in recent years, a robot system must implement additional safety controls to allow researchers with no mechanical or electrical engineering background to use the robot while keeping the risk of damage to the physical system low. The system must also be capable of working securely over a network. Finally, the system must not be confined to a single programming language, allowing the user to use "the right tool for the right job."

To keep compatibility with the legacy software, direct control is implemented by commanding the MDS Robot over the Controller Area Network (CAN) bus via a dedicated real-time daemon. A process based controller approach is used for the control system using a "first-in-last-out" (FILO) non-head-of-line blocking ring buffer type of shared memory. This allows for the use of multiple programming languages in the same system. This also gives each controller the ability to *read the newest data first* which is typically of most importance to real-world robot controllers. SSH tunneling is used when secure connections between network connected controllers is required. Real-time collision avoidance with a robust inverse kinematics solution is implemented within the MDS-Ach system allowing for a multitude of types of users to safely control the robot. In the following sections, we describe our methodology and provide usage examples of the system. We also include, in the appendix, a description of the tools and the implemented API. Please note that this paper is heavily based on and is an addition to the authors previous work in Lofaro et. al. [13]. This work adds detailed instructions and examples of how to utilize the MDS-Ach system.

2 Background

There are currently many implementations of middleware for robot systems. The most notable one is the Robot Operating System which is commonly known

as ROS[1]. ROS is a middle-ware which is TCP-based. It allows for communication between different controllers other over "topics." This is typically done to connect the hardware systems of the robot such as the sensors and actuators to the logic and control. Its biggest strength is large ecosystem. Additionally, ROS is most useful for systems that require many controllers but do not require real-time capabilities. Currently ROS 2.0 [3] is being developed. ROS 2.0 will add real-time capability to the system. Because of the latter MDS-Ach is written specifically to be compatible with both ROS 1.0 and 2.0.

The YARP system, or Yet Another Robot Platform, is a C/C++ based middle-ware. The purpose of YARP is to connect control processes, sensors, and actuators. YARP is tested on Windows, Linux, and OSX [6]. It uses shared memory, over TCP when needed, for communication between the YARP server and clients. Non-blocking and latest data first reading is implemented via double and triple buffers. YARP is currently limited to a C/C++ API.

OpenRDK is an open source framework for robotics [4]. It uses socket communication and shared memory to implement its thread based architecture. This impressive control system utilized linking techniques and blackboard-based communication to allow for input/output data port conceptual system design. OpenRDK is a thread based design. Our desired system is processed based not thread based.

Joint Architecture for Unmanned Systems, also known as JAUS [11], was originally an initiative started in 1998 by the United States Department of Defense (DoD) to develop an open architecture for the domain of unmanned systems. JAUS was formerly known as Joint Architecture for Unmanned Ground Systems (JAUGS) and is built on five principles: vehicle platform independence, mission isolation, computer hardware independence, technology independence, and operator use independence. Still in use by the DoD, JAUS communicates with other systems over TCP and/or UDP. Though a formidable system, the public ecosystem is relatively small when compared to competitors. A comparison of the above-mentioned middleware, but also other robotic frameworks can be seen in Table 1.

The *x-Ach* system is based on the Ach IPC, or inter process communication [12]. Current implementation

of *x-Ach* include Hubo-Ach [12] for the Hubo series of robots, Android-Ach for phones, Shoko-Ach [14] for the underwater legged robot AquaShoko, MDS-Ach (this work) for the MDS Robot, and more. The *x-Ach* system is lightweight with non-head-of-line blocking like OpenRDK, however it uses processes for each controller instead of threads. Like YARP, we use the idea of newest data first. Like OpenRDK, we use shared memory and offer a choice between TCP and UDP depending on need. *x-Ach* is compatible with multiple languages including C/C++, Python, Java, etc. ROS currently has the largest robot controller ecosystem, however it is not real-time; ROS 2.0 will be real-time when completed. To leverage ROS' ecosystem, *x-Ach* is written to be compatible and easily integrated with either ROS versions.

3 Methodology

This section details the methodology and implementation of the MDS-Ach system and is based on our original paper [13].

3.1 Controller Area Network Communication

A key goal of this work is to keep compatibility with older robot's legacy software and well-defined calibration procedures. In this case the specific robot in question is the MDS Robot. We created the MDS-Ach system to connect directly to the robot via the CAN (Controller Area Network) bus. This is the communications bus that is concurrently used to control each actuator on the MDS Robot and is also used with the legacy system, in this case the MDS Motion Server. The CAN bus is specifically designed for multiple devices/controllers to communicate over it. The latter allows all of the legacy software, utilities, and tools to run, monitor, and calibrate the robot while allowing for integration with the state of the art robotic software. The direct communicating via the CAN bus results in keeping compatibility with the original software and tools without having to modify, recompiled, or in any way adjust these tools in order for the MDS-Ach system to run.

3.2 x-Ach

The *x-Ach* system has a process based architecture. This means that it runs individual controllers as independent, synchronous and/or asynchronous, processes. Each process communicates with each other over the IPC called Ach which is a circular buffer and is non-head-of-line-blocking [12].

Ach was chosen because it is low-latency (key for real-time control) and has a *first-in-last-out* (FILO) buffer. This allows controllers to get the newest information first while retaining the ability of reading older information at a later time if needed. The latter is

very important to real-time robotic systems. Packed c-structs are used for messages types in order to keep the system architecture agnostic. This means that *x-Ach* controllers running on different platforms (i.e. x86, amd64, ARM, and other systems) can communicate with each other despite the different memory block sizes. It is important to note that all of the memory types are well defined within the packed c-struct. For example a (int32_t) is used instead of (int) to ensure data congruence between different architecture.

Controllers communicate with each other over Ach channels. Each controller has a standardized input channel called "reference" (*ref*), a standardized processed reference output channel called "processed reference" (*p-ref*), and a standardized output channel called "state" (*state*). Details on the reference (*ref*) channels can be found in (Section 3.2.1) and details on the state (*state*) channel can be found in (Section 3.2.2).

3.2.1 Reference Channels

Each process based controller has two reference channels. One is a standardized input channel called "reference" (*ref*) and one is a standardized processed reference output channel called "processed reference" (*p-ref*). The latter two channels are where other controllers can write, or publish, to and read from, or subscribe to, respectively. It is important to note that the reference (*ref*) channel does not bind. This means that multiple controllers can write, or publish, simultaneously to the same reference channel. The controller will only use the most recent message received and only utilize the other older messages if it is specifically needed.

The process can synchronize its control loop to the incoming reference (*ref*) input or be asynchronous (i.e. free running). The latter is useful for use and development of synchronous and asynchronous controller systems and controllers. Both of the reference channels, the input reference (*ref*) and the process output reference (*p-ref*), are identical in structure.

As stated in Section 3.2, the reference channels consist of packed c-structs. The size of the structure is dependent on the number of degrees of freedom (DOF) of the robot of the given robot. The reference structure contains joint-space references (*JS-ref*) and work-space (Cartesian space) references (*WS-ref*) (see Figure 5). This is in the same structure to allow for less complexity in the controllers' number and types of sources and sinks.

3.2.2 State Channel

Each MDS-Ach controller has one channel for the *state* what is real-only by other processes. Other controllers can read the most up to date state of the robot/controller by reading the *state* channel even if the given robot/controller is currently updating the state. For the implementation on the MDS a daemon is created called the *MDS-Ach daemon* (see Section 3.3). This daemon publishes the most recent joint-space states including, but not limited to, actual position of the

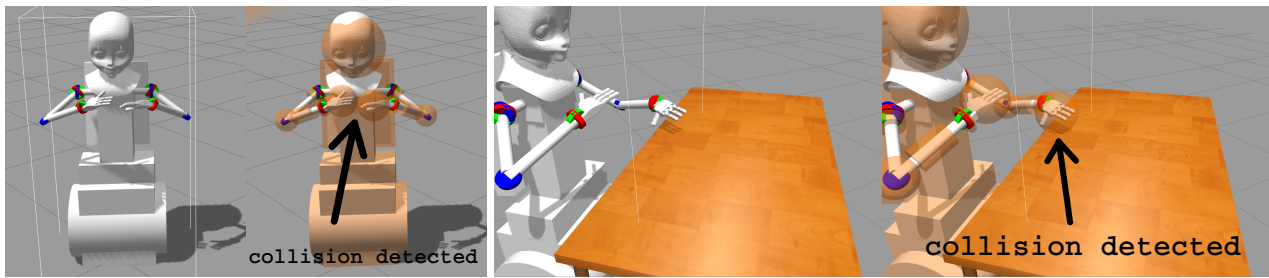


Figure 2: High-resolution MDS Robot model (**LEFT**) next to the high-resolution MDS Robot model with the low-resolution collision models overlapped (**CENTER LEFT**). The collision models are simplifications of the high resolution geometry and are denoted in orange. The collision model is “over-sized” in order to help the system detect a collision before an impact actually occurs. *Self Collision Avoidance Test*: Robot does not collide when each hand is told to go to the same location. (**RIGHT**) *World Collision Avoidance Test*: Robot does not collide when the hand is commanded to a position that will collide with an object in its workspace. In this example the hand stops before it hits the table.

joint, last received reference/commanded position, joint load/current, etc. Like the reference channels the state channel is a packed c-struct. The size of the structure is dependent on the number of DOFs of the robot. An example use case for the state channel is the collision detection daemon in Section 3.4.2. This daemon reads the *state* channel and applies those values to its internal model to check for collisions in real-time. The daemon then outputs collision state of the robot on its state channel. This is done without disrupting the real-time performance of the *MDS-Ach Daemon*.

3.3 Daemon

The MDS-Ach daemon is the bridge between the CAN bus, which commands the MDS Robot, and the process based x-Ach controllers. The goal of all x-Ach daemons, including the MDS-Ach daemon, is to be the “*driver*” for the given robot. The CAN bus is half-duplex running at a rate of 1.0 *Mbps*. Joint-space control and sensor feedback (i.e. reference and state information) is sent over the CAN bus to and from the daemon. The MDS-Daemon is specifically calibrated to keep the load on the CAN bus at approximately 60% of its bandwidth saturation. The latter is done to help guarantee real-time performance and on-time data delivery. The x-Ach daemons have been implemented and tested on multiple different types of robots with different CAN packet structures and a varying amount of communication buses. 200 *hz* real-time performance was achieved when utilizing multiple CAN buses (two) along with a specialized packet structure and the utilization of the PREEMPT_RT linux kernel [15] as seen in our previous work [12]. The MDS requires the use of only one CAN bus with a BAUD rate of 1*Mbps* and state-full packet structure. All of the latter limitations require us to run at 10 *hz* to guarantee real-time performance, with sub *ms* accuracy, for this specific robot.

The MDS-Ach daemon has an optional first-order real-time joint-space position smoothing filter. This filter is applied to the input from the reference (*ref*) channel (Section 3.2.1) before being applied to the control and sent to the robot over the CAN bus. The first-

order real-time joint-space position smoothing filter converges to within 95% of the reference input within 4.0 seconds and is enabled by default. This filter was added in order to reduce acceleration and jerk of each joint without limiting the maximum velocity.

The MDS-Ach daemon reads the reference command (as described in Section 3.2.1) in real-time and sends the command data over the CAN bus. This process is asynchronous in reference to the robot’s actuators control loop. If multiple commands are sent within one cycle only the newest one is read and processed. There is a zero order hold if there are no new commands between given control cycles. The exception to this is during setup phases such as “*homing*”, resetting/error correcting, and other hardware setup/configuration specific commands. Additionally, during each cycle the MDS-Ach daemon requests and reads the information from the sensors via the CAN bus and writes it to the *state* channel (see Section 3.2.2).

3.4 Input Pipeline

The Main Controller receives the user-level commands in either joint-space or Cartesian-space, over the ACH shared memory. The latter input goes through the pipeline described in section below before it is sent to the Daemon (Section 3.3).

3.4.1 Joint Mux

The Joint Mux is an event-based process that takes in joint-space references from multiple processes. It updates the primary joint-space reference, which is sent to the MDS-Ach daemon, only with the joint-space references that are controlled by and updated by a given controller while preserving the ones that they controller does not have write permission to and/or were not modified. The MDS-Ach daemon is then sent the resulting consolidated (muxed) reference command. The purpose of this is to allow multiple controllers to update individual joints without conflicting with (overwriting) other joint commands while preserving a common message type.

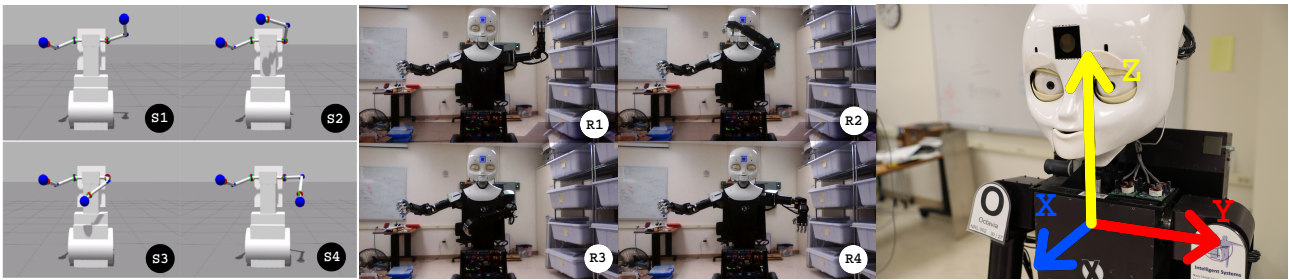


Figure 3: Example of the MDS-Robot (MIDDLE) and its simulation (LEFT) utilizing the ik solver, joint smoother, and self-collision avoidance to move its hand in a square pattern while keeping all rotational degrees of freedom in the null space. (RIGHT) Coordinate system for the MDS Robot when using MDS-Ach.

3.4.2 Real-Time Collision Detection

The fully integrated real-time collision detection system utilizes the opensource simulator GazeboSim and ODE [16, 17] to ensure safe system operation. To guarantee the real-time performance, the system must detect self-collisions as well as collisions with objects in the environment within one time step of the MDS-Ach daemon (as defined in 3.3). To maintain compliance with the real-time deadline unneeded features of the simulator, such as *physics*, are disabled. This results in a reduced computational load. Additionally a geometrically simplified/reduced model of the MDS robot is created. This model is comprised of basic shapes such as cylinders, boxes, and spheres. This simplified model is used to create the collision model to further reduce computational complexity and improve performance of the system. The performance improves because detecting collisions between two spheres only requires comparing the euclidean distance between the spheres centers and to the sum of the spheres' radii. There is no collision as long as the sum of the radii is less than the euclidean distance from the centers. The same can be said for cylinders if the conditional measurement is made from the center of the sphere to the closest point on the axial line of the cylinder. Figure 2 shows the MDS Robot model with the collision model overlapped.

The collision state is written in state the message format, as described in Section 3.2.2, to the *Main Controller*. If the reference position is free of collisions, a joint-space reference is sent to the the MDS-Ach daemon via the Joint Mux. The MDS-Ach daemon will execute the motion on the physical robot. If a collision is detected, the most recent safe joint-space reference values will be used for all the joints of the affected arm(s). "Expert" user is able to overwrite this behavior, if needed, by sending a joint-space command from the main controller.

3.4.3 Joint-Space Smoothing

When the robot is operating in situations where it is moving between joint-space configurations or required to stop its motion to avoid collision/self-collision (see Section 3.4.2), the safety of the robot's joints needs to be ensured. Torque due to high joint-space acceleration

is one of the primary causes of robot joint damage. We need to reduce the torque applied to the joints without causing joint-space overshoot. Furthermore this needs to be done in real-time and on-line. The motion at this level can not be pre-planned so it can be used in real-time tasks such as servoing or world interaction. This section shows how we reduce the acceleration, which reduces torque, on the joints while reducing the overshoot. The latter is done by applying the filter shown in (1).

$$\theta_n = \frac{(\theta_{n-1}L - 1) + \theta_{des}}{L} \quad (1)$$

Where θ_n is the output of the filter which is the new reference position (angle) the joint will be commanded to at step n ; θ_{n-1} is the reference position to the joint from the previous time step, i.e. $n - 1$; L is the weight of the filter (defined by its integer length); and θ_{des} is the desired reference position in joint-space that the joint is requested to go at time step n .

The position as recorded from the joints' encoders (θ_{enc}) are used to add joint-space compliance to the system. This is done by replacing θ_{n-1} in (1) with θ_{enc} as seen in (2). The use of the measured angle allows us to take advantage of the natural compliance in the system and magnify it. When the filter is applied it results in a pose "sag" due to gravity.

$$\theta_n = \frac{(\theta_{enc}L - 1) + \theta_{des}}{L} \quad (2)$$

3.4.4 Inverse Kinematics

We utilize the "Inverse Jacobian" method for our on-line inverse kinematics solver. This is used to the joint-space angles for the commanded Cartesian-space (work-space) positions [18, 19]. The joint limits are utilized when constructing the the Jacobian during each cycle. During each step of the search the resulting pose is checked for self-collisions via ODE (see Section 3.4.2). A new *intermediate* goal position is created if a collision is found. The process is then repeated. The joint-space configurations found from the Jacobian IK is then passed through the filter described in (1) to prevent large steps in joint-space. Each step of the resulting filtered positions are then checked in real-time by the collision checker (Section 3.4.2) before being sent to the MDS-Ach daemon.

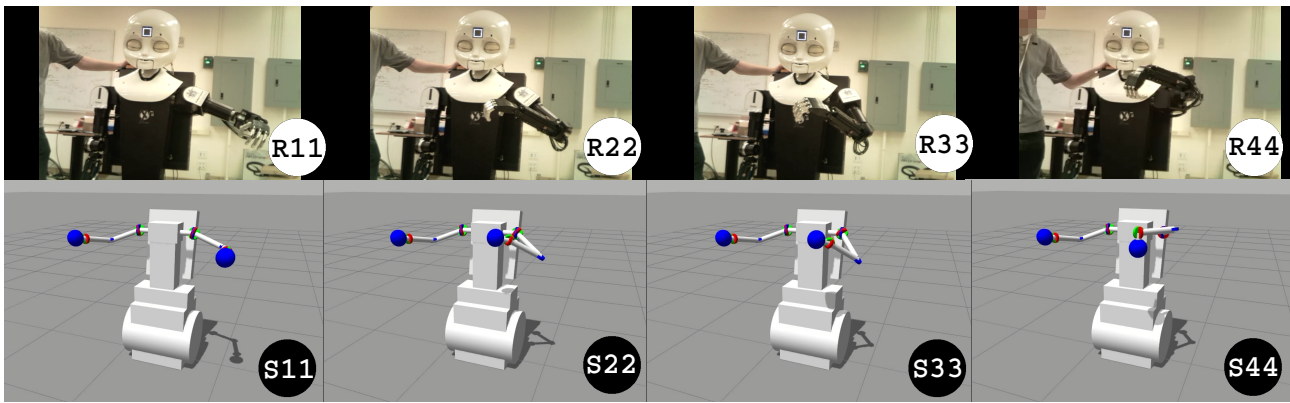


Figure 4: Example of the MDS-Robot (TOP) and its simulation (BOTTOM) utilizing the IK solver, joint smoother, and self collision avoidance to move its hand in a box with the coordinates $(x, y, z, \theta_y) = (0.35m, 0.45m, 0.25m, null) \rightarrow (0.35m, 0.10m, -0.05m, null) \rightarrow (0.45m, 0.10m, -0.05m, null) \rightarrow (0.45m, null, null, 90^\circ)$ while keeping some linear and some rotational degrees of freedom in null space.

The Inverse Jacobian IK method is used because it allows control the required work-space degrees of freedom while the other degrees of freedom remain in *null space*. The MDS Robot is used for world interaction tasks such as opening doors (via pushing), grabbing cups, pushing buttons etc. thus having work-space degrees of freedom in *null space* is needed. In the door opening example (via pushing) the robot only degree of freedom that strictly matters is the x value (out of the robot’s chest - see Figure 3). The x value defines the distance the robot pushes the door open. The height (z) and the left/right distance (y) does not matter as much and thus can be left in null space. Furthermore the orientation (all three degrees of freedom) of the hand also belongs in the *null space*.

Solving for the required degrees of freedom for the given task, and leaving all the others in the *null space*, allows less iterations/computations for the solver resulting in a faster solving time and the ability to run in real-time. For MDS Robot, real-time constraints are satisfied if the joint-space values for a given Cartesian-space (joint-space) position are calculated in less than 0.5 sec. This is achieved when the desired Cartesian-space position can be described with $DOF \leq 4$. The higher order of the required position requirements (i.e. the less degrees of freedom in the *null space*) the more time (on average) is required to solve the joint-space solution. A measured ~ 0.7 sec is required to solve for a 5-DOF position and ~ 1.3 sec for a 6-DOF on our contemporary computers using a single core. It is important to note that the latter calculated times are averages, also a solution is not guaranteed to be found even if one exists.

Figure 3 shows the MDS-Robot and its simulation utilizing the inverse kinematics daemon. The figure shows the MDS-Robot commanding its end-effector to move in a box pattern in real-time. A detailed description of the test can be found in Section 4.1.

MDS-Ach System Diagram

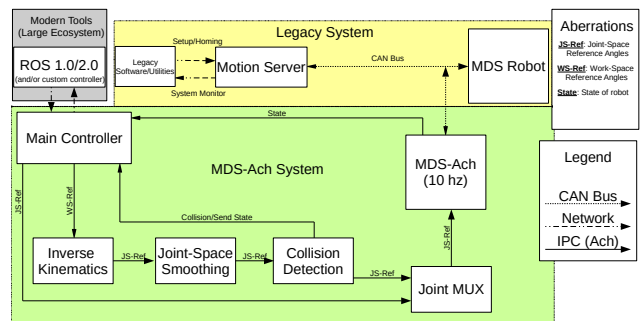


Figure 5: MDS-Ach middleware: A real-time process based control system used to extend the research life of the MDS Robot.

3.4.5 Network Daemon

The network daemon utilizes *achd*, as a part of the Ach library, to share data over a “socket” supported network. The *achd* daemon pushes the state data from the primary control computer on the robot to the external “user-level” computer controller when new state data is available. When a reference command is updated by the user-level computer controller it is pushed to the primary control computer over the network by the *achd* daemon. Updates are only pushed or pulled when there is new state or reference data in order to conserve bandwidth. UDP is used by default for data transport to help facilitate tighter real-time performance. TCP or TCP over a SSH tunnel can be used by the user depending on timing, reliability, and security requirements. More details can be found in our previous work [12].

3.4.6 Security

Intercepting and spoofing network traffic is a major threat to any robot system and the threat persists even on a properly configured Unix/Linux computer. These “man-in-the-middle” attacks allow a third party to control the robot using the existing controllers if they can

spoof the feedback data [20]. To prevent a man-in-the-middle attack over the network, a secure SSH tunnel with pre-shared keys is used to transmit and receive state and reference data between the main controller and the user-level controller. Since SSH tunnel operates over TCP, the network throughput between the user-level controllers and the Main Controller is reduced, and may even affect the real-time performance.

3.5 External Framework Bridge

The ROS 1.0 and ROS 2.0 bridge allows us to make use of the extensive ROS ecosystem and serves as proof of concept for extensibility of this middleware. This bridge talks directly with the Main Controller via the Ach shared memory or the network daemon (3.4.5). This controller converts the state data published by the MDS-Ach system into ROS messages to be published again on specified ROS topics. To reduce latency (1) the state topic of ROS is synchronous with the state channel of MDS-Ach, and (2) the reference channel is synchronous with the state topic of ROS (seen Fig. 5).

4 Testing

At the conclusion of the MDS-Ach middleware development, we tested the software to confirm its functionality. We present the testing procedures and results of the inverse kinematics system (Section 4.1), the *live* self-collision detection system (Section 4.2), and the *live* external-collision detection system (Section 4.3).

4.1 Inverse Kinematics Test

To validate the usability of the inverse kinematics system, we performed couple tests to show its ability to move the end-effector to a position (4.1.1) and an orientation (4.1.2) in the Cartesian-space (work-space) while keeping other variables in the null-space. The coordinate system used for the IK tests is shown in Figure 3; the origin is the orthogonal projection of the base of the neck onto the line defined by the rotations points of the shoulders. x - y plane is horizontal. x - z plane extends in front and behind the robot while y - z plane extends to the sides. All signs of rotation follow the “right hand rule.”

4.1.1 Position Control IK Test

This section presents results showing that the end-effector can be moved between multiple linear Cartesian-space (work-space) locations (x, y, z) without constraining the end-effector orientation. The left arm is commanded and moved to the locations shown in (3) where S1, S2, S3, and S4 are the coordinates for the simulated robot and R1, R2, R3 and R4 are the coordinates for the real robot.

$$\begin{aligned} S1 = R1 &= (0.40m, 0.45m, 0.25m) \\ S2 = R2 &= (0.40m, -0.05m, 0.25m) \\ S3 = R3 &= (0.40m, -0.05m, -0.25m) \\ S4 = R4 &= (0.40m, 0.45m, -0.25m) \end{aligned} \quad (3)$$

The screenshots of the resulting motion can be found in Fig. 3.

4.1.2 Orientation Control IK Test

This section presents results showing that the end-effector can be moved between multiple Cartesian-space poses (x, y, z, θ_y) while keeping the remaining orientations in the null space. The left arm is commanded and moved to the locations shown in (3) where S11, S22, S33, and S44 are the coordinates for the simulated robot and R11, R22, R33 and R44 are the coordinates for the real robot. In the case of the last motion (R44 and S44), θ_y is set to 90° and a desired x to a value of $0.45m$ while all other degrees of freedom remain in the null space. The full set of coordinates for this test can be found in (4).

$$\begin{aligned} S11 = R11 &= (0.35m, 0.45m, 0.25m, null) \\ S22 = R22 &= (0.35m, 0.10m, -0.05m, null) \\ S33 = R33 &= (0.45m, 0.10m, -0.05m, null) \\ S44 = R44 &= (0.45m, null, null, 90^\circ) \end{aligned} \quad (4)$$

This shows that the IK process can solve for some degrees of freedom while keeping the others in null space (see Figure 4). This improves system performance when higher fidelity solutions are not required for system operation.

4.2 Self-Collision Detection and Avoidance Test

To test the self-collision system as described in Section 3.4.2, we drove the hands to the same location using the IK system described in Section 3.4.4. Multiple instances of this test were performed. In all the runs, the hands stopped before colliding. Figure 2 shows one example of the multiple self-collision tests. In this instance of the test, both the left and the right hands were told to go to the (x, y, z) coordinates $(0.3m, 0.2m, 0.0m)$. As expected, the hands did stop when the two parts of the collision model touched as shown in Fig. 2.

4.3 World-Collision Detection and Avoidance Test

To test the world-collision avoidance system described in 3.4.2, we drove the hand from the position stated in Section 4.2 out to an x value of $0.5m$ towards the table. We placed a model of a table in the simulated world where the real table would be. From there we drove the hand down in z . Similar to the results in 4.2 the hand would not move farther than the collision point between the collision model of the hand and the collision model of the table (see Figure 2). This was done for multiple objects and multiple collision locations.

5 Usage

This section documents how to start, stop, and use different MDS-Ach utilities. The MDS-Ach controls the MDS Robot via the CAN bus. It provides smoothing/filtering, multi-process control architecture, multi-language support, inverse kinematics for right and left arm, and more.

5.1 Daemon Control: mds-ach

The MDS-Ach daemon is the process running the background that controls communications between the physical (or simulated) robot and the controllers. The daemon controls the CAN bus when connected to the physical robot. Section 5.1 describes the different input options and modes of the MDS-Ach daemon. A full description of the options for the console input for the MDS-Ach daemon can be found in Appendix A.

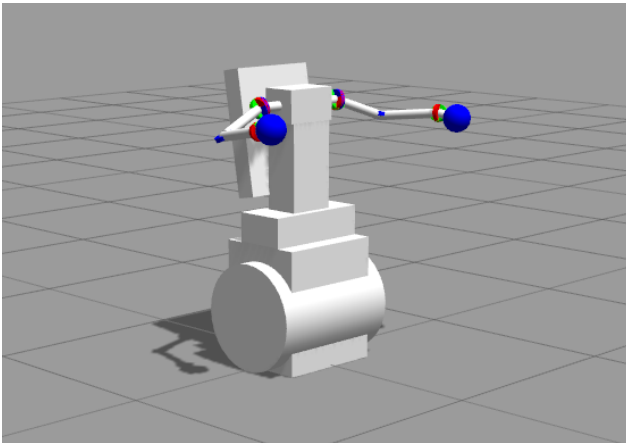


Figure 6: MDS Robot simulated in Gazebo.

5.2 Startup Procedure

This section documents the startup procedure for the MDS-Ach system.

5.2.1 Initialize

We preserved the original MDS setup procedure provided by MDS manufacturer, Xitome. Once setup is complete, which include a fairly involved homing procedure for all the joints, the user is able to keep the Xitome subsystem running. User may **not** issue commands to any of the joints using the Xitome system once MDS-Ach is started.

5.2.2 Start Daemon

It is important that the MDS robot is in its homed configuration when starting the MDS-Ach system. This is because MDS-Ach system assumes the robot's starting position to initialize the IK and the self-collision systems. If the robot is not in its "homed" position, than some joints will get a step input to go to the starting position. All starting positions are read from the

anatomy.xml configuration file located in `/etc/mds-ach`. This file is identical to that used by the Xitome system for configuration.

The MDS-Ach system can run on the robot (Option 1) and on a simulated robot (Option 2). Option 1 sends all of the commands over the CAN bus while Option 2 sends all commands to GazeboSim[16]. Both Option 1 and Option 2 have an x-Ach[13] abstraction layer between their respective communication buses and the controller. A full description of the implementation procedure can be found in Appendix B.

5.3 Examples

Examples of how to do basic operations using the MDS-Ach system on both the real robot and the simulator can be found in Appendix C. The examples show how to start the MDS-Ach daemon on the physical and simulated robot. An image of the simulated robot can be seen in Figure 6. All example code/software can be found in Lofaro et. al. [21].

6 Utilities

This section describes the utilities available for use with the MDS-Ach system. All utilities work seamlessly with the physical and simulated robot.

6.1 MDS-Ach Console

The MDS-Console utility allows the user to read and set individual joint angle values via the command-line. It also allows the user to read the Cartesian-space pose (6 DOF) of the end-effector and set the desired target pose (3, 4, 5, and 6 DOF). The console provides control interface for all the joints listed in Table 2. We included examples of single joint interactions, but also end-effector control, in Appendix D.

Table 2: Joint abbreviations (short and long) with definitions

Definition	< joint > (short option)	< joint > (long option)
Right Elbow Pitch	REP	RightElbowFlex
Right Shoulder Yaw	RSY	RightUpperArmRoll
Right Shoulder Roll	RSR	RightShoulderAbd
Right Shoulder Pitch	RSP	RightShoulderExt
Left Elbow Pitch	LEP	LeftElbowFlex
Left Shoulder Yaw	LSY	LeftUpperArmRoll
Left Shoulder Roll	LSR	LeftShoulderAbd
Left Shoulder Pitch	LSP	LeftShoulderExt
Right Wrist Roll	RWR	RightWristFlex
Right Wrist Yaw	RWY	RightWristRoll
Left Wrist Roll	LWR	LeftWristFlex
Left Wrist Yaw	LWY	LeftWristRoll
Torso Yaw	WST	TorsoPan
Neck Roll	NKR	HeadRoll
Neck Pitch (lower)	NKP1	HeadPitch
Neck Pitch (upper)	NKP2	NeckPitch
Neck Yaw	NKY	HeadPan

6.2 MDS-Ach Read

The MDS-Read utility allows the user to view the joint space references, state, and address of each active joint. This utility updates the information at $\sim 20\text{Hz}$. It may be started or stopped at any time without affecting the overall system.

6.2.1 Prerequisites

Since MDS-Read is a monitoring tool, it must be started after the MDS-Ach is already running.

6.2.2 Startup

To start the MDS-Read utility, run the command below. The expected terminal can be seen in Figure 7.

Bash/MDS-Ach Console:

```
$ mds-ach read
```

Joint	Enabled	address	Commanded Pos	Actual Pos
REP	1	0x0048	-1.25000	0.00000
RSY	1	0x0049	0.15000	0.00000
RSR	1	0x004b	-1.18000	0.00000
LEP	1	0x004c	-2.21719	0.00000
LSY	1	0x004d	-0.44890	0.00000
LSR	1	0x004f	0.98192	0.00000
RWR	1	0x0050	-0.00000	0.00000
RWY	1	0x0052	-0.00000	0.00000
LWR	1	0x0054	-0.31819	0.00000
LWY	1	0x0056	0.05389	0.00000
RSP	1	0x005b	-0.44500	0.00000
WST	1	0x005c	-0.00000	0.00000
LSP	1	0x005f	-0.52020	0.00000

Figure 7: Expected window for (`$ mds-ach read`)

- **Note 1:** If you are running with the simulator the state (Actual pos) will always be zero. When running on the physical robot the real position (in radians) will be shown.
- **Note 2:** If you make more joints “active” in the `/etc/mds-ach/configs/mdsach.xml` configuration file they will automatically show up in MDS-Read.

6.3 Software Interface

The MDS-Ach system currently works with the C/C++ and Python programming languages. Appendix E describes the required libraries for each of the latter languages.

7 Operating in Joint-Space

This section documents how to use MDS-Ach to control the robot in joint-space. To move one, or multiple joint/joints the following steps must be followed:

- **Open Ach Channels** - This is the ring buffer shared memory that this controller communicates with MDS-Ach with.

- **Create Required Data Structures** - Standardized (packed) data structure that contains all of the reference and state data for the robot.
- **Get Joint ID** - Determine the IDs of the joint/joints to be controlled (See Table 2).
- **Queue New Motor Position** - Put the desired motor positions in the reference structure at the index position defined by the Joint IDs.
- **Set Motor Position (put)** - Write the reference structure that has been updated with the desired position for the desired joint to the reference Ach channel.
- **Close Ach Channels on Exit** - Though not required it is good practice to close the unused Ach channels upon exiting the controller.

Figure 8 shows the expected simulator window when following the above steps and setting the Left Elbow Pitch (LEP) and the Right Shoulder Pitch (RSP) to -0.2 rad and 0.1 rad respectively.

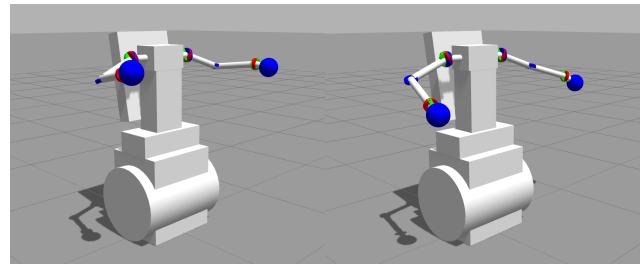


Figure 8: Expected simulator window for (`$. /mds-simple-demo`). (LEFT) Before running. (RIGHT) After running.

It is important to note that Appendix F is an depth explanation and example of how to setup the Ach channels for communication with the MDS-Ach system and how to control the robot in joint-space while using the smoothing filter process. Examples are given in C/C++ and Python.

8 Operating in Cartesian-Space

In this section we discuss end-effector operations in Cartesian-space which utilize forward and inverse kinematics controllers of the MDS robot.

8.1 Built-in IK Controller (Python)

The MDS-Ach system has a built-in inverse kinematics (IK) controller for both the right and left arms. This controller automatically runs when the MDS-Ach system is started. The controller solves for one arm at a time.

The system begins to solve the IK equations automatically, when you post a new desired position on the IK reference channel. Below is a step by step of how you do this using Python.

8.1.1 Python imports

The following are required imports for the MDS-Ach system while using python: `mds_ach`, `ach`, and `mds`. Other imports are for the given controller implementation.

Python:

```
import mds_ach as mds
import ach
import time
import sys
import os
import math
```

8.1.2 Create Open Ach Channel for IK

Ach channels are how you communicate with MDS-Ach. You simply write data to the channel and the robot can read the data in newest to oldest order. The section below shows you how to open an Ach channel. Please note that `ach.Channel()` takes a string as an input. Here we opened one channel. This is a different channel than found in previous sections because it is only for the IK controller.

Python:

```
# Open ACH Channel for IK
k = ach.Channel(mds.MDS.CHAN_IK_NAME)
```

8.1.3 Make IK Structure

Similar to controlling the robot in joint-space you need to set a reference structure to the desired work-space position. For this you need to initialize the structure. See below for the initialization of the work-space structure.

Python:

```
# Make new IK structure
ikc = mds.MDS_IK()
```

8.1.4 Setting the DOF controlled

When using the built-in IK controller you need to set the number of DOF that you are controlling. With this controller you are required to set the DOF in the following order: $p_x, p_y, p_z, \theta_x, \theta_y, \theta_z$ where p_n is the position on axis n and θ_n is the rotation about axis n . For example if you DOF is set to 4 you are controlling p_x, p_y, p_z , and θ_x . If you are controlling 2 you will control p_x and p_y . In the example below we are controlling 3, i.e. p_x, p_y , and p_z . This order can also be seen in Table 5.

Python:

```
# Set the amount of DOF you want to control
dof = 3
```

We set these to the values in Table 3.

Table 3: Inverse Kinematic Values Set for Example 8.1.4

Param #	Definition	Abbreviation	Value (rad)
1	Position in X	p_x	0.3
2	Position in Y	p_y	0.2
3	Position in Z	p_z	0.0
4	Rotation in X	θ_x	Null Space
5	Rotation in Y	θ_y	Null Space
6	Rotation in Z	θ_z	Null Space

θ_x, θ_y , and θ_z are in the Null Space because we do not care where they are as long as the first three parameters are met. We can set these values to what ever we want and they will be ignored. In this case we set them to zero.

Python:

```
# Set values for work-space in
# [x, y, z, rx, ry, rz] order
eff = [0.3, 0.2, 0.0, 0.0, 0.0, 0.0]
```

8.1.5 Choosing Arm for IK

Here we pick the arm for the IK. Our options are set as enums in the `mds_ach.py` and `mds.h` includes. All options for right and left arms can be found in Table 4.

Table 4: Definitions for left and right arms using `mds_ach.py` in Python

Arm	Python Definition
Left	<code>mds.LEFT</code>
Right	<code>mds.RIGHT</code>

Here we set the arm to the left arm.

Python:

```
# set arm
armi = mds.LEFT
```

8.1.6 Set IK Structure

Just as in the joint-space method, we need to the values in our structure before we send it to the robot. Here we set all of the parameters from above to the structure `ikc` that we created.

Python:

```
# Put setting into ik structure
ikc.move = armi
ikc.arm[armi].ik_method = dof
ikc.arm[armi].t_x = eff[0]
ikc.arm[armi].t_y = eff[1]
ikc.arm[armi].t_z = eff[2]
ikc.arm[armi].r_x = eff[3]
ikc.arm[armi].r_y = eff[4]
ikc.arm[armi].r_z = eff[5]
```

8.1.7 Command the Robot

Just as with the joint-space controller you need to “put” the data on the ACH channel. Unlike the joint-space controller it will not move as soon as you send it. The controller will first have to solve the IK. Upon finding the solution the controller will send it to the robot. If a solution is found it will take anywhere between 0.1 and 5.0 seconds. If there is no solution found the robot will not move. Note: the robot limits its self to 1000 search iterations for an IK solution.

Python:

```
# put on to ACH channel
k.put(ikc)
```

8.1.8 Running the Code

To run the code do the following within the **mds-ach/examples** folder.

Bash:

```
$ python ms_ik_one.py
```

The expected terminal out can be seen in Figure 9.

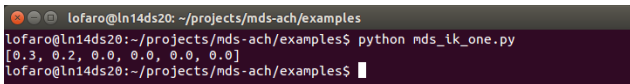


Figure 9: Expected window for (\$ python ms_ik_one.py)

To check that the IK worked you can run the FK in the MDS-Ach console and/or run the simulator. The before and after of the MDS-Ach console is found in Figure 10 and Figure 11, respectively.

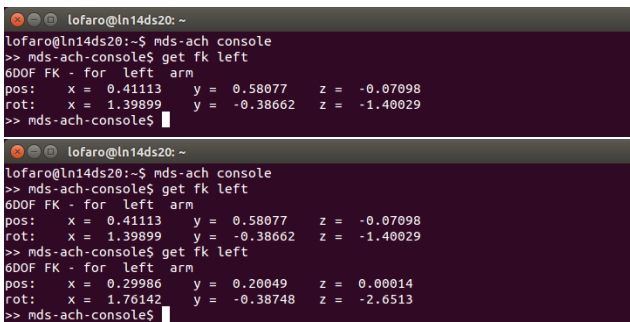


Figure 10: Expected MDS-Ach Console window for (>> mds-ach-console\$ get fk left). (TOP) Before running. (BOTTOM) After running.

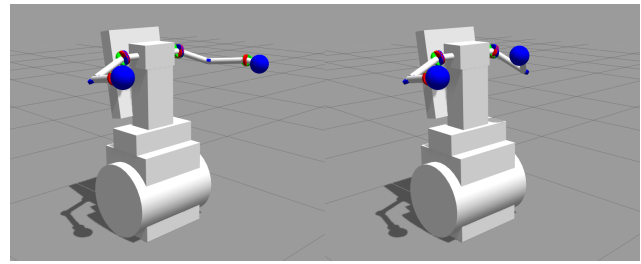


Figure 11: Expected simulator window for (\$ python mds_ik_one.py). (LEFT) Before running. (RIGHT) After running.

8.1.9 Full Code

The full example can be found below as well in the file:

mds-ach/examples/mds_ik_one.py

8.2 Built-in IK Controller (C/C++)

The MDS-Ach system has a built-in inverse kinematics (IK) controller for both the right and left arms. This controller automatically runs when the MDS-Ach system is started. The controller solves for one arm at a time.

The system works by starting to solve the IK equations when you post a new desired position on the IK reference channel. Appendix G has a step by step of how you do this using C/C++ which is identical in methodology to the Python implimentaiton in Section 8.1.

8.3 Making a Box using Inverse Kinematics (Python)

This section shows you how to control the MDS robot via the built in IK module. This example shows us using solving a 3 DOF IK. The method used can be expanded to any DOF between 1 and 6.

The example given is the robot moving its left hand in a 0.5 m box 0.4 m away from the origin. The hand will not move in the x plane, only in the y and z.

8.3.1 Making the Box

The following points will be hit in order: (0.4,0.45,0.25) → (0.4,-0.05,0.25) → (0.4,-0.05,-0.25) → (0.4,0.45,-0.25) with all units in meters. Example Python code for state flow. Note state 0 (0.3,0.2,0.0) is the initial state and will not be returned to.

Python:

```

if ii == 0:
    c = [0.3, 0.2, 0.0]
    ii = 1
elif ii == 1:
    c = [0.4, 0.45, 0.25]
    ii = ii+1
elif ii == 2:
    c = [0.4, -0.05, 0.25]
    ii = ii+1
elif ii == 3:
    c = [0.4, -0.05, -0.25]
    ii = ii+1
elif ii == 4:
    c = [0.4, 0.45, -0.25]
    ii = 1

```

8.3.2 Select Arm

The following selects the arm used for the IK. You may use “left” or “right” for the left and right arms respectively.

Python:

```

arm = 'left'
armi = -1
if arm == 'left':
    armi = mds.LEFT
if arm == 'right':
    armi = mds.RIGHT

```

8.3.3 Parse Desired Position

Parse the desired position for the end-effector to 6 DOF array.

Python:

```

dof = 3
eff = [0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
for i in range(0,dof):
    eff[i] = c[i]

```

8.3.4 Set Desired Position

Sets the desired position for the end-effector. `ikc.arm[i].ik_method` is the number of DOF that you will be controlling. `ikc.arm[i].t_*` and `ikc.arm[i].r_*` is the desired position and rotation of the end effector respectively.

Python:

```

if armi >= 0:
    ikc.move = armi
    ikc.arm[armi].ik_method = dof
    ikc.arm[armi].t_x = eff[0]
    ikc.arm[armi].t_y = eff[1]
    ikc.arm[armi].t_z = eff[2]
    ikc.arm[armi].r_x = eff[3]
    ikc.arm[armi].r_y = eff[4]
    ikc.arm[armi].r_z = eff[5]

```

8.3.5 Send Desired Position to be Solved

The following sends the desired position and ik method to the IK controller to attempt a solution.

Python:

```
k.put(ikc)
```

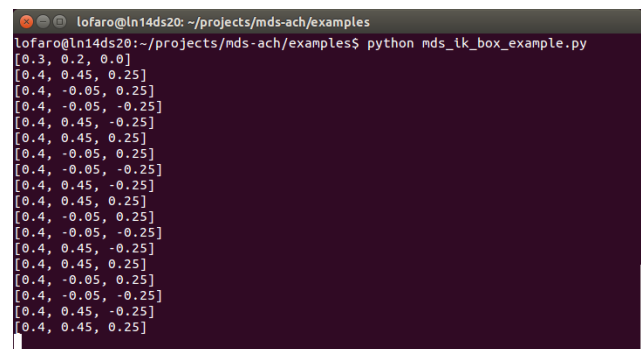
8.3.6 Running the Code

To run the code do the following from within the `mds-ach/examples` directory

Bash:

```
$ python examples/mds_ik_box_example.py
```

The expected terminal output can be seen in Figure 12.



```

lofaro@ln14ds20:~/projects/mds-ach/examples
lofaro@ln14ds20:~/projects/mds-ach/examples$ python mds_ik_box_example.py
[0.3, 0.2, 0.0]
[0.4, 0.45, 0.25]
[0.4, -0.05, 0.25]
[0.4, -0.05, -0.25]
[0.4, 0.45, -0.25]
[0.4, 0.45, 0.25]
[0.4, -0.05, 0.25]
[0.4, -0.05, -0.25]
[0.4, 0.45, -0.25]
[0.4, 0.45, 0.25]
[0.4, -0.05, 0.25]
[0.4, -0.05, -0.25]
[0.4, 0.45, -0.25]
[0.4, 0.45, 0.25]
[0.4, -0.05, 0.25]
[0.4, -0.05, -0.25]
[0.4, 0.45, -0.25]
[0.4, 0.45, 0.25]
[0.4, -0.05, 0.25]
[0.4, -0.05, -0.25]
[0.4, 0.45, -0.25]
[0.4, 0.45, 0.25]

```

Figure 12: Expected window for (`$ python examples/mds_ik_box_example.py`). Time order is left to right, top to bottom.

The expected robot pose can be seen in Figure 13. The latter figure shows the virtual robot not the physical robot.

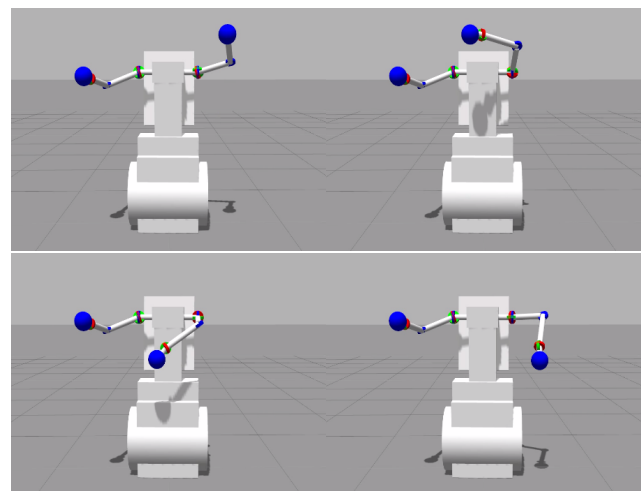


Figure 13: Expected robot poses for (`$ python examples/mds_ik_box_example.py`)

8.3.7 Full Code

The full example can be found in the file:

`mds-ach/examples/mds_ik_box_example.py`

8.4 Inverse Kinematics (IK) API

The Inverse Kinematic (IK) API utilizes the Inverse Jacobian method to solve for the joint-space values given an end-effector position of each of the arms. The IK solver is capable of solving for any and all of the six degrees of freedom of the end-effector i.e. translation (x, y, z) and rotation $(\theta_x, \theta_y, \theta_z)$ while keeping the non-constrained position/rotations in Null space. The number of steps and error range can be specified by the user.

Appendix H gives an example of how to get and set the forward and inverse kinematics via the `mds_ik` API. The goal of this tutorial is to solve for the joint-space values given the desired work-space values as found in Table 3.

9 Conclusion

In conclusion we have made a middleware called MDS-Ach that enables the legacy MDS Robot to be used with modern day robot software, thus extending its life as a research robot. Low-latency non-head-of-line blocking FILO shared memory and network connectivity is used to share data between real-time processes. SSH tunneling is used if a secure network connection between controllers is required. Built-in collision avoidance, inverse kinematics, and support for multiple programming languages was implemented to expand usability to our non-hardware-focused partners. Finally, a ROS interface was developed with specific focus on making it ROS 2.0 compatible to enable the use of the extensive ROS ecosystem. These combined contributions allowed MDS-Ach to significantly extend the research life of the MDS Robot.

Acknowledgment This work was performed in part at the Naval Research Laboratory under the project Adaptive Real-Time Algorithms for Multiagent Cooperation in Adversarial Environments. The views, positions and conclusions expressed herein reflect only the authors opinions and expressly do not reflect those of the Naval Research Laboratory, nor those of the Office of Naval Research.

References

- [1] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, 2009.
- [2] C. Breazeal, M. Siegel, M. Berlin, J. Gray, R. A. Grupen, P. Deegan, J. Weber, K. Narendran, and J. McBean, "Mobile, dexterous, social robots for mobile manipulation and human-robot interaction," in *International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2008, Los Angeles, California, August 11-15, 2008, New Tech Demos Proceedings*. ACM, 2008, p. 27. [Online]. Available: <http://doi.acm.org/10.1145/1401615.1401642>
- [3] Y. Maruyama, S. Kato, and T. Azumi, "Exploring the performance of ros2," in *Proceedings of the 13th International Conference on Embedded Software*, ser. EMSOFT '16. New York, NY, USA: ACM, 2016, pp. 5:1–5:10. [Online]. Available: <http://doi.acm.org/10.1145/2968478.2968502>
- [4] D. Calisi, A. Censi, L. Iocchi, and D. Nardi, "OpenRDK: a framework for rapid and concurrent software prototyping," in *Proceedings of Int. Workshop on System and Concurrent Engineering for Space Applications (SECESA)*, Nov. 2008.
- [5] T. H. J. Collett and B. A. Macdonald, "Player 2.0: Toward a practical robot programming framework," in *Proc. of the Australasian Conference on Robotics and Automation (ACRA)*, 2005.
- [6] G. Metta, P. Fitzpatrick, and L. Natale, "Yarp: Yet another robot platform," *International Journal of Advanced Robotics Systems, special issue on Software Development and Integration in Robotics*, vol. 3, no. 1, 2006.
- [7] C. Ct, Y. Brosseau, D. Ltourneau, C. Raevsky, and F. Michaud, "Robotic software integration using marie," *International Journal of Advanced Robotic Systems*, vol. 3, no. 1, pp. 055–060, March 2006. [Online]. Available: <http://www.ars-journal.com/International-Journal-of-Advanced-Robotic-Systems/Volume-3/055-060.pdf>
- [8] N. Ando, T. Suehiro, K. Kitagaki, T. Kotoku, and W.-K. Yoon, "RT-Component Object Model in RT-Middleware Distributed Component Middleware for RT (Robot Technology)," pp. 457–462, Jun. 2005.
- [9] A. Makarenko and A. Brooks, "Orca: Components for robotics," in *In 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, 2006.
- [10] M. Henning, "A new approach to object-oriented middleware," *IEEE Internet Computing*, vol. 8, no. 1, pp. 66–75, Jan 2004.
- [11] S. Rowe and C. R. Wagner, "An introduction to the joint architecture for unmanned systems (jaus)," *Ann Arbor*, vol. 1001, p. 48108.
- [12] N. Dantam, D. Lofaro, A. Hereid, P. Oh, A. Ames, and M. Stilman, "The ach library: A new framework for real-time communication," *Robotics Automation Magazine, IEEE*, vol. 22, no. 1, pp. 76–85, March 2015.
- [13] D. M. Lofaro, C. Ward, M. Bugajska, and D. Sofge, "Extending the life of legacy robots: Mds-ach, a real-time, process based, networked, secure middleware based on the x-ach methodology," in *15th International Workshop on Advanced Motion Control (IEEE-AMC2018)*, March 2018.
- [14] A. Perez, M. Orsag, and D. Lofaro, "Design, implementation, and control of the underwater legged robot aquashoko for low-signature underwater exploration," in *2018 15th IEEE International Conference on Ubiquitous Robots (UR)*, 2018.
- [15] J. Brown and B. Martin, "How fast is fast enough? Choosing between Xenomai and Linux for real-time applications," in *Twelfth Real-Time Linux Workshop on October 25 to 27, in Nairobi, Kenya*, 2010.
- [16] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai, Japan, Sep 2004, pp. 2149–2154.
- [17] R. Smith, "Open dynamics engine," 2008. [Online]. Available: <http://www.ode.org/>
- [18] S. R. Buss, "Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods."
- [19] D. Lofaro, R. Ellenberg, P. Oh, and J. Oh, "Humanoid throwing: Design of collision-free trajectories with sparse reachable maps," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, Oct 2012, pp. 1519–1524.
- [20] D. M. Lofaro, "Secure robotics," in *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, Aug 2016, pp. 311–313.
- [21] D. Lofaro, "Mds-ach," in <https://github.com/thedancomplex/mds-ach>, 2015.

10 APPENDIX

A Daemon Console

This section shows the specific console inputs available for the MDS-Ach system.

Console Input: `$ mds-ach` - Shows command options

Console Input: `$ mds-ach start` - Start all channels and processes and console

- (no-arg) : Starts MDS-Ach system on CAN Bus 0 "CAN0"
- nocan : Starts MDS-Ach system with no output to real CAN Bus. Will output to the virtual can "VCAN42" instead

Console Input: `$ mds-ach console` - Starts the human interface console for MDS-Ach

Console Input: `$ mds-ach stop` - Close all channels and processes

Console Input: `$ mds-ach make` - makes all the MDS channels

Console Input: `$ mds-ach kill` - Emergency kill the daemon process

Console Input: `$ mds-ach killall` - Emergency kill the daemon process and removes all ACH channels

Console Input: `$ mds-ach resetbus` - Resets the Bus

Console Input: `$ mds-ach remote` - Starts a remote connection to xxx.xxx.xxx.xxx via achd

Console Input: `$ mds-ach sim` - Starts the sim in gazebo

- (no-arg) : Starts the sim in gazebo
- kill : Kills gazebo sim

Console Input: `$ mds-ach changerobot` - Changes the robot's configuration file/anatomy

- (no-arg) : No Change
- isaac : Changes to Isaac's anatomy
- lucas : Changes to Lucas' anatomy
- octavia : Changes to Octavia's anatomy

B Startup Procedure

To start the MDS-Ach daemon select one of the two options. Note: Both Option 1 and Option 2 will start the following processes:

- mds-daemon : *primary control for the MDS robot*
- mds-filter : *filtering process to allow for step inputs*
- mds_ik_module (python) : *inverse kinematics (ik) controller for the MDS*

B.1 (Option 1) With physical MDS robot:

Bash:

```
$ mds-ach start
```

B.2 (Option 2) With virtual MDS robot:

Bash (Start with no CAN Bus):

```
$ mds-ach start nocan
```

Bash (Start Simulator):

```
$ mds-ach sim
```

The simulator can be run with either (Option 1) or (Option 2) above.

C Examples

The following are examples of how to do basic operations using the MDS-Ach system on both the real robot and the simulator.

Run MDS-Ach Daemon:

Bash:

```
$ mds-ach start
```

The resulting terminal windows should look like Figure 14.

```
lofaro@ln14ds20: ~
re-started bus-errors arbit-lost error-warn error-pass bus-off
0 0 0 0 0 0
RX: bytes packets errors dropped overrun mcast
0 0 0 0 0 0
TX: bytes packets errors dropped carrier collsns
0 0 0 0 0 0
7: can0: <NOARP,UP,LOWER_UP,ECHO> ntu 16 qdisc pfifo_fast state UNKNOWN mode DEF
AULT group default qlen 100000
link/can promiscuity 0
can state ERROR-ACTIVE restart-ms 100
bitrate 1000000 sample-point 0.750
tq 125 prop-seg 2 phase-seg1 3 phase-seg2 2 sjw 1
kvaser_usb: tseg1 1..16 tseg2 1..8 sjw 1..4 brp 1..64 brp-inc 1
clock 8000000
re-started bus-errors arbit-lost error-warn error-pass bus-off
1 0 0 0 0 0
RX: bytes packets errors dropped overrun mcast
8 1 0 0 0 0
TX: bytes packets errors dropped carrier collsns
0 0 0 0 0 0
nohup: appending output to 'nohup.out'
lofaro@ln14ds20:~$ nohup: appending output to 'nohup.out'
nohup: redirecting stderr to stdout
```

Figure 14: Expected window for (`$ mds-ach start`)

Run MDS-Ach Daemon with simulator only:

Bash:

```
$ mds-ach start nocan
```

This command will start the MDS-Ach daemon. This should be run on the computer where the simulator is located. The MDS-Ach daemon (mds-daemon) will run in the background even if the terminal session is closed. This will run with the simulator and not with the real robot. Use this mode if you only want to use the simulator.

The resulting terminal windows should look like Figure 15.

```
lofaro@ln14ds20: ~
lofaro@ln14ds20:~$ mds-ach start nocan
No CAN mode set
Device "can0" does not exist.
can0: ERROR while getting interface flags: No such device
cannot find device "can0"
cannot find device "can0"
cannot find device "can0"
cannot find device "can0"
can0: ERROR while getting interface flags: No such device
Device "can0" does not exist.
vcn42 set
lofaro@ln14ds20:~$ nohup: appending output to 'nohup.out'
nohup: appending output to 'nohup.out'
nohup: redirecting stderr to stdout
```

Figure 15: Expected window for (`$ mds-ach start nocan`)

Run MDS-Ach simulator:

Bash:

```
$ mds-ach sim
```

Once the MDS-Ach Daemon is started (see above) you can run the simulator. The simulator is open loop in the respect that it does not feed back information to the MDS-Ach system. It is used as a visual representation of the robot for debugging and initial controller testing.

The resulting terminal windows should look like Figure 16.

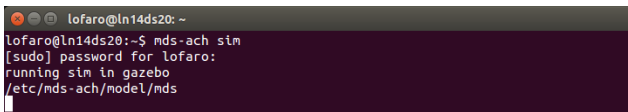


Figure 16: Expected window for (\$ mds-ach sim)

The resulting simulator windows should look like Figure 6.

D MDS-Ach Console

The MDS-Console utility allows the user to get and set joint space values via the command line. It also allows the user to get the work-space end-effector position (6 DOF) and set the work-space end-effector position (3, 4, 5, and 6 DOF).

Prerequisites:

The MDS-Ach system must be running prior to running the MDS-Console

Startup:

Bash:

```
$ mds-ach console
```

To start the MDS-console run the above command. The expected terminal can be seen in Figure 17.

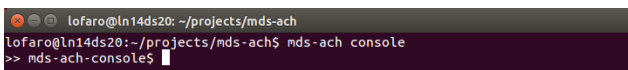


Figure 17: Expected window for (\$ mds-ach console)

Commands:

This section shows the commands available to the MDS-Console. Note: when in MDS-Console the console will show the following in the terminal:

Bash/MDS-Ach Console:

```
>> mds-ach-console$
```

Function: goto:

The goto command will tell the joint (in joint space) what position (in radians) where to go. The usable joints and abbreviations can be found in Table 2.

Bash/MDS-Ach Console:

```
>> mds-ach-console$ goto <joint> <value>
```

This will set the Right Shoulder Pitch (RSP / RightShoulderExt) to a value of -0.123 rad. The expected terminal can be seen in Figure 18.

Bash/MDS-Ach Console (Example):

```
>> mds-ach-console$ goto RSP -0.123
```

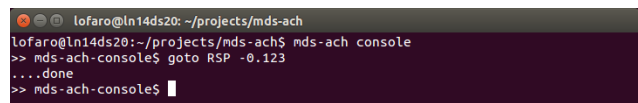


Figure 18: Expected window for (>> mds-ach-console\$ goto <joint> <value>) using example (>> mds-ach-console\$ goto RSP -0.123)

The “get” command gets the joint space position of the <joint> in radians. The usable joints and abbreviations can be found in Table 2.

Function: get:

Bash/MDS-Ach Console:

```
>> mds-ach-console$ get <joint>
```

This will get the reference and state of the Torso yaw (WST / TorsoPan). The expected terminal can be seen in Figure 19.

Bash/MDS-Ach Console (Example):

```
>> mds-ach-console$ get WST
```

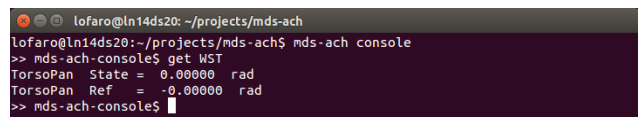


Figure 19: Expected window for (>> mds-ach-console\$ get <joint>) using example (>> mds-ach-console\$ get WST)

Function: get fk:

The “get fk” command gets the work-space position of the left or right end-effector in 6 DOF coordinates (meters and radians) with the origin being the intersection of the robot’s neck and shoulder. The <arm> options are: **left** and **right** for the left and right arm respectively.

Bash/MDS-Ach Console:

```
>> mds-ach-console$ get fk <arm>
```

Bash/MDS-Ach Console (Example):

```
>> mds-ach-console$ get fk left
```

This will get the 6 DOF work space position of the left end-effector. The expected terminal can be seen in Figure 20.

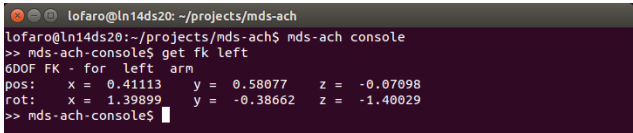


Figure 20: Expected window for (>> mds-ach-console\$ get fk < arm >) using example (>> mds-ach-console\$ get fk left) where (pos) is the position in meters and (rot) is the rotation about x,y,z in radians.

Function: ik:

Bash/MDS-Ach Console:

```
>> mds-ach-console$ ik <arm> <dof> <param 1>
... <param N>
```

The “ik” command utilizes the inverse kinematic controller (mds_ik_module) to solve 1-6 DOF inverse kinematic solutions for the MDS robot’s left or right end-effector. This module uses the Inverse Jacobian Inverse Kinematic solver method. If the desired work-space location is too far away such that the IK controller cannot reach it in 1000 iterations, it will discontinue attempting to find a solution and return without a reply. You may then try another point closer to that of the current end-effector point.

< arm > : has the options of “left” and “right.” and denote the left and right end-effector respectively.

< dof > : denotes the number of degrees of freedom you will be controlling using the inverse kinematic controller.

< param 1 > ... < param N > : denotes the positions and orientations for the < arm >. The number of parameters must equal that of < dof >. The order must be as follows (Table 5):

Table 5: Inverse Kinematic Parameter Order

Param #	Definition	Abbreviation
1	Position in X	p_x
2	Position in Y	p_y
3	Position in Z	p_z
4	Rotation in X	θ_x
5	Rotation in Y	θ_y
6	Rotation in Z	θ_z

Bash/MDS-Ach Console (Example):

```
>> mds-ach-console$ ik left 3 0.3 0.2 0.0
```

This will find a joint space solution using IK methods for the desired end-effector position to be (0.3 m, 0.2 m, 0.0 m) in (x,y,z). The angle about all axes is in the null space.

The expected output (with running get fk left before and after to show the effect) can be seen in Figure 21.

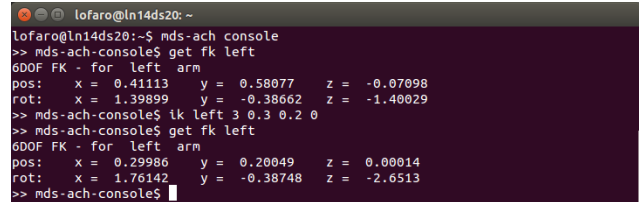


Figure 21: Expected window for (>> mds-ach-console\$ ik < arm > < dof > < param 1 > ... < param N >) using example (>> mds-ach-console\$ ik left 3 0.3 0.2 0.0)

E Software Interface

The MDS-Ach system currently works with the C/C++ and Python programming languages. This section describes the required libraries for each of the latter languages.

Python imports:

The following are required imports for the MDS-Ach system while using python: mds_ach, ach, and mds. Other imports are for the given controller implementation.

Python:

```
#!/usr/bin/env python
import mds_ach as mds
import ach
```

C/C++ includes:

The following are required imports for the MDS-Ach system while using python: mds_ach, ach, and mds. Other imports are for the given controller implementation.

C/C++:

```
#include <mds.h>
#include <ach.h>
```

Required library:

MakeFile:

```
-lach
```

C/C++ MakeFile example with required libraries:

The following is an example make file for a C implementation of a controller. Please note it utilizes the required -lach library.

MakeFile:

```

default: all
CFLAGS := -I./include -g -std=gnu99
CC := gcc
BINARIES := mds-simple-demo
LIBS := -lach -lrt -lm

all : $(BINARIES)

mds-simple-demo: src/mds-simple-demo.o
    $(CC) -o $@ $(LIBS)

%.o: %.c
    $(CC) $(CFLAGS) -o $@ -c $<

clean:
    rm -f $(BINARIES) src/*.o

```

F Operating in Joint-Space

This section shows you how to setup the Ach channels for communication with the MDS-Ach system and how to control the robot in joint-space while using the smoothing filter process.

F.1 Control one joint/DOF (Python)

This section shows you how to setup the Ach channels for communication with the MDS-Ach system and how to control one joint of the robot in joint-space while using the smoothing filter process. Specifically we will set the joint-space values as those seen in Table 6. Note that to run this example you will need the libraries for MDS-Ach as seen in Section E.

Table 6: Set the joint space values for the following joint

Name	Alternate Name	Value (rad)
LSP	LeftShoulderExt	-0.123

Open Ach Channels:

Ach channels are how you communicate with MDS-Ach. You simply write data to the channel and the robot can read the data in newest to oldest order. The section below shows you how to open an Ach channel. Please note that `ach.Channel()` takes a string as an input. Here we opened two channels:

- `s` : state channel
- `r` : reference channel

Python:

```

s = ach.Channel(mds.MDS_CHAN.STATE_NAME)
r = ach.Channel(mds.MDS_CHAN.REF.NAME)

```

Create Required Data Structures:

C-Type data structures are used to pass data between our controllers. Below we create three well defined structures for the state and the reference channels. These structures are defined in `mds.ach.py` and `mds.h` which is located in your python and include paths.

- `state` : state channel of type `MDS_STATE`
- `ref` : reference channel of type `MDS_REF`

Python:

```

state = mds.MDS_STATE()
ref = mds.MDS_REF()

```

Get Joint ID:

To command a joint you must get the ID of the joint. The ID numbers are defined in the `anatomy.xml` configuration file. You can use the joint abbreviations in Table 18 to find the ID number.

Python:

```

# Get address of LSP
jntn = mds.getAddress('LSP', state)
print 'Address_{}_{}'.format(jntn, state)

```

Queue New Motor Position:

You can set a new desired motor angle by setting the reference channel. In this case we are using the filter channel which is the safest one to use due to the velocity and acceleration limiting. Please note that this does NOT send the command to the motor. It queues the values for them to be sent to the motors. They are only sent to the motors after they are “put” on the ach channel.

In the example below we are setting the ‘LSP’ by using the ID number from above to a joint-space angle of -0.123 rad.

Python:

```

# Set LSP reference to -0.123 using the
# filter controller
ref.joint[jntn].ref = -0.123

```

Set Motor Position (put):

Once all joints are set in the structure you can “put” it on the proper ach channel. Please note that even if you did not set a motor value it will still be put on the channel along with the rest of the data structure. It is best practice to read the latest channel, then modify what you want to change, then put the modified structure on the channel. This will help with not sending the robot to unintended configurations.

Python:

```
# Command motors to desired references
# (post to ach channel)
r.put(ref)
```

Close Ach Channels on Exit:

Though not required it is good practice to close your unused ach channels upon exit of your controller.

Python:

```
# Close channels
s.close()
r.close()
```

Running the Code: To run the code do the following from within the *mds-ach/examples* directory

Bash:

```
$ python mds_simple_demo_python_1_DOF.py
```

The expected terminal out can be seen in Figure 22.

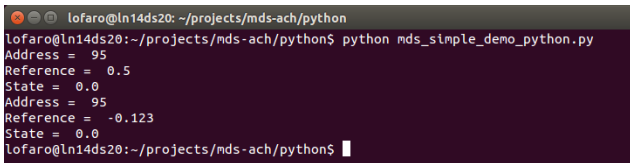


Figure 22: Expected window for the python 1-DOF example including additional print statements

You can also monitor the change using the MDS-Ach Read utility and/or the simulator/real robot.

Full Code:

The full example can be found below as well in the file:

`mds-ach/examples/mds_simple_demo_python_1_DOF.py`

E.2 Control Two joints/DOF (C/C++)

This section you will set 2 separate DOFs to two different values and see the results. Specifically we will set the joint-space values as those seen in Table 7. Note that to run this example you will need the libraries for MDS-Ach as seen in Section E. We will set the following:

Table 7: Set the joint space values for the following joint

Name	Alternate Name	Value (rad)
LEP	LeftElbowFlex	-0.2
RSP	RightShoulderExt	0.1

Open Ach Channels:

Ach channels are how you communicate with MDS-Ach. You simply write data to the channel and the robot can read the data in newest to oldest order. The section below shows you how to open an Ach channel. Please note that `ach.Channel()` takes a string as an input. Here we opened three channels.

- `chan_state` : state channel
- `chan_ref` : reference channel

C/C++:

```
int r = ach_open(&chan_ref,
                MDS_CHAN_REF_FILTER_NAME, NULL);
assert(ACHLOK == r);

r = ach_open(&chan_state,
            MDS_CHAN_STATE_NAME, NULL);
assert(ACHLOK == r);
```

Create Required Data Structures:

C-Type data structures are used to pass data between our controllers. Below we create three well defined structures for the state and the reference channels. These structures are defined in `mds_ach.py` and `mds.h` which is located in your python and include paths.

- `state` : state channel of type `MDS_STATE`
- `ref` : reference channel of type `MDS_REF`

C/C++:

```
mds_ref_t H_ref;
mds_state_t H_state;
memset(&H_ref, 0, sizeof(H_ref));
memset(&H_state, 0, sizeof(H_state));
```

Get Up-To-Date Reference:

It is important that you set your initial reference structure with the current values of the reference channel. This is because when you do command the joints you command them all at once, even if you did not change their value. See below for how to get the latest reference values.

C/C++:

```
int r = ach_get(&chan_ref, &H_ref,
               sizeof(H_ref), &fs,
               NULL, ACHLO_LAST);
if(ACHLOK != r) {
    if(debug) {
        fprintf(stderr, "Ref_r_ = %s\n",
                ach_result_to_string(r));
    }
    else{assert(sizeof(H_ref) == fs);}
}
```

Get Joint ID:

For the moment when using C/C++ the joint IDs are hard coded in the mds.h include file. You may use normal humanoid acronyms or NRL legacy acronyms.

Queue New Motor Positions:

You can set a new desired motor angle by setting the reference channel. In this case we are using the filter channel which is the safest one to use due to the velocity and acceleration limiting. Please note that this does NOT send the command to the motor. It queues the values for them to be sent to the motors. They are only sent to the motors after they are “put” on the ach channel.

In the example below we are setting the joints as defined in Table 7.

C/C++:

```
H_ref.joint[LEP].ref = -0.2;
H_ref.joint[RSP].ref = 0.1;
```

Set Motor Position (put):

Once all joints are set in the structure you can “put” it on the proper ach channel. Please note that even if you did not set a motor value it will still be put on the channel along with the rest of the data structure. It is best practice to read the latest channel, then modify what you want to change, then put the modified structure on the channel. This will help with not sending the robot to unintended configurations.

C/C++:

```
/* Write to the feed-forward channel */
ach_put( &chan_ref, &H_ref, sizeof(H_ref));
```

Running the example:

To run the example compile then run the resulting executable mds-simple-demo.

Bash:

```
$ ./mds-simple-demo
```

The resulting output can be seen using the MDS-Ach Read utility (Figure 7) and/or the simulator/real robot (Figure 8).

Full Code:

The full example can be found in the file:

```
mds-simple-demo/src/mds-simple-demo.c
```

G Built-in IK Controller (C/C++)

The MDS-Ach system has a built-in inverse kinematics (IK) controller for both the right and left arms. This controller automatically runs when the MDS-Ach system is started. The controller solves for one arm at a time.

The system works by starting to solve the IK equations when you post a new desired position on the IK reference channel. Below is a step by step of how you do this using C/C++.

C/C++ Includes:

The following are required imports for the MDS-Ach system while using python: mds.h and ach.h. Other imports are for the given controller implementation.

C/C++:

```
// for mds
#include <mds.h>
// for ach
#include <ach.h>
```

Create Open Ach Channel for IK:

Ach channels are how you communicate with MDS-Ach. You simply write data to the channel and the robot can read the data in newest to oldest order. The section below shows you how to open an Ach channel. Please note that ach.open() takes a string as an input. Here we opened one channel. This is a different channel then found in previous sections because it is only for the IK controller.

C/C++:

```
// open ik chan
int r = ach_open(&chan_ik,
                MDS_CHAN_IK_NAME, NULL);
assert( ACHLOK == r);
```

Make IK Structure:

Similar to controlling the robot in joint-space you need to set a reference structure to the desired work-space position. For this you need to initialize the structure. See below for the initialization of the work-space structure.

C/C++:

```
// Make new IK structure
mds_ik_t H_ik;
memset( &H_ik, 0, sizeof(H_ik));
```

Setting the DOF controlled:

When using the built-in IK controller you need to set the number of DOF that you are controlling. With this controller you are required to set the DOF in the following order: $p_x, p_y, p_z, \theta_x, \theta_y, \theta_z$ where p_n is the position on axis n and θ_n is the rotation about axis n . For example if you DOF is set to 4 you are controlling p_x, p_y, p_z , and θ_x . If you are controlling 2 you will control p_x and p_y . In the example below we are controlling 3, i.e. p_x, p_y , and p_z . This order can also be seen in Table 5.

C/C++:

```
// Set the amount of DOF you want to control
dof = 3
```

We set these to the values in Table 3.

$\theta_x, \theta_y,$ and θ_z are in the Null Space because we do not care where they are as long as the first three parameters are met. We can set these values to what ever we want and they will be ignored. In this case we set them to zero.

C/C++:

```
/* Set values for work-space in
[x, y, z, rx, ry, rz] order */
double eff[6] = {0.3, 0.2, 0.0, 0.0, 0.0, 0.0};
```

Choosing Arm for IK:

Here we pick the arm for the IK. Our options are set as enums in the mds.h include. All options for right and left arms can be found in Table 8.

Table 8: Definitions for left and right arms using mds_ach.py in Python

Arm	C/C++ Definition
Left	LEFT
Right	RIGHT

Here we set the arm to the left arm.

C/C++:

```
// set arm
int arm = LEFT;
```

Set IK Structure:

Just as in the joint-space method, we need to the values in our structure before we send it to the robot. Here we set all of the parameters from above to the structure ikc that we created.

C/C++:

```
// Put setting into ik structure
H_ik.move = arm;
H_ik.arm[arm].ik_method = dof;
H_ik.arm[arm].t_x = eff[0];
H_ik.arm[arm].t_y = eff[1];
H_ik.arm[arm].t_z = eff[2];
H_ik.arm[arm].r_x = eff[3];
H_ik.arm[arm].r_y = eff[4];
H_ik.arm[arm].r_z = eff[5];
```

Command the Robot:

Just as with the joint-space controller you need to “put” the data on the ACH channel. Unlike the joint-space controller it will not move as soon as you send it. The controller will first have to solve the IK. Upon finding the solution the controller will send it to the robot. If a solution is found it will take anywhere between 0.1 and 5.0 seconds. If there is no solution found the robot will not move. Note: the robot limits its self to 1000 search iterations for an IK solution.

C/C++:

```
// put on to ACH channel
ach_put( &chan_ik, &H_ik, sizeof(H_ik));
```

Running the Code:

To run the code do the following within the **mds-simple-demo-ik** folder.

Bash:

```
$ make clean
$ make
$ ./mds-simple-demo-ik
```

To check that the IK worked you can run the FK in the MDS-Ach console and/or run the simulator. The before and after of the MDS-Ach console is found in Figure 23 and Figure 24 respectively.

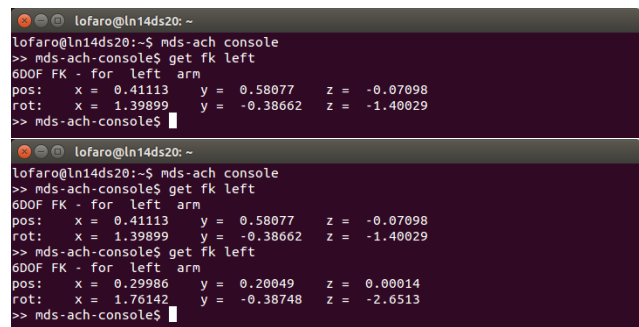


Figure 23: Expected MDS-Ach Console window for (>> mds-ach-console\$ get fk left). (TOP) Before running. (BOTTOM) After running.

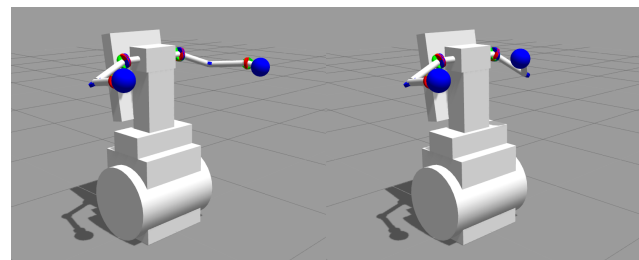


Figure 24: Expected simulator window for (\$./mds-simple-demo-ik) and/or (\$ python mds_ik_solver_example.py). (LEFT) Before running. (RIGHT) After running.

Full Code:

The full example can be found in the file:

```
mds-simple-demo/src/mds-simple-demo-ik.c
```

H Inverse Kinematics API

This section will show you an example of how to get and set the forward and inverse kinematics via the mds_ik API. The goal of this tutorial is to solve for the joint-space values given the desired work-space values

as found in Table 3. This API is written for use with Python but also has hooks for C/C++.

Required Imports:

Below are the required imports for using the forward/inverse kinematics portion of the mds API. These will be located in the python 2.7 path.

Python:

```
import mds_ach as mds
import ach
import mds_ik as ik
import mds_ik_include as ike
from mds_ach import *
```

Open Ach Channels:

You will need the state and the reference (filtered) channels to solve and set the IK using the API. The state is used to get the current location of the robot. The reference is used to set the joint-space values once they are found.

Python:

```
# Open Ach Channels
r = ach.Channel(mds.MDS_CHAN_REF_FILTER_NAME)
s = ach.Channel(mds.MDS_CHAN_STATE_NAME)
```

Make Structures:

Make the structures for the state and the reference channels. This will hold your latest state data and the latest set reference to the reference filter channel.

Python:

```
# Make Structs
state = mds.MDS.STATE()
ref = mds.MDS.REF()
```

Get Latest Data:

Get the latest data from the state and reference channels.

Python:

```
# Get the latest on the channels
[status, framesize] = s.get(state, wait=False, last=True)
[status, framesize] = r.get(ref, wait=False, last=True)
```

Set Desired Work-Space Position:

Set the desired work-space position as found in Table 3. We set all 6 DOF here however only the first three are required due to our desired work-space position. Later sections removes the extra DOFs.

Python:

```
# Desired position
eff_end = np.array([0.3, 0.2, 0.0, 0.0, 0.0, 0.0])
```

Set Desired Arm:

The ik.getIK() function takes a string to define the left or right arm. Thus we set our “arm” as a string. The options for the latter can be found in Table 9.

Table 9: Definitions for left and right arms using mds_ik.py in Python

Arm	MDS IK (Python) Definition
Left	'left'
Right	'right'

Python:

```
# Define Arm
arm = 'left'
```

Get Current Joint-Space Pose:

Next we have to get the current joint-space pose of the robot and putting it into a 6 DOF array. This is done by looking for the index of each joint and then putting them in an array. The array must be in the order defined in Table 10.

Python:

```
# Get current joint space pose of arm
jntn = mds.getAddress('LSP', state)
j0 = state.joint[jntn].ref
jntn = mds.getAddress('LSR', state)
j1 = state.joint[jntn].ref
jntn = mds.getAddress('LSY', state)
j2 = state.joint[jntn].ref
jntn = mds.getAddress('LEP', state)
j3 = state.joint[jntn].ref
jntn = mds.getAddress('LWY', state)
j4 = state.joint[jntn].ref
jntn = mds.getAddress('LWR', state)
j5 = state.joint[jntn].ref
eff_joint_space_current=[j0, j1, j2, j3, j4, j5]
```

Table 10: Arm joint-space pose order.

Array Index	Joint Definition	Short Name (left)	Short Name (right)
0	Shoulder Pitch	LSP	RSP
1	Shoulder Roll	LSR	RSR
2	Shoulder Yaw	LSY	RSY
3	Elbow Pitch	LEP	REP
4	Wrist Yaw	LWY	RWY
5	Wrist Roll	LWR	RWR

Set DOF and Order of Desired Work-Space Position:

We now set the DOF of the desired work-space (acceptable numbers are 1-6) and the order in which we

input them. Using the MDS IK API we can input the desired work-space coordinates in any order we desire. The order is defined by an array of string that we call ‘order’. The valid inputs for this array can be found in Table 11.

Table 11: Valid Inputs for “order” Array

Definition	Input to Array (String)
Position in x (p_x) (meters)	'p-x'
Position in y (p_y) (meters)	'p-y'
Position in z (p_z) (meters)	'p-z'
Rotation about x (θ_x) (radians)	't-x'
Rotation about y (θ_y) (radians)	't-y'
Rotation about z (θ_z) (radians)	't-z'

Here we have 3 DOF with an order of position in x , y , then z . Note: This order MUST match that of the order in your desired work-space position array in Section H.

Python:

```
# set the dof and the order (dof = 3)
dof = 3
order = ['p-x', 'p-y', 'p-z']
```

Set Error:

Here we set the error for the IK solver. The error is a 3D array with the attributes found in Table 12.

Table 12: Error Array Index Definitions

Index	Definition
0	Max change in angular position (θ) per iteration (rad)
1	Max change in linear position (xyz) per iteration (m)
2	Max linear Error ($m^2 + rad^2$)

The max change in angular position is the maximum distance in radians that the solver will move per iteration. The max change in translational position is the maximum distance translated in x, y, z per iteration in meters. The max linear error (e_{max}) is the linear distance in N DOF that disregards units. The error in the end-effectors actual position vs its desired position (e_{eff}) can be found via the equation below.

$$e_{eff} = \left(\sum_{i=0}^{N-1} (d_{des}[i] - d_{act}[i])^2 \right)^{\frac{1}{2}} \quad (5)$$

where d_{des} is the desired pose and d_{act} is the actual pose. The system will say the IK has been solved if $e_{eff} \leq e_{max}$.

To set the desired errors make the 3D array as seen below. If no error is input the default values of [0.01, 0.01, 0.01] will be used.

Python:

```
# set the allowable error
err = np.array([0.01, 0.01, 0.01])
```

Set Max Solving Iterations:

Due to the nature of the IK method utilized the system can fall into local minimum or a hysteresis. This will cause the system to attempt to solve indefinitely. To avoid this use add a maximum number of iterations to try when solving the IK. This number is set below. If not set it will use the default value of 1000.

Python:

```
# Set solving step number max
stepNum = 1000
```

Limit input to desired DOF:

Here you limit the size of the desired end-effector position vector to the ‘dof’ that we defined in Section H.

Python:

```
# Desired position for only the dof we want
eff_end = eff_end[:dof]
```

Solving the IK:

Now we can solve the IK via the use of ik.getIK() in the mds_ik python module.

Python:

```
# Solve IK
jnt_return = ik.getIK(eff_joint_space_current,
                    eff_end, order, arm, err,
                    stepNum)
```

The output of this is a 2D array where index 1 returns a -1 if there is no IK solved. Index 0 is the array of the end-effectors joint-space values for the given desired work-space position. The order of the joint-space values are the same as what is found in Table 10.

Python:

```
# returns in the following order
# Joint space return =
# [LSP, LSR, LRY, LEP, LWY, LWR]
eff_joint_space_current = jnt_return[0]
```

Putting the IK solution on the robot:

To put the IK solution on the robot we map the received joint space values on the reference (ref) structure of the robot’s MDS-Ach system. Again the order of the joint-space values are the same as what is found in Table 10. In this case we are setting the left arm.

Python:

```
# Set to joint space
jntn = mds.getAddress('LSP', state)
ref.joint[jntn].ref=eff_joint_space_current[0]
jnt n= mds.getAddress('LSR', state)
ref.joint[jntn].ref=eff_joint_space_current[1]
jntn = mds.getAddress('LSY', state)
ref.joint[jntn].ref=eff_joint_space_current[2]
jnt n= mds.getAddress('LEP', state)
ref.joint[jntn].ref=eff_joint_space_current[3]
jntn = mds.getAddress('LWY', state)
ref.joint[jntn].ref=eff_joint_space_current[4]
jntn = mds.getAddress('LWR', state)
ref.joint[jntn].ref=eff_joint_space_current[5]
```

Commanding the robot:

Just as in previous sections, we have to “put” the resulting structure on the reference (filtered) channel before the robot will move. Please note that we are using the filtered channel the the robot will not “jerk” during the joint-space step input operation.

Python:

```
# Send to the robot
r.put(ref)
```

Getting the Forward Kinematics from joint-space pose:

We can get the work-space position of the arm via the forward kinematics (FK) by utilizing the joint-space pose the the ik.getFkArm() function of the mds_ik module.

Python:

```
# get FK of arm
A = ik.getFkArm(eff_joint_space_current , arm)
eff_end_ret=ik.getPosCurrentFromOrder(A, order)
```

ik.getFkArm() will return a 4x4 matrix which includes the rotation and translations components. The ik.getPosCurrentFromOrder() returns the work-space position and orientation in the same order as requested by “order”. This will return a 1xN array where N is equal to the length of your “order”. You can find the error by doing the following.

Python:

```
# find different in desired vs actual pos
eff_end_dif = eff_end - eff_end_ret
```

Running the code:

To run the code enter the “examples” directory of the mds-ach project and run the following command:

Bash:

```
$ python mds_ik_solver_example.py
```

The resulting console output should look like Figure 25.

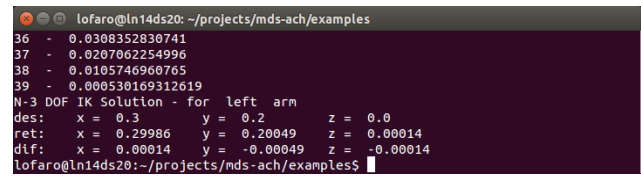


Figure 25: Expected window for (\$ python mds_ik_solver_example.py)

The resulting simulator output should look like Figure 24.

Full Code:

The full example can be found in the file:

mds-ach/examples/mds_ik_solver_example.py

State Estimation based Echolocation Bionics and Image Processing based Target Pattern Recognition

David Kondru^{*1}, Mehmet Celenk¹, Xiaoping A. Shen²

¹School of Electrical Engineering and Computer Science, Ohio University, Athens, Ohio-45701, USA

²Department of Mathematics, Ohio University, Athens, Ohio-45701, USA

ARTICLE INFO

Article history:

Received: 28 November, 2018

Accepted: 08 January, 2019

Online : 20 January, 2019

Keywords:

Bionics

Kalman filter

Detection

ABSTRACT

This paper deals with the theoretical aspect of bat echolocation and bionics, and image processing-based target recognition and identification methods. The state estimation methods utilizing the linear rustic filters such as fixed gain and Kalman filters are studied and implemented for echolocation bionics for estimating the LOS distance. A complete mathematical modeling and simulation of bat dynamics and its prey are presented upon which the relative LOS distance is reconstructed with state estimators and less RMS errors. Also, target recognition and identification using Optical, IR Digital Night Vision and Thermal camera is studied and implemented at different environmental conditions to demonstrate the superiority of thermal camera.

1. Introduction

Since the dawn of the Strategic Defense Initiative (SDI), different types of threats and the complex defense system architectures have always been a confrontation to the target detection and tracking. In this context, detection is always prior to tracking without which the aerospace guidance and control is practically impossible [1]. In fact, this is true for civilian sensors such as Primary and Secondary Surveillance Radars to control incoming and departing airplanes, as well as, for military sensors to perform a sequence of successful tasks [2]. The degree of target detection and tracking varies from single sensor to multiple sensors which led to multi sensor fusion for multi object tracking. Intelligence, Surveillance and Reconnaissance (ISR), missile guidance and control, remote sensing and oceanography, computer vision and robotic applications are the few wide variety of applications utilizing from single sensor to multiple sensor detection and tracking [3]. Based on the type of threat and nature of attack many integrated solutions have been introduced to confront with the threat. One such an example is U.S. National Missile Defense system. Aurgus and Cerebrus are the FLIR designed integrated systems for border surveillance with 24/7 situational awareness [4]. Apart from RADAR, Optical and Thermal technology, Bat acoustic detection system, a new technology was introduced for target localization and identification. Echolocation bionics [5], [6] based on the bat technology utilizing the acoustic echo is a process of determining the relative range and velocity, bearing angle and

size of the target. In fact, echolocation bionics has the superior advantages and success rate compared to other detection and tracking technologies that made it helpful in civil and military applications. However, irrespective of any given dynamic situation and the type of technology, the designated track has to be followed by a closed loop system tracker for airborne interception or ground attack that requires high performance data association algorithms. The purpose of data association and processing algorithms is to mitigate the high degree of uncertainty associated with the target motion and the environmental conditions resulting from system noise and measurement noise [7] [8]. With these insights, this paper presents the theoretical advancements in echolocation bionics and, tracking and filtering of bat acoustic signal that could enhance the performance of echolocation bionic sensors. On the other hand, this paper also focuses on application of image processing techniques using Finite Impulse Response filters as an added extension to the echolocation bionics. Two-dimensional FIR filters are the most profoundly used filters for image data processing from a given particular optical device. Image data processing techniques such as image sharpening and smoothing, detecting edges in a particular image clustered with many features, and contrast improvement are necessary to apply for agiven 2D image data in the presence of noisy environments. All these techniques are applied to visualize the data as being the key objective in modern computational sciences. Data visualization is not an easy operation as how it has been in a perfect daylight conditions. In poor visibility, capturing and processing of image data and its

^{*}David Kondru, Ohio University, Email: k.rajusolomon@gmail.com

visualization is quite daunting and a difficult task. However, there are optical and thermal sensors that can enhance processing of an image even in poor visibility conditions. In this paper, the image of a drone is captured by a digital night vision camera with built-in IR mode and a thermal camera to examine the effects with a 2D FIR filter. This paper presents the superiority and the performance of individual sensor in detection and enhancement, and recognize the patterns in the captured images. Section 2 presents the echolocation bionics and state estimation. Section 3 discusses the sensor specifications and test images of target. Section 4 demonstrates the image processing based target recognition and identification methods.

2. Echolocation and Tracking

Usually, the mouth (nose) of bat is for broadcasting echolocation and its ears were used as receiving antennas. The emission system of bat is an adaptive waveform where each pulse consists of eight signals. Four long-constant frequency (CF) and four short constant frequency modulated (FM) harmonics [9]. The speed information is obtained by CF component and the FM component determines the detection and imaging. Whenever a bat gets closer to its stationary prey, due to Doppler Effect the echo frequency becomes higher than the emitted pulse [10]. At the same time, the pulse density becomes larger and larger which is shown in Figure 1 describing the attack process of a bat and the adaptive waveform conversion on the top left.

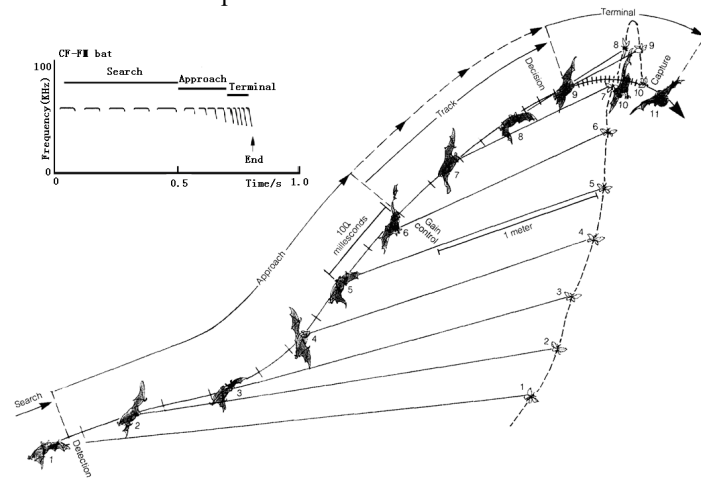


Figure 1. Bat intercepting its prey [10]

A three dimensional simulation of a bat sonar beam while attacking its prey, and the two dimensional spectrogram sequence of the bat echolocation relative to the target [11] is shown in Figure 2. Initially the bat searches for its prey using its directional beam and once the prey is detected the target is locked with the sonar beam. The bat while hunting to its prey decreases the number of echoes it emits using automatic gain control.

For the purpose of simulation and based on the literature available as represented in the Figures 1 and 2, the three dimensional motion of the bat is considered as a spiral motion and the prey or the target is stationary. The position and the velocity of the target are almost negligible when compared to the motion of the bat. It is assumed that the bat has already received the passive observation of the target and the simulation shows the hunting of the bat toward its prey. The relative position which is also known as line of sight

(LOS) between the bat and the prey is considered for tracking using different state estimators and the performance evaluation of each estimator in terms of RMS error measure. Using the flat earth approximation [12], the target and its prey are described in a Spherical Coordinate system [13] upon which the position is illustrated in Cartesian coordinate system for 3D target tracking. Therefore, the finite dimensional representation [14] for the target and the prey model, as well as, the LOS system measurements [15] are modeled by

$$[x \dot{x} \ddot{x} \ y \ \dot{y} \ \ddot{y} \ z \ \dot{z} \ \ddot{z}]^T \tag{1}$$

$$r = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \tag{2}$$

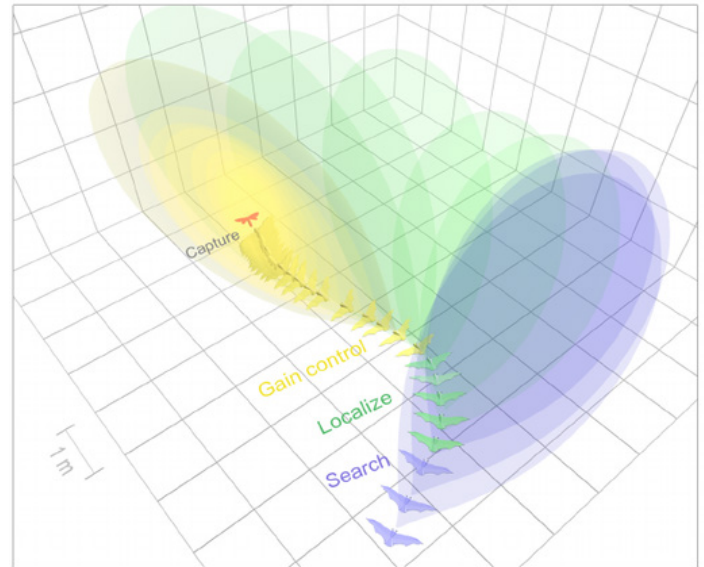


Figure 2. Approach methods of bat [11]

Equation 1 describes the 3D motion characteristics of bat and its prey in terms of position, velocity and acceleration. Equation 2 presents the relative position or the line of sight distance between the bat and its target. Figure 3 shows the motion characteristics [16] that are utilized in state variable form and all the motions are simulated in MATLAB.

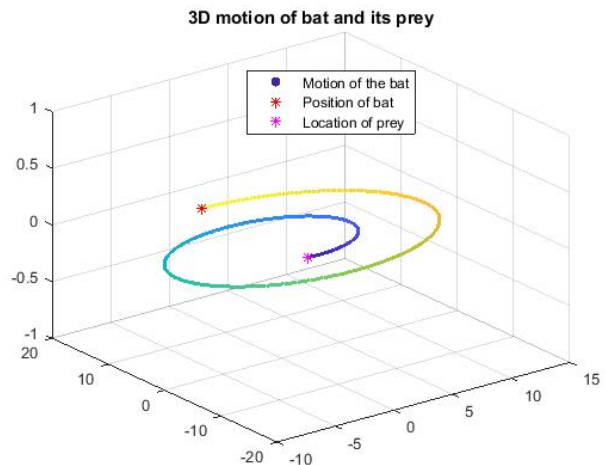


Figure 3. Bat its prey motion dynamics

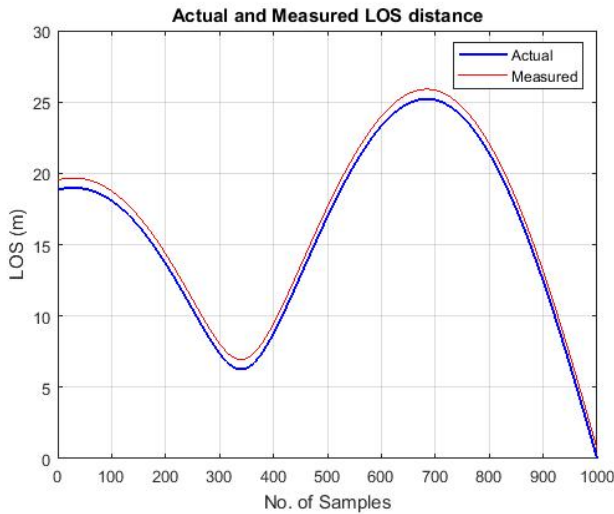


Figure 4. Relative LOS distance

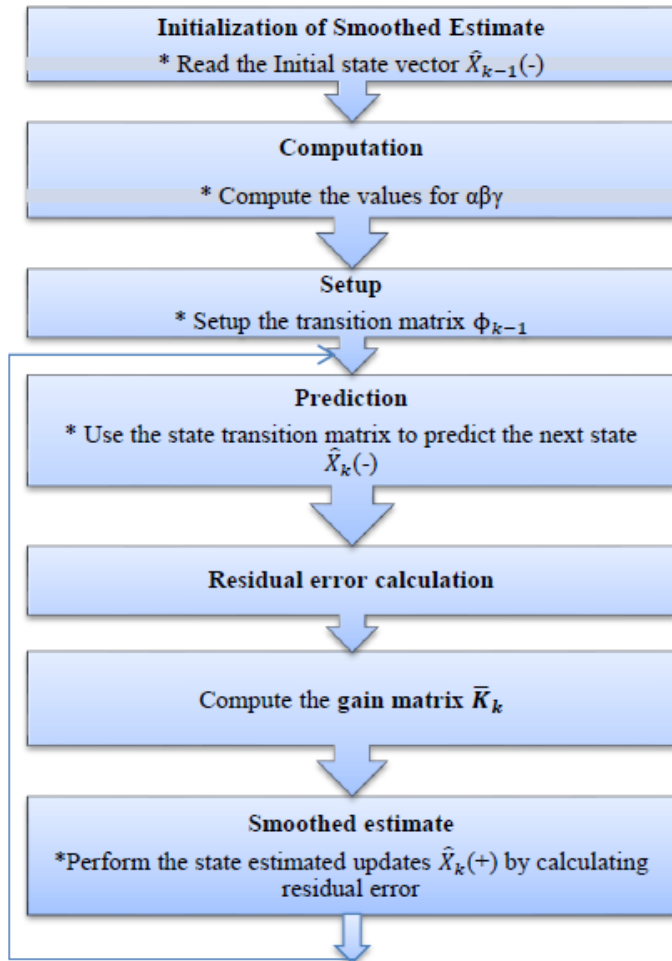


Figure 5. Fixed gain filter recursive flow diagram [19]

2.1. State Estimation of Line of Sight

The difference in position between the bat and its prey is considered as the line of sight (LOS) distance for which a popular finite-dimension approach known as state estimation is utilized in this paper for predicting and correcting the LOS distance. Therefore, the generic representation of a kinetic model [17] is given by

$$\hat{X}_k(-) = F\hat{X}_{k-1}(+) + Gw_{k-1} \quad (3)$$

$$Z_k = H\hat{X}_{k-1}(+) + v_k \quad (4)$$

Two different filters namely fixed-gain and Kalman filters are studied and implemented for LOS estimation. Before examining each filter, the underlying assumption made is portraying the measurement acoustic noise as zero-mean Gaussian (ZMG) as shown in the Figure 4 with a known standard deviation.

2.2. Linear Rustic Filters

Assuming that the system is a linear time invariant [18], the first class of rustic filters namely fixed gain and Kalman filters are implemented for which the process flow diagram for each filter is shown in Figures 5 and 6. A one dimensional third order filter is adopted for the LOS estimation.

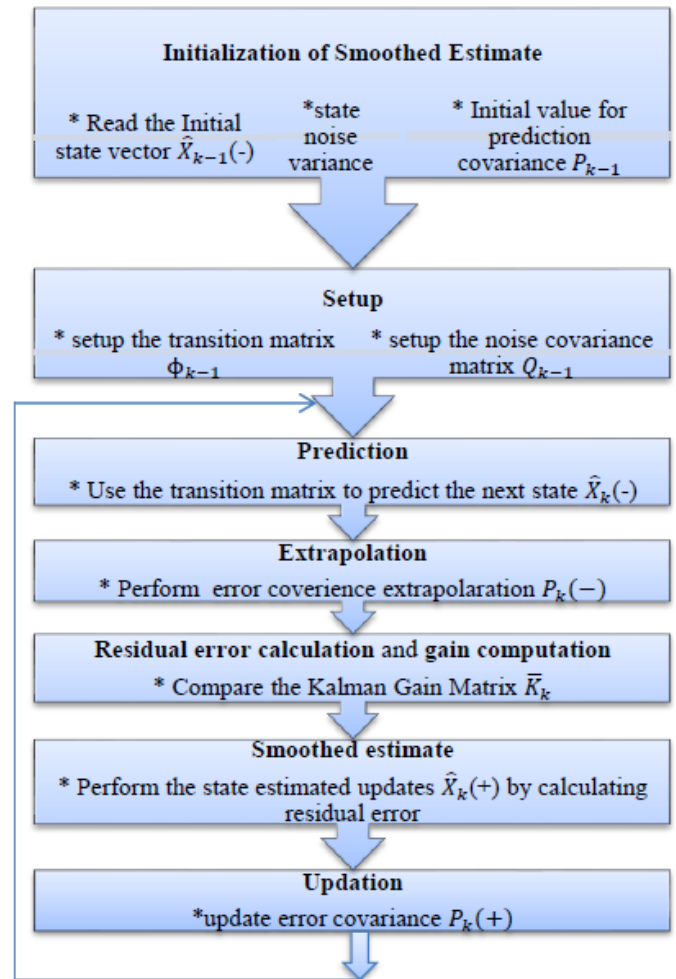


Figure 6. Kalman filter recursive flow diagram [19]

Table 1 Simulation Parameters

Sampling Interval (T)	0.01 sec
Measurement Noise Variance w_K	0.035
Process Noise Variance v_k	0.5

2.3. Simulation Analysis

For the purpose of simulation, it is assumed that both bat and prey are in the vicinity with a radius of maximum of 25 meters. The simulation parameters are shown in Table 1.

The initial values of prediction covariance and transition matrix are employed as follows:

$$P_{k-1} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}; \quad \phi = \begin{bmatrix} 1 & T & T^2/2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix} \quad (5)$$

To ensure the performance of each estimator the error analysis in LOS distance is computed using RMS error given by the Equation 6.

$$RMS\ Error_{pos,ang} = \sqrt{\frac{1}{N} \sum_{i=1}^N \frac{(x_{true} - x_{esti})^2 + (y_{true} - y_{esti})^2 + (z_{true} - z_{esti})^2}{3}} \quad (6)$$

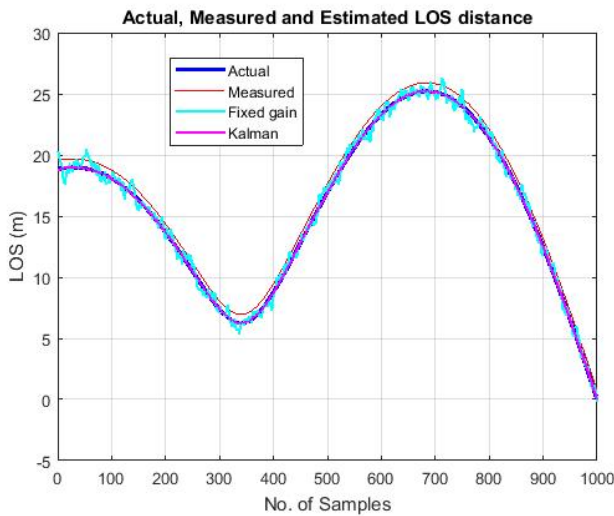


Figure 7. Actual, Measured and LOS distances

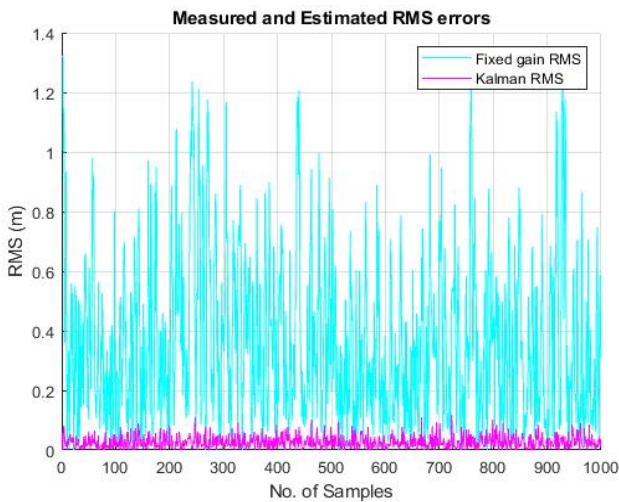


Figure 8. RMS Errors

Figure 7 shows the tracking performances for LOS distance. The principal function of these trackers is for better signal reconstruction by reducing measurement noise with less residual.

The performance of each filter estimates is shown from Figure 8. It is found that the efficiency of fixed gain filter is highly reliable upon the filter gain coefficients which are solely depend upon the smoothing coefficient that varies from zero to one. A value of 1 is chosen for heavy smoothing and low value for less smoothing. Whereas the Kalman filter has a better estimate than fixed gain filter because the Kalman gains are computed dynamically. In other words the Kalman gain matrix is determined from the state error variance as well as from measurement and noise variances adjusts adaptively. The Kalman filter computes its own state error uncertainty estimates while observing a new measurement. Also as it is mentioned [20] if the new measurement noise error variance is bigger than the state error variance, the Kalman filter will place less emphasis on the new measurement than if the state error is larger than the measurement error. On the other hand the simple fixed gain filter computes the filter gains based on assigned smoothing coefficient. This effect and the superiority of Kalman filter over fixed gain estimator can be seen in RMS errors shown in Figure 8. The performance of the fixed gain and Kalman filters serves as best linear estimators for reducing mean squared errors in LOS distance. The next section deals with the target recognition and identification using image processing techniques. From the Figure 8 and using the equation 6 the RMS errors for measured, fixed gain and Kalman filter is found to be 1.5350 m, 0.3119 m and 0.0206 m, respectively, and are shown in Table 2.

Table 2: Thermal Camera Specs

LOS RMS errors	
Filter type	RMS error
Measured	1.5350
Fixed gain filter	0.3119
Kalman filter	0.0206

Table 3: IR Digital Night Vision Camera Specs

Sensor Type	Technical Specs	
Digital Night Vision	Sensor	CMOS Sensor, Max. 24 MP
	Lens	Fixed lens, F/3.2, f=7.5mm
	Focus Range	Normal: 1m ~ infinity
	Image	Resolution: FHD, 3Mp – 24 Mp
	Video	Resolution: FHD (1920*1080) - VGA
	Night Vision range	<3.0 m
IR LED Illuminator	No. of LED	198
	IR range	300-400 feet (100-130 meters)
	Light beam angle	45-60 degree
	Wavelength	850nm

3. Sensor Specifications

3.1. Night Vision Camera

A night vision digital video camera is employed for capturing the images to post process the images in the MATLAB. The night vision camera selected for this experiment belongs to a class of image intensifier (I²) device. Also, a CMVision IR infrared

illuminator is added to the night vision camera for more illumination of the target drone. The important technical specifications [21] of the two devices are given in Table 3.

3.2. Thermal Camera

A Seek CompactXR thermal imaging camera designed for smart phone is utilized in this research work. The purpose of using a thermal camera is based on their ability to detect and track in extreme pitch black conditions. They are more sensitive to temperature variations. Designed with over 32,000 thermal pixels, the CompactXR will sense the temperature illumination and displays on the smart phone installed with the Seek thermal app. The important technical features [22] of the thermal sensor are given in Table 4.

Table 4: Thermal Camera Specs

Technical Specs	
Sensor	206*156 thermal sensor
Field of View (FoV)	20° narrow FoV
Operating distance	1,800 ft (548 m)
Detection range	-40°F to 626°F
Wavelength	Long Wave Infrared (7.2 - 13 Microns)

3.3. Target Drone and Platform Setup

A 4 channel X5C 2.4GHz remote control quadcopter is used as a target drone. The size of the drone is 16.5 x 12.2 x 3.8 inches. Although, this drone is designed with built-in 6 axis gyroscope with 360° 3D eversion and throwing flight function, the target drone was made to hang on a wall as well as tree for the sake of simplicity in simulation. A tripod mounted with different sensors shown in Figure 9 is used to capture the images of the drone at different light conditions.

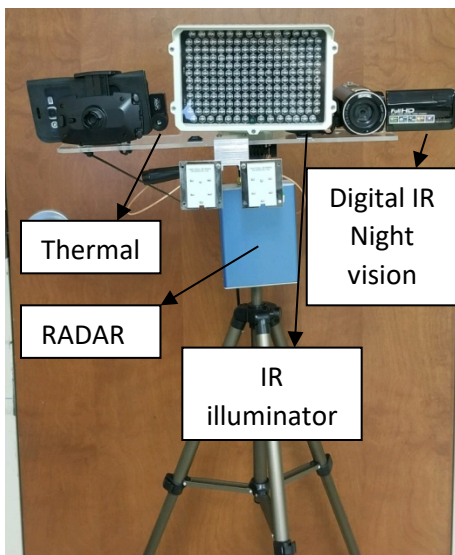


Figure 9. Sensors Tripod

To better understand the image processing techniques, two different environmental backgrounds are considered and the performance of each device is observed. The first set of images is

taken with a dark background and the second set of images is taken as the target drone blended to the outside environment.

A. First set of Images

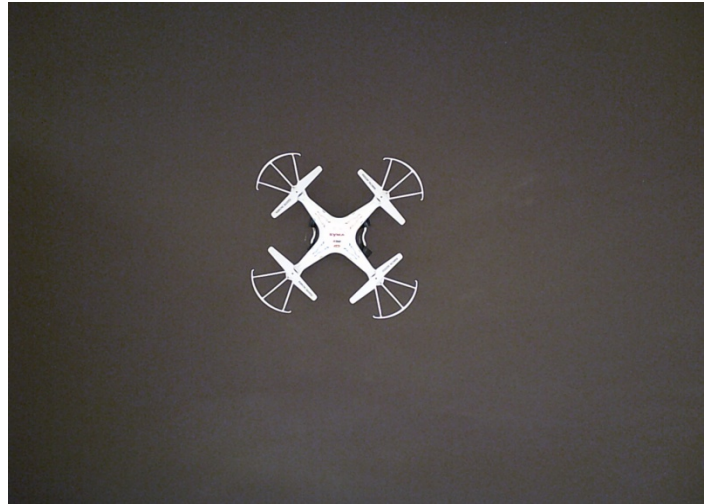


Figure 10. Test image with dark background

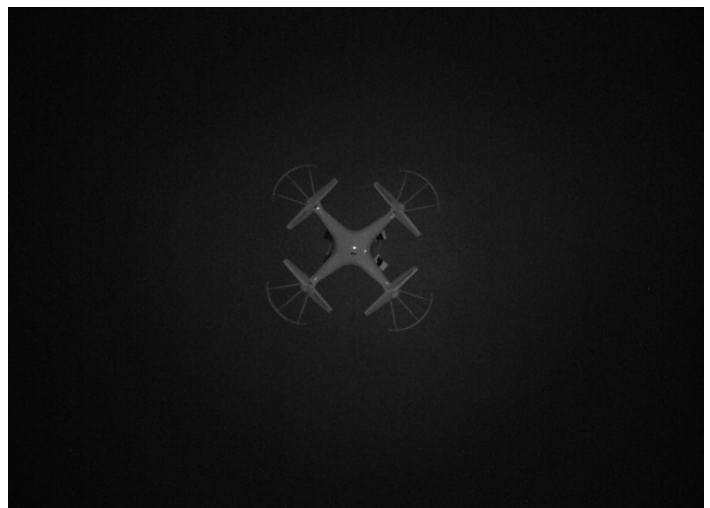


Figure 11. Test image in night vision

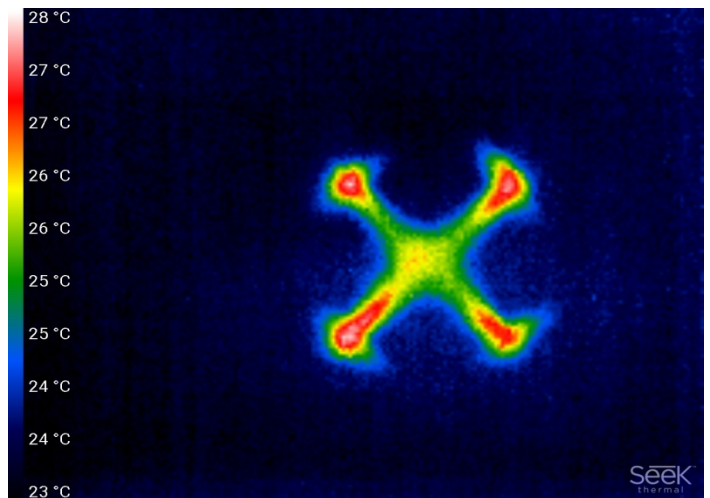


Figure 12. Thermal image

B. Second Set of Images



Figure 13. Test image when target blended to the background



Figure 14. Figure: Test image in night vision

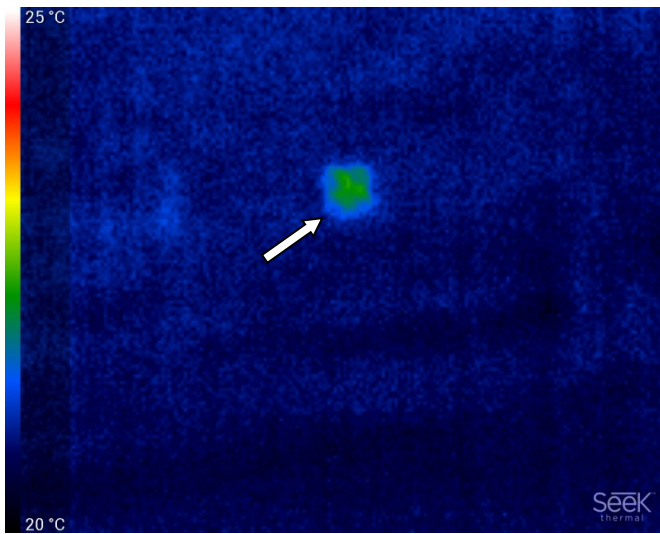


Figure 15. Thermal image

The heat dissipated by the motors of the drone can be easily seen in Figures 12 and 13. The next section deals with the numerical www.astesj.com

simulation analysis by utilizing the edge detection, image sharpening, smoothing, and contrast enhancement techniques [20].

4. Numerical Simulation Analysis

Each pixel element representing the brightness on the corresponding location in an image is an 8-bit number and can be combined with the adjacent pixels through simple weighted average. Also, the texture of the image as well as sharpness, resolution and contrast can be altered. Moreover, the importance of feature extraction using a signal processing filter known as template detector can also be applied to an image to extract the features like edges of a drone when blended into the surrounding clutter. For the regions where there is a perfect match between the template and pixel, the template detector produces high output for those regions. In the neighborhood of diversity of drones, the computer will be able to recognize one drone from the other using vision feature in which the 2D digital filter or the convolver has become the main processing tool. The core function of a 2D filter is to combine the local pixels with the weights to output the filtered image brightness. It can be expressed as

$$B'(x, y) = \sum_{i=-N}^{+N} \sum_{j=-M}^{+M} w_{i,j} B(x + i, y + j) \quad (7)$$

where, $B'(x, y)$ is the processed image brightness and $w_{i,j}$ are the weights. Based on the solid foundation of 2D FIR filters, the moving average technique has originated as a fundamental process for brightness computation that utilizes weighted average of the neighborhood pixels. The visual effects such as alteration of image focus, image contrast and sharpness can also be implemented with 2D filters. They can enhance and reduce the texture features in an image. One of the simple illustrations of a 2D FIR filtering process is pixilation or pixel decimation where all the pixels are replaced by the average of the pixels of original higher resolution image. This technique is often applied in television broadcast to obscure some offensive images and blur the selected areas.

4.1. Edge Detection

One of the wider applications of a 2D FIR filter is edge detection to enhance boundaries or the shape of an object in an image. As 2D FIR filtering being a prominent application in computer pattern recognition, determining the edges of a target in a particular image can extract important features such as its size and shape. The target considered here is a drone in the image captured by digital night vision and thermal camera. A gradient filter which is an example of an edge detector is applied in the direction of positive x-gradient. Since it is undesirable to produce an output image offset by half a pixel in the desired gradient direction, a symmetric finite-difference approximation is typically used to avoid the unwanted offsets. It is given by the equation

$$B'(x, y) = -\frac{\partial}{\partial x} B(x, y) = -[B(x + 1, y) - B(x - 1, y)] \quad (8)$$

The effect of the negative sign in front of the equation 8 is shown in Figures 16, 18 and 19. These are the observations performed by

an edge detector using a gradient with directional derivatives. The disadvantage of these directional derivatives is the presence of the inherent noise. As this method is solely based on a finite difference method, there will always be a random noise associated because of the difference computation between the two pixels. The finite difference method tends to amplify the noise and averaging the pixels tends to smooth out the image noise. However, apart from these deficiencies, the noise can be suppressed by adding more pixels in the filter. This can be accomplished by simply averaging the derivatives based on the choice of direction. If it is in x-direction, the derivatives in the adjacent rows will be averaged, else the adjacent columns in y-direction. The choice of filter weights when the gradient is in the negative x-direction are given by the equation

$$-\nabla_x = \begin{bmatrix} w_{-1,+1} & w_{0,+1} & w_{-1,+1} \\ w_{-1,0} & w_{0,0} & w_{+1,0} \\ w_{-1,-1} & w_{0,-1} & w_{+1,-1} \end{bmatrix} = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} \quad (9)$$

Equation is known as kernel operator that has an ease of rotation based on the desired gradient direction. In essence, the negative kernel operator in upward (positive) y-direction is given by equation

$$-\nabla_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \quad (10)$$

There are different 2D based gradient kernel operators available with respect to size and complexity, and finite-difference approximation based higher order filters. Any two-dimensional image data with gradient information is useful for automatic detection to identify and extract the geometric features. These directional derivative gradients applied either in the x (toward right) and y (upward) axis are essential for detecting the edges in a

particular orientation of an image. However, the complete geometric features of an image in all directions, i.e., all edges of a particular object in an image are needed for the application of this paper. Indeed, by detecting all edges simultaneously, the important features such as shape and size, as well as orientation can be determined. A widely used conventional edge detection technique known as Sobel edge detection is utilized in this paper for identifying the target object which is the drone in the test images. A Sobel edge detector first calculates the spatial derivatives in each x-and-y directions. Thereby, it sums their squares and computes the square root. This complex process can be avoided by Kirsh operator which is very less complex in nature. Instead of computing the square root of the sum, the Kirsh operator estimates the neighborhood gradients and calculates the maximum absolute value to produce an edge detected output. Figures 16, 17 and 18 show the application of Sobel edge detection combined with Kirsh operator approximation depicting the significance of edge detection for feature extraction. Moreover, apart from the various features of these images that are useful for pattern recognition analysis, these edge detection techniques are also applicable to enhance the image visual quality.

4.2. Image Enhancement

Typically, these edge detectors are considered as high-pass filters through which the random changes in spatial brightness could be seen. On the other aspect, the sharpness of an image to its highest acuity can be enhanced by amplifying the higher frequencies and attenuating the lower frequencies. A rotational invariant Laplacian operator is applied for the test images shown in Figure 10, 11, 13 and 14 to perform this operation and is given by

$$-\nabla^2 B = \frac{\partial^2 B}{\partial x^2} + \frac{\partial^2 B}{\partial y^2} \quad (11)$$

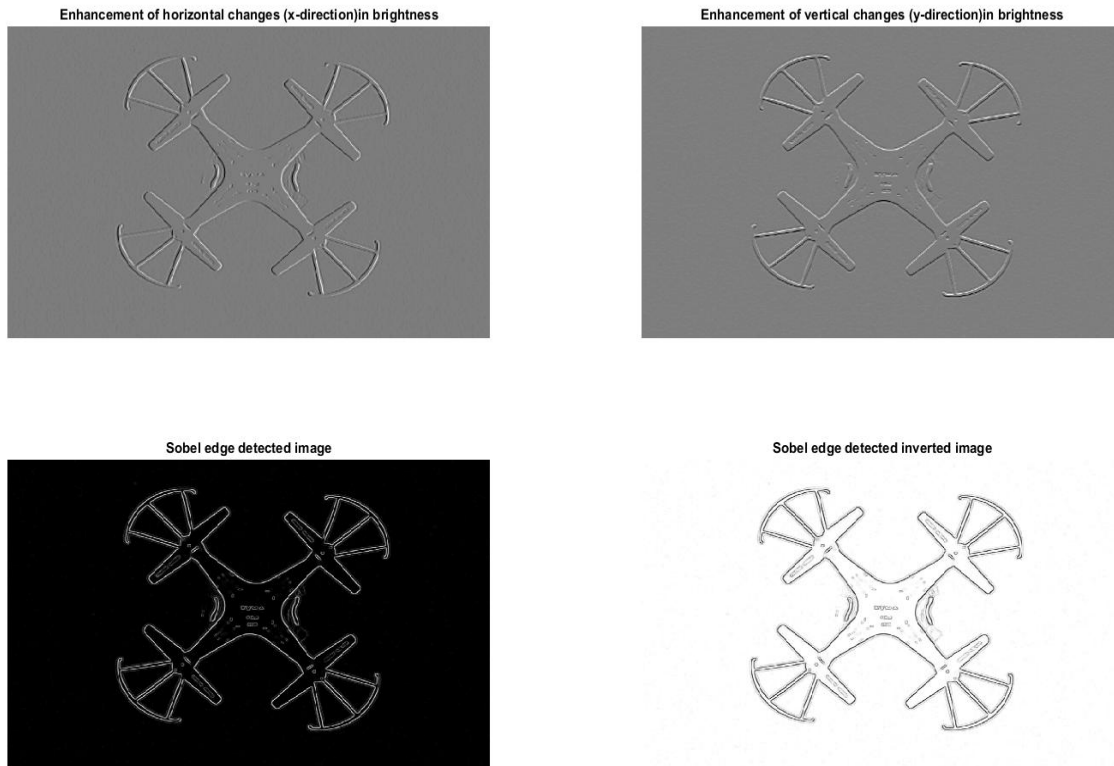


Figure 16. Enhancement and edge detection of optical image

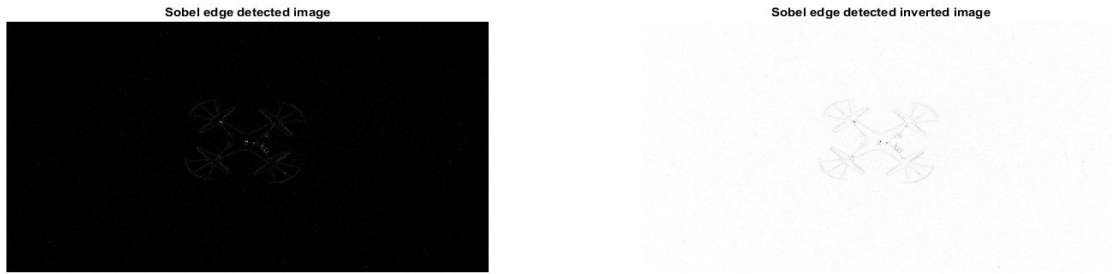


Figure 17. Edge detection of night vision image

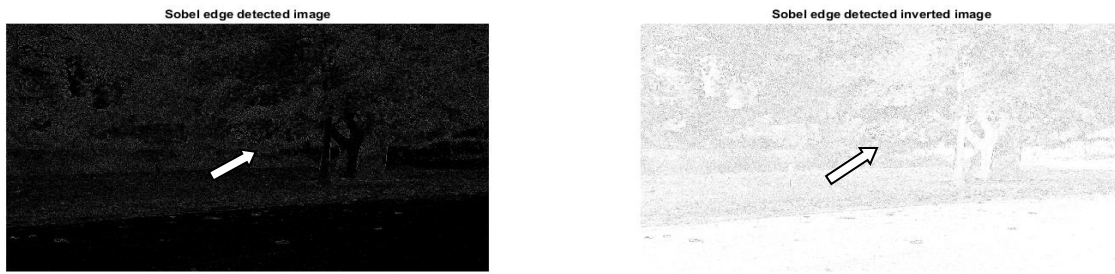


Figure 18. Edge detection of optical image



Figure 19. Edge detection of night vision image

The Laplacian operator employs finite difference method in summing up the x-direction and negative (second) derivatives in y-direction expressed as

$$-\nabla^2 B = \begin{bmatrix} 0 & -1 & 0 \\ -1 & +4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad (12)$$

In order to reduce noise by considering all eight neighboring pixels as is the case in the edge detector, the addition in the diagonal components are done to negative Laplacian given by

$$-\nabla^2 B = \begin{bmatrix} -1 & -1 & -1 \\ -1 & +8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \quad (13)$$

Based on the finite differences, equations 11, 12 and 13 are known as normalized operators for approximating a negative Laplacian operator. The normalization was done in such a way that the average brightness in the image is not shifted. The advantage with the Laplacian operator is that it will cancel the varying brightness however, it makes the image noisier as the random variations are amplified by the operator from pixel to pixel. Moreover, the edge enhancement and the image sharpening are done by adding the output image of the negative Laplacian operator to its original image. Although, it adds noise, the high frequency spatial components are enhanced by image sharpening filter, though the

filtered image doesn't look like the original image. Hence, in this case an unsharp filter can be utilized to improve the sharpness of an original image by adding or subtracting some percentage to the original image.

4.3. Image Sharpening

The digital un-sharp filter is done either by adding or subtracting the Laplacian filtered image to the original or by subtracting the low-pass-filtered image from the original. The image which is added or subtracted to the original in place of the Laplacian image is considered as the high frequency image. Another less noisy approach is the Laplacian of Gaussian (LOG) filter in which first the image is low pass filtered to remove noises and Laplacian is applied. In essence, as the 2D convolution operation of the filter is associative, the LOG smoothing filter result can be used as the 2D FIR filter. The LOG function is given by

$$\tilde{N}^2 G = \frac{-1}{\pi\sigma^4} \left(1 - \frac{x^2+y^2}{2\sigma^2}\right) e^{-\frac{(x^2+y^2)}{(2\sigma^2)}} \quad (14)$$

where, x and y can vary from -1 to +1 pixel. Assuming $\sigma=0.66$, the unsharp operator is given by

$$\nabla^2 G = \begin{bmatrix} -1 & -1 & -1 \\ -1 & +8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \quad (15)$$

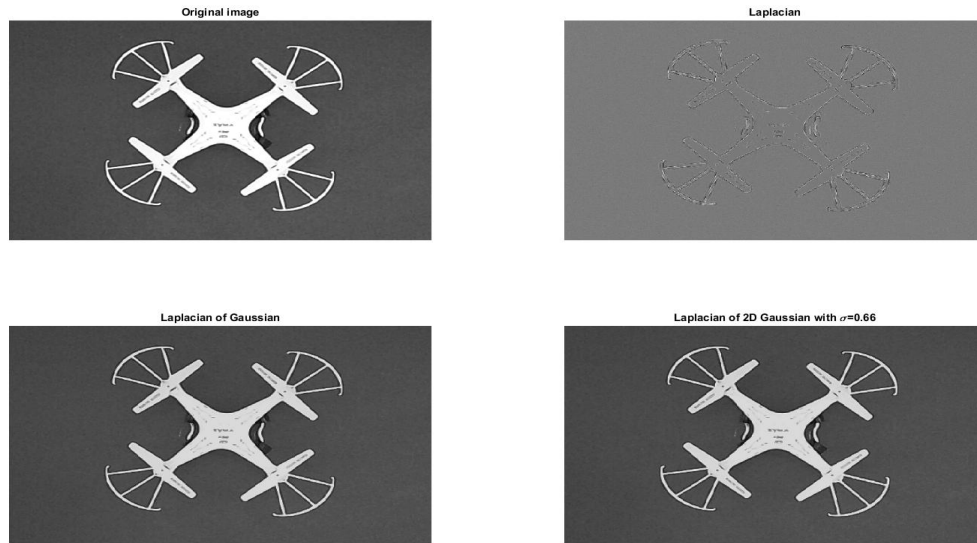


Figure 20. Image enhancement of optical image

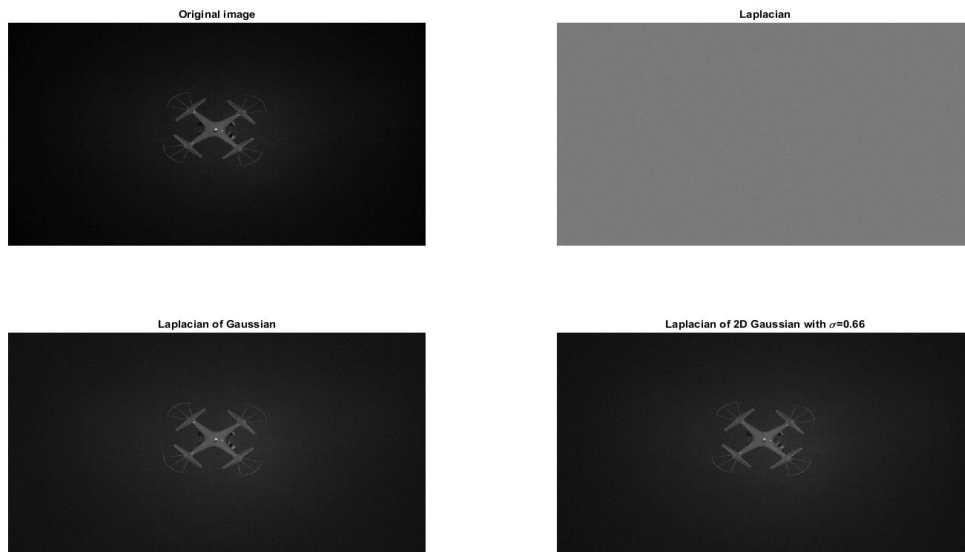


Figure 21. Image enhancement of night vision image

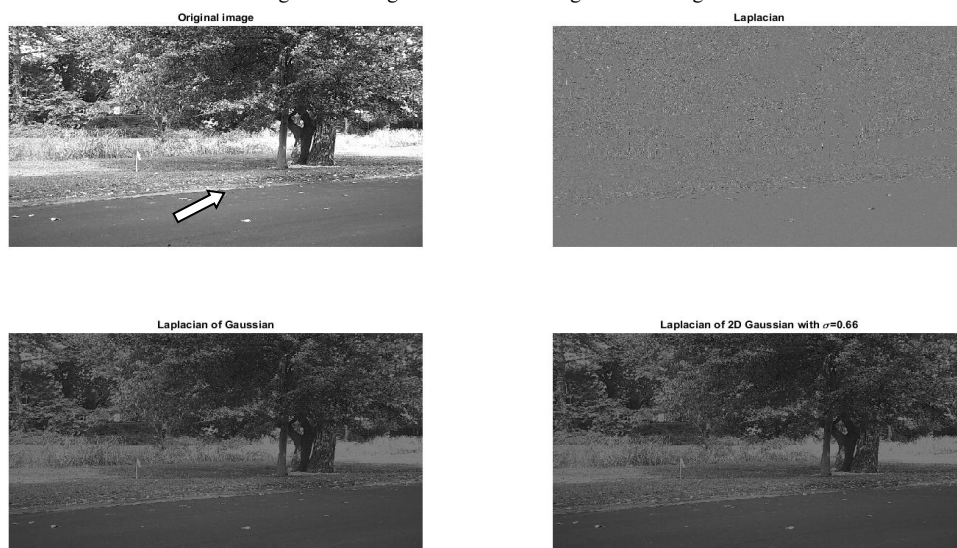


Figure 22. Image enhancement of optical image

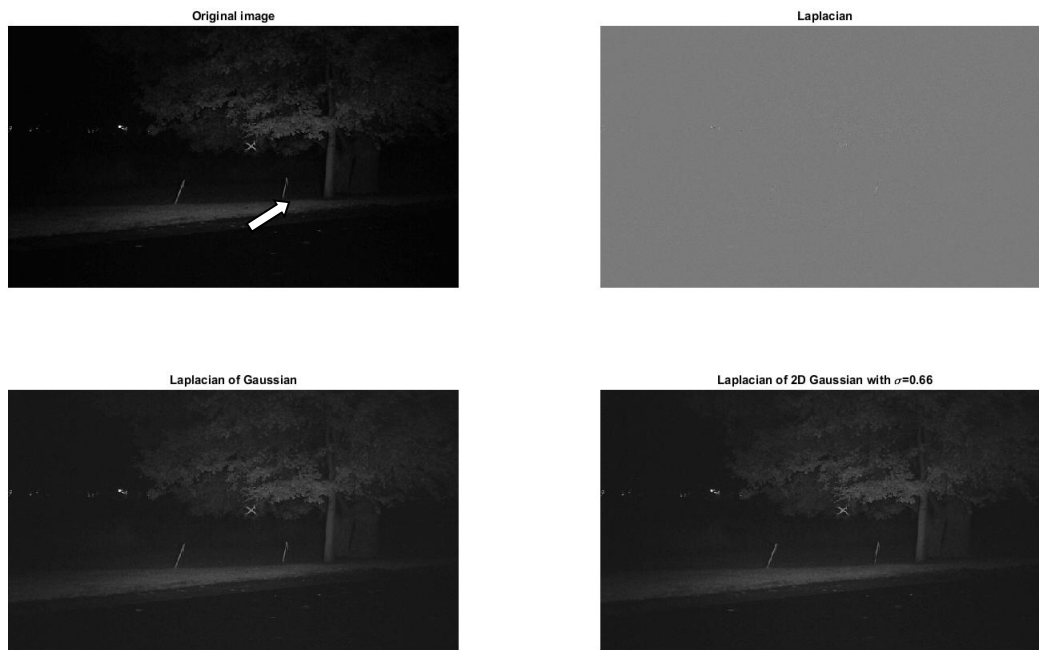


Figure 23. Image enhancement of night vision image

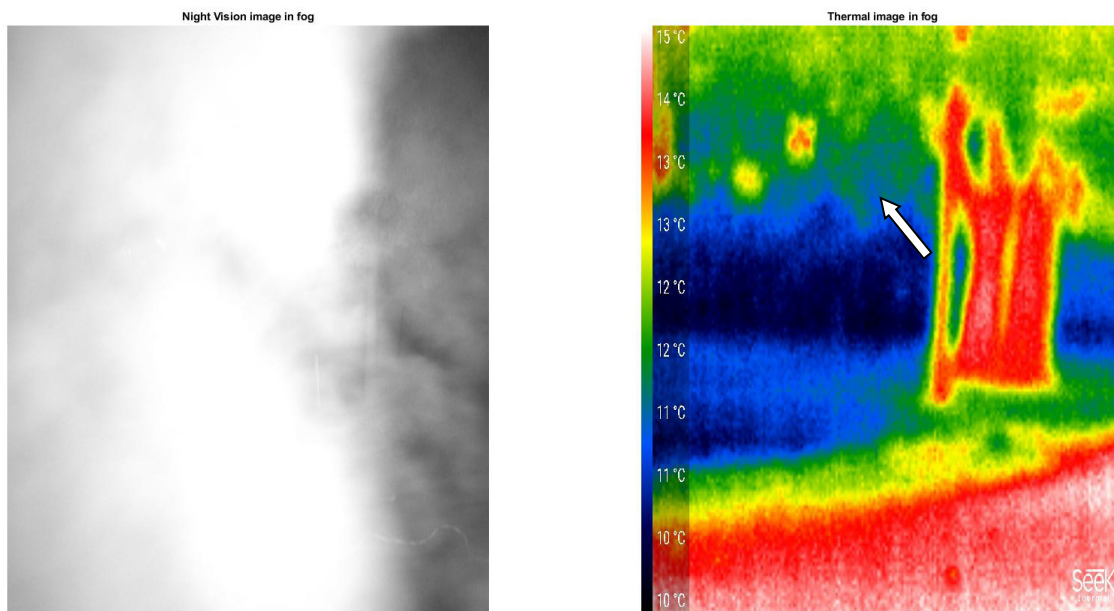


Figure 24. Comparison of night vision and thermal vision in dense fog

Conclusion

This paper deals with the advancement of echolocation bionics and its importance for analyzing and preying upon the target. The mathematical modeling and simulation of bat and its prey is developed upon which the tracking of LOS distance using different state estimators is presented. The RMS error shows the superiority of state estimators for better reconstruction of the signal with less residual. Also, the second half of the paper presents the image processing based techniques to identify the target drone in different places at different environmental situations using multiple sensors. Figure 24 demonstrates the advantages of thermal sensor over digital IR night vision cameras in the midst of fog showing that the thermal sensors can detect the heat signatures, whereas, night vision cannot penetrate through fog.

References

- [1] Bassem R. Mahafza, Radar Systems Analysis and Design Using MATLAB, Chapman & Hall/CRC, 2016.
- [2] Ba-Ngu Vo, Mahendra Mallick, Yaakov Bar-Shalom, Stefano Coraluppi, Richard Osborne, III, Ronald Mahler, Ba-Tuong Vo, "Multitarget Tracking", Wiley Encyclopedia of Electrical and Electronics Engineering, Wiley, Sept. 2015.
- [3] Kursad Yildiz, "Electronic Attack and Sensor Fusion Techniques for Boost-Phase Defense against Multiple Ballistic Threat Missiles", Master Thesis, Naval Postgraduate School Monterey, 2005.
- [4] CommandSpaceCerberus Datasheet, www.FLIR.com, FLIR systems Inc., 2015
- [5] Sergio Miranda, Chris Baker, and Karl Woodbridge, "Knowledge-based resource management for multifunction radar", IEEE Signal Processing Magazine, 23(1), 66-76, 2006.
- [6] Joseph R. Guerci, Next Generation Intelligent Radar. 2007 IEEE Radar Conference, 2007, 24(1):7-10.

- [7] Ji Zhang, Yu Liu, "Single Maneuvering Target Tracking in Clutter Based on Multiple Model Algorithm with Gaussian Mixture Reduction", Journal of Applied Research and Technology, Vol. 11, October 2013.
- [8] David Kondru, Mohan Krishna., "Kalman Filter based Target Tracking for Track While Scan Data Processing", 2nd IEEE International Conference on Electronics and Communication Systems (ICECS), 2015, Pages: 878 – 883.
- [9] Binbin Cheng, Research on Bionic Processing for Auto-Adaptive Radar Waveform, Doctor Thesis, Tsinghua University, 2009.
- [10] Cheng Bin-bin, Zhang Hai, Zhang Xiaoping and Li Hesheng, "Bats' Acoustic Detection System and Echolocation Bionics", 2012 IEEE Radar Conference, 08 June 2012, Atlanta, GA, USA.
- [11] Corcoran AJ, Wagner RD, Conner WE (2013) Optimal Predator Risk Assessment by the Sonar-Jamming Arctiine Moth *Bertholdiatrigona*. *PLoS ONE* 8(5):e63609. doi: 10.1371/journal.pone.0063609.
- [12] Paul Zarchan., Tactical and Strategic Missile Guidance, AIAA, Vol.239, 2012.
- [13] Andeggs, "3D Spherical Coordinates", https://en.wikipedia.org/wiki/Spherical_coordinate_system#/media/File:3D_Spherical.svg, 4th Aug 2009.
- [14] Mohinder S. Grewal and Angus P. Andrews, Kalman filtering theory and practice using MATLAB, 2nd Edition, 2001 Wiley & Sons, Inc
- [15] Robert M. Rogers, Applied Mathematics in Integrated Navigation Systems, AIAA Education Series, July 2000.
- [16] Weisstein, Eric W. "Helix." Math World-A Wolfram Web Resource. <http://mathworld.wolfram.com/Helix.html>
- [17] Nahum Shimkin, Estimation and Identification in Dynamical Systems, Lecture Notes, Fall 2009, Department of Electrical Engineering, Technion-Israel Institute of Technology
- [18] Pierre Minvielle, "Decades of Improvements in Re-entry Ballistic Vehicle Tracking", IEEE Aerospace and Electronic Systems Magazine, Volume 20, Number 8, August 2005.
- [19] David Kondru, Mohan Krishna., "Kalman Filter based Target Tracking for Track While Scan Data Processing", 2nd IEEE International Conference on Electronics and Communication Systems (ICECS), 2015 Pages: 878 – 883.
- [20] David C. Swanson, Signal Processing for Intelligent Sensor Systems with MATLAB®, pp. 95-101, Second Edition, CRC Press, 2011 Taylor and Francis Group.
- [21] User's Manual, Digital Camcorder with IR Night Vision, WEILIANTE Co., Ltd.
- [22] Compact Series User Manual, Seek Thermal, https://www.thermal.com/uploads/1/0/1/3/101388544/compact_series_user_manual-web.pdf

Web Authentication: no Password; Listen and Touch

Viorel Lupu*

Research & Development, Online Services SRL, Buzău, 120191, Romania

ARTICLE INFO

Article history:

Received: 15 December, 2018

Accepted: 22 January, 2019

Online : 26 January, 2019

Keywords:

Web

Security

Authentication

Multi-Factor

Image-Password

Voice notification

Mobile phone

Password

Disabilities

Smart-Headphones

ABSTRACT

Just as electricity has an essential role in our lives, the internet network and especially web services have become of vital importance nowadays. Without security service layers, apparently small things like checking a child's school schedule on web may turn the daily routine into a nightmare. Web services users are still required to use many combinations of usernames and passwords. Despite technologically advances that bring many benefits to those owning top of range smartphones, complex combinations of identifiers and passwords are still required for basic security. Top tier smart mobile phones also add device-specific vulnerabilities to the risk of misuse or may expose sensitive data like biometrics. To meet users' expectations, authentication systems must be safe, fast, efficient, intuitive and easy to use, especially on mobile phones. User satisfaction, reduced fraudulent authentication issues, increased security, reduced management costs, regulatory compliance are main goals for the new advanced web technologies systems. This paper presents some real-time multi-factor authentication methods that uses voice calls to communicate random passwords to registered users. The ultimate goal is to relieve web service users from the stress of memorizing complex combinations, or copying text strings for user identifiers and passwords from paper or external memory devices like mobile phones. The new features are presented for a web service after migration from the traditional authentication system to the one with the proposed new method. This work is an extension of the paper entitled "Securing Web Accounts by Graphical Password and Voice Notification" published in 2018 IEEE International Conference on Engineering, Technology and Innovation (ICEITMC).

1. Introduction

Any authentication service must be designed, built and operated on a basis of user centric vision [1]. These design standards, associated with data protection requirements [2, 3] may imply costs, time, technologies and human resources that are beyond reach for many small companies which provide simple and useful web sites like school timetables. Users must obey the site's rules and the method of authentication. They must use given usernames and long, maybe complex or random passwords. There are several options, such as saving the login data in the browser or writing them elsewhere. If the web service is provided by an authority such as a bank or a governmental agency, the importance of a proper user access goes higher as risks of identity frauds are also high [3]. Every website that is part of our digital life, which requires and maintains personal data is equally important; without adequate security measures, any of these websites can be a weak bastion after which we might become cyber-victims.

In terms of security, when talking about Internet web services, appropriate means of authentication must be provided so that only authorized users have access to them. At the same time, Internet web services have to be easily accessible. New vulnerabilities or changes of standard requirements may occur during operations affecting the security environment completely.

Web authentication is a process that relies most on human-computer interaction. The authentication system directly influences the quality of Internet web services, in terms of usability. There is a tradeoff between the complexity of the authentication system and the service usability [4, 5].

Internet web services based on the usage of sensitive data, require high quality security systems, often a combination of SSL protocol enabled connections and multi-factor authentication systems. Low-cost single-factor authentication systems based on text strings must be reconsidered. They are no longer feasible neither for user identity nor for passwords. The RFC 2196

*Viorel Lupu, +40744545600, viorel.lupu@onlineservices.ro

recommendations have posed new restrictions on the user, as more vulnerabilities have been revealed [4, 5, 6 and 7].

On the other hand, it is difficult for the human user to remember long or complex identifiers and passwords, not to mention in stressful environments or in physically challenged contexts, in which such authentication methods are proved to be inappropriate. As a result, human user often resorts to storing identifiers and passwords in browsers memories or in easily accessible, visible files. They tend to use predictable graphic passwords and, sometimes, the same identifiers and passwords are in use for multiple websites [4, 6].

Also, we would argue that text-based authentication systems are not suitable for smartphones. Often smartphone users find it difficult to remember, read and type text during the authentication process, such as for example a simple 'copy-paste' action of a unique text passcode received by SMS, an operation which may prove to be a challenging task on a mobile phone.

Multifactor authentication is considered the best practice for authentication [4] today. Users have to provide more and more information to the authentication system in order to authenticate successfully. Authentication standards like Universal Authentication Framework (UAF) or Universal Second Factor Authentication (U2F) proposed by Fast Identity Online Alliance (FIDO) are based on latest knowledge about public key cryptography [4, 5], biometrics, mobile devices, server technologies, best practices in the field etc. in order to provide an open and public accessible framework [8].

Internet services authentication systems have to assure that secure protocols run properly: digital certificates need to be valid [5], software updates need to be at the latest version etc. System time is also very important. Web servers keep time in different ways and more or less accurate. Synchronizing the web server local time with an external time server over the Internet it is a standard practice nowadays. If the time source is manipulated or not updated, everything may go astray mostly because digital certificates, the basis for secure web communications, contain time references. Cloudflare deploys a new authenticated time service called *Roughtime*, in an effort to secure certain timekeeping services. The publicly available service is based on an open-source project of the same name that was started by Google [9, 10].

Authentication failures result in inestimable damage and there are plenty of examples: Sony Pictures Entertainment corporation (announced on December the 3rd 2014), Yahoo (on December the 15th 2016, where 3 billion accounts were compromised in a series of incidents), Equifax (on September the 7th 2017, where 148 million data sets with personal data were stolen, including social security number, with a recovery cost of \$400 million dollars) or for Marriott International (announced on November the 30th 2018, where about 500 million guests, including full names, date of birth, passport information, payments, preferences etc.). [11]. It is estimated that a third of U.S. businesses have had a customer information breached in 2017, including the information needed to authenticate their customers [12]. All these events might create a spiral of more and more security incidents that would make use of this data.

Authentication failure has new dimensions such as social engineering, mobile phone thefts and most important, revealed by www.astesj.com

recent trends, SIM-Swap or cloning the GSM Subscriber Identity Module (SIM) card. Cybercriminals look for future victims using social media websites. On such websites, it seems easy to find personal details of the victim (e.g. the date of birth, the e-mail address, the mobile phone number, etc.). Finally, it is not so hard to convince the GSM operator to make an emergency transfer of the GSM number to a new SIM (in the criminal hands). Then the authentication system works as expected. But not in favor of real users [13, 14], as the real user just discovers that his/her mobile phone is dead. From a technical point of view these fraudulent actions seem beyond the scope of the authentication process.

The security system should not rely entirely on the security components of the operating system. Recent security vulnerabilities (the case of Meltdown and Spectre) leak data as encryption keys, passwords [15], security identifiers, images etc., data that should be protected by the operating system. Instead, it is disclosed without user consent. The process of solving those vulnerabilities is a long one, as it will have to deal with hardware architecture redesign [15].

Nowadays, security is a serious concern for the entire society, being the result of technology, people and policies working together. In this sense, the new GDPR have been imposed in the EU since May the 25th 2018 [2]. In Europe, "companies must alert government authorities within 72 hours of a known breach and may be fined up to 4 percent of their global revenue under data protection laws" [17].

This study proposes a different solution for real-time, multi-factor authentication: image selection [17, 18 and 19] guided by voice via phone call notifications [20, 21]. This paper presents a detailed analysis, results, related works and comments on the implementation decisions, as well as on user reactions. Our aim is to improve the authentication process with better user experience and comfort. Also, we intend to extend its usability in high stress conditions (public hospitals, police and justice departments, public administration) and for the elderly or people with disabilities.

2. Background

Forms are commonly and efficiently filed with the use of secure web forms, be it small or medium size enterprises (SMEs), corporations or government entities. In the situation analyzed in this paper, the web forms in use run on a national authority's centralized web services application due to the need for real-time sensitive medical records such as those used in the field of human organ transplants. The operators reporting the data, taking into account the strict medical specialization and the legal implications, are medics accredited for the respective field. Each medical unit enrolled in the national transplant system has at least one accredited medic on staff. During the analyzed period, there were over 60 accredited medics in the program. Reports were filed on a daily basis, but also following certain unforeseen events.

The secure web forms are filled in using computer terminals placed in the operative staff common rooms of the respective hospitals, as terminals used for patient registration. Terminal access is permitted to all the medical staff in the hospital. Thus, the transplant database system is potentially at risk of exposure to unauthorized personnel. The credentials were disseminated to high school level medical personnel, which were operating data. Many times, the Internet address, user name and password were written

on post-it note stuck on the terminal' screens, further allowing access for curious individuals to presumably protected data. Even though such post-it's were not used in all locations, there were notebooks or unprotected files containing these credentials.

The described situation does not represent an exception; it is often met in other environments, where the legal implications may be less serious. There were also security conscious operators who used their private computer terminals or mobile devices for these reports. In the effort to limit the data exposure, several technologies and methodologies were evaluated, being clear that the use of text strings as passwords is difficult to use on mobile devices [4, 5 and 22].

An attractive and convenient alternative is the use of personal mobile devices in comparison to investing in new equipment. Nowadays, the personal mobile telephone is ubiquitous, being a smart device, permanently connected to the internet and provided with an array of sensors used for increased security. High resolution video cameras, near field communication (NFC) readers, fingerprint readers, geo-location (GPS) receivers are already embedded in current smart phones, open the opportunity to use multi-factor security solutions. Starting with 2009, the smart phone market absorbed more than 173 million units, out of which almost 2 million have an embedded NFC reader (www.statista.com). NFC terminals have the ability to read NFC chips embedded in many types of supports such as implants, ID cards, labels and rings.

The NFC technology embedded into mobile devices, based on the Near Field Communication Data Exchange Format (NDEF) allows the automatic reading of a secure web site address and users identifiers. According to the NDEF specification, a URI record for the secure website using https://domain/application/UserId format [20] allows the user to open automatically the web application by approaching the NFC support to the smart mobile device. Thus, increasing security, user comfort, though the potential problems could be even greater in case these NFC supports are lost or stolen.

We have also analyzed the possibility of sending one-time, unique random passwords to authorized users, based on preregistered personal mobile phone numbers. The option of sending the password through Short Message System (SMS) is almost intuitive. Banks use this system for two-factor authentication for several decades. In 2005 it was demonstrated that the use of this system can be compromised since the SMS as a communication technology is quite vulnerable [19, 22]. This vulnerability is inherited from the telephony signaling system, also called SS7, developed in 1975 and in use ever since. Due to this, the SMS is no longer a recommended communication channel in authentication systems.

Cybercriminals rapidly penetrate weak security policies on mobile devices using a combination between social engineering, Internet available advanced software technologies and the use of sniffers for text strings sent on clear text communication channels. Their target is to clone SIM card in order to take over the victim's mobile telephone number, build a replica of the security environment on a malicious device for later use in hijacking all victim's Internet accounts where the phone number is registered for multifactor authentication (e.g. bank account). Methods that were unimaginable several years ago are now available to anyone who wants to dabble with them [13, 14].

Images or sounds can be an alternative to text passwords. Image-based authentication seems to be promising especially due www.astesj.com

to the fact that the human mind retains images and image associated actions better in comparison to written text strings [4, 6, 7, 17 and 23]. The replacement of the text password with a graphic one does not increase the level of security of the authentication method, being equivalent to the use of a four-digit PIN number [23].

The Interactive Voice Response (IVR) systems have been around in call centers since the 1970s and are presently used to transmit voice codes. For example, Microsoft runs an international system for software license activation using a similar IVR technology that receives and transmits the activation codes to be typed by the user.

The system analyzed in this paper implemented the use of a graphic password, which is communicated through the telephone network.

3. A new method for web authentication

The suggested authentication method aims to increase system security and user satisfaction. Taking into consideration that users easily recall images [4, 6 and 18] and image-related actions, the proposed authentication system displays images and processes user actions (as screen touch or clicks). The images are selected to be meaningful to the users. With voice guidance through a telephone call, the users have to choose between them. Processing the user's actions, the authentication system assembles an indicator and compares it with the random one-time password constructed into voice indications.

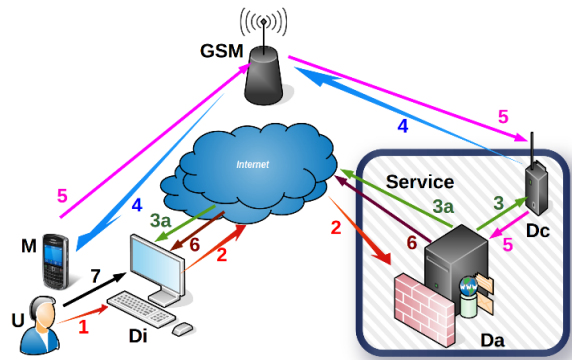


Figure 1: Web authentication system

The proposed authentication method (Figure 1) has two main phases. The first phase consists of the user's identification and starts the authentication session. The second randomly generates the password, initiates a phone call and authenticates the user. If the authentication system receives a known user identity, then the second phase is automatically started.

The first phase is of a special importance, derived from the fact that there are a lot of devices capable of initiating secure web sessions, from simple, classic personal computers to the latest mobile smart-devices with biometrics. The classic old-style computer browser provides options to save website address and user identifiers. Specially crafted computers like laptops may be equipped with additional specialized hardware to read fingerprints, NFC cards or to process images and sounds. With latest

smartphones, user identifiers may be stored and better protected e.g. by fingertips and encrypted file systems. The authentication system has to detect from the authentication request who is the potential user and from what device the authentication request is made. If that device is already registered and all secure attributes are up to date, then the authentication procedure goes forward.

The authentication process is carried out following several steps [21] as illustrates in Figure 1:

1. A secure web session (2) is established with the server system after the user's (U) action (1) upon the Internet terminal (Di). The user's action (1) is carried out either by typing the URI (Da), or by using the device's video camera visual to recognize a QR code, or entering proximity of the terminal with an NFC ID card issued by the organization running the web service, if the respective terminal has an embedded NFC reader.

2. When the SSL session is opened (2), it provides a number of details about the Internet terminal (Di) that initiated the session and about the particular user (U). If those details are read off the NFC card's chip or the browser's memory, then the system automatically continues to the next step. As such, the user has to provide an identifier related to a user recognized by the web system (Da) and a password to confirm the authentication request. The request confirmation password can be in the form of a text string, graphic image set or an audible sound, depending on the particular security policy.

3. If the secure web session is initiated with the user's identifier and a known confirmation password, the web service generates a unique one-time random password and directs (3) the telephone call initiation device (Dc) to establish a connection with the user's (U) telephone number (4) on record. Simultaneously with the telephone connection, the web system (Da) will securely update (3a) the display on the requesting terminal (Di).

4. The communication device (Dc) initiates the telephone call using the public telecom network (4) (i.e. GSM). The web service (Da) monitors the status of the telephone call (5) using the communication device (Dc). The system waits for the call to be answered before starting dictation. The telecom network can also transmit other events regarding the status of the call such as if the line is busy or the mobile phone is outside the coverage area etc. Once the user answers, the communication device (Dc) will play the voice indications appropriate for the respective unique password.

5. The user (U) answers the call or not. Events detected by the communication device (Dc) in the public telecom network are sent in real-time (5) to the web system (Da). The web system will close the web session if the telephone call fails for any reason (i.e. line is busy).

6. The web system (Da) securely communicates the status of the call with the user's terminal. Thus, when the user answers the call (5), the terminal receives and displays the set of images (6). Figure 2 is an example of a set of images displayed on the user's terminal.



Figure 2: Example of a set of images for one-time-password

5. Following the indications received in the call (5), the user selects (7) the relevant ones amongst the other images (Fig. 2)

displayed on the internet terminal, thus communicating the unique password to the web servers. The system validates the password and authenticates the user if during the call the latter has selected the appropriate images in the instructed order. Afterwards, the authenticated user may continue the web session. Any deviation from these restrictions leads to the closing of the web session by the web service.

The aforementioned web authentication method has the following advantages:

The password used to start the whole process (step 1) does not directly authenticate the user. It is merely a password confirming the user's intent to authenticate into the system. If unauthorized individuals use the respective password, the process cannot be continued since they do not possess the authorized user's mobile phone terminal. The initiation of every web session generates a telephone call and in consequence, the real user receives unsolicited calls, knowing that his account is being abused. The real user can take steps and change the confirmation password. In case of malicious authentication requests, the system eliminates abuse before an imposter can access the system.

Using NFC ID cards, rings or implants, the user is relieved of the need to memorize identifiers or confirmation passwords, which can be constructed of very long strings that are practically impossible to memorize, leading to increased system security.

In the absence of NFC support on protected access devices (i.e. mobile telephone terminals with biometric sensors), the identification strings may be stored in the browser's memory or in password manager type applications. The terminal's web browser (Di) may also save the Internet address and credentials of the secure website. The website address may be displayed in the form of an icon on the display of the Internet terminal, as to ease the user's task to memorize such details. The secure web session once initiated using the icon can also provide the user's identifier and confirmation password.

If public access web terminals are used, the confirmation password can be image-based thus avoiding the possibility that the respective terminals may memorize the confirmation password in the browser or password manager applications. Still, this type of password may be visually exposed to people in close proximity or to video surveillance systems.

The images used by the user to build the password based on the telephone call instructions are displayed on the internet terminal in the step 6, only if the user answers the call. This workflow reduces the exposure of the images to unauthorized users, increasing the security of the web service.

The telephone calls can be processed by the operating system (OS) of the smartphone. Thus, we can apply policies based on time periods or in correlation to other data such as geo-fencing (GPS position), GSM network name, Wi-Fi network SSID etc. as to semi-automatically reject unwanted calls.

The communication device (Dc) can produce the sounds necessary for voice calls through synthesis (text to speech) or by combining pre-recorded sounds. The pre-recorded sounds can be the result of professional recordings. Moreover, by mixing voice with background noise, the resulting sounds can be recognized by the users as authentic web service instructions.

This authentication method is terminal OS (operating system) and web browser independent (e.g. Linux, Windows, Android,

IOS with Chrome, Firefox, Edge, Opera etc.). Suggested authentication method does not rely entirely on the operating system security components. It is crucial that the Secure Socket Layer (SSL) based on digital certificates etc. to be fully functional.

The implemented and studied system is being improved continuously. The public access terminals available to all operative hospital staff have been fitted with specialized USB connected NFC readers. With these readers and NFC ID cards, the authorized medical staff have quicker access to secured web forms used to report sensitive data, at the same time being less exposed to different security risks such as phishing attacks. As a result, the system became safer, more comfortable and much more productive.

4. Findings

The proposed web authentication service has been implemented [20] and runs continuously since December 2016 to serve the central transplant agency and authorized hospitals across one European Union member state. Beginning with September 2017 the usage of the new authentication system has become mandatory. Every authorized hospital is represented by an accredited specialist who daily reports medical events and data to the national authority, 24/7.

Web service usage data were collected continuously in log files. Data collected for this analysis is based on the regular system usage by 60 authorized operators in the timeframe between September 2017 and April 2018. In this time span, the number of operators increased. The analysis in this paper is based on all the web traffic for 20.000 successful authenticated sessions.

The proposed authentication method was implemented on short notice and without staff training, taking into account that it was virtually impossible to stop activities in hospitals only for this specific training. Thus, each operator has discovered the new authentication method when a new medical event or new data had to be reported to the central entity or when, in some cases, a lot of unsolicited phone calls from the web service (a phone-call for every authentication request) have been received. These unsolicited phone calls forced the specialist to change his/her password which as a consequence of the new authentication method become a confirmation password. Many specialists have learned to use private web sessions and not to store the user name and password in Internet browsers when using computers available to all the department staff. Subsequently, they received NFC ID cards and NFC readers. The Internet web service was re-engineered for a special non-public sign-in web page which started the secure web session and the authentication process by automatically reading the NFC link. This administrative procedure reduces web traffic in general, but did not affect the selected sessions for this analysis, because the use of the NFC ID card automatically generated a secure web session for a known user identifier (if it is still valid) to start the phone call process.

For this analysis all web site logs were saved and processed. The web server logs any web request, writing in the logs the client's IP address, protocol and the time moment when the web request was made, as well as the files that were requested on the server, the details of the browser, etc. [28]. Telephone communication logs contains telephone numbers, call events and timestamps. In this analysis all this data is correlated from web site

logs, telephone communication logs and data from the web site database to filter user authentication requests and necessary details about the authentication process.

Web server logs contain a large number of unauthorized sessions because of the web services' public exposure. The current analysis does not take into account all unauthorized sessions. For any user, the authentication process model starts with the first GET (a method from the HTTP protocol [29]) web request for the web sign-in page retrieved from the web server log. These GET web requests are filtered for those requests that send valid user identifiers, accepted confirmation passwords and are continued with phone calls.

Then data is filtered and consolidated by web sessions having a unique client set of IP address, user identifier, session timeframe and phone call. In this stage, the user identity and telephone number are associated from the central database. Telephone calls log complete the analyzed authentication process model with call events. All sessions with an unsuccessful phone call are discarded i.e. sessions when the phone calls ended with busy signals, rejected calls, dial error, busy network or not answered at all, including phone calls with out of network coverage signals or if the user hanged-up before the end of the voice message. Then, searching the web server logs for session authentication failure or success events all necessary information is collected and processed. Now the entire authentication process is modelled. Resulted authentication process models are then ordered by user and timestamps. Figure 3 illustrates sample data processed by this model for one user (where real user identifier was obfuscated with *UserName* for security reasons).

As illustrated in figure 3, the user's authentication activity pattern reveals the fact that this particular user has tried to authenticate many times without success, especially because he/she was too slow to select the indicated images within the required amount of time. The first three lines of data reveal the fact that he/she has tried to authenticate three times on the 29th of September 2017 starting from 9:37 AM with no authentication success. Then, there is a new failed try after nearly two hours, same day at 11:17 AM. The next day, on the 30th of September the user has tried to authenticate two times from 9:36 AM and succeed the authentication one hour later, at 10:32 AM. On the 2nd of October, two days later, the user was able to succeed from the first try, at 9:58 AM. The IP address reveals that, starting with the 2nd of October, the web sessions are initiated from another network device.

Moment	Userid	IP	Call
29.09.2017 09:37	UserName	213.233.88.223	Notified
29.09.2017 09:38	UserName	213.233.88.223	Notified
29.09.2017 09:38	UserName	213.233.88.223	Notified
29.09.2017 11:17	UserName	213.233.88.223	Notified
30.09.2017 09:36	UserName	213.233.88.223	Notified
30.09.2017 10:32	UserName	213.233.88.223	Notified
30.09.2017 10:32	UserName	213.233.88.223	Ok
02.10.2017 09:58	UserName	86.124.60.4	Notified
02.10.2017 09:59	UserName	86.124.60.4	Ok
03.10.2017 06:44	UserName	86.124.60.4	Notified
03.10.2017 06:44	UserName	86.124.60.4	Ok
03.10.2017 10:49	UserName	86.124.60.4	Notified
03.10.2017 10:50	UserName	86.124.60.4	Notified
03.10.2017 10:50	UserName	86.124.60.4	Ok

Figure 3: Authentication process modeled from log files

Calculating the averages for all the users, the average number of failed authentications attempts is 250% higher as compared to the authenticated ones, in the first month of use. In the following months, similar to the phenomenon described in Figure 3, the average values decrease. The value is 70% in the second month of use. Notice the decrease in the following months: 45% in the third month, 10% in the fourth month and 4% in the fifth month. These average values form the graph in Figure 4, confirm the fact that in time the users learn the new authentication mechanism and become productive.

The graph in Figure 4 also contains failed authentication attempts launched from public access terminals. The data analysis shows on one hand that the public access system was exposed to unauthorized use and, on the other hand, that part of the medical staff who knew the system and accessed it without respecting the legal procedures, have finally understood that that it was no longer possible to do so. These authentication attempts have been reduced considerably by implementing the new method.

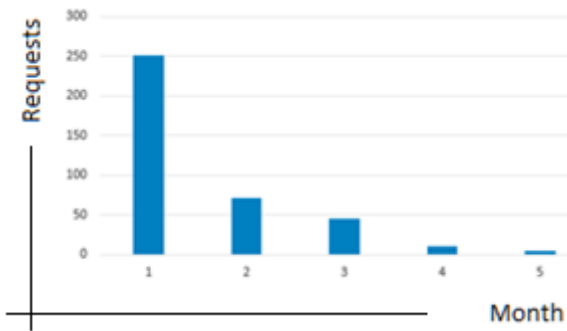


Figure 4: Learning cycle (5 months)

From analyzing the use of data, it has been revealed that users need an average of 30 seconds to authenticate, the difference being calculated between the authentication request timestamp and the start of the secure web session. The 30 second interval is composed of the time needed to place the telephone call (10 seconds) and 20 seconds needed for the user to answer the call and to select the images indicated verbally during the call. Those 30 seconds are the price paid for an increased security system and this time frame is comparable, as well as less time consuming than multifactorial authentication by other means (by text message, where the user must wait to receive the message, to open the application and must memorize the received code in order to later type it into the web interface).

While running the new system we also found that users are creative in finding new solutions to circumvent strict rules. As such, they requested that their personal telephone number to be changed in the database to the one publicly accessible in the department or have temporarily redirected their calls to other telephone numbers, of work colleagues usually. These practices are barely legal. However, we found fewer examples of such behavior as compared to the previously mentioned cases with post-it notes placed on the terminal's display in the operative medical staff room, that facilitated access to anyone who would breach the system. In order to counteract these use cases, we needed better organization frameworks, better user training and the use of advanced technologies that, for example, can detect automatic call forwarding.

One of the measures we applied, was to expedite the implementation of equipment necessary for the use of NFC technologies, such as the implementation of specialized NFC readers and user NFC ID cards. Thus, the users no longer knew the authentication web page URI, the unique user identifier nor the confirmation password. In order to initiate the session, they just needed to have the NFC ID in close proximity to the reader. The NFC ID cards improve the user comfort, they contribute to the increased system security, but, at the same time, they increase the implementation costs. The NFC ID cards were readily accepted by the medical staff, which later realized how easy they are to use.

Still, call redirection, unsecured mobile telephone theft, SIM card cloning are still major vulnerabilities for the system, especially if the voice notification subsystem uses an internationally recognized language and the images are directly indicated.

5. Related work

Since the user mentally establishes the connection between sounds and images, indirect indication is possible. For example, images can display animals while sounds describe their favorite food. In the set of animal icons, the rabbit will be chosen when the word "carrot" is heard.

The use of mother tongues can improve system security. A flower is indicated by vocalizing the expression "smells good". It is difficult for an unauthorized, non-native speaker to choose the flower icon when hearing *lukter godt* or *miroase bine* while this comes natural for a native speaker.

Indirect indications require special attention when it comes to expressing emotions. It might be harder for an unauthorized person to guess the correct images, when those images are selected in relevance to the experiences and according to the expected emotions of the actual user. Voice indications may sound cryptic because the meaning behind any symbol needs previous user initiation to understand and act in the required timeframe. An example of this kind of set of images are illustrated in figure 5.

These methods may have a difficulty level so high that unauthorized users are unable to guess correct image passwords even if they control both the authentication workstation and the user's phone. This method of authentication emphasizes the fact that the user's prior knowledge is an important element to be considered in multifactor authentication processes.



Figure 5: Example of a set of images

Security may also be enhanced within the proposed method of authentication using some traps as a result of some simple changes to the set of the displayed images or to the voice indications. This kind of changes require targeted training to user groups. This way, users will be instructed not to select any image following the first voice indication. The first voice indication in the call is a trap for users that do not know this detail. Another simple change is to use trap images. Trained users can easily avoid them even if they are urged to select them. In the image set from Figure 5, a trap image may be the icon of the key. A user with no training might touch that icon, following the voice indication and the authentication would fail.

Language that is familiar to the user (i.e. the local dialect) as well as false verbal instructions and/or trap images can provide insurmountable obstacles for unauthorized users. These methods have the advantage of being categorized as "what the user knows" [4, 5] and are not equipment-dependent (since they are compatible with fixed-line, analogue telephony) and with some necessary adjustments they can be used as an authentication method for users with severe deficiencies (such as visually impaired persons).

The web service interface must be carefully designed and implemented to hide the correlation between a trap and a failed authentication. Repeated tries must not help the unauthorized user to discover the reason why the authentication fails.

No technique is infallible and, in the domain area of security, authentication, or content protection, continuous efforts must be made to be at least one step ahead of cybercriminals. The suggested authentication method was developed on the basis that the users, the most important entity of the system, make mistakes. In order to reduce the effects of human mistakes and to help users in their overall effort to maintain the data safe, a simple security device has been added to the system. Figure 6 presents the updated authentication system; the added security device is labeled CA in the form of audio headphones.

The enhanced voice notification flow (figure 6) contains a smart-headphone (CA) as compared to the original flow (figure 1). These smart-headphones must be capable to function as regular headphones connected to the mobile phone and smart enough to detect special encoded data sub-channel in the audio stream. When such a stream is detected, the smart headphones automatically converts encoded data into comprehensible sounds according to the user's language. In other words, the headset hides a security component that users need to have in order to succeed in the authentication process.

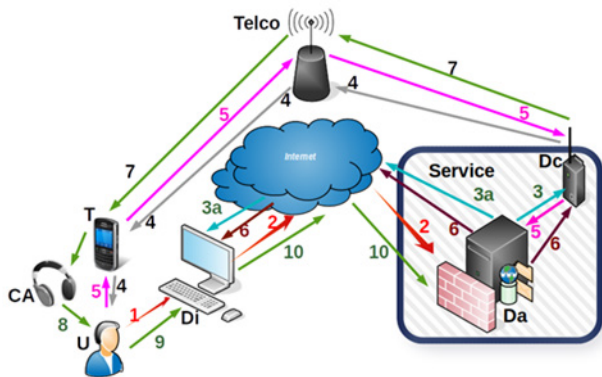


Figure 6: Enhanced web authentication system

In the enhanced authentication method, the telephone voice call contains an encrypted audio stream sub-channel, unintelligible to humans. This way the enhanced authentication method is more robust if the phone call is redirected, the mobile phone is lost, cloned etc. keeping the one-time random passwords voice indications just for certified users. One must have the smart-headphones connected to be able to listen the voice indications.

These especially designed smart headphones may be subject to the addition of more security features like biometrics or other security features that restrict smart headphone use outside of the particular scope (e.g. the user must touch the headphones to allow

decryption of the audio stream, a gesture unknown to a malicious person who manages to have both the smartphone and the smart headphones, since the shape of the smart headphones and their functionality seem normal/common).

Following these ideas, a prototype of these smart headphones was made. The voice notification system has been parameterized to transmit data using well-known dual tone multi-frequency signaling (DTMF) analogue telephony technology. DTMF streams are transmitted over phone calls, simultaneously and in the same frequency band as the voice channel. DTMF audio streams are easily decoded as bit streams using low power CMOS chips (e.g. MT8870). Bit streams are transmitted to a connected low power microsystem built as a System on a Chip (SOC). SOC processes all data streams and finally produces voice indications into the headphone speakers.

The proposed enhanced authentication process flow (Fig. 6, Fig. 7) is started by the user (1) in interaction with a browser. The browser sends the user identifier, the user confirmation password and the details about the device that initiates browser request (e.g. NFC device identifier) to the server, over an SSL Internet connection. The server prepares a random one-time-password (OTP) and engages (3) the modem to call the known user and, at the same time, instructs (3a) the browser to display relevant information.

The modem tries to call the known user's phone number (4). If user's phone is busy or if there is a telephone network congestion or if the user answers or rejects the call (5) then the server either closes the web session or continues it by sending to the modem (6) a code derived from the OTP and to the browser a web page to display the set of images (6a).

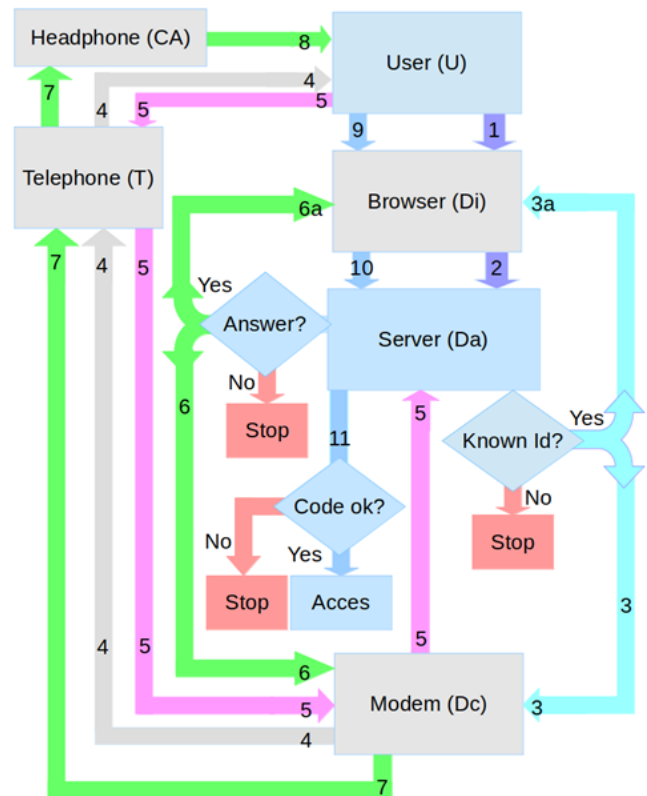


Figure 7: Enhanced authentication process flowchart

The modem starts the audio playback and mixes the DTMF encoded indications in the audio stream (7). The DTMF encoded string is detected, decoded and played back (8) by the smart headphones. The user acts (9) on displayed images based on his knowledge and according to the voice indications received through the smart headphones. The browser sends the user actions (10) (click or touch events) back to the server, in real-time. The server analyses the user web session continuously and verifies if all the required conditions are met (11) in order to authenticate the current user successfully.

Although the presented architecture (Fig. 6) contains a hidden security element in an audio headset, with virtualization technologies everything may become software based in a carefully architected mobile system. User comfort increases with virtualized devices, but the system management layer will have to address new security risks. These technological options must be thoroughly studied before being implemented as a public service.

Conclusion

As the analysis presented in this paper shows, the new authentication methods redefine the user interaction with the web system during the authentication process. These new methods create an independent channel of communication between the user and the system, through the public telephone network in order to dictate a unique, random single-use password. The presented methods remove some important security vulnerabilities such as SIM card cloning, used to take fraudulent control of web accounts.

We have analyzed in detail the operation of a medical system that implements the new manner of web authentication, presenting a series of data and use models which on one hand confirm the usefulness of these methods and on the other hand stand at the basis of future improvements. Before implementation, the analyzed medical system was plagued with unauthorized access sessions by individuals using public access terminals available in hospitals. After the implementation we have found a drastic reduction of unauthorized access coming from this type of terminals.

The central authority running the analyzed web service implemented the use of NFC ID cards and distributed specialized NFC readers for the necessary terminals. The use of NFC ID cards has become standard, noticing increased speed in the authentication process. Now, users no longer need to memorize Internet web site address, personal identifiers nor passwords in order to initiate the secured web sessions used to signal the authentication intent of the user.

Even though the paper presents a system based on random single use passwords (OTP) that are equivalent to 4-digit PIN numbers, they do not represent a restriction. The system analyzed in this paper uses such passwords because they are sufficient to obtain the desired effect, without disrupting unnecessarily the fragile balance between security, functionality and comfort.

In accordance to the presented method, the authentication process produces plenty of other useful data besides that collected during the password creation phase. The continuous analysis of the system's identification data using a Deep Learning algorithmic process could increase security as a whole by applying custom user profile policies. The user profiles contain data that refers to the timeframe needed for the voice indications to be recognized and

actions to be taken for successful authentication. These ideas have not been implemented nor analyzed yet, but are considered for further development.

As the latest advances in computing (including better, more powerful processors, large and very fast storage systems e.g. solid-state drives (SSD), artificial intelligence, deep learning technologies) are accessible to the public but also to cybercriminals, further efforts must to be undertaken to enhance every component of the security system in order to fight cyber-crime. This work tries to emphasize the role of people and culture in the authentication process. The main ideas presented in this paper have been implemented / tested and analyzed in detail. As one would expect, the subject open to future research, as knowledge, innovation, technological advances in Information Technologies, communications, miniaturization and low energy consumption devices will bring new, more secure and user-friendly solutions.

Conflict of Interest

The author declares no conflict of interest.

References:

- [1] ISO 9241 Ergonomics of Human-System Interaction Standard, <https://www.iso.org>, 2018.
- [2] EU General Data Protection Directive (GDPR), <https://www.eugdpr.org/>, 2018
- [3] Paul A. Grassi, Michael E., Garcia James L. Fenton, "Digital Identity Guidelines", NIST Special Publication 800-63-3, <https://doi.org/10.6028/NIST.SP.800-63-3>, 2018
- [4] D. Dasgupta, A. Roy, A. Nag, "Advances in User Authentication", ISSN:2363-6149, 2017
- [5] Richard E. Smith. "Authentication – From Passwords to Public Keys", ISBN: 0-201-61599-1, 2002
- [6] D. Charreau, S.M. Furnell & P.S. Dowland, "PassImages : An alternative method of user authentication" in *Advances in Networks, Computing and Communications 2*, ISBN: 9781841021409, 2004
- [7] S. Chiasson, A. Forget, R. Biddle, P.C. van Oorschot, "User interface design affects security: Patterns in click-based graphical passwords" <https://link.springer.com/article/10.1007/s10207-009-0080-7>, 2009
- [8] FIDO Alliance, <https://fidoalliance.org/>, 2018
- [9] Lily Hay Newman, "Cloudflare and Google Will Help Sync the Internet's Clocks - and Make You Safer", <https://www.wired.com/story/cloudflare-google-rough-time-sync-clocks-security/>, retrieved Nov. 2018
- [10] Roughtime Project, <https://rough.time.google.com/rough-time>, 2018
- [11] Nicole P., Amie T., Adam S., "Marriott Hacking Exposes Data of Up to 500 Million Guests", *The New York Times*, <https://www.nytimes.com/2018/11/30/business/marriott-data-breach.html>, accessed 2018-12-01
- [12] State of Authentication Report 2017, FIDO Alliance, <https://fidoalliance.org/>, 2018
- [13] "Silicon Valley Execs Targeted in 'SIM Swap' Hacking, \$1 Million in Crypto Stolen", <https://www.newsbt.com/2018/11/22/silicon-valley-exec-targeted-in-sim-swap-hacking-1-million-in-crypto-stolen/>, 2018
- [14] "\$14 Million in Cryptocurrency Allegedly Stolen By SIM Swappers, Authorities Report", <https://www.cryptoglobe.com/latest/2018/09/14-million-in-cryptocurrency-allegedly-stolen-by-sim-swappers-oklahoma-city-authorities-report/>, 2018
- [15] Paul Kocher et al., "Spectre Attacks: Exploiting Speculative Execution" <https://spectreattack.com>, 2018
- [16] N. Perloth, A. Tsang, A. Santano, Marriott Hacking Exposes Data of Up to 500 Million Guests <https://www.nytimes.com> Nov. 30, 2018
- [17] Z. Zhao, G. Ahn, J. Seo, H. Hu, "On the Security of Picture Gesture Authentication" 22nd USENIX Security Symposium ISBN: 978-1-931971-03-4, 2013
- [18] Arti Bhanushali, et al., "Comparison of graphical password authentication techniques" *International Journal of Computer Applications (0975 – 8887)* Volume 116 – No. 1, 2015
- [19] William E. Burr et al. "Electronic authentication guideline" NIST Special Publication 800-63-2, 2013
- [20] NFC Forum, "Essentials for Successful NFC Mobile Ecosystems", 2008

- [21] Online Services srl, "Constructive assembly and method for granting authorized access to an Internet service platform", PCT/RO2017/000002 Patent pendig, 2016
- [22] Entersekt, "OTP security past its expiration date" <https://www.entersekt.com>, 2014
- [23] Entersekt, "Securing the mobile banking channel" <https://www.entersekt.com>, 2014
- [24] Sonia Chiasson, "Usable Authentication and Click-Based Graphical Passwords" ISBN: 978-0-494-47475-4, 2009
- [25] M. Mathuri Pandi, A. Valarmathi, "A secured graphical password authentication system" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 5 ISSN: 2278-0181, 2013
- [26] Amish Shah, Parth Ved, Avani Deora, Arjun Jaiswal, Mitchell D'silva, "Shoulder-surfing resistant graphical password system" Procedia Computer Science 45, pp. 477 – 484, 2015
- [27] Vinit Khetani, Jennifer Nicholas, Anuja Bongirwar, Abhay Yeole, "Securing web accounts using graphical password authentication through watermarking" International Journal of Computer Trends and Technology (IJCTT) – volume 9 number 6 ISSN: 2231-2803, 2004
- [28] "Logging in W3C httpd". World Wide Web Consortium. 1995-10-12. Retrieved 2018-11-20.
- [29] R. Fielding et. al., "Hypertext Transfer Protocol – HTTP/1.1", The Internet Society (1999).

Hypervolume-Based Multi-Objective Reinforcement Learning: Interactive Approach

Hiroyuki Yamamoto, Tomohiro Hayashida*, Ichiro Nishizaki, Shinya Sekizaki

Graduate School of Engineering, Hiroshima University, 739-8524, Japan

ARTICLE INFO

Article history:

Received: 25 October, 2018

Accepted: 13 January, 2019

Online : 26 January, 2019

Keywords:

Multi-objective optimization

Reinforcement learning

Pareto optimal solution

Interactive method.

ABSTRACT

In this paper, we propose a procedure of interactive multi-objective reinforcement learning for multi-step decision problems based on the preference of a decision maker. The proposed method is constructed based on the multi-objective reinforcement learning which is applied to multi-step multi-objective optimization problems. The existing literature related to the multi-objective reinforcement learning indicate that the Hypervolume is often effective to select an action from the Pareto optimal solutions instead of determining the weight of the evaluation for each objective. The experimental result using several benchmarks indicate that the proposed procedure of interactive multi-objective reinforcement learning can discover a certain action which is preferred by the decision maker through interactive.

1. Introduction

In Robot that autonomously decides action based on surrounding environmental information such as disaster rescue robot and robot cleaner needs to optimize multiple objectives simultaneously, such as moving as fast as possible to the target location, increasing safety, reducing consumption of fuel and batteries, etc. However, in multi-criteria decision making, objectives often conflict with each other, and in such cases there does not exist complete optimal solutions that simultaneously minimize or maximize all the objectives. Instead of a complete optimal solution, a solution concept, called Pareto optimality, is introduced in multi-objective optimization, and many efforts are accumulated to find a set of Pareto optimal solutions [1]. At a Pareto optimal solution, in order to improve a certain objective function, we have to sacrifice one or more other objective functions. Moreover, when there exist two or more Pareto optimal solutions, a decision maker selects the most preferred solution from among the Pareto optimal solution set based on his or her own preference. However, a complex procedure is required to identify the preference structure of the decision maker [2]. If we interpret each Pareto optimal solution as a candidate and try to select one solution out of the Pareto optimal solution set, we do not necessarily need to identify the preference structure of the decision maker. Then, we can employ an interactive decision making method that derives a so-called preference solution of the decision maker by using the local preference information obtained from an interactive process with the decision maker [3-5].

As a solution method for linear or convex optimization problems, it is possible to apply mathematical solution methods finding an exact optimal solution such as the simplex method, the successive quadratic programming, and the generalized reduced gradient method. On the other hand, for non-convex or discontinuous optimization problems, we have to employ some approximate optimization methods such as evolutionary computation methods including genetic algorithm, genetic programming, and evolution strategy. Such evolutionary technologies include swarm intelligence such as particle swarm optimization, ant colony optimization and so on. Effectiveness of such attempts for difficult optimization problems have been expected [6-8]. However, it is difficult to apply these evolutionary computation methods to multi-step optimization problems like a chase problem [9]. For multi-step problems, trial-and-error methods using multi-agent systems is suitable, and for the learning mechanism for artificial agents reinforcement learning with bootstrap type estimation is often employed [10].

This paper, we propose an interactive multi-objective reinforcement learning method for choosing actions based on the preference of a decision maker for multi-step multi-objective optimization problem. In previous studies, for applying reinforcement learning to multi-objective optimization problem, after a multi-objective optimization problem is reformulated into a single-objective optimization problem by using a scalarization method with weighting coefficients, usual (i.e. single-objective) reinforcement learning is employed [11] [12]. However, it is difficult to determine the weight of the evaluation for each objective beforehand. Therefore, van Moffaert et al. proposed

*Tomohiro Hayashida, Email: hayashida@hiroshima-u.ac.jp

hypervolume-based multi-objective Q-learning (HBQL) [13] and Pareto Q-learning (PQL) [14] to evaluate Pareto optimal solution set with three indices of hypervolume, cardinality, and Pareto relation, and demonstrated effectiveness of these methods. The hypervolume [15] employed in HBQL is an index for evaluating a Pareto optimal solution set which means the size of a region dominated by obtained Pareto optimal solutions and limited by a certain reference point.

In this paper, focusing on a property that a value of the hypervolume increases as the number of Pareto optimal solutions is small in the neighborhoods, we propose an interactive method reflecting the preference of the decision maker for multi-objective reinforcement learning in which the hypervolume is used for efficiently finding diverse Pareto optimal solutions, not for selecting the preferred solution from among Pareto optimal solution set.

This paper is organized as follows. In section 2, we review related works of hypervolume-based multi-objective reinforcement learning. We propose interactive multi-objective reinforcement learning method in section 3. In section 4, numerical experiments are conducted to verify the effectiveness of the proposed method. Finally, in section 5, we summarize the result of the paper and discuss the future research directions.

2. Related Works

In multi-step multi-objective optimization problem, it is necessary to select appropriate action at each step in order to simultaneously optimize multiple objectives. Furthermore, when making decisions, it is also necessary to reflect the preferences of the decision maker (DM). In this paper, we consider these issues in multi-step multi-objective optimization problem into three aspects, "Multi-objective optimization", "Multi-step", and "Reflecting the preference of the decision maker".

For "Multi-objective optimization", instead of optimizing by converting to a single-objective optimization problem by assigning appropriate weights to each objective, the proposed method searches for Pareto optimal solution set. For "Multi-step", we use reinforcement learning, which can comprehensively evaluate actions of several steps by feeding back the obtained reward. Furthermore, we adopt an interactive approach for "Reflecting the preference of the decision maker". This section, we briefly introduce relevant research in each of "Multi-objective optimization", "Multi-step", and "Reflecting the preference of the decision maker".

2.1. Multi-objective optimization

Conventional research on multi-objective optimization uses a procedure of scalarization such as conversion to single-objective optimization problem using weighted sum of each objective and optimization. The weight of each objective is set taking into consideration the preference structure of DM, and based on this, the scalarized single objective function is optimized. DM evaluates the solution, and if necessary, readjusts the weights of each objective and redoes learning. Then, these procedures are repeated. In this method, if the objective number is large, accurate evaluation of the tradeoff relationship of each objective, that is, evaluation of

the weight is difficult, and it is not necessarily an appropriate method.

On the other hand, in the multi-objective optimization problem, one solution preferred by the decision maker should be reasonably selected from the Pareto optimal solution set which is executable and not dominated by other solutions. Various methods such as Multi-objective Genetic Algorithm (MOGA) [16] and Multi-objective Particle Swarm Optimization (MOPSO) [17] have been proposed so far to obtain a Pareto optimal solution.

However, considering the multi-step optimization problem like a chase problem [9], the agent needs to acquire the multi-step action decision rule by trial and error, and it is difficult to apply these methods. Multi-objective optimization by reinforcement learning [10] which adopts bootstrap type learning is suitable for such a problem.

2.2. Multi-step

Reinforcement learning

Reinforcement learning[10] is a framework for agents to learn policies and achieve goals from the interaction between the environment and agents. In addition, it is a method to acquire optimum action (policy) by learning value function trial and error. Here, The agent perform learning and action selection, and objects that are composed of all outside of this agent and whose agents respond are called environments. Since policy, which is an agent's action decision rule in reinforcement learning, decides action based on the current environment, it is desirable that the environment be described by Markov Decision Process (DMP).

Markov Decision Process

In the stochastic state transition model, the distribution of the next state depends only on the current state. In other words, when the state transition does not depend on history such as past state transition, such property is called Markov property. Let $s_t = s$ and $a_t = a$ be the current (t_{th} period) state and chosen action of the agent, then the transition probability $P_{ss'}^a$, of each possible next state to s' is described as follows:

$$P_{ss'}^a = Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \quad (1)$$

Based on the transition probability (1), the expected reward $R_{ss'}^a$, is can be calculated as follows:

$$R_{ss'}^a = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\} \quad (2)$$

In addition, this paper uses Q learning, which is a kind of reinforcement learning, to acquire agent strategies.

Q-Learning

In Q-learning, the action value function is called the Q value, and the Q value is updated as follows based on the immediate reward r .

$$Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{a'} Q(s', a') - Q(s, a)] \quad (3)$$

Here, α is the learning rate and γ is the discount rate parameter, which the designer decides when using this algorithm.

Q-Learning algorithm

- 1: Initialize $Q(s, a)$
- 2: Repeat for each episode:
- 3: Initialize the state s
- 4: Repeat for each step of the episode:
- 5: Select action a with s using a strategy derived from Q (e.g. ϵ -greedy strategy based on the Q table)
- 6: Execute the action a and observe the reward r and the next state s'
- 7: $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
- 8: $s \leftarrow s'$;
- 9: If s is not the terminate state, repeat.

Figure 1: Q-Learning algorithm

The idea of updating the current estimator based on another estimator such as estimate of value of following state as described above is called *bootstrap*. In the multi-step optimization problem, the bootstrap type learning updates the evaluation value of the current state based on the estimated value of the evaluation value of the next state. Therefore, for example, when the evaluation value is interpreted as the expected value of the reward, even if the problem is complicated, the approximate value can be obtained by repeating the learning.

2.3. Reflecting the preference of the decision maker

In the multi-objective optimization problem, generally there are multiple Pareto optimal solutions, so it is desirable to rationally select a solution from the Pareto optimal solution set based on the preference structure of the decision maker. On the other hand, it is difficult to model the preference structure of the decision maker mathematically and strictly, so approximate modeling methods have been applied. Representative methods of multi-criteria decision making method include the Analytic Hierarchical Process (AHP)[18], the Analytic Network Process (ANP)[18] and the like. AHP is a method for evaluating relative importance of all standards by decision-makers subjectively making one-pair comparison of pairs of criteria, and so far application to many decision-making problems has been reported. In addition, ANP is a method in which AHP that is a hierarchical structure is expanded to a network structure. However, conventional methods such as AHP and ANP are not mathematically and strictly evaluating the preferences of decision makers, and the obtained solutions are not necessarily the most preferred by decision makers. On the other hand, Multi-Attribute Utility Analysis (MAUT)[18] has been proposed as a method for strictly evaluating and modeling the preference structure of decision makers, but the procedure is complicated and the burden on decision makers is also great.

In this paper, we adopt an interactive method that selects criteria that DM wants to improve against the selected solution, adjusts weight coefficient, and repeats solution improvement until DM accepts as a method to select solutions most preferred by DM among executable Pareto optimal solutions.

The interactive method

It is difficult to identify a function that properly defines and expresses the preference structure of the decision maker. In such a

case, by using the interactive method[5] based on limited preference information obtained through dialogue between decision makers and analysts called a *preferred solution*, it is possible to select the most preferred solution without expressing the preference structure of the decision maker.

In the following multi-objective linear programming problem,

$$\left. \begin{array}{l} \text{minimize } z_1(\mathbf{x}) = \mathbf{C}_1\mathbf{x} \\ \text{minimize } z_2(\mathbf{x}) = \mathbf{C}_2\mathbf{x} \\ \vdots \\ \text{minimize } z_k(\mathbf{x}) = \mathbf{C}_k\mathbf{x} \\ \text{subject to } \mathbf{x} \in X = \{\mathbf{x} \in R^n | A\mathbf{x} \leq \mathbf{b}, \mathbf{x} > 0\} \end{array} \right\} \quad (4)$$

Let w_1, w_2, \dots, w_k be weights of the conflicting objects $\mathbf{z}(\mathbf{x}) = (z_1(\mathbf{x}), z_2(\mathbf{x}), \dots, z_k(\mathbf{x}))^T$ which the decision maker determines subjectively. The multi-objective linear programming problem is converted into a single objective linear programming problem as follows.

$$\left. \begin{array}{l} \text{minimize } \sum_{i=1}^k w_i z_i(\mathbf{x}) \\ \text{subject to } \mathbf{x} \in X \end{array} \right\} \quad (5)$$

The analyst interactively updates the weights w_1, w_2, \dots, w_k until the decision maker satisfies with the solution obtained by solving this problem.

3. Hypervolume-Based Multi-Objective Reinforcement Learning

By extending MDP mentioned in Section 2.1 as multi-objective Markov decision process (MOMDP), reinforcement learning can be applied to multi-objective problems. In MOMDP, rewards are given as vectors defined by multi-dimension rather than scalar. That is, given the m objectives o_1, o_2, \dots, o_m the rewards given are m dimensional vector $\vec{r} = (r_1, r_2, \dots, r_m)^T$. Here, T represents transpose.

3.1. A solution discovery method based on a scalarized objective function

The inner product of the reward vector $\vec{r} = (r_1, r_2, \dots, r_m)^T$ given from the environment and the weight coefficient vector \vec{w} is as follows.

$$\begin{aligned} r &= w_1 r_1 + w_2 r_2 + \dots + w_m r_m \\ &= \sum_{i=1}^m w_i r_i \end{aligned} \quad (6)$$

By using this, it is possible to apply single-objective Q learning to MOMDP. As a method of determining weighting coefficients, AHP, ANP, an interactive method, and the like are conceivable.

3.2. Multi-objective optimization and Hypervolume

Let n sets of solutions be $S_i = (x_i^1, x_i^2, \dots, x_i^m) (i = 1, 2, \dots, n)$. Hypervolume [15] represents the volume of the region between a reference point and a region dominated by the solution set in the objective function space. Specifically, it is the surrounded by the boundary formed by solution set and the reference point $\mathbf{r} = (x_r^1, x_r^2, \dots, x_r^m)$ necessary to limit the area, therefore Hypervolume can be shown as follows.

$$[\mathbf{r}, \mathbf{S}_i]^m \equiv [x_r^1, x_i^1] \times [x_r^2, x_i^2] \times \dots \times [x_r^m, x_i^m] \quad (7)$$

Here, $volume(\mathbf{r}, \mathbf{S}_i)$ represents the size of the area defined based on reference point \mathbf{r} and Pareto solution \mathbf{S}_i , and $volume(\cup_i[\mathbf{r}, \mathbf{S}_i]^m)$ represents the region formed by the reference point and the Pareto optimal solution set. Here, $\cup_i[\mathbf{r}, \mathbf{S}_i]^m$ represents the union of the areas of Pareto optimal solution set.

Figure 2 shows the solution set of the 4 solutions $\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_4$ in the two-object maximization problem and Hypervolume by reference point \mathbf{r} . Let $volume(\mathbf{r}, \mathbf{S}_1)$ be the area surrounded by thick lines. Hypervolume with four solutions and reference points can be represented by the union of gray area and hatched area.

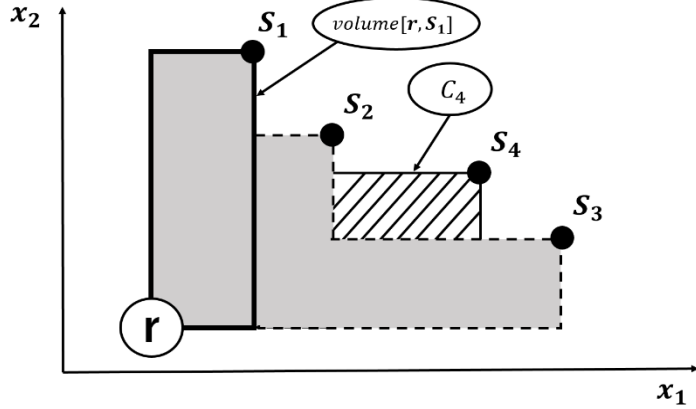


Figure 2: Example of Hypervolume

Let consider a new solution \mathbf{S}_{k+1} a set of current solution set $S = (\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_4)$. Here, let the increased area, that is, the increment of Hypervolume, be the *contribution* of Pareto solution \mathbf{S}_{k+1} . Figure 2, the *contribution* of the Pareto solution \mathbf{S}_4 is the area of C_4 which is indicated by the hatched portion. In the objective function space, the Pareto solution in the region where the Pareto solutions are dense is lower the *contribution* and the *contribution* becomes higher as the Pareto solution in the region where the Pareto solution are sparse. Therefore, by selecting solutions with high *contribution*, it is possible to acquire a wide range of Pareto solution sets without paternity of Pareto solution to some areas.

In this paper, optimization is performed using multi-objective Q-learning which adopts ϵ -greedy strategy as an action decision policy. In order to obtain a sufficient number of solution candidates, it is necessary to repeat learning according to a decision policy including random action choice such as ϵ -greedy decision policy. Such that an agent chooses an action with the highest Q-value with probability $(1-\epsilon)$, and the agent randomly choose an action with a low probability ϵ . However, since the solution that the decision maker most desires is not always the solution with the highest *contribution*, a decision-making method that selects a solution rationally from the Pareto optimal solution set is necessary other than the method selecting the solution with the highest Hypervolume.

In this paper, we propose a rational decision making method based on the preference structure of decision makers by interactive method from discovered Pareto optimal solution set of the target multi-objective optimization problem. Next section describes the proposed method in detail.

4. Interactive Multi-Objective Reinforcement Learning

In this paper, in order to acquire a sufficient number of Pareto optimal solutions as an optimization method for the multi-objective optimization problem, this paper constructs an efficient solution search method by hypervolume-based multi-objective reinforcement learning. After searching solutions, a solution is selected based on the decision makers' preference from Pareto optimal solution set by using interactive method. Weight vectors corresponding to the objectives o_1, o_2, \dots, o_m used in the proposed method are defined as follows.

$$\vec{w} \equiv (w_1, w_2, \dots, w_m)^T \quad (8)$$

Figure 3.

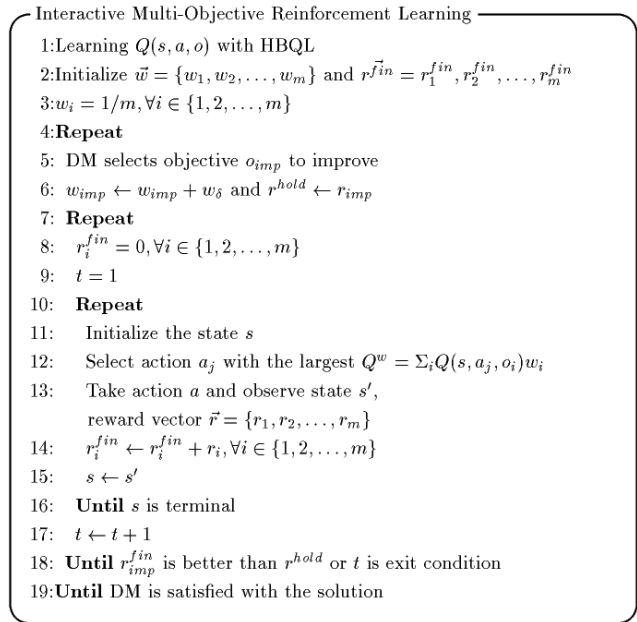


Figure 3: The algorithm of the proposed method

By combining such interactive method with hypervolume-based multi-objective reinforcement learning, it became possible to efficiently derive a Pareto optimal solution reflecting the preference of decision makers in multi - step multi - objective optimization problem.

5. Numerical experiments

5.1. Benchmark Problems and Objectives

This paper conducts numerical experiments using path finding problems as benchmarks of multi-objective optimization problems in continuous value environment in order to confirm the effectiveness of the proposed method. The numerical example is defined on the two-dimensional grid map to which numerical values are assigned to each grid. The numerical value assigned to each grid is a scalar value or a multidimensional vector, and the numerical value at the coordinate $\mathbf{x} = (x_1, x_2)$ is defined by $f_i(\mathbf{x}), i = 1, 2, \dots, m - 1$ for each objective. Also, the total of numerical differences between grids obtained when an agent moves to an adjacent grid is defined as a cost. Specifically, when an agent moves from the grid \mathbf{x}^1 to the grid \mathbf{x}^2 directly, the numerical difference obtainable for each objective is $|f_i(\mathbf{x}^1) - f_i(\mathbf{x}^2)|$. The total numerical difference obtained by the agent

during single episode is defined as a cost of the corresponding episode. The agent goes on the 2-dimensional grid map and makes a goal, taking into account the following objectives.

- (1) Fewer steps of an episode.
- (2) Smaller cost of an episode.

Detailed settings of the experiments are as follows.

- The agent recognizes the numerical value of the current position and the grid in the eight directions adjacent to the current position.
- The agent obtains a positive reward r_{goal} only when the agent reaches a goal.
- Each step the agent moves, it acquires a negative reward r_{step} and a non-positive reward $\mathbf{r} = (r_1, r_2, \dots, r_{m-1})$ based on the numerical difference.
- When the agent tries to go outside the grid map, it acquires a negative reward r_{wall} as a penalty.
- The maximum number of movement of an agent is 10000 steps. If it is not possible to reach the goal during such limitation, learning is interrupted and the next episode is started.
- The maximum number of updates of the weights is 100, and the solution derived by the weights at the time when the number of updates of the weights reaches 100 times, a solution is suggested to the decision maker based on the weights before updating.

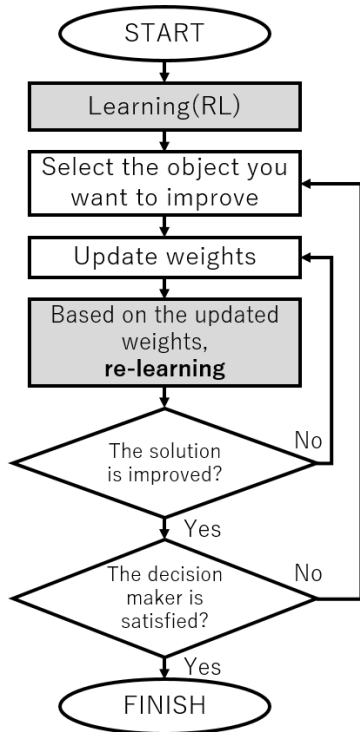


Figure 4: Construction of the conventional method

5.2. The conventional method

In the numerical experiments of this paper, in order to confirm the effectiveness of the proposed method, the experimented result

of the proposed method are compared with the result of the conventional method. As a conventional method, single-objective reinforcement learning based on the weighting coefficient method is adopted. This is a method of converting the reward function of the multi-objective optimization problem into a single objective optimization problem by scalarizing the reward function with a weight vector and applying a single-objective Q-learning[19]. Therefore, when the weight vector for scalarizing the reward function is changed, it is necessary to re-learn the Q-value. The flowchart of the proposed method and the comparison method are shown Figure 4 and Figure 5. The differences between these two methods are shown by the gray part.

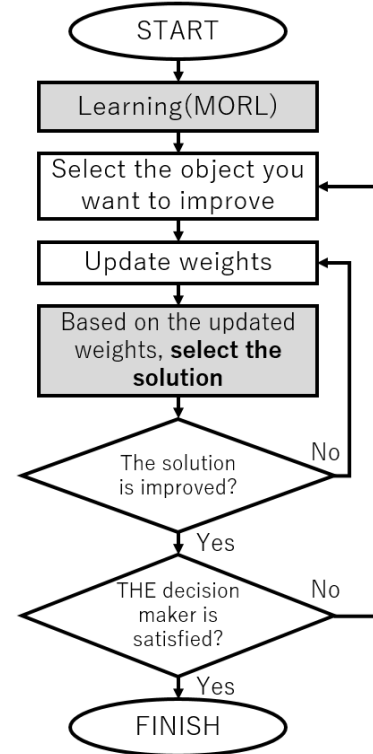


Figure 5: Construction of the proposed method

5.3. Result and Discussion

Figure 6, ..., Figure 8. The lower left green colored cell is the start and the upper right blue colored cell is the goal. Multidimensional numerical values can be allocated on one cell by overlaying these grid maps in Figure 6, ..., Figure 8. Table 1 outlines the numerical experiments of this paper such as combination of grid maps.

Table 1: Outline of Numerical Experiment

Route finding Problem	Minimization objective
#1	(step, $f_1(x)$, $f_2(x)$)
#2	(step, $f_2(x)$, $f_3(x)$)
#3	(step, $f_3(x)$, $f_1(x)$)
#4	(step, $f_1(x)$, $f_2(x)$, $f_3(x)$)

The environment and parameters of this experiment are shown below:

- OS: windows 10 Home 64bit
- RAM: 8.00GB
- CPU: intel core-i3 (3.90GHz, 2core, 4thread)

parameters

- $\alpha = 0.10$
- $\gamma = 0.95$
- $w_\delta = 0.1$
- $r_{goal} = 130$
- $r_{wall} = -130$
- Number of learning episodes = 10000
- At the start of learning, $\epsilon = 0.99$. In addition, after 1000 episodes, $\epsilon = \epsilon \times 0.97$ is calculated every 100 episodes to reduce the value of ϵ .

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1
1	2	4	4	4	4	4	4	4	4	4	4	4	4	2	1
1	2	4	7	7	7	7	7	7	7	7	7	7	4	2	1
1	2	4	7	11	11	11	11	11	11	11	11	7	4	2	1
1	2	4	7	11	16	16	16	16	16	11	7	4	2	1	1
1	2	4	7	11	16	22	22	22	16	11	7	4	2	1	1
1	2	4	7	11	16	22	29	22	16	11	7	4	2	1	1
1	2	4	7	11	16	22	29	22	16	11	7	4	2	1	1
1	2	4	7	11	16	16	22	16	16	11	7	4	2	1	1
1	2	4	7	11	11	11	16	11	11	11	7	4	2	1	1
1	2	4	7	7	7	7	7	11	7	7	7	4	2	1	1
1	2	4	4	4	4	4	7	4	4	4	4	4	2	1	1
1	2	2	2	2	2	2	4	2	2	2	2	2	2	1	1
1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1

Figure 6: Map of $f_1(x)$

4	4	4	4	4	4	4	1	1	1	1	1	1	1	1	1
4	10	10	10	10	10	4	1	3	3	3	3	3	3	3	1
4	10	19	19	19	10	4	1	3	7	7	7	7	7	3	1
4	10	19	31	19	10	4	1	3	7	13	13	7	7	3	1
4	10	19	31	19	10	4	1	3	7	13	13	7	7	3	1
4	10	10	19	10	10	4	1	3	7	13	7	7	7	3	1
4	4	4	10	4	4	4	1	3	3	7	3	3	3	3	1
1	1	1	4	3	1	1	1	1	1	3	4	1	1	1	1
1	3	3	3	7	3	3	1	4	4	4	10	4	4	4	4
1	3	7	7	13	7	3	1	4	10	10	19	10	10	4	4
1	3	7	13	13	7	3	1	4	10	19	31	19	10	4	4
1	3	7	13	13	7	3	1	4	10	19	31	19	10	4	4
1	3	7	7	7	7	3	1	4	10	19	19	19	10	4	4
1	3	3	3	3	3	3	1	4	10	10	10	10	10	4	4
1	1	1	1	1	1	1	1	4	4	4	4	4	4	4	4

Figure 7: Map of $f_2(x)$

30	30	30	30	30	60	60	60	60	60	80	1	1	1	1	1
30	80	80	80	30	60	100	100	100	100	80	1	40	40	40	1
30	80	100	80	30	60	100	180	100	100	80	1	40	60	40	1
30	80	100	80	30	60	100	180	100	100	80	1	40	60	40	1
30	30	80	30	30	80	80	100	80	80	1	1	40	1	1	1
60	60	60	60	80	1	1	1	1	1	1	60	60	60	60	60
60	100	100	100	80	1	40	40	40	1	60	100	100	100	100	60
60	100	180	100	80	1	40	60	40	1	60	100	180	100	100	60
60	100	180	100	80	1	40	60	40	1	60	100	180	100	100	60
80	80	100	80	80	1	1	40	1	1	60	60	100	60	60	60
1	1	1	1	1	60	60	60	60	60	60	30	30	30	30	30
1	40	40	40	1	60	100	100	100	100	60	30	80	80	80	30
1	40	60	40	1	60	100	180	100	100	60	30	80	100	80	30
1	40	60	40	1	60	100	180	100	100	60	30	80	100	80	30
1	1	40	1	1	60	60	100	60	60	30	30	80	30	30	30

Figure 8: Map of $f_3(x)$

Table 2, ..., Table 9 show the flow of solution improvement of the proposed method and comparative method in each problem. The solution displayed in the top row of each table is the initial solution. An improved solution to the objective that the decision maker has requested improvement is shown in the next section.

Comparing Table 2 and Table 3, it can be seen that the number of times of improvement is better for the solution improvement by the proposed method shown in Table 2 than for the solution improvement by the conventional method shown in Table 3.

This experiment is based on the premise that the answer of the decision maker is not contradictory to its own preference structure. However, in reality, since decision makers cannot grasp their own preference structure, if the number of responses increases, the decision maker answers incorrectly and there is a high possibility that inconsistency will occur. In addition, as the objective number increases, its possibility appears to be remarkable. Therefore, by reducing the number of responses by decision makers like the proposed method, the accuracy of the answer increases and the possibility of inconsistency can be lowered.

In addition, it can be seen that the improvement of the solution by the proposed method shown in Table 2 is shorter in execution time than the solution improvement by the conventional method shown in Table 3. This is because the proposed method finds Pareto optimal solution sets collectively regardless of the decision maker's preference, and then selects one by dialogue, so that re-learning for each interactive is unnecessary. In the conventional method, since the scalarized reward function changes according to the decision request of the decision maker by interaction, the Q value by the new reward function must be re-learned. In addition, it is known that as the number of objects, the number of states, and the number of actions increase, the execution time increases exponentially. By improving such an efficient solution, the burden on decision makers can be reduced.

Similarly for the other problems, by comparing Table 4 with Table 5 and comparing Table 6 with Table 7, it was confirmed that the proposed method is shorter than the conventional method and the number of times of improvement is smaller than the conventional method, regarding the execution time required to improve the solution..

Also, in Problem 4, the proposed method is able to obtain satisfactory solutions for decision makers, however it cannot be obtained with the conventional method. In the experimental result of proposed method shown in Table 8, efficient solution search using Hypervolume is possible, however in the conventional method in Table 9, since the search of the solution is repeated by updating the solution at once, overlapping search operations are performed. By such inefficient search, it is considered that the satisfying solution of the decision maker could not be found by the conventional method.

When considering application to realistic multi-step multi-objective optimization problems such as rescue robots and robot cleaners, it was necessary to set weight coefficients of each objective beforehand as in the conventional method so far, it was difficult to select action reflecting the preference of the decision maker. However, by using the method proposed in this paper, it was possible for decision makers to efficiently find the most satisfactory Pareto optimal solutions in the multi - step multi - objective optimization problem.

6. Conclusion

This paper, proposes a multi-objective reinforcement learning with interactive approach to hypervolume-based multi-objective reinforcement learning and confirmed its effectiveness against multi-objective optimization problem in continuous value environments. Conventionally, a method of converting a multi-objective optimization problem into a single-objective problem and applying a single-objective reinforcement learning is common, however, it is difficult to decide the weighting coefficient when converting to a single-objective problem. Also, in order to consider the preference of the decision maker in the weighting coefficient, it is necessary to adjust by single-objective reinforcement learning each time. Such procedure is inefficient in the means of computational cost. Compared to such conventional methods, the proposed method proved to be able to efficiently find Pareto optimal solutions considering the preference of decision makers. As a future task, it is necessary to verify by practical experiment using real machine.

Table 2: Improvement by proposed method (Problem #1)

Target	Suggested Solution			Execution time [sec]
	step	cost _{f1}	cost _{f2}	
	18	30	18	3.824
cost _{f1}	19	20	22	0.010
Finish			Total	3.834

Table 3: Improvement by conventional method (Problem #1)

Target	Suggested Solution			Execution time [sec]
	step	cost _{f1}	cost _{f2}	
	17	30	26	0.888
cost _{f2}	18	30	18	3.427
cost _{f1}	21	20	22	27.953
step	19	38	38	0.852
cost _{f1}	20	30	26	0.849
cost _{f1}	25	28	34	4.241

step	17	30	26	0.862
cost _{f1}	19	20	22	0.829
Finish			Total	39.901

Table 4: Improvement by proposed method (Problem #2)

Target	Suggested Solution			Execution time [sec]
	step	cost _{f2}	cost _{f3}	
	23	54	0	45.220
step	20	54	118	0.013
step	18	48	196	0.010
Finish			Total	45.243

Table 5: Improvement by conventional method (Problem #2)

Target	Suggested Solution			Execution time [sec]
	step	cost _{f2}	cost _{f3}	
	23	54	0	1.009
step	21	54	78	15.976
step	20	34	398	6.274
step	19	54	196	24.766
step	18	48	196	0.867
Finish			Total	48.892

Table 6: Improvement by proposed method (Problem #3)

Target	Suggested Solution			Execution time [sec]
	step	cost _{f3}	cost _{f1}	
	23	1	30	46.693
Step	20	118	30	0.012
step	17	236	56	0.013
cost _{f1}	20	118	30	0.010
Step	17	236	56	0.009
step	14	354	56	0.011
cost _{f1}	18	196	42	0.012
cost _{f1}	20	118	30	0.010
Step	18	196	42	0.011
step	15	314	42	0.010
cost _{f1}	17	236	30	0.010
Finish			Total	46.801

Table 7: Improvement by conventional method (Problem #3)

Target	Suggested Solution			Execution time [sec]
	step	cost _{f3}	cost _{f1}	
	23	0	30	0.997
Step	22	238	26	5.546
Step	21	196	42	8.470
Step	20	118	30	20.603
Step	18	196	42	28.715
Step	17	274	42	5.228

cost f_1	22	276	40	11.209
cost f_1	22	118	30	4.360
Step	19	196	42	1.692
Step	18	196	42	0.879
step	17	274	42	7.727
cost f_1	18	354	40	11.889
Step	15	314	42	2.513
cost f_1	19	356	38	14.884
cost f_1	17	236	30	7.655
Finish			Total	132.367

Table 8: Improvement by proposed method (Problem #4)

Target	Suggested Solution				Executi on time [sec]
	step	cost f_1	cost f_2	cost f_3	
	23	30	54	0	346.790
step	21	30	54	78	0.015
step	20	30	54	118	0.015
step	20	30	54	118	0.017
step	18	42	48	196	0.014
cost f_1	20	30	54	118	0.014
step	18	42	48	196	0.013
cost f_2	17	42	36	354	0.015
cost f_3	18	42	48	196	0.011
cost f_3	20	30	54	118	0.013
cost f_1	20	30	54	118	0.016
step	20	30	54	118	0.016
step	17	30	54	236	0.015
cost f_3	20	30	54	118	0.012
cost f_2	18	30	42	276	0.014
cost f_2	18	30	38	276	0.014
Finish			Total	347.004	

Table 9: Improvement by conventional method (Problem #4)

Target	Suggested Solution				Executi on time [sec]
	step	cost f_1	cost f_2	cost f_3	
	23	30	54	0	1.103
step	22	30	50	158	1.937
step	20	42	48	156	4.887
step	19	30	42	276	37.351
step	18	30	54	196	16.966
cost f_2	21	30	50	276	0.945
step	20	30	54	118	1.835
step	19	42	54	196	1.802
step	18	42	48	196	8.227
cost f_1	21	36	70	316	1.852
step	20	22	38	238	0.919

step	18	42	48	196	0.919
cost f_1	26	40	66	316	3.661
cost f_1	20	30	54	118	0.920
cost f_2	20	38	38	238	0.934
step	18	42	48	196	1.827
:					

References

- [1] E. Zitzler and L. Thiele, "Multi-objective optimization using evolutionary algorithms-A comparative case study," in 5th Int. Conf. Parallel Problem Solving from Nature (PPSN-V), A. E. Eiben, T. B'ack, M. Schoenauer, and H.-P. Schwefel, Eds. Berlin, Germany: Springer- Verlag, 1998, pp. 292-301.
- [2] R. L. Keeney, "Decision analysis: An overview", Operations research 30 (1982) 803-838. <https://doi.org/10.1287/opre.30.5.803>
- [3] H.-S. Kim and S.-B. Cho, "Application of interactive genetic algorithm to fashion design," Engineering Applications of Artificial Intelligence, vol. 13, no. 6, pp. 635-644, 2000. [http://dx.doi.org/10.1016/S0952-1976\(00\)00045-2](http://dx.doi.org/10.1016/S0952-1976(00)00045-2)
- [4] H. Takagi, "Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation." Proceedings of the IEEE, 89(9):1275-1296, September 2001. ISSN 0018-9219. Invited Paper. <https://doi.org/10.1109/5.949485>
- [5] Masatoshi Sakawa , Hitoshi Yano, "An interactive fuzzy satisficing method for multi-objective linear fractional programming problems", Fuzzy Sets and Systems, v.28 n.2, p.129-144, November 1988. [https://doi.org/10.1016/0165-0114\(89\)90017-1](https://doi.org/10.1016/0165-0114(89)90017-1)
- [6] Tetsuya Higuchi, Yong Liu, and Xin Yao, editors. "Evolvable Hardware". Springer, New York, 2006.
- [7] Sakanashi, H., Iwara, M., Higuchi, T.: "Evolvable hardware for lossless compression of very high resolution bi-level images". IEEE Proc.-Comput. Digit. Tech., vol. 151, no. 4, (July 2004) 277-286. <https://doi.org/10.1049/ip-cdr:20040015>
- [8] "Lossy/Lossless Coding of Bi-level Images", ISO/IEC 14492:2001.
- [9] H. Tamakoshi, S. Ishii, "Multi-agent reinforcement learning applied to a chase problem in a continuous world", Artif. Life Robot., vol. 5, no. 4, pp. 202-206, 2001.
- [10] R.S. Sutton and A.G. Barto, "Reinforcement Learning", MIT Press, 1989.
- [11] M. M. Drugan, "Multi-objective optimization perspectives on reinforcement learning algorithms using reward vectors," proceedings of ESANN, 2015.
- [12] S. Natarajan and P. Tadepalli. "Dynamic preferences in Multi-Criteria Reinforcement Learning." Inproc. of the 22nd International Conference on Machine Learning, 601-608, 2005. <https://doi.org/10.1145/1102351.1102427>
- [13] K. Van Moffaert, M. M. Drugan, and A. Nowe, "Hypervolume-based multi-objective reinforcement rearning," 7th International Conference on Evolutionary Multi-Criterion Optimization, 2015.
- [14] k. Van Moffaert, A Nowe, "Multi-Objective Reinforcement Learning using Sets of Pareto Dominating Policies." J. Mach. Learn. Res. 15(1), 3483-3512(2014)
- [15] J. Bader and E. Zitzler, "A hypervolume-based optimizer for high-dimensional objective spaces", *Lecture Notes in Economics and Mathematical Systems*, Springer, 2009. http://dx.doi.org/10.1007/978-3-642-10354-4_3
- [16] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.
- [17] C. A. Coello Coello and M. S. Lechuga, "MOPSO: A proposal for multiple objective particle swarm optimization," in Proc. Congr. Evolutionary Computation (CEC'2002), vol. 1, Honolulu, HI, May 2002, pp. 1051–1056.
- [18] G.-H. Tzeng and J.-J. Huang, Multiple Attribute Decision Making: Methods and Applications (CRC Press, 2011).
- [19] M. Sakawa, I. Nishizaki, H. Katagiri, "Fuzzy Stochastic Multi-objective Programming", *Springer Science Business Media LLC*, 2011.

Optimal Designs of Constrained Accelerated Life Testing Experiments for Proportional Hazards Models

Xiaojian Xu^{*1}, Wanyi Huang²

¹Department of Mathematics and Statistics, Brock University, L2S 3A1, Canada

²XPO Logistics, Changhua 500, Taiwan

ARTICLE INFO

Article history:

Received: 30 November, 2018

Accepted: 06 January, 2019

Online : 29 January, 2019

Keywords:

Multiple step-stress proportional hazards model

Fisher's information

Optimal ALT design

Design constraint Asymptotic covariance matrix

Exact design

D-optimal design

A-optimal design

Q-optimal design

ABSTRACT

This paper investigates the methods of optimal design construction for step-stress accelerated life testing (ALT) when a Cox's hazards model is adopted with either a linear or a quadratic baseline hazard function. We discuss multiple step-stress plans for time-censored ALT experiments. The maximum likelihood method is utilized for estimating the model parameters. The information matrices have been derived for both models. The optimal stress-changing times and optimal stress levels are determined simultaneously under three different optimality criteria. In order to demonstrate the performance of the resulting designs, a simulation procedure is also provided. The efficiencies of our resulting optimal three-step-stress ALT plans are compared with their competitors using both practical examples and a simulation study. The efficiency comparison results have shown that the three-step-stress designs obtained with two optimal stress changing times and an optimal middle stress level are most efficient, compared to the corresponding optimal two-step-stress designs and to the optimal three-step-stress designs with a conveniently chosen middle stress. Furthermore, the efficient gains are most significant for hazard rate prediction for both cases when either a linear or a quadratic baseline hazard is assumed. Additionally, such efficiency gain is much greater for the case when the baseline function being quadratic than the case when that being simple linear.

1. Introduction

Reliability significantly influences the quality of product. Thus, many manufacturers make great effort to enhance the product reliability that largely determines the product competitiveness. Such importance has brought practitioners' attention to reliability evaluation. In life data analysis, we have to collect the failure time of a product under normal design conditions to quantify life characteristics of the product. However, such failure data for lifetime are very difficult to obtain in many situations, especially for the product with high reliability. Nowadays, lifetimes of many products are too long and the life testing period between design and release is limited; so, the tests under normal design conditions are too lengthy to get any failures. To overcome this problem, accelerate life testing (ALT) has been developed.

Since ALT can shorten the lifetime of a product, we often adopt it in order to obtain the failure information quickly within a limited time frame. In an ALT experiment, the test units are generally subjected to the stress levels that are higher than normal design level. The commonly used stress factors for failure acceleration include temperature, vibration, voltage, and pressure. Both the use of specific accelerated stress factors and the range of stress levels for a particular material or product are often suggested by engineering practice. Then, the failure data obtained at accelerated conditions have to be extrapolated through a proper model so that the characteristics of life distribution at normal design conditions can be estimated. A number of different types of stress loading schemes (such as constant, cyclic, step, progressive, and random stress loading) are available in practice when performing an ALT. By the relation between the stress levels and testing time, these stress loading schemes can be classified into two categories: time-independent and time-dependent stress loadings. When the stress loading is time-independent such as constant stress loading, the

^{*}Corresponding Author: Xiaojian Xu, 1812 Sir Isaac Brock Way, St. Catharines, ON, L2S 3A1 Canada; 905-688-5550 ext. 3300; xxu@brocku.ca.

stress level applied to each of the test units stays the same during the entire testing period. One of the demerits of using constant stress loading in ALT is that it could still take too long to run for the test experiment to observe sufficient failures when the inappropriate testing stress levels are applied. To address this problem, a time-dependent stress loading scheme is preferred to assure quick failures. When a time-dependent stress loading, such as step-stress, is adopted, test units are subjected to a stress level that is changing over testing time. In this paper, we consider step-stress ALT where all test units are subjected to the stress levels increasing by steps.

This paper is an extension of our previous work [1] originally presented in The Third IEEE International Conference on World Computing and Big Data Analysis. In literature, almost all the previous works done by others for Cox's proportional hazards (PH) based ALT models are very limited on simple step-stress assuming the baseline hazard function is simple linear. However, [2] have indicated that the optimal multiple step-stress ALT may further improve the quality of the reliability prediction for a parametric model. Therefore, this paper broadens the results of [1] where has only addressed the optimal ALT plans for PH models with a linear baseline function and moves on to discussing multiple step-stress optimal designs for ALT when adopting a PH model with either a simple linear or a quadratic baseline hazard function. In this paper, both optimal stress levels and optimal stress-changing times for three-step-stress ALT designs are derived under three different optimality criteria.

2. Literature Review and Preliminaries

2.1. Optimal designs for step-stress ALTs with a non-PH model

Miller and Nelson [3] first discussed Q-optimal designs for simple step-stress ALT tests with complete failure data assumed to be exponentially distributed. Their optimal designs were attained by minimizing the asymptotic variance (AVAR) of the maximum likelihood estimator (MLE) of the mean lifetime at the normal design stress level. Then, their work was extended to censored data by [4] who obtained the optimal simple step-stress ALT designs incorporating time-censoring. For some products or material, their failure times often follows a Weibull distribution. Assuming a Weibull distribution with a constant scale parameter, both [5] and [6] constructed optimal simple step-stress ALT designs for time-censoring. Bai and Kim in [5] obtained the optimal low stress level and optimal stress-changing time in order to minimize the AVAR of the MLE of a specific quantile of the product's lifetime distribution at the normal design stress level whereas [6] obtained their optimal stress-changing time in order to minimize the AVAR of the MLE for reliability prediction instead. In addition, Hunt and Xu in [7] further investigated optimal simple step-stress ALT plans for a Weibull distribution; however, they assumed both the shape and scale parameters were functions of the stress levels. Their resulting optimal designs chosen the stress-changing time in order to minimize AVAR of the MLE of reliability prediction at the normal design stress level and at a pre-specified time. They also reviewed the research work on optimal designs for step-stress ALT. Please see the references therein. We note that all these work previously done had provided the design construction methods only for simple step-stress ALT plans.

Moreover, Ma and Meeker in [8] extended the research by [5] to provide a general method for multiple step-stress ALTs assuming a log-location-scale family of distributions. They discussed an approach to calculate the large-sample approximate variance of the MLE for a percentile of the failure time distribution at normal design conditions when the failure data were observed from a step-stress ALT. By adopting a cumulative exposure model, their approach allowed for both multiple step-stress loading and censoring. Their results also showed that depending on the values of the model parameters and certain percentile of interest, one of the three test plans proposed could be the most preferable in terms of optimum variance. For a Weibull lifetime distribution, however, with possible inaccuracy in the assumed log-linear life-stress [2] investigated the optimal stress-changing time for simple step-stress ALT plans in order that the asymptotic mean squared error of the underlying reliability estimator could be minimized, and the robust choices of three-step-stress plans were also discussed with the awareness of possible imprecision in the assumed life-stress relationship by minimizing the asymptotic squared bias.

2.2. Optimal designs for constant stress or simple step-stress ALTs with a PH model

Jiao in [9] first investigated the optimal design problem for a PH model when a constant-stress ALT experiment being planned, and then developed the optimal designs for reliability prediction by optimally choosing both stress levels and proportion of units allocated to each stress level in order to attain the most accurate reliability estimate at normal design conditions. Moreover, when a step-stress ALT experiment was planned, [9] discussed the simple step-stress ALT plan for reliability prediction and obtained the optimal stress level by minimizing the variance of the MLE of hazard rate at the normal design stress level and over a pre-specified time period. In addition, [9] also provided an algorithm for solving the constrained nonlinear optimization problems.

On the other hand, Elsayed and Zhang in [10] revealed an optimal simple step-stress ALT plan so as to obtain the most accurate reliability function estimates at normal design conditions. They also formulated a nonlinear programming problem to minimize the asymptotic variance of the hazard rate estimator over a prespecified the period at the normal design stress level. More recently, Hu, Plante, and Tang in [11] briefly discussed the optimal low and high stress levels in a simple step-stress ALT in order to minimize the mean squared error of the estimated upper confidence bound for the cumulative failure probability of a product at normal design conditions, with a given stress-changing time. In sum, all these existing works provided the methods of optimal design construction only for simple step-stress ALT plans. Therefore, we expand the previous work of others and investigate the optimal designs of multiple step-stress ALT for PH models in this paper.

2.3. Optimal designs for general proportional hazards models

There is broader literature of optimal designs available for general proportional hazards models, which are not necessary with consideration of step-stress ALT plans. To name a few, Becker, McDonald, and Khoo in [12] constructed D-optimal designs for proportional hazards models with one or two parameters when its baseline hazard function was specified. They developed the

method of minimum variance design construction when various censoring schemes were adopted. Dette and Sahm in [13] provided a standardized maximum variance design criterion which could be applied to obtain the optimal designs whereas McGree and Eccleston in [14] created compound criteria so that optimal designs could be derived for multi-objective scenarios. L'opez-Fidalgo, Rivas-L'opez, and Del Campo in [15] also proposed an algorithm to find optimal designs for typical Cox regression models incorporating censoring. More recently, Konstantinou, Biedermann, and Kimber investigated the general maximin D-optimal designs for a class of models and discussed the application of their design construction method to the proportional hazards models in [16].

2.4. Our model and some preliminaries

One of the most commonly used means for predicting the lifetime of a product is Cox's PH model since it provides sufficient flexibility for identifying the effects of covariates on the failure rate. In this paper, a PH model with hazard ratio being independent of time is considered. Therefore, the hazard rate of a product can be expressed

$$\lambda(t; \mathbf{s}) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{s}), \quad (1)$$

where $\lambda_0(t)$ is a baseline hazard function, $\boldsymbol{\beta}$ is a column vector of unknown parameters, and \mathbf{s} is a column vector of the covariates (applied or transformed stresses) for an ALT experiment, and that are independent of the baseline hazard. Presumably, the stresses are having multiplicatively effects on the hazard rate in this model.

We conduct ALT with step-stress loading where all the test units begin at a prespecified stress level. After a time fraction, the stress level is changed to a higher stress level. The stress level can be raised more than once before the test completes. A simple step-stress ALT, also called a two-step-stress ALT, only uses two stress levels during an ALT experiment. In contrast, in multiple step-stress ALT, more than two stress levels are needed and the stress level is raised at least twice before the test ends. In this paper, we discuss multiple step-stress ALT experiments. We denote by s_D the normal design stress level. For a three-step-stress, we signify the low, middle and high stress levels by s_1 , s_2 , and s_3 , the stress-changing times by τ_1 and τ_2 , and censoring time by c .

In this paper, we assume $\lambda_0(t)$ to be in either a linear or a quadratic form. For either case, we investigate the optimal three-step-stress ALT design construction under D-optimality, A-optimality, and Q-optimality, respectively. We obtain D-optimal designs in order to minimize the determinant of covariance matrix of the estimators for the model parameters (or equivalently maximize the determinant of Fisher's information matrix), A-optimal designs in order to minimize the trace of covariance matrix of the estimators for the model parameters, and Q-optimal design to minimize the asymptotic variance of the estimator for a specific quantity of interest. The major quantity of interest in this paper is the average hazard rate over a particular period of time under normal design conditions.

For the case of $\lambda_0(t)$ being simple linear, the main part of the optimal design construction was presented in [1]. Therefore, we

will only restate the necessary notations being continuously used here and also summarize the previous results for comparison purpose in this paper. For the case of $\lambda_0(t)$ being a quadratic function, we will present the discussion and derivation in full. Furthermore, because there was no simulation study done previously for either case, we will demonstrate simulation and comparison studies for both cases in the present paper.

The rest of this paper is organized as follows: in Section 3, the optimal three-step-stress ALT design results have been summarized for the case when the baseline hazard function $\lambda_0(t)$ in (1) is assumed to be a simple linear function. In Section 4, the optimal three-step-stress ALT design have been derived for the case when $\lambda_0(t)$ is considered to be a quadratic function. The method of the optimal design construction involves the minimization for a nonlinear objective function with nonlinear constraints, and some practical examples are used to illustrate the proposed method for the construction of constrained optimal designs in Sections 3 and 4. We have also evaluated the performance of the resulting designs obtained in both Sections 3 and 4 through simulations and comparisons in Section 5. Some concluding remarks are presented in Section 6.

3. Optimal designs when $\lambda_0(t)$ being a simple linear function

Xu and Huang in [1] first focused on determining the optimal three-step-stress ALT designs for a PH model with time-censoring in the case when the baseline hazard function is considered being a simple linear function. They have provided detailed derivation for the information matrix and the optimal choices of both the middle stress level and stress-changing times for all cases considered there. For comparison reasons, we keep the same notations which are presented in Subsection 3.1, and their main results are summarized in this section.

3.1. Notation

Although the method of development for multiple step-stress ALT designs can be provided in general, we formulate our design construction using three-step-stress ALT experiments for its simplicity. A three-step-stress ALT experiment with time-censoring involves a predefined censoring time c , and three test stress levels s_1 , s_2 , and s_3 , which satisfies $s_D < s_1 < s_2 < s_3$. Assuming there are n test units available for the ALT experiment, they are all first placed at low stress level s_1 for a time interval $[0, \tau_1]$. Afterwards, the test units survived by time τ_1 are subjected to a middle stress level s_2 for the next time interval $(\tau_1, \tau_2]$. Next, the remaining units by time τ_2 are subjected to the highest stress level s_3 for the last time interval $(\tau_2, c]$. Then, the test ends at censoring time c . We consider Model (1), under a stress level s , where $\lambda_0(t)$ being independent of s .

We make use of the notations $F(t; s)$, $f(t; s)$, $\Lambda(t; s)$ $R(t; s)$

t , respectively, at a given stress level s . We also employ the cumulative exposure model (CEM), please see [17] for details, to address the changes

of $F(t; s)$, in lifetime t due to the stress step-ups in a step-stress ALT experiment. We denote the cdf under s_i by

$$F_i(t) = F(t | s = s_i), i = 1, 2, \text{ and } 3.$$

Then, the stepwise cdf with respect to t can be expressed by:

$$F_{CEM}(t) = \begin{cases} F_1(t) & \text{if } t \leq \tau_1, \\ F_2(a+t-\tau_1) & \text{if } \tau_1 \leq t \leq \tau_2, \\ F_3(b+t-\tau_2) & \text{if } \tau_2 \leq t \leq c, \\ P(t > c) & \text{if } t > c, \end{cases}$$

Where a and b satisfy $F_1(\tau_1) = F_2(a)$ and

$F_2(a + \tau_2 - \tau_1) = F_3(b)$. Namely, $a = F_2^{-1}[F_1(\tau_1)]$ and

$$b = F_3^{-1}[F_2(a + \tau_2 - \tau_1)].$$

We also make use of the following three indicators, each as a function of stress change times τ_1 , τ_2 , or censoring time c , and failure time t :

$$I_1 = I_1(t \leq \tau_1) = \begin{cases} 1 & \text{if } t \leq \tau_1, \\ 0 & \text{if } t > \tau_1, \end{cases} \quad I_2 = I_2(t \leq \tau_2) = \begin{cases} 1 & \text{if } t \leq \tau_2, \\ 0 & \text{if } t > \tau_2, \end{cases}$$

$$I_3 = I_3(t \leq c) = \begin{cases} 1 & \text{if } t \leq c, \\ 0 & \text{if } t > c, \end{cases}$$

where $0 < \tau_1 < \tau_2 < c$. Taking a three-step-stress design ξ with three stress levels being s_1 , s_2 , and s_3 , the log-likelihood function of an observed lifetime t can be written as:

$$\begin{aligned} \ln L(t; \xi) = & I_1 I_2 I_3 \left[\ln(\gamma_0 + \gamma_1 t) + \beta s_1 - \left(\gamma_0 t + \frac{\gamma_1}{2} t^2 \right) \exp(\beta s_1) \right] \\ & + (1 - I_1) I_2 I_3 \left[\ln(\gamma_0 + \gamma_1 x) + \beta s_2 - \left(\gamma_0 x + \frac{\gamma_1}{2} x^2 \right) \exp(\beta s_2) \right] \\ & + (1 - I_2) I_3 \left[\ln(\gamma_0 + \gamma_1 y) + \beta s_3 - \left(\gamma_0 y + \frac{\gamma_1}{2} y^2 \right) \exp(\beta s_3) \right] \\ & - (1 - I_3) \left[\left(\gamma_0 d + \frac{\gamma_1}{2} d^2 \right) \exp(\beta s_3) \right], \end{aligned}$$

where $x = a + t - \tau_1$, $y = b + t - \tau_2$, and $d = b + c - \tau_2$. (2)

As assumed the baseline hazard rate being independent of the covariates, according to [10], the Fisher's information matrix, F , for the full sample with size n can be obtained as

$$F = n \begin{bmatrix} E \left\{ -\frac{\partial^2 \ln L}{\partial \gamma_0^2} \right\} & E \left\{ -\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_1} \right\} & 0 \\ E \left\{ -\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_1} \right\} & E \left\{ -\frac{\partial^2 \ln L}{\partial \gamma_1^2} \right\} & 0 \\ 0 & 0 & E \left\{ -\frac{\partial^2 \ln L}{\partial \beta^2} \right\} \end{bmatrix},$$

and the covariance matrix, Σ , of the maximum likelihood estimators (MLE) of γ_0 , γ_1 , β is the inverse matrix of F :

$$\Sigma = F^{-1} = \begin{bmatrix} Var(\hat{\gamma}_0) & Cov(\hat{\gamma}_0, \hat{\gamma}_1) & 0 \\ Cov(\hat{\gamma}_0, \hat{\gamma}_1) & Var(\hat{\gamma}_1) & 0 \\ 0 & 0 & Var(\hat{\beta}) \end{bmatrix}.$$

The elements of the Fisher's information matrix for t are the negative expectations of the corresponding second partial derivatives. We denote the non-zero elements of Fisher's information by F_{00} , F_{11} , F_{01} , and F_{β} . In [1], Xu and Huang have derived the expressions of these elements as follows:

$$\begin{aligned} F_{00} = & E \left\{ -\frac{\partial^2 \ln L}{\partial \gamma_0^2} \right\} \\ = & \exp(\beta s_1) \int_0^{\tau_1} \frac{1}{\lambda_0(t)} \exp[-\Lambda(t; s_1)] dt \\ & + \exp(\beta s_2) \int_a^{a+\tau_2-\tau_1} \frac{1}{\lambda_0(x)} \exp[-\Lambda(x; s_2)] dx \\ & + \exp(\beta s_3) \int_b^{b+c-\tau_2} \frac{1}{\lambda_0(y)} \exp[-\Lambda(y; s_3)] dy, \end{aligned} \tag{3}$$

$$\begin{aligned} F_{11} = & E \left\{ -\frac{\partial^2 \ln L}{\partial \gamma_1^2} \right\} \\ = & \exp(\beta s_1) \int_0^{\tau_1} \frac{t^2}{\lambda_0(t)} \exp[-\Lambda(t; s_1)] dt \\ & + \exp(\beta s_2) \int_a^{a+\tau_2-\tau_1} \frac{x^2}{\lambda_0(x)} \exp[-\Lambda(x; s_2)] dx \\ & + \exp(\beta s_3) \int_b^{b+c-\tau_2} \frac{y^2}{\lambda_0(y)} \exp[-\Lambda(y; s_3)] dy, \end{aligned} \tag{4}$$

$$\begin{aligned} F_{01} = & E \left\{ -\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_1} \right\} \\ = & \exp(\beta s_1) \int_0^{\tau_1} \frac{t}{\lambda_0(t)} \exp[-\Lambda(t; s_1)] dt \\ & + \exp(\beta s_2) \int_a^{a+\tau_2-\tau_1} \frac{x}{\lambda_0(x)} \exp[-\Lambda(x; s_2)] dx \\ & + \exp(\beta s_3) \int_b^{b+c-\tau_2} \frac{t}{\lambda_0(y)} \exp[-\Lambda(y; s_3)] dy, \end{aligned} \tag{5}$$

and

$$\begin{aligned} F_{\beta} = & E \left\{ -\frac{\partial^2 \ln L}{\partial \beta^2} \right\} = s_1^2 \exp(\beta s_1) \int_0^{\tau_1} \lambda_0(t) \Lambda(t; s_1) \exp[-\Lambda(t; s_1)] dt \\ & + s_2^2 \exp(\beta s_2) \int_a^{a+\tau_2-\tau_1} \lambda_0(x) \Lambda(x; s_2) \exp[-\Lambda(x; s_2)] dx \\ & + s_3^2 \exp(\beta s_3) \int_b^{b+c-\tau_2} \lambda_0(y) \Lambda(y; s_3) \exp[-\Lambda(y; s_3)] dy + s_3^2 \Lambda(d; s_3). \end{aligned} \tag{6}$$

We note that the derivation of these elements in this subsection and the expressions of the loss functions in next subsection have been omitted, please see [1] for details.

3.2. Loss functions and design constraints

In this paper, we consider the following three different design criteria: D-optimality, A-optimality, and Q-optimality. In [1], Xu and Huang has derived the corresponding three loss functions as

$$L_D = \frac{1}{(F_{00}F_{11} - F_{01}^2)F_\beta}, \tag{7}$$

$$L_A = \frac{F_{00} + F_{11}}{F_{00}F_{11} - F_{01}^2} + \frac{1}{F_\beta}, \tag{8}$$

and

$$L_Q = \frac{F_{11}T - F_{01}T^2 + \frac{1}{3}F_{00}T^3}{F_{00}F_{11} - F_{01}^2} + \frac{s_D^2(\gamma_0^2T + \gamma_0\gamma_1T^2 + \frac{1}{3}\gamma_1^2T^3)}{F_\beta}, \tag{9}$$

respectively.

In ALT practice, there are often some requirement needed on the minimum number of failures at each stress level. Similarly to [10] for designing optimal simple step-stress ALTs, we take certain practical constraints into consideration in our optimal design construction process for multiple step-stress ALTs. In this paper, we consider three design constraints as below:

- The minimum expected number of failures (MNF) at stress level s_1 is required as w_1 :

$$n \Pr[t \leq \tau_1 | s_1] \geq w_1; \tag{10}$$

- The MNF at stress level s_2 is given as w_2 :

$$(n - n_1) \Pr[a + t - \tau_1 \leq \tau_2 | s_2] \geq w_2; \tag{11}$$

- The MNF at stress level s_3 is given as w_3 :

$$(n - n_1 - n_2) \Pr[b + t - \tau_2 \leq c | s_3] \geq w_3, \tag{12}$$

where n_i is the number of failures under each stress level s_i , $i = 1, 2, 3$.

With these constraints, the optimal decision variables (τ_1 , τ_2 , s_2) can be determined by minimizing each of the above-mentioned loss functions. Specifically, we choose the optimal designs in order to minimize (7), (8), or (9) with respect to (τ_1 , τ_2 , s_2) subject to all three constraints (10), (11), and (12). We denote the optimal designs corresponding to the three optimality criteria by $^D\xi$, $^A\xi$, and $^Q\xi$, which can be expressed as

$$^D\xi = \arg \{ \min(L_D | \text{given (10), (11), and (12)}) \}, \tag{13}$$

$$^A\xi = \arg \{ \min(L_A | \text{given (10), (11), and (12)}) \}, \tag{14}$$

$$\text{and } ^Q\xi = \arg \{ \min(L_Q | \text{given (10), (11), and (12)}) \}. \tag{15}$$

3.3. Optimal designs when the middle stress being fixed

In [10], Elsayed and Zhang discussed an example of conducting a two-step-stress ALT experiment for metal oxide semiconductor capacitors and estimating the hazard rate over 10 years at design temperature $50^\circ C$. The total number of test units was $n = 200$ and the test was censored by $c = 300$ hours. In order to avoid any unanticipated change in failure mechanisms during the ALT experiment, the maximum testing temperature was determined to be $250^\circ C$ by the engineering experimenters. The initial values of the model parameters was taken as $\gamma_0 = 0.0001$, $\gamma_1 = 0.5$, and $\beta = -3800$. The Q-optimal low accelerated stress level was found to be $145^\circ C$.

In this subsection, we use this example to compare the optimal two-step-stress ALT designs obtained in [10] with our proposed three-step-stress ALT designs. Previously, the optimal two-step-stress ALT designs are having $s_1 = 145^\circ C$ and $s_2 = 250^\circ C$. For our three-step-stress ALT plans, we keep the number of test units, censoring time, and the lowest and highest stress levels all the same. Namely, $n = 200$, $c = 300$, $s_1 = 145^\circ C$ and $s_3 = 250^\circ C$. The initial values for the model parameters are kept the same as well. Conveniently, the experimenters often take s_2 as the average of s_1 and s_3 . Therefore, we discuss two scenarios for the choice of s_2 : (a) being $197.5^\circ C$, which is the average of $s_1 = 145^\circ C$ and $s_3 = 250^\circ C$; and (b) being optimally chosen, together with stress changing times, in order to minimize a specified loss function. We note that when temperature is taken as a stress factor appeared in a PH model for ALT, absolute temperature is often used as its measurement unit for model fitting.

Table 1. Constraint cases for two-step-stress ALT

Notation	Constraint parameters	MNF
C_{12}	$w_1 = 40, w_2 = 30$	70
C_{34}	$w_1 = 30, w_2 = 40$	70
C_{55}	$w_1 = 40, w_2 = 20$	60
C_{66}	$w_1 = 20, w_2 = 40$	60

Table 2. Constraint cases for three-step-stress ALT

Notation	Constraint parameters	MNF
C_1	$w_1 = 40, w_2 = 20, w_3 = 10$	70
C_2	$w_1 = 40, w_2 = 15, w_3 = 15$	70
C_3	$w_1 = 30, w_2 = 30, w_3 = 10$	70
C_4	$w_1 = 30, w_2 = 20, w_3 = 20$	70
C_5	$w_1 = 40, w_2 = 10, w_3 = 10$	60
C_6	$w_1 = 20, w_2 = 20, w_3 = 20$	60

In Scenario (a), the optimal stress changing times, τ_1 and τ_2 , can be chosen in order to minimize L_D , L_A , or L_Q under some specific constraints. We take several practical constraint cases with given w_1 , w_2 and w_3 , and these cases accompanied by their notations are recorded in Table 1 for two-step-stress ALT designs, and in Table 2 for three-step-stress ALT designs.

We denote the D-, A- and Q-optimal designs obtained for two-step-stress ALT under a given constraint case C_k by ${}^D \xi_{C_k}^2$, ${}^A \xi_{C_k}^2$, and ${}^Q \xi_{C_k}^2$, where $k = 12, 34, 55, 66$, and the D-, A- and Q-optimal designs obtained for three-step-stress ALT under a given constraint C_k by ${}^D \xi_{C_k}^{3(i)}$, ${}^A \xi_{C_k}^{3(i)}$, and ${}^Q \xi_{C_k}^{3(i)}$, where $k = 1, \dots, 6$, and $i = 1, 2$ with i referring to the different scenario of s_i ($i = 1$ for Scenario (a), and $i = 2$ for Scenario (b)). Moreover, we define the asymptotic D-, A-, Q-efficiencies of $\xi_{C_k}^{3(i)}$ relative to $\xi_{C_k}^2$ as

$${}^D \text{eff}(3(i), 2) = \frac{L_D({}^D \xi_{C_k}^2)}{L_D({}^D \xi_{C_k}^{3(i)})}$$

$${}^A \text{eff}(3(i), 2) = \frac{L_A({}^A \xi_{C_k}^2)}{L_A({}^A \xi_{C_k}^{3(i)})}, \text{ and}$$

$${}^Q \text{eff}(3(i), 2) = \frac{L_Q({}^Q \xi_{C_k}^2)}{L_Q({}^Q \xi_{C_k}^{3(i)})}, \text{ for } i = 1, 2;$$

and the asymptotic D-, A-, Q-efficiencies of $\xi_{C_k}^{3(i)}$ relative to $\xi_{C_k}^{3(j)}$ as

$${}^D \text{eff}(3(i), 3(j)) = \frac{L_D({}^D \xi_{C_k}^{3(j)})}{L_D({}^D \xi_{C_k}^{3(i)})}$$

$${}^A \text{eff}(3(i), 3(j)) = \frac{L_A({}^A \xi_{C_k}^{3(j)})}{L_A({}^A \xi_{C_k}^{3(i)})}, \text{ and}$$

$${}^Q \text{eff}(3(i), 3(j)) = \frac{L_Q({}^Q \xi_{C_k}^{3(j)})}{L_Q({}^Q \xi_{C_k}^{3(i)})}, \text{ for } i, j = 1, 2.$$

The optimal stress-changing time(s), for both two-step-stress plans and the three-step-stress ALT plans with s_2 being fixed at $197.5^\circ C$, are obtained by minimizing L_D , L_A , or L_Q , respectively for all the constraint cases considered. We note that the resulting optimal stress-changing times are all the same for these three different optimal criteria although their corresponding relative efficiencies are not the same. The optimal stress-changing times and relative efficiencies for all six constraint cases (as listed in Table 2) are presented in Tables 3 and 4. The overall average efficiency gain after unitizing an optimal three-step-stress plan is 5.44% when s_2 being fixed at $197.5^\circ C$.

4. Optimal designs when baseline hazard is a quadratic function

4.1. Preliminary

We have considered the PH based ALT with the case where the baseline hazard function is a simple linear function in Section 3. In

many practical situations, this simple model can be under-fitted. Therefore, in this section, we focus on the PH model when the baseline hazard function is in a quadratic form. Namely, we adopt Model (m1) where the baseline hazard function is rather being

$$\lambda_0(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2. \tag{16}$$

Table 3. D-, A-, Q-optimal stress-changing times, when $s_2 = 197.5^\circ C$

3-step-stress ALT			2-step-stress ALT	
optimal designs	τ_1	τ_2	optimal designs	τ_1
${}^{D,A,Q} \xi_{C_1}^{3(1)}$	178.2	223.6	${}^{D,A,Q} \xi_{C_{12}}^2$	183
${}^{D,A,Q} \xi_{C_2}^{3(1)}$	178.0	209.2	${}^{D,A,Q} \xi_{C_{12}}^2$	183
${}^{D,A,Q} \xi_{C_3}^{3(1)}$	162.9	219.4	${}^{D,A,Q} \xi_{C_{34}}^2$	156
${}^{D,A,Q} \xi_{C_4}^{3(1)}$	162.1	194.8	${}^{D,A,Q} \xi_{C_{34}}^2$	156
${}^{D,A,Q} \xi_{C_5}^{3(1)}$	198.0	227.5	${}^{D,A,Q} \xi_{C_{55}}^2$	201
${}^{D,A,Q} \xi_{C_6}^{3(1)}$	162.1	194.8	${}^{D,A,Q} \xi_{C_{66}}^2$	156

Table 4. D-, A-, Q- relative efficiencies, when $s_2 = 197.5^\circ C$

	C_1	C_2	C_3	C_4	C_5	C_6
${}^D \text{eff}(3(1), 2)$	1.06	1.04	1.10	1.06	1.04	1.06
${}^A \text{eff}(3(1), 2)$	1.06	1.03	1.09	1.06	1.04	1.06
${}^Q \text{eff}(3(1), 2)$	1.04	1.03	1.08	1.05	1.03	1.05

Consequently, the cdf, pdf, the cumulative hazard function, and the reliability function of failure time t , at a given stress level s , become

$$F(t; s) = 1 - \exp \left[- \left(\gamma_0 t + \frac{\gamma_1}{2} t^2 + \frac{\gamma_2}{3} t^3 \right) \exp(\beta s) \right],$$

$$f(t; s) = \lambda_0(t) \exp(\beta s) \exp \left[- \left(\gamma_0 t + \frac{\gamma_1}{2} t^2 + \frac{\gamma_2}{3} t^3 \right) \exp(\beta s) \right],$$

$$\Lambda(t; s) = \left(\gamma_0 t + \frac{\gamma_1}{2} t^2 + \frac{\gamma_2}{3} t^3 \right) \exp(\beta s), \text{ and}$$

$$R(t; s) = \exp \left[- \left(\gamma_0 t + \frac{\gamma_1}{2} t^2 + \frac{\gamma_2}{3} t^3 \right) \exp(\beta s) \right]. \tag{17}$$

Then, the log-likelihood function of t , under a three-step-stress design ξ with stress levels being s_1 , s_2 , and s_3 , become

$$\ln L(t; \xi) = I_1 I_2 I_3 \left[\ln \left(\gamma_0 + \gamma_1 t + \gamma_2 t^2 \right) + \beta s_1 - \left(\gamma_0 t + \frac{\gamma_1}{2} t^2 + \frac{\gamma_2}{3} t^3 \right) \exp(\beta s_1) \right]$$

$$+ (1 - I_1) I_2 I_3 \left[\ln \left(\gamma_0 + \gamma_1 x + \gamma_2 x^2 \right) + \beta s_2 - \left(\gamma_0 x + \frac{\gamma_1}{2} x^2 + \frac{\gamma_2}{3} x^3 \right) \exp(\beta s_2) \right]$$

$$+ (1 - I_2) I_3 \left[\ln \left(\gamma_0 + \gamma_1 y + \gamma_2 y^2 \right) + \beta s_3 - \left(\gamma_0 y + \frac{\gamma_1}{2} y^2 + \frac{\gamma_2}{3} y^3 \right) \exp(\beta s_3) \right]$$

$$- (1 - I_3) \left[\left(\gamma_0 d + \frac{\gamma_1}{2} d^2 + \frac{\gamma_2}{3} d^3 \right) \exp(\beta s_3) \right],$$

where $x, y,$ and d are defined as in (2). Its first partial derivatives with respect to the model parameters $\gamma_0, \gamma_1,$ and β are the same as in (3) but with $\lambda_0(t)$ being defined as in (16) instead, and the one with respect to γ_2 is

$$\frac{\partial \ln L}{\partial \gamma_2} = I_1 I_2 I_3 t^2 \left[\frac{1}{\lambda_0(t)} - \frac{t}{3} \exp(\beta s_1) \right] + (1 - I_1) I_2 I_3 x^2 \left[\frac{1}{\lambda_0(x)} - \frac{x}{3} \exp(\beta s_2) \right] + (1 - I_2) I_3 y^2 \left[\frac{1}{\lambda_0(y)} - \frac{y}{3} \exp(\beta s_3) \right] - (1 - I_3) \frac{d^3}{3} \exp(\beta s_3).$$

Table 5. D-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			D-relative efficiency	
Optimal design	τ_1	τ_2	${}^D \text{eff}(3(2), 2)$	${}^D \text{eff}(3(2), 3(1))$
${}^D \xi_{C_1}^{3(2)}$	152.65	247.44	1.13	1.07
${}^D \xi_{C_2}^{3(2)}$	148.33	226.85	1.09	1.05
${}^D \xi_{C_3}^{3(2)}$	152.93	247.44	1.19	1.09
${}^D \xi_{C_4}^{3(2)}$	143.25	211.74	1.12	1.06
${}^D \xi_{C_5}^{3(2)}$	179.40	245.78	1.08	1.04
${}^D \xi_{C_6}^{3(2)}$	161.43	211.03	1.11	1.05

Table 6. A-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			A-relative efficiency	
Optimal design	τ_1	τ_2	${}^A \text{eff}(3(2), 2)$	${}^A \text{eff}(3(2), 3(1))$
${}^A \xi_{C_1}^{3(2)}$	155.20	247.47	1.13	1.07
${}^A \xi_{C_2}^{3(2)}$	147.99	226.84	1.09	1.06
${}^A \xi_{C_3}^{3(2)}$	155.13	247.47	1.19	1.09
${}^A \xi_{C_4}^{3(2)}$	142.41	211.73	1.11	1.05
${}^A \xi_{C_5}^{3(2)}$	155.18	247.47	1.09	1.05
${}^A \xi_{C_6}^{3(2)}$	142.39	211.73	1.11	1.05

Table 7. Q-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			D-relative efficiency	
Optimal design	τ_1	τ_2	${}^Q \text{eff}(3(2), 2)$	${}^Q \text{eff}(3(2), 3(1))$
${}^Q \xi_{C_1}^{3(2)}$	156.19	247.47	1.09	1.05
${}^Q \xi_{C_2}^{3(2)}$	149.81	226.86	1.07	1.04
${}^Q \xi_{C_3}^{3(2)}$	156.71	247.47	1.14	1.06
${}^Q \xi_{C_4}^{3(2)}$	144.20	211.75	1.09	1.04
${}^Q \xi_{C_5}^{3(2)}$	156.45	247.47	1.06	1.03
${}^Q \xi_{C_6}^{3(2)}$	143.98	211.75	1.09	1.04

Its second partial derivatives $\frac{\partial^2 \ln L}{\partial \gamma_0^2}, \frac{\partial^2 \ln L}{\partial \gamma_1^2}, \frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_1},$ and $\frac{\partial^2 \ln L}{\partial \beta^2}$ can be expressed as the same as in Xu and Huang (2018) but with their λ_0 and Λ function being defined as in (16) and (17), and their corresponding elements of the Fisher's information matrix for a single failure time t (the expected value of negative second derivatives) are as the same as in (3), (4), (5), and (6). The remaining second derivatives are

$$\frac{\partial^2 \ln L}{\partial \gamma_2^2} \approx -\frac{I_1 I_2 I_3 t^4}{\lambda_0^2(t)} - \frac{(1 - I_1) I_2 I_3 x^4}{\lambda_0^2(x)} - \frac{(1 - I_2) I_3 y^4}{\lambda_0^2(y)},$$

$$\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_2} \approx -\frac{I_1 I_2 I_3 t^2}{\lambda_0^2(t)} - \frac{(1 - I_1) I_2 I_3 x^2}{\lambda_0^2(x)} - \frac{(1 - I_2) I_3 y^2}{\lambda_0^2(y)},$$

$$\frac{\partial^2 \ln L}{\partial \gamma_1 \gamma_2} \approx -\frac{I_1 I_2 I_3 t^3}{\lambda_0^2(t)} - \frac{(1 - I_1) I_2 I_3 x^3}{\lambda_0^2(x)} - \frac{(1 - I_2) I_3 y^3}{\lambda_0^2(y)},$$

and their corresponding elements of the Fisher's information matrix for a single t can be derived as below:

$$F_{22} = E \left\{ -\frac{\partial^2 \ln L}{\partial \gamma_2^2} \right\} = \int_0^{\tau_1} \frac{t^4}{\lambda_0^2(t)} f(t; s_1) dt + \int_a^{a+\tau_2-\tau_1} \frac{x^4}{\lambda_0^2(x)} f(x; s_2) dx + \int_b^{b+c-\tau_2} \frac{y^4}{\lambda_0^2(y)} f(y; s_3) dy$$

$$= \exp(\beta s_1) \int_0^{\tau_1} \frac{t^4}{\lambda_0(t)} R(t; s_1) dt + \exp(\beta s_2) \int_a^{a+\tau_2-\tau_1} \frac{x^4}{\lambda_0(x)} R(x; s_2) dx + \exp(\beta s_3) \int_b^{b+c-\tau_2} \frac{y^4}{\lambda_0(y)} R(y; s_3) dy,$$

$$F_{02} = E \left\{ -\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_2} \right\} = \int_0^{\tau_1} \frac{t^2}{\lambda_0^2(t)} f(t; s_1) dt + \int_a^{a+\tau_2-\tau_1} \frac{x^2}{\lambda_0^2(x)} f(x; s_2) dx + \int_b^{b+c-\tau_2} \frac{y^2}{\lambda_0^2(y)} f(y; s_3) dy = \exp(\beta s_1) \int_0^{\tau_1} \frac{t^2}{\lambda_0(t)} R(t; s_1) dt + \exp(\beta s_2) \int_a^{a+\tau_2-\tau_1} \frac{x^2}{\lambda_0(x)} R(x; s_2) dx + \exp(\beta s_3) \int_b^{b+c-\tau_2} \frac{y^2}{\lambda_0(y)} R(y; s_3) dy,$$

and

$$F_{12} = E \left\{ -\frac{\partial^2 \ln L}{\partial \gamma_1 \gamma_2} \right\} = \int_0^{\tau_1} \frac{t^3}{\lambda_0^2(t)} f(t; s_1) dt + \int_a^{a+\tau_2-\tau_1} \frac{x^3}{\lambda_0^2(x)} f(x; s_2) dx + \int_b^{b+c-\tau_2} \frac{y^3}{\lambda_0^2(y)} f(y; s_3) dy = \exp(\beta s_1) \int_0^{\tau_1} \frac{t^3}{\lambda_0(t)} R(t; s_1) dt + \exp(\beta s_2) \int_a^{a+\tau_2-\tau_1} \frac{x^3}{\lambda_0(x)} R(x; s_2) dx + \exp(\beta s_3) \int_b^{b+c-\tau_2} \frac{y^3}{\lambda_0(y)} R(y; s_3) dy,$$

with R function being defined as in (17).

As indicated in Section 3.1, the correlations between the stress coefficient β and baseline parameters γ_0, γ_1 and γ_2 are equal to zero. Consequently, the covariance matrix Σ , of the MLEs of $\gamma_0, \gamma_1, \gamma_2, \beta$ can be expressed as

$$\Sigma = \mathbf{F}^{-1} = \begin{bmatrix} \text{Var}(\hat{\gamma}_0) & \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_1) & \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_2) & 0 \\ \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_1) & \text{Var}(\hat{\gamma}_1) & \text{Cov}(\hat{\gamma}_1, \hat{\gamma}_2) & 0 \\ \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_2) & \text{Cov}(\hat{\gamma}_1, \hat{\gamma}_2) & \text{Var}(\hat{\gamma}_2) & 0 \\ 0 & 0 & 0 & \text{Var}(\hat{\beta}) \end{bmatrix},$$

where \mathbf{F} is the Fisher's information matrix of the full sample with size n and

$$\mathbf{F} = n \begin{bmatrix} E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0^2}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0 \partial \gamma_1}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0 \partial \gamma_2}\right\} & 0 \\ E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0 \partial \gamma_1}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_1^2}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_1 \partial \gamma_2}\right\} & 0 \\ E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0 \partial \gamma_2}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_1 \partial \gamma_2}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_2^2}\right\} & 0 \\ 0 & 0 & 0 & E\left\{-\frac{\partial^2 \ln L}{\partial \beta^2}\right\} \end{bmatrix} = n \begin{bmatrix} F_{00} & F_{01} & F_{02} & 0 \\ F_{01} & F_{11} & F_{12} & 0 \\ F_{02} & F_{12} & F_{22} & 0 \\ 0 & 0 & 0 & F_{\beta} \end{bmatrix}.$$

4.2. Loss functions and design constraints

Now the loss functions under D-, A-, and Q-optimality become

$$L_D = \det n \begin{bmatrix} \text{Var}(\hat{\gamma}_0) & \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_1) & \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_2) & 0 \\ \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_1) & \text{Var}(\hat{\gamma}_1) & \text{Cov}(\hat{\gamma}_1, \hat{\gamma}_2) & 0 \\ \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_2) & \text{Cov}(\hat{\gamma}_1, \hat{\gamma}_2) & \text{Var}(\hat{\gamma}_2) & 0 \\ 0 & 0 & 0 & \text{Var}(\hat{\beta}) \end{bmatrix} = \det \begin{bmatrix} F_{00} & F_{01} & F_{02} & 0 \\ F_{01} & F_{11} & F_{12} & 0 \\ F_{02} & F_{12} & F_{22} & 0 \\ 0 & 0 & 0 & F_{\beta} \end{bmatrix}^{-1} = \frac{1}{(F_{00}F_{11}F_{22} + 2F_{01}F_{02}F_{12} - F_{00}F_{12}^2 - F_{11}F_{02}^2 - F_{22}F_{01}^2)F_{\beta}}, \tag{18}$$

$$L_A = \text{tr} n \begin{bmatrix} \text{Var}(\hat{\gamma}_0) & \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_1) & \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_2) & 0 \\ \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_1) & \text{Var}(\hat{\gamma}_1) & \text{Cov}(\hat{\gamma}_1, \hat{\gamma}_2) & 0 \\ \text{Cov}(\hat{\gamma}_0, \hat{\gamma}_2) & \text{Cov}(\hat{\gamma}_1, \hat{\gamma}_2) & \text{Var}(\hat{\gamma}_2) & 0 \\ 0 & 0 & 0 & \text{Var}(\hat{\beta}) \end{bmatrix} = \text{tr} \begin{bmatrix} F_0 & F_{01} & F_{02} & 0 \\ F_{01} & F_1 & F_{12} & 0 \\ F_{02} & F_{12} & F_2 & 0 \\ 0 & 0 & 0 & F_{\beta} \end{bmatrix}^{-1} = \frac{F_{00}F_{11} + F_{00}F_{22} + F_{11}F_{22} - F_{01}^2 - F_{02}^2 - F_{12}^2}{F_{00}F_{11}F_{22} + 2F_{01}F_{02}F_{12} - F_{00}F_{12}^2 - F_{11}F_{02}^2 - F_{22}F_{01}^2} + \frac{1}{F_{\beta}}, \tag{19}$$

and

$$L_Q = \exp(-2\beta s_D) \int_0^T n \text{Var}[\lambda(t; s_D)] dt = \exp(-2\beta s_D) \int_0^T n \text{Var}\left[\left(\hat{\gamma}_0 + \hat{\gamma}_1 t + \hat{\gamma}_2 t^2\right) \exp \beta s_D\right] dt = \exp(-2\beta s_D) \int_0^T \left[\frac{\partial \lambda}{\partial \gamma_0} \frac{\partial \lambda}{\partial \gamma_1} \frac{\partial \lambda}{\partial \gamma_2} \frac{\partial \lambda}{\partial \beta} \right] \mathbf{F}^{-1} \left[\frac{\partial \lambda}{\partial \gamma_0} \frac{\partial \lambda}{\partial \gamma_1} \frac{\partial \lambda}{\partial \gamma_2} \frac{\partial \lambda}{\partial \beta} \right]^T dt = \int_0^T \left[\frac{(F_{11}F_{22} - F_{12}^2) - 2t(F_{01}F_{22} - F_{02}F_{12}) + t^2(2F_{01}F_{12} - 2F_{11}F_{02} + F_{00}F_{22} - F_{02}^2) - 2t^3(F_{00}F_{12} - F_{01}F_{02}) + t^4(F_{00}F_{11} - F_{01}^2)}{F_{00}F_{11}F_{22} + 2F_{01}F_{02}F_{12} - F_{00}F_{12}^2 - F_{11}F_{02}^2 - F_{22}F_{01}^2} + \frac{s_D^2(\gamma_0 + \gamma_1 t + \gamma_2 t^2)^2}{F_{\beta}} \right] dt = \frac{(F_{11}F_{22} - F_{12}^2)T - (F_{01}F_{22} - F_{02}F_{12})T^2 + (2F_{01}F_{12} - 2F_{11}F_{02} + F_{00}F_{22} - F_{02}^2)T^3 - \frac{1}{2}(F_{00}F_{12} - F_{01}F_{02})T^4 + \frac{1}{3}(F_{00}F_{11} - F_{01}^2)T^5}{F_{00}F_{11}F_{22} + 2F_{01}F_{02}F_{12} - F_{00}F_{12}^2 - F_{11}F_{02}^2 - F_{22}F_{01}^2} + \frac{s_D^2\left(\frac{\gamma_0^2}{2}T + \gamma_0\gamma_1T^2 + \frac{\gamma_1^2 + 2\gamma_0\gamma_2}{3}T^3 + \frac{\gamma_2^2}{2}T^4 + \frac{\gamma_2^3}{3}T^5\right)}{F_{\beta}}, \tag{20}$$

respectively.

The optimal decision variables (τ_1, τ_2, s_2) are chosen by minimizing the loss function (18), (19), or (20) with some design constraints. We also keep the three constraints (10), (11), and (12) as the same as those in Section 3. The corresponding designs $^D \xi, ^A \xi$ and $^Q \xi$ can also be described as (13), (14), and (15).

4.3. Optimal three-step-stress designs

In this subsection, we revisit the example presented in Section 3, and discuss the optimal ALT designs when the fitting model is a PH model with a quadratic baseline hazard function. Suppose from previous experience, the initial values for the model parameters are $\gamma_0 = 0.0001, \gamma_1 = 0.5, \gamma_2 = 0, \beta = -3800$. All other values of the design parameters in the example remain the same. Thus, the accelerated stress levels remain as $s_1 = 145^\circ C, s_3 = 250^\circ C$. Here we also consider the six constraint cases as used in Section 3.

When s_2 is conveniently chosen as $197.5^\circ C$, the resulting two-step-stress and three-step-stress optimal designs under D-, A, and Q-optimality criteria appeared to be the same again. Nonetheless, the relative efficiencies are not always better than their corresponding peers, the optimal two-step-stress designs. These designs and their relative efficiencies under D-, A-, and Q-optimality are displayed in Tables 8 and 9, respectively.

Table 8. D-, A-, Q-optimal stress-changing times, when $s_2 = 197.5^\circ C$

3-step-stress ALT			2-step-stress ALT	
optimal designs	τ_1	τ_2	optimal designs	τ_1
$^{D,A,Q} \xi_{C_1}^{3(1)}$	178.2	223.6	$^{D,A,Q} \xi_{C_{12}}^2$	183
$^{D,A,Q} \xi_{C_2}^{3(1)}$	178.0	209.2	$^{D,A,Q} \xi_{C_{12}}^2$	183
$^{D,A,Q} \xi_{C_3}^{3(1)}$	162.9	219.4	$^{D,A,Q} \xi_{C_{34}}^2$	156
$^{D,A,Q} \xi_{C_4}^{3(1)}$	162.1	194.8	$^{D,A,Q} \xi_{C_{34}}^2$	156
$^{D,A,Q} \xi_{C_5}^{3(1)}$	198.0	227.5	$^{D,A,Q} \xi_{C_{55}}^2$	201
$^{D,A,Q} \xi_{C_6}^{3(1)}$	162.1	194.8	$^{D,A,Q} \xi_{C_{66}}^2$	156

Table 9. D-, A-, Q- relative efficiencies, when $s_2 = 197.5^\circ C$

	C_1	C_2	C_3	C_4	C_5	C_6
${}^D eff(3(1),2)$	0.93	0.89	1.14	1.0	0.94	1.0
${}^A eff(3(1),2)$	1.06	1.03	1.09	1.06	1.04	1.06
${}^Q eff(3(1),2)$	0.88	0.86	1.04	0.94	0.91	0.94

Table 10. D-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			D-relative efficiency	
Optimal design	τ_1	τ_2	${}^D eff(3(2),2)$	${}^D eff(3(2),3(1))$
${}^D \xi_{C_1}^{3(2)}$	178.05	245.70	1.91	2.05
${}^D \xi_{C_2}^{3(2)}$	175.60	220.01	1.42	1.60
${}^D \xi_{C_3}^{3(2)}$	145.65	247.10	2.80	2.46
${}^D \xi_{C_4}^{3(2)}$	154.65	210.94	1.79	1.79
${}^D \xi_{C_5}^{3(2)}$	179.37	245.61	1.47	1.90
${}^D \xi_{C_6}^{3(2)}$	154.65	210.94	1.79	1.79

Table 11. A-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			A-relative efficiency	
Optimal design	τ_1	τ_2	${}^A eff(3(2),2)$	${}^A eff(3(2),3(1))$
${}^A \xi_{C_1}^{3(2)}$	155.20	247.47	1.13	1.07
${}^A \xi_{C_2}^{3(2)}$	147.99	226.84	1.09	1.06
${}^A \xi_{C_3}^{3(2)}$	155.13	247.47	1.19	1.09
${}^A \xi_{C_4}^{3(2)}$	142.41	211.73	1.11	1.05
${}^A \xi_{C_5}^{3(2)}$	155.18	247.47	1.09	1.05
${}^A \xi_{C_6}^{3(2)}$	142.39	211.73	1.11	1.05

Very similar to the results found in Section 3 when the baseline hazard function being simple linear, the resulting designs with an optimal middle stress level provide much more efficiency gains than those with a fixed middle stress level when the baseline hazard function being quadratic. The optimal stress-changing times and the optimal middle stress levels, under D-, A, and Q-optimality criteria, for Model (1) with (16) are displayed in Tables 10, 11, and 12, respectively. These tables also include the efficiencies of the resulting optimal three-step-stress designs relative to both their corresponding

optimal two-step-stress designs and optimal three-step-stress designs listed in Table 8, under D-, A-, and Q-optimality.

Table 12. Q-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			Q-relative efficiency	
Optimal design	τ_1	τ_2	${}^Q eff(3(2),2)$	${}^Q eff(3(2),3(1))$
${}^Q \xi_{C_1}^{3(2)}$	178.15	245.96	1.71	1.94
${}^Q \xi_{C_2}^{3(2)}$	177.96	224.53	1.39	1.62
${}^Q \xi_{C_3}^{3(2)}$	148.51	247.29	2.36	2.27
${}^Q \xi_{C_4}^{3(2)}$	162.04	210.97	1.62	1.72
${}^Q \xi_{C_5}^{3(2)}$	198.03	241.97	1.37	1.51
${}^Q \xi_{C_6}^{3(2)}$	162.04	210.97	1.62	1.72

We note that these three-step-stress ALT designs with optimal middle stress level have largely reduced the loss function for all cases. From Tables 10-12, the efficiency gains among all the cases considered are of a minimum of 9% and a maximum of 180% with respect to the optimal two-step-stress designs. The overall average efficiency gain is as high as 55.4% over the optimal two-step-stress designs and 59.7% over the optimal three-step-stress designs with a fixed middle stress. Such efficiency gains are much higher than the efficiency gains when adopting a PH model with a simple linear baseline function. The resulting optimal stress-changing times vary under three criteria, but all the optimal middle stress levels are equal being the lower bound of s_2 . For all three criteria, the most efficient design occurs when the constraint case is C_3 . Therefore, if the experimenter is uncertain which constraints should be applied, we would recommend to apply optimal designs ${}^D \xi_{C_3}^{3(2)}$ or ${}^A \xi_{C_3}^{3(2)}$ for better model parameter estimation and ${}^Q \xi_{C_3}^{3(2)}$ for more accurate hazard rate prediction.

5. Simulations

In order to demonstrate the performance of resulting designs obtained in Sections 3 and 4 with a given sample size, we carry out a simulation study. We first provide a procedure to simulate data from a given step-stress ALT experimental design, then we examine and compare the performances of the optimal designs constructed by the previous sections.

5.1. A simulation procedure

The following four steps describe our simulation procedure:

Step 1: Generating the number of failures and the failure times under the low stress level, s_1 :

Given the initial values of true model parameters with the stress levels and stress-changing times of an optimal design, we can use a binomial distribution to simulate the data. We divide $[0, \tau_1]$, τ_1 , into m_1 subintervals: $[t_{10}, t_{11}]$, $(t_{11}, t_{12}]$, ..., $(t_{1, m_1-1}, t_{1, m_1}]$, where

$0 = t_{10} < t_{11} < \dots < t_{1m_1} = \tau_1$. Let X_i be the failure number over the i th subinterval $(t_{1,i-1}, t_{1i}]$, we assume that the distribution of X_i is a binomial distribution; namely,

$$X_i \sim B\left(n - \sum_{j=1}^{i-1} X_j, p_{1i}\right), i = 1, \dots, m_1,$$

where p_{1i} is the failure rate within the i th interval $(t_{1,i-1}, t_{1i}]$ under the low stress level, s_1 . Then, p_{1i} can be derived as follows:

$$p_{1i} = P\{t_{1,i-1} < t \leq t_{1i} | s_1\} = R(t_{1,i-1}; s_1) - R(t_{1i}; s_1)$$

$$= \exp\left[-\left(\gamma_0 t_{1,i-1} + \frac{\gamma_1}{2} t_{1,i-1}^2 + \frac{\gamma_2}{3} t_{1,i-1}^3\right) \exp(\beta s_1)\right] - \exp\left[-\left(\gamma_0 t_{1i} + \frac{\gamma_1}{2} t_{1i}^2 + \frac{\gamma_2}{3} t_{1i}^3\right) \exp(\beta s_1)\right]$$

Note that when the baseline function is a simple linear function, we let $\gamma_2 = 0$. Then, we generate X_i failure times randomly from a uniform distribution within each subinterval $(t_{1,i-1}, t_{1i}]$. Therefore, the total number of failure times generated under s_1 is $n_1 = \sum_{i=1}^{m_1} X_i$.

Step 2: Generating the number of failures and the failure times under the middle stress level, s_2 :

Similarly to Step 1, we divide the second time period $(\tau_1, \tau_2]$ into m_2 subintervals: $(t_{20}, t_{21}]$, $(t_{21}, t_{22}]$, ..., $(t_{2,m_2-1}, t_{2m_2}]$, where $\tau_1 = t_{20} < t_{21} < \dots < t_{2m_2} = \tau_2$. Let Y_i be the failure number over the i th interval $(t_{2,i-1}, t_{2i}]$, the distribution of Y_i can be assumed as

$$Y_i \sim B\left(n - n_1 - \sum_{j=1}^{i-1} Y_j, p_{2i}\right), i = 1, \dots, m_2,$$

where p_{2i} is the failure rate within the i th interval $(t_{2,i-1}, t_{2i}]$ under the middle stress level, s_2 . Then, p_{2i} can be computed by

$$p_{2i} = P\{a + t_{2,i-1} < t \leq a + t_{2i} | s_2\} \\ = \exp\left[-\left[\gamma_0(a + t_{2,i-1}) + \frac{\gamma_1}{2}(a + t_{2,i-1})^2 + \frac{\gamma_2}{3}(a + t_{2,i-1})^3\right] \exp(\beta s_2)\right] \\ - \exp\left[-\left[\gamma_0(a + t_{2i}) + \frac{\gamma_1}{2}(a + t_{2i})^2 + \frac{\gamma_2}{3}(a + t_{2i})^3\right] \exp(\beta s_2)\right],$$

where $a = F_2^{-1}[F_1(\tau_1)]$. Then, we generate Y_i failure times randomly from a uniform distribution within $(t_{2,i-1}, t_{2i}]$ for each i . Therefore, the total number of failure times generated under s_2 is $n_2 = \sum_{i=1}^{m_2} Y_i$.

Step 3: Generating the number of failures and the failure times under the high stress level, s_3 :

Now, we divide the third time period $(\tau_2, c]$ into m_3 subintervals: $(t_{30}, t_{31}]$, $(t_{31}, t_{32}]$, ..., $(t_{3,m_3-1}, t_{3m_3}]$, where $\tau_2 = t_{30} < t_{31} < \dots < t_{3m_3} = c$. Let Z_i be the failure number over the i th interval $(t_{3,i-1}, t_{3i}]$. The distribution of Z_i can be assumed as

$$Z_i \sim B\left(n - n_1 - n_2 - \sum_{k=1}^{i-1} Z_k, p_{3i}\right), i = 1, \dots, m_3,$$

where p_{3i} is the failure rate within the i th interval $(t_{3,i-1}, t_{3i}]$ under the high stress level, s_3 . Then, p_{3i} can be derived as

$$p_{3i} = P\{b + t_{3,i-1} < t \leq b + t_{3i} | s_3\} \\ = \exp\left[-\left[\gamma_0(b + t_{3,i-1}) + \frac{\gamma_1}{2}(b + t_{3,i-1})^2 + \frac{\gamma_2}{3}(b + t_{3,i-1})^3\right] \exp(\beta s_3)\right] \\ - \exp\left[-\left[\gamma_0(b + t_{3i}) + \frac{\gamma_1}{2}(b + t_{3i})^2 + \frac{\gamma_2}{3}(b + t_{3i})^3\right] \exp(\beta s_3)\right],$$

where $b = F_3^{-1}[F_2(a + \tau_2 - \tau_1)]$. Then, we generate Z_i failure times randomly from a uniform distribution within $(t_{3,i-1}, t_{3i}]$ for each i , and finally the total number of failure times generated under s_3 is $n_3 = \sum_{i=1}^{m_3} Z_i$.

Step 4: Estimating the parameters:

For simplicity, we keep the length of all subinterval equal to q hours in this paper. The log-likelihood function can be expressed as

$$l(\beta; \xi) = n_1 \beta s_1 + \sum_{i=1}^{m_1} \left[X_i \left(\ln(\gamma_0 + \gamma_1(qi)) - \left(\gamma_0(qi) + \frac{\gamma_1}{2}(qi)^2 \right) \exp(\beta s_1) \right) \right] \\ + n_2 \beta s_2 + \sum_{i=1}^{m_2} \left[Y_i \left(\ln(\gamma_0 + \gamma_1(a + qi)) - \left(\gamma_0(a + qi) + \frac{\gamma_1}{2}(a + qi)^2 \right) \exp(\beta s_2) \right) \right] \\ + n_3 \beta s_3 + \sum_{i=1}^{m_3} \left[Z_i \left(\ln(\gamma_0 + \gamma_1(b + qi)) - \left(\gamma_0(b + qi) + \frac{\gamma_1}{2}(b + qi)^2 \right) \exp(\beta s_3) \right) \right] \\ - (n - n_1 - n_2 - n_3) \left[\gamma_0(b + c - \tau_2) + \frac{\gamma_1}{2}(b + c - \tau_2)^2 \right] \exp(\beta s_3). \tag{21}$$

For each simulation run, we can compute the maximum likelihood estimates of the model parameters by maximizing (21).

Let r be the number of simulation runs. We may use these r estimates to calculate the simulated squared bias (SBIAS²), simulated variance (SVAR), and simulated mean squared error (SMSE) of each parameter estimator. We define SBIAS², SVAR, and SMSE for an estimator $\hat{\theta}$ as:

$$\text{SBIAS}^2(\hat{\theta}) = \left(\frac{1}{r} \sum_{j=1}^r \hat{\theta}_j - \theta_0 \right)^2, \\ \text{SVAR}(\hat{\theta}) = \frac{1}{r-1} \sum_{j=1}^r \left(\hat{\theta}_j - \frac{1}{r} \sum_{j=1}^r \hat{\theta}_j \right)^2, \\ \text{and SMSE}(\hat{\theta}) = \frac{1}{r} \sum_{j=1}^r (\hat{\theta}_j - \theta_0)^2, \tag{22}$$

where θ_0 is the given true parameter value, $\hat{\theta}_j$ is the estimate of θ_0 from the j th simulation run. By these definitions, we can compute SBIAS², SVAR, and SMSE respectively for $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\beta}$ using the resulting designs from the example obtained in Sections 3 and 4. We can also compute SBIAS², SVAR, and SMSE of the MLE of hazard rate over T under normal design conditions for our resulting designs. The true hazard function over T under normal design stress level s_D can be expressed as

$$g(\alpha) = \int_0^T \lambda(t; s_D) dt = \left(\gamma_0 T + \frac{\gamma_1}{2} T^2 + \frac{\gamma_2}{3} T^3 \right) \exp(\beta s_D) \quad (23)$$

with $\alpha = [\gamma_0, \gamma_1, \gamma_2, \beta]^T$. By the invariance property of MLEs, the MLE of $(g|true)$ is

$$\hat{g} = \left(\hat{\gamma}_0 T + \frac{\hat{\gamma}_1}{2} T^2 + \frac{\hat{\gamma}_2}{3} T^3 \right) \exp(\hat{\beta} s_D). \quad (24)$$

Then, SBIAS², SVAR, and SMSE of \hat{g} can also be computed by (22).

5.2. Performance of the optimal designs obtained in Sections 3 and 4

Taking $n = 200$, $c = 300$ hours, and $r = 1000$, the length of all subintervals to be 5 hours, we use the simulation procedure introduced in Section 5.1 to evaluate the performance of our resulting designs. We adopt the constraint case C_1 for a demonstration.

For the PH model with a simple linear baseline function, from Section 3.3, $^{D,A,Q} \xi_{C_1}^{(1)}$ having $[\tau_1 = 178.2, \tau_2 = 223.6, s_2 = 197.5]$, is the D-, A-, and Q-optimal design with fixed middle stress at $s_2 = 197.5^\circ C$. After we update s_2 to the optimal middle stress level, we get new D-, A-, and Q-optimal designs for C_1 . Since the resulting D-, A- and Q-optimal designs are similar, we only present the simulation result for our Q-optimal design. The Q-optimal design we obtained in Section 3.4 for C_1 is $^Q \xi_{C_1}^{(2)}$ with $[\tau_1 = 156.19, \tau_2 = 247.47, s_2 = 155]$. We denote the efficiencies of a design ξ_A relative to another design ξ_B in terms of SBIAS², SVAR, and SMSE as $eff_B(\xi_A, \xi_B)$, $eff_V(\xi_A, \xi_B)$, and $eff_M(\xi_A, \xi_B)$, respectively. Based on these 1000 simulation runs, we compute all the SBIAS², SVAR, and SMSE of the MLE for $\gamma_0, \gamma_1, \beta$, and g when each of the two different designs, $^Q \xi_{C_1}^{(2)}$ and $^Q \xi_{C_1}^{(1)}$, is adopted. Table 13 shows the efficiencies of $^Q \xi_{C_1}^{(2)}$ relative to $^Q \xi_{C_1}^{(1)}$ in terms of SBIAS², SVAR, and SMSE of $\hat{\beta}, \hat{\gamma}_0, \hat{\gamma}_1$, and (24) respectively. We note that since these are Q-optimal designs and therefore the efficiency gains appear more for estimating g . This is consistent with the previous finding of the asymptotic efficiency gains as discussed in Section 3. The simulation results indicate that the most efficiency

gains appear in variance reduction with an extreme for γ_0 . The reason behind is that the optimal designs constructed was aiming to minimize the asymptotic variances.

In Section 3, the Q-optimal designs are obtained by minimizing the asymptotic variance of the \hat{g} . As expected, when $^Q \xi_{C_1}^{(2)}$ was adopted, we have gained the most efficiency (as high as 76.57%) in terms of the SVAR(\hat{g}) compare to $^Q \xi_{C_1}^{(1)}$. The SVAR of MLE of the model parameters are all very much reduced, and their efficiency gains of $^Q \xi_{C_1}^{(2)}$ relative to $^Q \xi_{C_1}^{(1)}$ in terms of SVAR($\hat{\gamma}_0$), SVAR($\hat{\gamma}_1$), and SVAR($\hat{\beta}$) are all higher than 100%. Moreover, SBIAS² and SMSE of \hat{g} are also being reduced. All the results confirm that $^Q \xi_{C_1}^{(2)}$ outperforms the design $^Q \xi_{C_1}^{(1)}$.

Table 13. The efficiencies of $^Q \xi_{C_1}^{(2)}$ relative to $^Q \xi_{C_1}^{(1)}$ in terms of SBIAS², SVAR, and SMSE

	$\hat{\beta}$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	\hat{g}
$eff_B(^Q \xi_{C_1}^{(2)}, ^Q \xi_{C_1}^{(1)})$	0.9824	0.9803	1.0	1.0175
$eff_V(^Q \xi_{C_1}^{(2)}, ^Q \xi_{C_1}^{(1)})$	2.2207	24508114.61	2.0038	1.7657
$eff_M(^Q \xi_{C_1}^{(2)}, ^Q \xi_{C_1}^{(1)})$	0.9838	0.9851	1.0	1.0345

For the PH model with a quadratic baseline function, the optimal design $^{D,A,Q} \xi_{C_1}^{(1)}$ $[\tau_1 = 178.15, \tau_2 = 223.57, s_2 = 197.5]$ (please see Table 8) is the D-, A-, and Q-optimal designs with a fixed middle stress at $s_2 = 197.5^\circ C$. When we simultaneously choose an optimal middle stress together with optimal stress-changing times, the D-, A-, and Q-optimal designs become $^D \xi_{C_1}^{(2)}$ having $[\tau_1 = 178.15, \tau_2 = 245.96]$, $^A \xi_{C_1}^{(2)}$ having $[\tau_1 = 155.20, \tau_2 = 247.47]$, and $^Q \xi_{C_1}^{(2)}$ having $[\tau_1 = 178.15, \tau_2 = 245.96]$, and they all are with $s_2 = 155$. We note that the D-optimal design and Q-optimal design are quite similar. Thus, we only present the simulation results for our resulting A- and Q-optimal designs in this example. Based on the 1000 simulation runs, we compute all SBIAS², SVAR, and SMSE of the MLE for $\gamma_0, \gamma_1, \gamma_2, \beta$ and g when each of the three different designs, $^{D,A,Q} \xi_{C_1}^{(1)}$, $^A \xi_{C_1}^{(2)}$, and $^Q \xi_{C_1}^{(2)}$, is adopted. Tables 14 and 15 display the efficiencies of $^A \xi_{C_1}^{(2)}$ and $^Q \xi_{C_1}^{(2)}$ relative to $^{D,A,Q} \xi_{C_1}^{(1)}$ in terms of SBIAS², SVAR, and SMSE of $\hat{\beta}, \hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$, and (24) respectively. The simulation results have shown that the designs, $^A \xi_{C_1}^{(2)}$ and $^Q \xi_{C_1}^{(2)}$, by optimally selecting the middle stress level and stress-changing time simultaneously can reduce SVAR(\hat{g}) by 282% and 183% compared to $^{D,A,Q} \xi_{C_1}^{(1)}$. We also note that not all of SVARs of MLE of model parameters have been reduced much by using $^A \xi_{C_1}^{(2)}$ or $^Q \xi_{C_1}^{(2)}$. Only SVAR($\hat{\gamma}_0$) and SVAR($\hat{\gamma}_2$) have significantly lessened among all SVARs.

Further, we uncover that both SBIAS² and SMSE of all $\gamma_0, \gamma_1, \gamma_2, \beta,$ and g are reduced by adopting either $A_{\xi_{C_1}^{3(2)}}$ or $Q_{\xi_{C_1}^{3(2)}}$. The efficiencies of $A_{\xi_{C_1}^{3(2)}}$ or $Q_{\xi_{C_1}^{3(2)}}$ in terms of SMSE(g) are more than 50 and 39 times higher than $D,A,Q_{\xi_{C_1}^{3(1)}}$. This indicates that unitizing $A_{\xi_{C_1}^{3(2)}}$ or $Q_{\xi_{C_1}^{3(2)}}$ can provide a great efficiency gain when experimenters are interested in estimating a hazard rate.

Conclusion

Optimal three-step-stress ALT designs for PH models, with either a linear or a quadratic baseline function, have been constructed in this paper. For a three-step-stress ALT, the practitioner often naturally set the average of high and low stress as the middle stress level. Nevertheless, from the results of both simulated and asymptotic efficiency comparison, we have revealed that the optimal three-step-stress ALT designs with both optimal stress-changing times and with an optimal middle stress level outperform the most among all the designs and for all the scenarios considered. Therefore, the three-step-stress plans with both optimal stress-changing times and optimal middle stress level are recommended especially when the hazard rate prediction is interested.

In Section 3, we have presented the resulting optimal designs for a practical ALT example when fitting a PH model with a simple linear baseline hazard function. Taking six different MNF constraint cases, we have found the optimal allocations of the stress-changing times and the optimal middle stress level that can minimize the loss function $L_D, L_A,$ or L_Q . Thus, we have solved the minimization problem for a nonlinear objective function with multiple nonlinear constraints (MNF at different stress levels), and obtained the constrained optimal designs under each of D-, A-, and Q-optimality. The resulting optimal designs under three different criteria are quite similar. In addition, we have also found that the middle stress level should be kept as close to the lower bound of the middle stress level as possible as long as the constraint condition is satisfied.

Table 14. The efficiencies of $A_{\xi_{C_1}^{3(2)}}$ relative to $D,A,Q_{\xi_{C_1}^{3(1)}}$ in terms of SBIAS², SVAR, and SMSE

	β	γ_0	γ_1	γ_2	g
$eff_B(A_{\xi_{C_1}^{3(2)}}, D,A,Q_{\xi_{C_1}^{3(1)}})$	1.0098	21.256 1	1.000 4	504.67 33	313.20 65
$eff_V(A_{\xi_{C_1}^{3(2)}}, D,A,Q_{\xi_{C_1}^{3(1)}})$	0.9713	11.492 4	0.027 7	420.46 43	2.8329
$eff_M(A_{\xi_{C_1}^{3(2)}}, D,A,Q_{\xi_{C_1}^{3(1)}})$	1.0097	14.017 3	1.000 4	42.756 5	40.255 4

In Section 4, we have derived the optimal designs when fitting a PH model with a quadratic baseline hazard function. The optimal stress-changing times and the optimal middle stress level are also chosen in order to minimize the loss functions under given nonlinear constraints. Similarly to the case with a linear baseline hazard, under these constraints, the optimal middle stress level can be located as close to the low stress level as possible as long as such constraints are satisfied. For each of D-, A-, and Q-optimality, six different constrained optimal

designs have been obtained. We also reveal that designing a three-step-stress ALT with a fixed middle stress seems only helping reduce the value of loss function under A-optimality. Thus, we suggest not conveniently taking the average of other two stress levels as the middle stress level, especially when a quadratic baseline hazard function is considered. We conclude that optimal three-step-stress designs have gained efficiency on an average of 68% for a hazard rate prediction compared to the corresponding optimal two-step-stress designs, which means that there is 7.56 times higher efficiency gain attained than the case when the baseline function is simple linear.

Table 15. The efficiencies of $Q_{\xi_{C_1}^{3(2)}}$ relative to $D,A,Q_{\xi_{C_1}^{3(1)}}$ in terms of SBIAS², SVAR, and SMSE

	β	γ_0	γ_1	γ_2	g
$eff_B(Q_{\xi_{C_1}^{3(2)}}, D,A,Q_{\xi_{C_1}^{3(1)}})$	1.027 1	31.459 9	1.000 4	199.040 5	280.627 9
$eff_V(Q_{\xi_{C_1}^{3(2)}}, D,A,Q_{\xi_{C_1}^{3(1)}})$	0.789 8	6.8388	0.014 9	309.285 5	3.8197
$eff_M(Q_{\xi_{C_1}^{3(2)}}, D,A,Q_{\xi_{C_1}^{3(1)}})$	1.025 8	9.8674	1.000 4	29.3142	51.1622

In Section 5, we have evaluated the performance of our resulting designs from both Sections 3 and 4 by simulations. The design for a three-step-stress ALT with an optimal middle stress level and two optimal stress-changing times has greatly increased the simulated efficiency of the hazard rate estimator. It is confirmed that the optimal designs with optimal middle stress significantly outperform those ones with the middle stress level fixed at the average of other two stress levels.

We note that although the design construction in this paper has been demonstrated by optimal designing three-step-stress ALT experiments when a single stress factor is involved, the method developed can be easily extended to conducting the optimal designs for more complicated multiple step-stress ALT, such as step-stress ALT with more than three steps, and/or ALT with multiple stress factors (but one of them engaged in conducting step-stress plans). We also notice that the inaccuracy of the assumed baseline hazard function can cause unavoidable prediction bias. Although the proposed designs perform very well when the model assumed is correct, they seem not helping much in reducing such bias when the assumed model is incorrect. If possible imprecision of the assumed PH model is suspected, then robust design approach should be considered. Some discussion on robust design for a PH model can be seen in [18].

Conflict of Interest

We declare that there is no any conflict of interest.

Acknowledgment

The research is supported by the Natural Sciences and Research Council of Canada.

References

[1] X. Xu, W. Huang, "Constrained Optimal Designs for Step-stress Accelerated Life Testing Experiments" *Proceedings of the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis*, 2018, 633-638.

- [2] X. Xu, S. Hunt, "Robust Designs of Step-Stress Accelerated Life Testing Experiments for Reliability Prediction" *Matematika*, 29(1), 203-212, 2013.
- [3] R. Miller, W. Nelson, "Optimum Simple Step-Stress Plans for Accelerated Life Testing" *IEEE Transactions in Reliability*, 32(1), 59-65, 1983.
- [4] D. S. Bai, M. S. Kim, S. H. Lee, "Optimum simple step-stress accelerated life tests with censoring" *IEEE Transactions in Reliability*, 38(5), 528-532, 1989.
- [5] D. S. Bai, M. S. Kim, "Optimum simple step-stress accelerated life tests for Weibull distribution and Type-I censoring" *Naval Research Logistics*, 40(2), 193-210, 1993.
- [6] N. Fard, C. Li, "Optimal simple step stress accelerated life test design for reliability prediction" *Journal of Statistical Planning and Inference*, 139(5), 1799-1808, 2009.
- [7] S. Hunt, X. Xu, "Optimal Design for Accelerated Life Testing with Simple Step-Stress Plans" *International Journal of Performability Engineering*, 8(5), 575-579, 2012.
- [8] H. Ma, W. Q. Meeker, "Optimum step-stress accelerated life test plans for log-location-scale distributions" *Naval Research Logistics*, 55(6), 551-562, 2008.
- [9] L. Jiao, "Optimal allocations of stress levels and test units in accelerated life tests." Ph.D Thesis. Rutgers University-New Brunswick, New Brunswick, New Jersey, 2001.
- [10] E. A. Elsayed, H. Zhang, "Design of Optimum Simple Step-Stress Accelerated Life Testing Plans" *Proceedings of the 2nd International IE Conference*, 1-14, 2004.
- [11] C. H. Hu, R. D. Plante, J. Tang, "Step-stress accelerated life tests: a proportional hazards--based non-parametric model" *IIE Transactions*, 44(9) 754-764, 2012.
- [12] N. Becker, B. McDonald, C. Khoo, "Optimal designs for fitting a proportional hazards regression model to data subject to censoring" *Austral. J. Statist.* 31, 449-468, 1989.
- [13] H. Dette, M. Sahm, "Minimax optimal designs in nonlinear regression models" *Statist. Sinica*, 8, 1249-1264, 1998.
- [14] J. M. McGree, J. A. Eccleston, "Investigating design for survival models" *Metrika*, 72, 295-311, 2010.
- [15] J. L'opez-Fidalgo, M. J. Rivas-L'opez, R. Del Campo, "Optimal designs for Cox regression" *Statistica Neerlandica* 63, 135-148, 2009.
- [16] M. Konstantinou, S. Biedermann, A. Kimber, "Optimal designs for two-parameter nonlinear models with application to survival models" *Statistica Sinica*, 24, 415-428, 2014.
- [17] W. Nelson, *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses*. Wiley. New York, 2004.
- [18] X. Xu, W. Huang, "Optimal Robust Designs for Step-stress Accelerated Life Testing Experiments for Proportional Hazards Models" *Mathematical and Computational Approaches in Advancing Modern Science and Engineering*, Eds: Bélair, J., Frigaard, I.A., Kunze, H., Makarov, R., Melnik, R., Spiteri, R.J., New York: Springer, 585-594, 2016.

Fuzzy Logic Implementation for Enhanced WCDMA Network Using Selected KPIs

Nosiri Onyebuchi Chikezie*, Onyenwe Ezinne Maureen, Ekwueme Emmanuel Uchenna

Electrical and Electronic Engineering Department, Federal University of Technology, Owerri, Nigeria

ARTICLE INFO

Article history:

Received: 25 November, 2018

Accepted: 18 January, 2019

Online : 29 January, 2019

Keywords:

Fuzzy Logic

KPI

QoS

MTN

AIRTEL

ABSTRACT

The paper focused on the implementation of fuzzy logic technology for improved Wideband Code Division Multiple Access (WCDMA) network using selected Key Performance Indicators (KPIs). Empirical and analytical methods were principally deployed for the study analyses. Empirical analyses were conducted on two designated networks which are MTN and AIRTEL observed with high network traffic to evaluate their network performances using the selected KPIs. Analytical method was further deployed to improve on the observed limited performance. Five (5) geographical locations within Owerri metropolis were selected for the measurements due to the perceived high density from end users; they include Aba road, FUTO road, Onitsha road, Orlu road and Wetheral road. Selected KPIs include Receive Signal Level (RXLEV), Call Setup Success Rate (CSSR), Call Drop Rate (CDR) and Call Completion Success Rate (CCSR) were used to evaluate the various performance characteristics of the networks based on the QoS. Results obtained from the field measurements computed using the selected KPIs could not meet up with any of the Nigeria Communication Commission (NCC) thresholds. A proposed Fuzzy Logic technique was introduced to the system while varying the congestion load characteristics for different environments using the following parameters; mean bit rate, mean burst rate, network statistics and retainability. An average of 10.2% increase in the system throughput was observed from the proposed system over the existing system.

1. Introduction

The discovery of Wideband Code Division Multiple Access (WCDMA) network has provided lots of ease in mobile transactions ranging from voice and data applications to internet surfing and other online operations. The importance attributed to this invention has led to the high demand of the network. In [1], observed the high traffic evident in the network was due to the increased number of subscribers which had contributed to incessant poor quality of service delivery. Due to this high influx of subscribers, the network performance of WCDMA systems began to deteriorate ranging from poor network coverage, constant block/drop calls (Poor call initialization and Handover), call/ network congestion and poor retainability/internet services. Hence, resulted to various degrees of complaints from the subscribers.

Owing to the observations, prompted the need to carry out extensive analysis on the performance of WCDMA network. The study considered basically two prominent networks viz MTN and Airtel in the South Eastern part of Nigeria (Owerri metropolis in

Imo State), characterized with greater number of subscribers. Evaluation of their network performances would be achieved in the study using a prominent criterion known as Key Performance Indicators (KPI). KPI is an important tool for network performance evaluation. It is classified as a minimum set of metrics used for tracking system progress towards a performance target [2]. The essential tool for the KPI assessment is the Quality of Service (QoS). Quality of Service is defined as the collective effect of service performance, which determines the degree of satisfaction of a user [3]. Some selected KPI's for the study include Receive Signal Level (RXLEV), Call Setup Success Rate (CSSR), Drop Call Rate (DCR), and Call Success Rate (CSSR). The Congestion control mechanism was another factor analyzed in the study. Analytical approach was also implemented using intelligent fuzzy logic model to enhance the QoS of WCDMA network using the selected KPI techniques.

2. Literature Review

In [4], compared the performance of various KPIs that were used by the Nigeria Communication Commission (NCC) for rating QoS using drive test approach. The results obtained were used to compare the NCC KPIs target. The outcome showed that

*Corresponding Author: Onyebuchi Nosiri; Email: buchinosiri@gmail.com

virtually all the networks could not meet the NCC target. The works of [5] studied the “compromise between network performance and end user satisfaction over UMTS Radio interface using an empirical investigation in Asaba Delta state”. The KPIs used were CDR, CSSR, HOSR, ESA and NRR. Five clustered ranging from 1-5 were considered in their analyses. Drive test was used to conduct measurements in the cellular network of study. The research work assessed the End User satisfaction by estimating the blocked calls probability (Pblock) and dropped calls probability (Pdrop) data by $P_{satisfied} = 1 - P_{unsatisfied}$ where $P_{unsatisfied} = P_{block} + P_{drop}$. The results obtained from the operators when compared with the NCC performance benchmark only made 20% in cluster 1-3 and 60% in cluster 4-5 respectively, which were generally below expectation and unsatisfactory. The authors of [6] carried out the study on “End-User Satisfaction Assessment Approach for Efficient Network Performance Monitoring in Wireless Communication Systems”. Nontrivial technique for extracting implicit and useful information from existing data sets were implemented. The KPIs used were CSSR, CDR. The results obtained showed that only one operator could meet the NCC threshold at a specified time period. The works of [1], developed software Engineering Approach in Mitigating QoS Challenges in Mobile Communication Networks in Nigeria”. The investigations showed minimal congestion amongst the four network studied. None of the operators met the NCC threshold for HOSR, most of the operators met the NCC target for Call Completion Rate (CCR) and in the overall performance, GLOBACOM was found to be the best. Furthermore, a Software engineering approach was developed for the system optimization. A Congestion Control in Asynchronous Transfer Mode (ATM) Network by [7] was proposed, introducing a KPI parameter (Call Completion Success Rate (CCSR)) into an existing fuzzy logic system to regulate network instability in the ATM network. CCSR was proposed as a parameter indicator which sends signal whenever congestion occurs in the system. The authors observed that if the incoming calls are 75%, the CCSR could send signal to fuzzy policer to retain 30% and allow 45% to pass. On the other hand, when the network is free, it signals the fuzzy policer to allow the number of 30% calls to pass. The introduction of CCSR parameter guaranteed network quality by controlling congestion. It does so by monitoring what happens at the end user or receiver end. And as such ensures that calls are made successfully without distortion. The proposed work of [7] provided a mitigation upgrade due to the introduction of one KPI (CSSR) parameter indicator. Hence,

this study traded the approach adopted by the authors of [7] and executed it in WCDMA networks realized using four (4) KPIs viz: RXLEV, CSSR, CDR and CCSR. The selected KPIs were relatively considered to evaluate the levels of congestion in the network.

QoS-Based Key Performance Indicators (KPIs)

Determination of the service quality parameters exposes operators to the following issues [8].

- Detecting probable errors in Base Station Subsystem (BSS) hardware and providing utilization of physical resources.
- Taking necessary actions in order to determine radio network problems and to optimize the network; such as frequency assignment, adjusting antenna tilt and changes in some other parameters.
- Observing system behaviors and changes in the system; such as traffic load, congestion, and successful attempts.
- Estimating the traffic load and network expansions, in parallel with increasing number of mobile users.
- Comparing the network with the competitors and providing better quality users.
- Comprehension of the condition of the market and following new technologies.

The relationship between QoS and KPIs is shown in Figure 1. Selected Key Performance Indicators for the study include [9]

1. **Call Success Setup Ratio (CSSR):** This is ratio of total number of completed calls to the total number of call attempts.

$$CSSR = \frac{\text{Successfully Completed Call Setups}}{\text{Call Setups Attempts}} \times 100 \tag{1}$$

2. **Call Completion Success Rate (CCSR):** This indicator can be derived either from network statistics or from drive test statistics. The indicator takes into account the fact that all failures are either drops or unsuccessful call set ups. It is a good parameter for evaluating the network accessibility and retainability as perceived by the customers. The indicator is derived using the expression:

$$CCSR = \frac{\text{Total Number of Completed Call}}{\text{Total Number of Call Attempted}} \times 100 \tag{2}$$

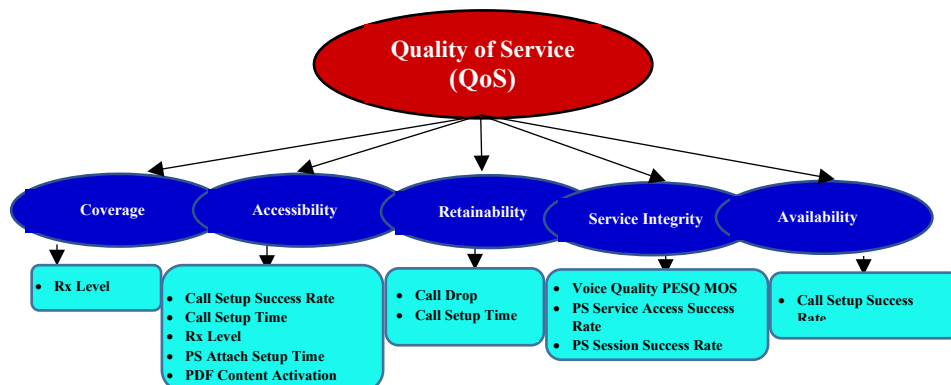


Figure 1 Relationship between QoS and KPIs [8]

Table 1: Threshold for Network performance KPI's [1, 10]

Key Performance Indicators	CDR %	CSSR %	CCSR %	CSFR %	HOSR %	HOFR %	RSCP (dBm)	Ec/Io (dBm)
NCC Threshold value in %	2	98	96	4	98	2	-85	-9

3. **Call Dropped Rate (CDR)** : It is the total number of calls dropped (not ending as desired by the user) or forced call disconnection by the network due to various reasons within the licensee's own network.

$$CDR = \frac{\text{Number of Dropped Calls}}{\text{Number of Successfully Completed Call Setup}} \times 100 \quad (3)$$

4. **Received Signal Level (RXLEV)**: This is referred to as the received signal level at the input of the mobile device. For WCDMA network it is in RSCP

$$RSCP = \frac{\text{Summation of all data that fall in } >-85\text{dBm}}{\text{Summation of all data}} \times 100 \quad (4)$$

5. **Network Accessibility Ratio (NAR)**: This is the probability that a mobile user will establish a successful voice communication between the two ends of the network within a given condition. It is also expressed as the call setup rate. It is represented as:

$$NAR = \frac{\text{Number of Successful Call Setup}}{\text{Number of Call Attempted}} \times 100 \quad (5)$$

6. **Network Retainability Ratio (NRR)**: It is the ratio between the number of successful calls and number of normally terminated calls. It is also the probability that an active call come to an end successfully in a network.

$$NRR = (1 - CDR) \times 100 \quad (6)$$

KPI's Benchmark for Network Providers by NCC

Table 1 shows some of the Network Performance KPI's and their percentage threshold levels set by Nigeria Communication Commission (NCC).

3. Methodology

The Empirical Analyses were carried out to ascertain the actual level of Network Performance based on observations and measurements taken from the studied networks end-user perspectives. An Analytical Approach was implemented as a mitigation measure using fuzzy logic technique.

3.1. Outdoor Measurement Analysis

Drive test was used to obtain the call statistics which was also used to calculate the KPI's. Statistic information of the network was obtained as relevant evaluation parameters for system optimization. The materials/tools used in the research study include Laptop, TEMS Phone(SONY ERRICSSON W995 TEMS POCKET), Data cable, Global Positioning System (GPS), TEMS Investigation Software, MapInfo Professional and MATLAB.

A HP laptop of core i3, 2GB, 500GB was used to install softwares such as MapInfo Professional, and TEMS Investigation used for the simulation analyses. The data cable, Sony Erricson Phone, and GPS were all connected on the the labtop. TEMS

Phone: Sony Erricsson W995 Tems Pocket was used. Tems Pocket is a phone-based test tool developed for measuring the performance and quality parameter of wireless networks. The tool collects measurements and event data for network monitoring. TEMS Phone was used to access the services of WCDMA from the end users to the provider. A data cable was used to access the basedband transmission of WCDMA service from the transmitter to the receivers. It was used to connect the Sony Errcson W995 TEMS pocket to the Labtop. The Global Positioning System (GPS) is a network of orbiting satellites that sends precise details of their position in space back to earth. The signals are obtained by GPS receivers, such as navigation devices and are used to calculate the exact position, speed and time at the vehicles location. It was used to determine the position of the mobile station to an accuracy of about 5 meters. TEMS Investegation Software was used to measure, analyze and optimize the mobile networks. It was also used to establish a good knowledge of the availability of the WCDMA network, and features about the quality of service. MapInfo Professional is a geographical information system software that was used for the mapping and location analysis of the network. It enabled us to visualize, analyze, interpret and output the WCDMA network data to reveal the quality of service relationships, pattern and trends. Figure 2 shows the MapInfo Professional. Matlab tool was chosen as a good programming software for the analysis.

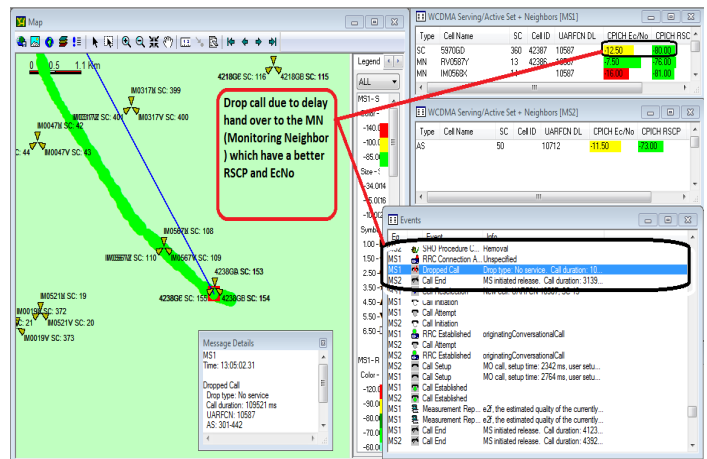


Figure. 2 MapInfo Professional

The flowchart representation of the proposed system is illustrated in Figure 3

3.2. Drive Test Set Up

All the components were connected appropriately to the laptop. The first tool to be connected to the laptop was the dongle which gives a license to the TEMS interface on the system. It should be noted that even though the TEMS interface could be opened without the dongle, a drive test cannot be carried out because the TEMS phone with which calls are made can never be

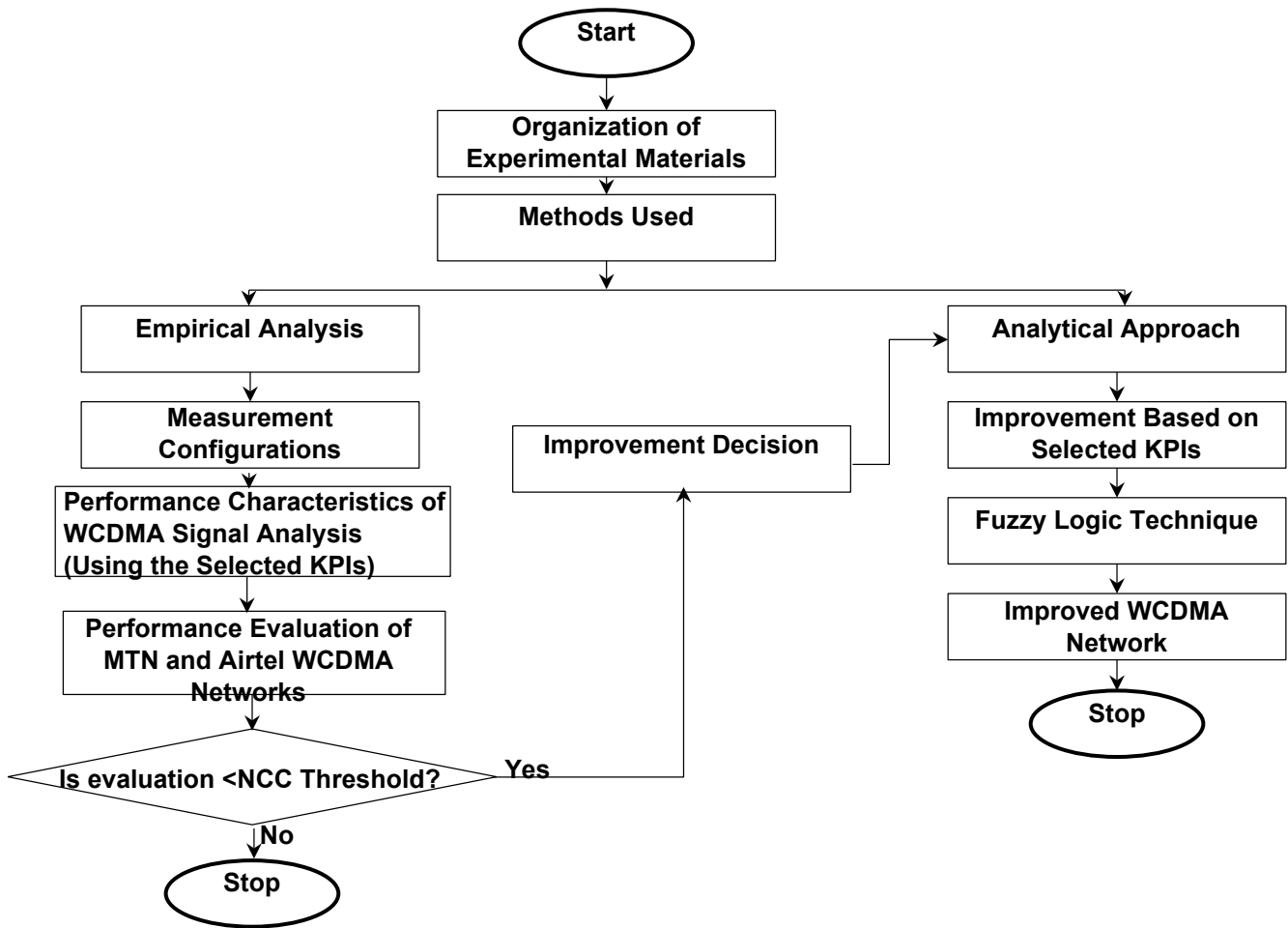


Figure 3 flow chart representation of proposed approach

Table 2: Call summary for MTN network
Call summary analysis for MTN network in one of the regions

Event	#[no.of]	Relationship	#Cell	#Log
Call Setup	82	-	1, 3, 5, 6, 10, 15	1
Call Attempt	94	-	1, 3, 5, 6, 10, 15	1
Call Attempt Retry	94	-	4, 6, 8, 10, 11	1
Call End	82	-	1, 3, 5, 6, 10, 15	1
Call Established	81	-	1, 3, 5, 6, 10, 15	1
Call Initiation	94	-	1, 3, 5, 6, 10, 15	1
Dropped Call	3	-	2	1
Missing WCDMA Intra-frequency Neighbor, based on DN reporting	9	-	3, 5, 6, 10, 15	1
RRC Connection Abnormal Release	1	-	2	1
Routing Area Update	1	-	2	1

Distributed graph of all logfiles
TEMS

Table 3: Call summary for AIRTEL network
Call summary analysis for AIRTEL network in one of the regions

Event	#[no.of]	Relationship	#Cell	#Log
Call Setup	91	-	1, 3, 5, 6, 10, 15	1
Call Attempt	93	-	1, 3, 5, 6, 10, 15	1
Call Attempt Retry	93	-	4, 6, 8, 10, 11	1
Call End	89	-	1, 3, 5, 6, 10, 15	1
Call Established	77	-	1, 3, 5, 6, 10, 15	1
Call Initiation	93	-	1, 3, 5, 6, 10, 15	1
Dropped Call	2	-	2	1
Missing WCDMA Intra-frequency Neighbor, based on DN reporting	9	-	3, 5, 6, 10, 15	1
RRC Connection Abnormal Release	1	-	2	1
Routing Area Update	1	-	2	1

Distributed graph of all logfiles
TEMS

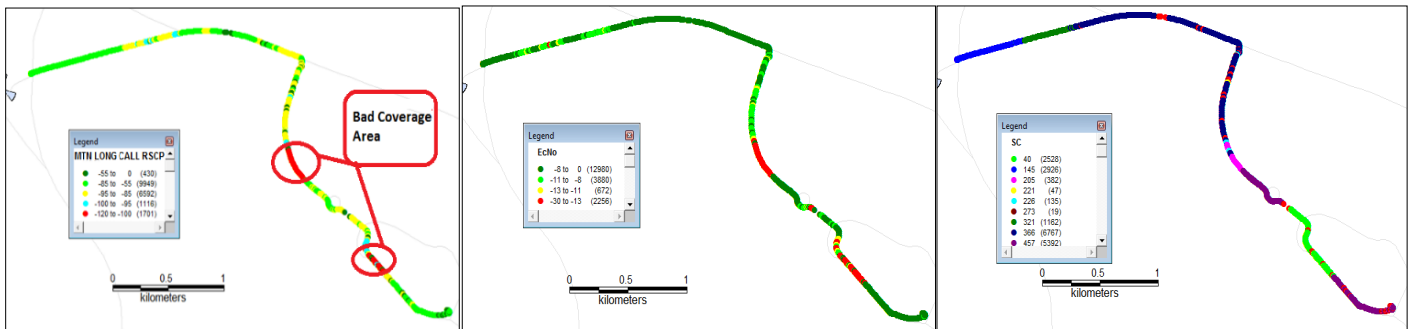


Figure 4: Represent MTN LONG calls on various parameters (RSCP, ECL0 and SC)

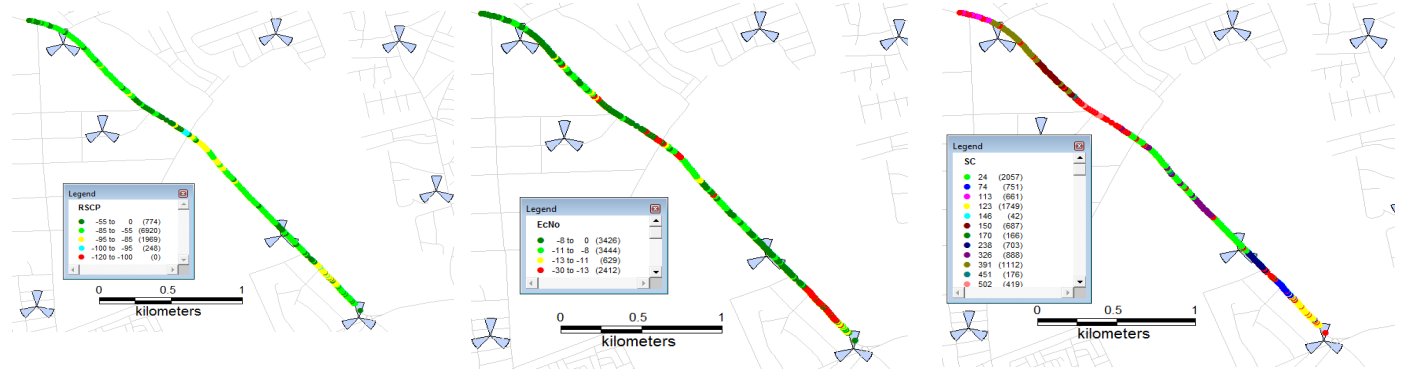


Figure 5: Represent AIRTEL LONG calls on various parameters (RSCP, ECNO and SC)

viewed and accessed. The first step is to power ON the laptop after which the dongle will be connected to it through one of the Universal Serial Bus (USB) ports. The next step requires connecting the TEMS phone through the phone’s USB cable and the GPS. The dongle allows accessibility to these two pieces of equipment. One of the good things about TEMS is the audio capability which helps to quickly detect any disconnected component. This is achieved with the encrypted voice in the TEMS which loudly interprets any of the components that has been disconnected. The laptop is always connected to an inverter which provides constant electricity for the laptop in other to overcome the battery drainage due to the number of connected components. A folder was created where the data will be saved (Log file) during the measurement. Tables 2 and 3 show call summary analyses of MTN and AIRTEL networks conducted in one of the regions while Figures 4 and 5 represent the long calls evaluations obtained from the measurements.

3.3. The System Algorithm

The algorithm for the proposed system are as follow:

- Step one: Detect call arrival rate of the system.
- Step two: Compute the status of queue bit rate of the system.
- Step three: Evaluate the queue capacity of the system with the arrival rate of the incoming calls.
- Step four: Estimate the state of the queue burst rate of the network.
- Step five: Pass or drop calls using the pass/ drop switch that is regulated by the Fuzzy policer.
- Step six: Store the drop calls on the buffer storage device.
- Step seven: Fuzzy congestion controller coordinates, analyzes and evaluates the performance analysis of the Network and the QoS using KPIs.

The proposed system deployed four KPIs which are RXLEV, CDR, CCSR, and CSSR as shown in figure 6, to be monitored at the receiver’s front end. This helps the policer to make a more robust decision in passing or dropping of calls.

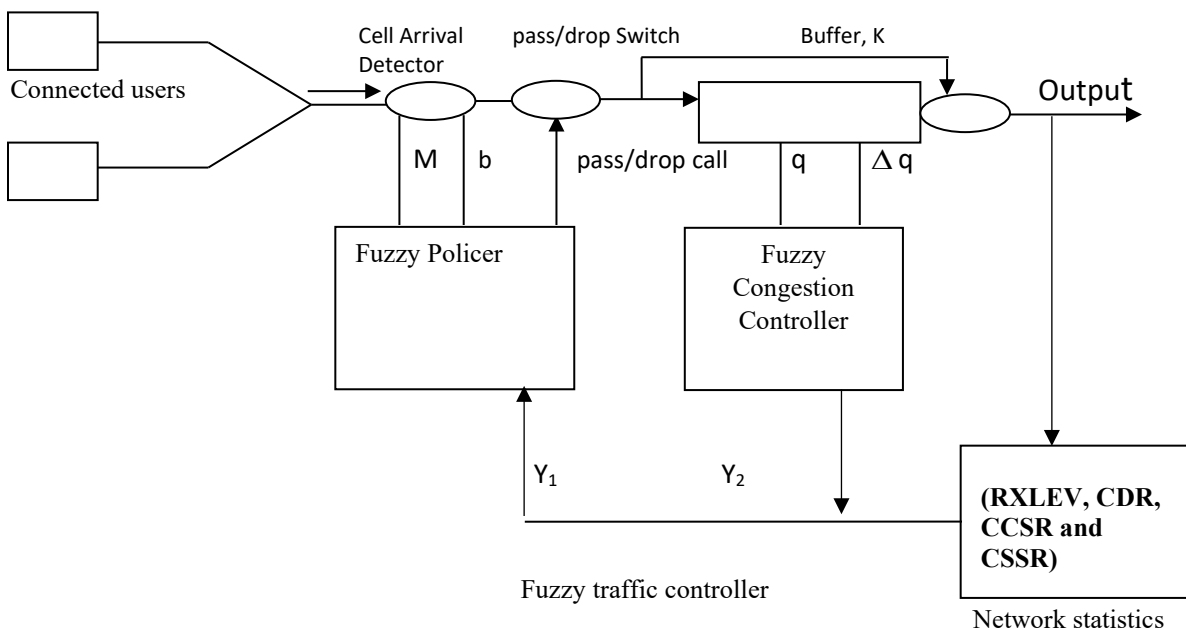


Figure 6 Structural Design of the proposed system

Components of the proposed Fuzzy Logic implemented

i. **Fuzzy Congestion Controller:** The Fuzzy Congestion Controller (FCC) goal is to simultaneously police mean rate and reject bursts while relieving and preventing congestion. Cells are passed or dropped based on an evaluation of conformance to the traffic contract and negotiated traffic parameters.

ii. **Cell arrival detector:** The objective of the Fuzzy Police (FP) is to police the mean rate and reject bursts. The Fuzzy Police (FP) performs its job by continuously evaluating the compliance/violation level of two parameters; i.e. the ratio of up-to-date mean bit rate to negotiated mean bit rate and ratio of up-to-date mean burst length to negotiated mean burst length. It then decides on the drop rate to be imposed on the cells based on the collective evaluation of the compliance/violation level of the two parameters.

iii. **Pass/drop switch:** This extension provides the Fuzzy Police (FP) with more information thereby enabling it to carry out more accurate decisions on passing or dropping cells.

iv. **Fuzzy Police:** The fuzzy police proposed in this study is a window-based control mechanism, in which the maximum number of cells that can be accepted in a specific window of certain length is a threshold that is dynamically updated by inference rules. The fuzzy police's task is to ensure that the source complies with negotiated mean rate over the duration of the connection.

Fuzzy variable input/output specifications

Fuzzy Input/output Specification for the existing system describes the term sets used to define each input parameter.

Input Specification

The input specifications of the system consists of five-inputs. The input values for the first network controller are as follow:

- a) ratio of up-to-date mean bit rate to negotiated mean bit rate (A_1)
- b) ratio of up-to-date mean burst length to negotiated burst length (A_2)
- c) State of the network (y):
 - i. Term set for queue length $T(q) = \{ \text{Empty (E), Full (F)} \}$
 - ii. Term set for queue-length change rate $T(\Delta q) = \{ \text{Negative (N), Positive (P)} \}$
- d) Term set for rate control $T(y) = \{ \text{Decrease (D), No Change (NC), Increase (I)} \}$

The system outputs are as follow:

- e) State of the network (y):
 - i. Call Drop (y_i). The up-to-date value is calculated by averaging all previous values up to the most current value.
 - ii. QoS based on KPIs (y):

Rules Set of Proposed Fuzzy Logic System

Rule one: if Bit rate is complying and burst length is complying and State of network is positive and QoS based on KPIs is positive then drop rate is Pass.

Rule Eighteen: if Bit rate is violate comply and burst length is **violate** and State of network is Negative and QoS based on KPIs is Negative then Drop rate is **Drop**.

The proposed flow mechanism is shown in Figure 7.

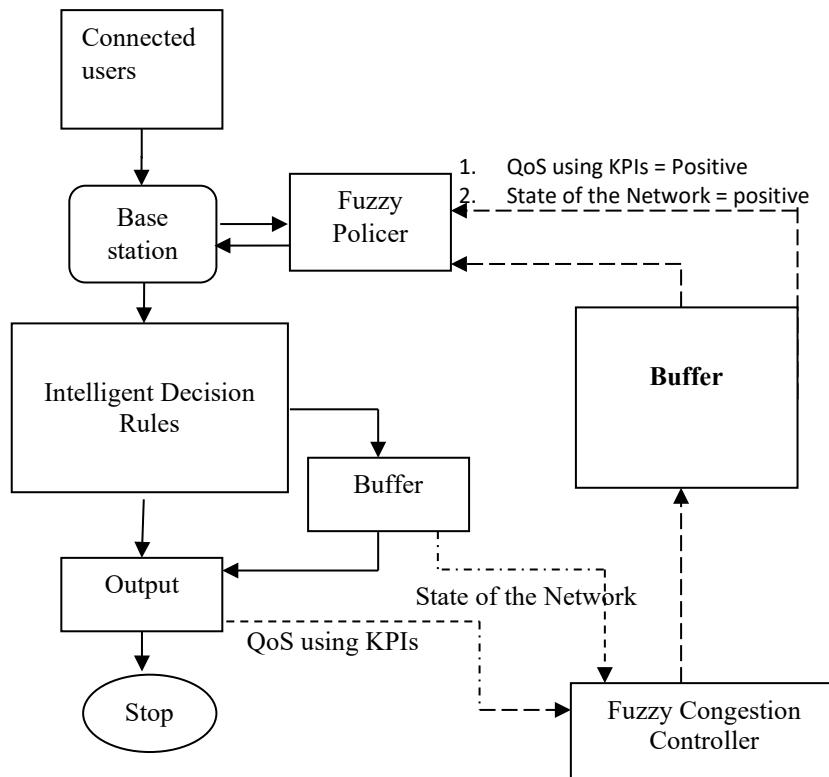


Fig 7: Flow Mechanism of the Proposed Fuzzy Logic System

Table 4: Fuzzy Input /Output Specifications

Input				Outputs
Term set for queue length T(q)	Term set for queue-length change rate T(Δq)	State of the network	QoS based on KPIs	Term set T(c)
Empty (E)	Negative (N),	Decrease (DC)	Staying on the network (SN)	Drop
Full (F)	Positive (P)	No Change (NC)	Not Staying on the network (NSN)	Between pass and Drop
		Increase (I)		Pass

The flow mechanism of the proposed system consists of the connected users, Base station, Fuzzy Police, Intelligent Decision Rules, buffer, and the fuzzy congestion controller. The connected users send and receive calls through the base station. The fuzzy police control the incoming calls in the base station. It ensures that the queue capacity in the base station is not exceeded. So it allows some calls to flow through and stores the unpermitted calls to the buffer. After which the cell is free to accommodate more calls, the stored calls in the buffer will be allowed to flow through the network. The fuzzy controller controls the stored calls in the buffer to prevent interference, collision and congestion. It also monitors the call retainability (CCSR, CSSR, CDR, RXLEV) of the system and gives a feedback of what the users are experiencing so that the system re-adjusts itself to correct the

The output is the membership function for the term set T(c) which are Drop (D), Between Pass & Drop (BPD) and Pass (P). Uncertainty in the network system may result to drop calls, calls between pass & drop and pass all calls. The value for Drop (D) calls will be set to zero for total drop of all calls, BPD set to a value within [0, 1] but closer to 1 for dropping a fraction of cells, and passing all calls set to 1. The simulation parameters are represented in Table 5.

Table 5 Simulation Parameters

Source Parameters	Packetized Voice
Peak rate, P	32 Kbps
Mean rate, m	11.2Kbps
Burst Size, b	26 cells
Silence period, μ	0.65 S
Burstiness, β	2.85
Fuzzy Parameters	Values
Mean Bit Rate	Range [1 1.5]
Mean Burst Length	Range [1 1.5]
Network State	Range [-1 1]
Call Completion Rate	Range [-1 1]
Call Drop Rate	Range [-1 1]
Call Setup Success Rate	Range [-1 1]
Receive Signal Level	Range [-1 1]
Queue Length	Range [0 1]
Queue Change Rate	Range [-1 1]
Out put	Range [0 1]
NumRules	24

4. Results

Figures 8 to 12 represent Bar Charts displayed for comparison of the following: Call completed success rate, Call Setup Success Rate, Call Network Accessibility, Service Retainability, Call Drop Rate along Aba road, FUTO road, Onitsha road, Orlu road, Wetheral road respectively and the NCC threshold.

problem by adjusting its transmission rate. The retainability is the overall network that indicates if the services are unreliable, reliable or unsatisfied and satisfied.

Table 4 shows the design specification for the system which consists of input and output designs. The Fuzzy input variable used the "term sets" to describe Term set for queue length T(q) = { Empty (E), Full (F) }; Term set for queue-length change rate T(Δq) = { Negative (N), Positive (P)}; The state of the network is represented with term set for rate control T(y) = {Decrease (DC), No Change (NC), Increase (I)}.The forth reused input is represented with the term set for network retainability rate as T(y2)= {Staying on the network (SN) Not Staying on the network (NSN)}.

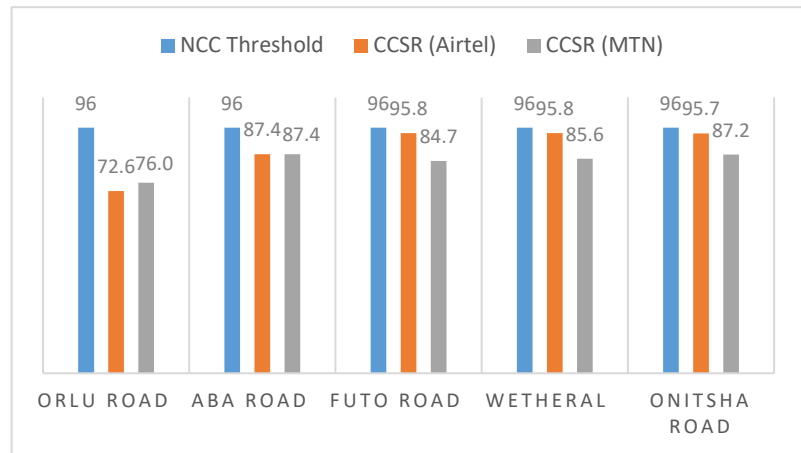


Figure 8 Bar Chart for comparison of call completed success

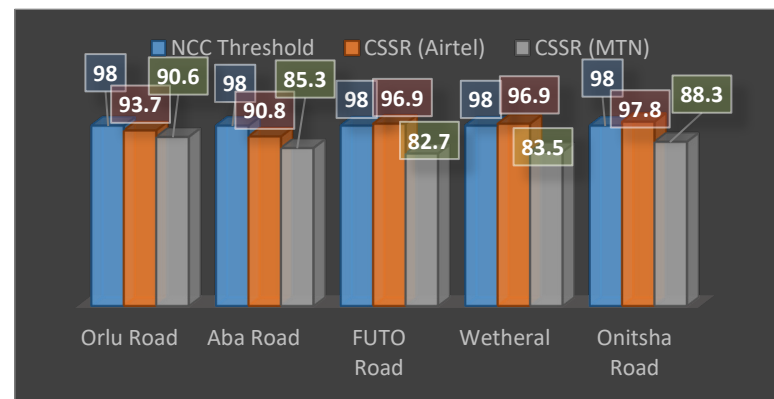


Figure 9 Bar chart representation of benchmarking of Call Setup Success Rate

The results obtained from the analyses of the drive test on the Received Signal Coded Power of the designated networks is shown in Table 6. Figure 13 illustrated the percentage coverage reliability of the RSCP.

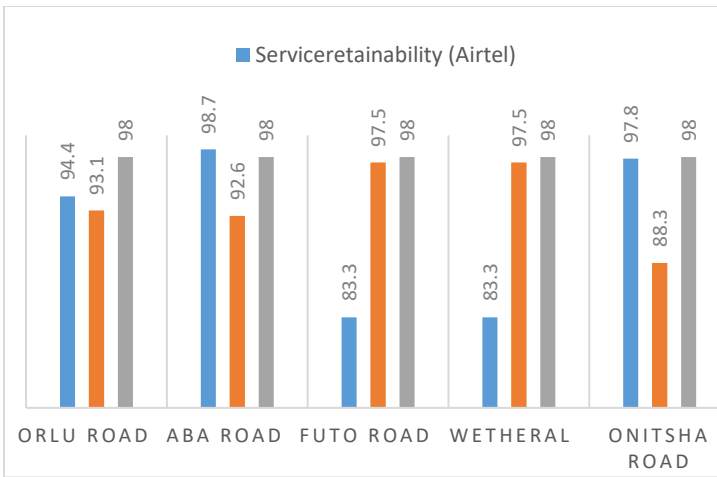


Figure 10 Bar chart representation of MTN and Airtel Call Network Accessibility

The summary result for network drive test conducted on the designated paths on coverage quality of service is shown in Table 7. Figure 14 demonstrated a graphical view of the system performance relative to the NCC threshold

4.1. Simulation Results Using Fuzzy Technique

A simulation was carried out to analyze the performance of the proposed Fuzzy Policer (FP) and Fuzzy Congestion Controller (FCC) in policing and controlling traffic flow within Virtual Channel Connections (VCCs) passing through a network node. The simulation interface of the proposed system is shown in Figure 15(a) and (b).

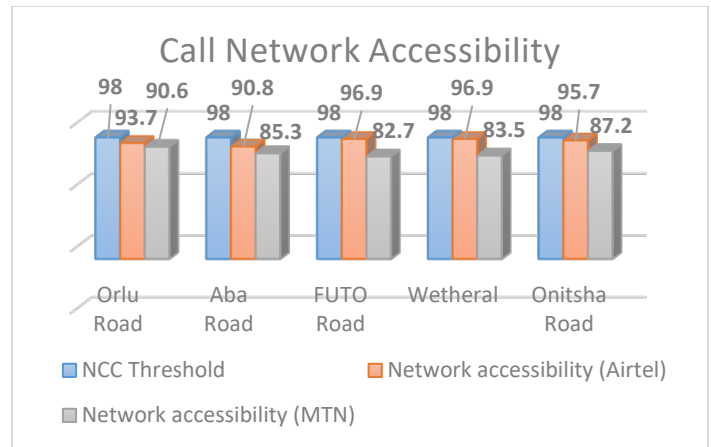


Figure 11 Bar chart representation of MTN and Airtel Service retainability

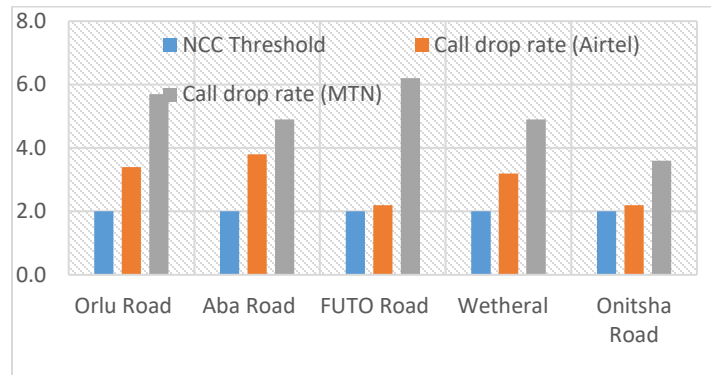


Figure 12 Bar chart representation of MTN and Airtel Dropped Call Rates

Table 6: Drive test results of the two WCDMA networks on the Received Signal Coded Power.

Ranges (dBm)	Orlu Road		Aba Road		FUTO Road		Wetheral Road		Onitsha Road	
	MTN	Airtel	MTN	Airtel	MTN	Airtel	MTN	Airtel	MTN	Airtel
(-55 to 0)	2.5	2.2	3.8	2.0	1.1	2.2	1.9	1.3	3.9	2.8
(-85 to -55)	44.9	35.4	46.2	33.9	25.1	31.6	45.3	38.9	34.9	44.1
(-95 to -85)	2.6	19.1	0.0	13.5	16.7	19.8	2.4	9.8	9.9	3.1
(-100 to -95)	0.0	0.9	0.0	0.6	2.8	2.7	0.4	0.0	1.3	0.0
(-120 to -100)	0.0	0.0	0.0	0.0	4.3	0.0	0.0	0.0	0.0	0.0

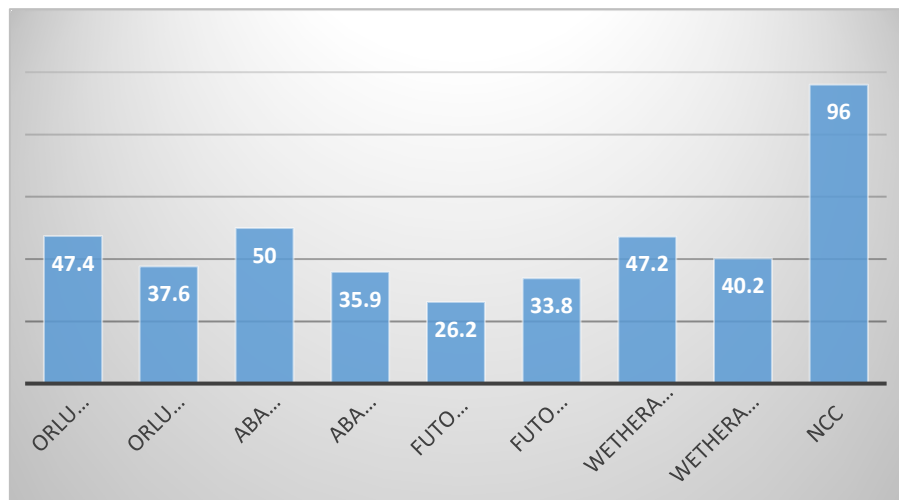


Figure 13 Bar chart of WCDMA Network coverage reliability of RSCP (%)

Table 7 Coverage Quality of Service for WCDMA Networks.

Ranges (dBm)	Orlu Road		Aba Road		FUTO Road		Wetheral Road		Onitsha Road	
	MTN	Airtel	MTN	Airtel	MTN	Airtel	MTN	Airtel	MTN	Airtel
(-8 to 0)	9.0	6.5	6.5	5.8	33.5	33.8	45.8	9.9	15.2	3.0
(-11 to -8)	32.7	45.6	24.2	51.3	39.3	46.1	28.8	32.3	32.8	22.4
(-13 to -11)	43.5	28.3	24.5	27.2	25.9	18.0	22.4	55.1	27.0	28.8
(-20 to -13)	14.8	19.5	44.8	15.6	1.4	2.1	3.0	2.6	25.0	45.8

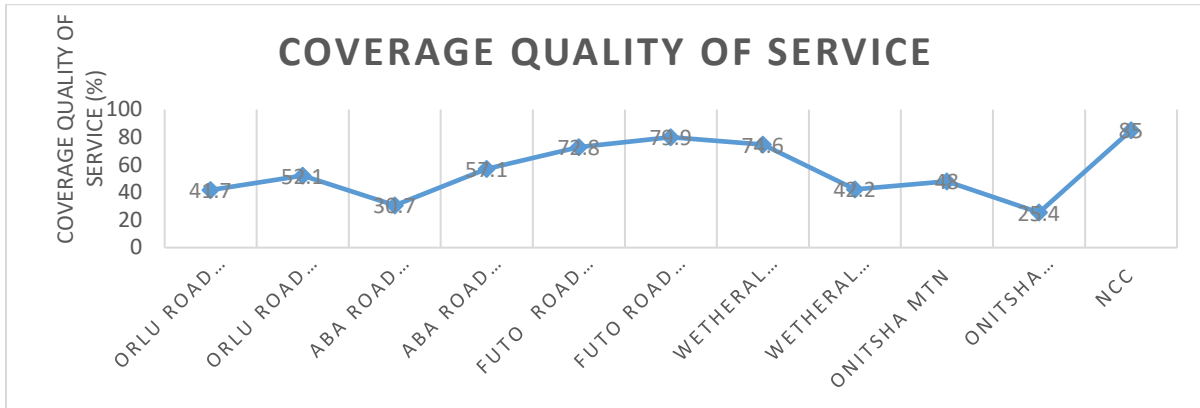
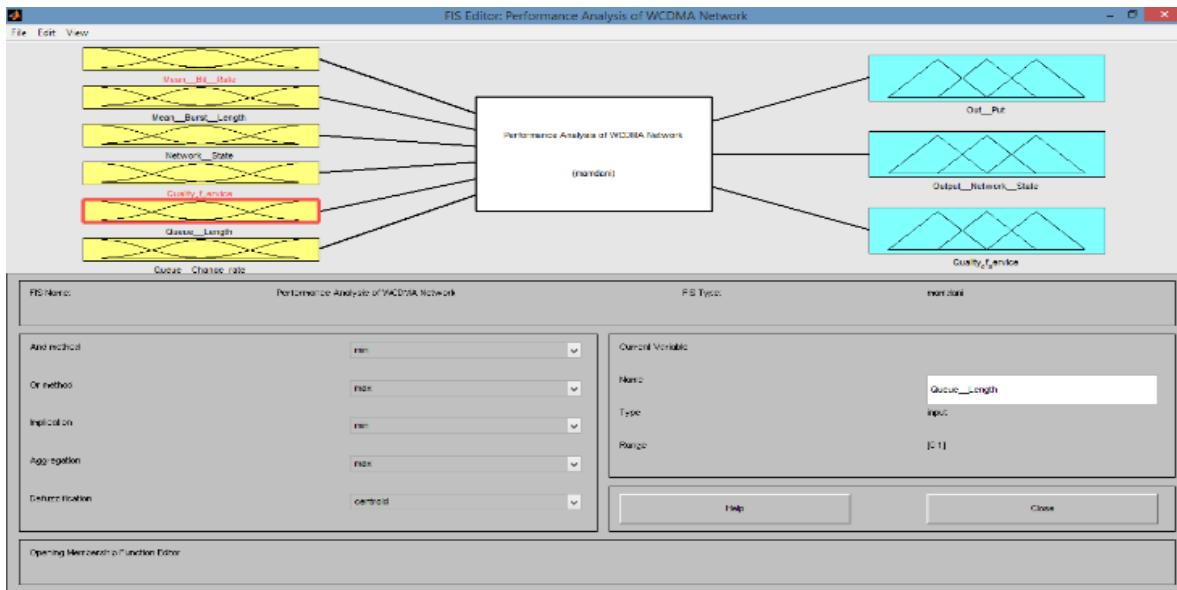
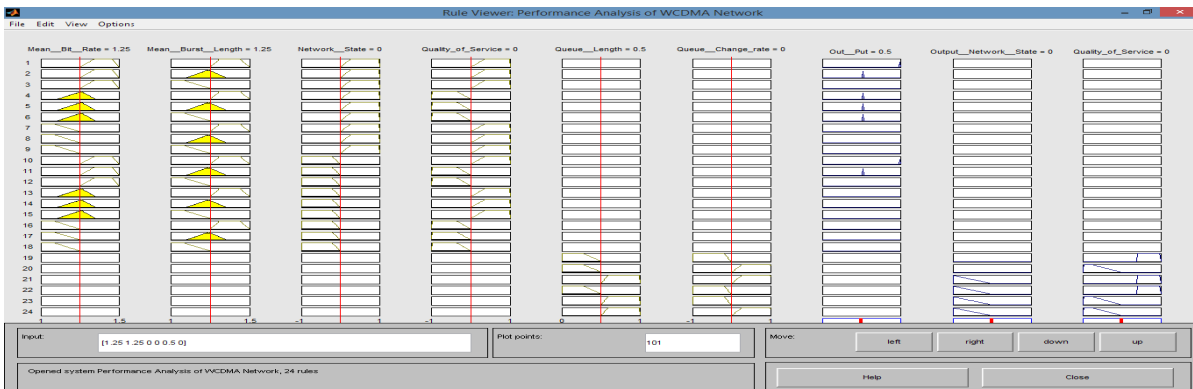


Figure 14 Coverage Quality of Service for WCDMA for MTN and Airtel Networks with the NCC benchmark.



(a)



(b)

Figure 15 the Simulation Interface

Testing was conducted on the proposed system by experimenting its ability to handle some variations of mean bit rate and violation of mean burst size at all stages of its congestion load environments.

4.2. Comparison of the performance characteristics of the improved WCDMA signals using MATLAB tools.

From Table 8, the congestion load environments of the existing system and the output is shown while Table 9 shows the

congestion load environments of the proposed system and the output. We implemented the same values (mean bit rate, mean burst length, network statistics and retainability) obtained in the conventional system, the output variations was actualized from the increased number of the input KPI parameters. The proposed system was compared with the existing system as represented in Figure 16. Figure 17 demonstrated the control surface of the fuzzy logic design system, showing the system axis dimensions: the system output, the mean burst rate and the mean bit rate.

Table 8 shows the congestion load environments for existing system [7]

Traffic load environments	Mean Bit Rate	Mean Burst L	Network S	Retianability	OUT_PUT Of the Existing System
congestion load environment 1	1.1	1.1	0.8989	0.9	0.5
congestion load environment 2	1.2	1.2	0.819	0.9222	0.5
congestion load environment 3	1.3	1.3	0.819	0.9222	0.761
congestion load environment 4	1.4	1.4	0.94	0.9222	0.99
congestion load environment 5	1.4	1.4	0.819	-0.0333	0.5

Table 9 shows the congestion load environments for proposed system

Traffic load environments	Mean Bit Rate	Mean Burst L	Network S	Retianability	OUT_PUT Of proposed System
congestion load environment 1	1.1	1.1	0.8989	0.9	0.63
congestion load environment 2	1.2	1.2	0.819	0.9222	0.69
congestion load environment 3	1.3	1.3	0.819	0.9222	0.831
congestion load environment 4	1.4	1.4	0.94	0.9222	0.99
congestion load environment 5	1.4	1.4	0.819	-0.0333	0.62

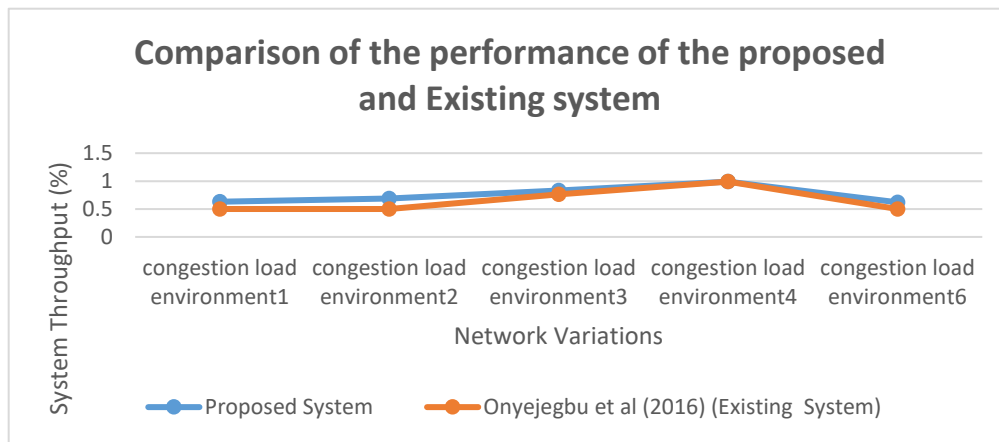


Figure 16 The proposed and existing systems comparison on congestion control efficiency.

4.3. Discussions

Table 6 presented the results obtained for the Received Signal Coded Power of the MTN and Airtel networks along the selected geographical regions. The ranges (dBm) -55 to 0 and -85 to -55 indicated very good and good signal strengths, while -95 to -85, -100 to -95, and -120 to -100 indicated fair, poor, and very poor signal strengths respectively. The addition of the signal strengths between the ranges gave the value of Coverage Reliability of the networks. The result showed that the coverage Reliability of MTN and Airtel networks are below the NCC targeted values refer to Figure 13. Also, the Coverage Quality of Service for the two networks shown in Table 7 indicated that the signal strengths

obtained are below the NCC targeted values as illustrated in Figure 14.

The linguistic inputs (Mean bit Rate, Mean burst Length, Network state, QoS based KPIs, queue length and queue change rate) of the proposed system was fuzzified. The inference mechanism of the fuzzy policer in the network structure extracts the fuzzified inputs to coordinate the flow of calls and ensures that the capacity of the queue length in the base station are not violated using proposed rule base. Efficient management of the queue capacity within the structured network could prevent congestion problems in the network. The fuzzy traffic controller was proposed to ensure concurrent monitoring of the state of the

network in the buffer and call retainability of the users. It communicates to the Fuzzy policer for efficient management of quality service delivery of the WCDMA network. The outputs of the system were defuzzified based on the appropriate decision making of the intelligent system. Tables 8 and 9 represented the congestion load environments for the existing and proposed systems while Figure 16 compared the congestion control efficiency. The proposed system demonstrated adequate congestion control improvement relative to the conventional system with better potentials to adjust the queue capacity within the Nodes ones there is evidenced network instability.

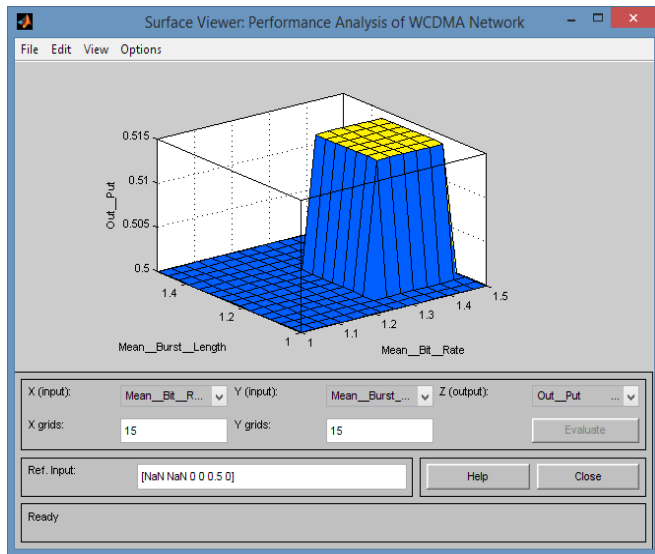


Figure 17 The control surface for the fuzzy logic system

Conclusion

The study was geared towards developing an improved technique to reduce the high level of network congestions observed in 3G wireless networks within Owerri metropolis. The research deployed an optimized fuzzy logic structured with some input parameters such as cell arrival detector, pass/drop switch, buffer, fuzzy policer, fuzzy congestion controller. Selected KPIs such as RXLEV, CDR, CCSR and CSSR were also introduced in the proposed system. Empirical analyses were conducted for adequate evaluation of the present operational conditions of the designated (MTN and AIRTEL) networks observed with high network traffic density within the selected regions. The Selected KPIs were used to evaluate the state-of-the-art conditions of the networks and compared it with the standard specified by the Nigerian Communication Commission (NCC). It was observed that none of the parametric values obtained from the KPIs met the NCC threshold which established the degraded nature of the networks. Analytical scheme known as fuzzy logic technique was introduced to proffer mitigation on the adverse existing congestion challenges. A conceptual framework was adopted from the literature which had similar featured characteristics but was limited with the number of KPIs introduced. The existing work failed to take into cognizance of the fact that congestion could affect other KPIs. The proposed Fuzzy Logic system featured with four KPIs was introduced to the system while varying the congestion load characteristics using the following parameters; mean bit rate, mean burst length, network statistics and retainability. Average of 10.2% increase in the system

throughput was obtained from the proposed system over the conventional system, which provided a significant improvement relative to the system congestion. The study had shown its prominence as to the need to ascertain the actual levels of quality of services provided to the end users. Meanwhile, the incorporation of the approach could enable the network providers to dynamically tackle the issue of network congestion.

References

- [1] Ugwuoke, F. N., Okafor, K. C., Onwusuru, I. M., & Udeze, C. C. "Using Software Engineering approach in mitigation QoS challenges in Mobile Communication Network in Nigeria". *Computing Information Systems Development Informatics & Allied Research Journal*, 15(1), 131-148, 2014.
- [2] Telecommunication management; Key Performance Indicators for UMTS and GSM (KPI). *3 GPP TS 32.410 0, 9.0*, 2009.
- [3] Terms and Definitions Related to Quality of Service and Network Performance Including Dependability. *ITU-T Rec. E.800*, 1993.
- [4] Ozovehe, A., & Usman, A. U. "Performance Analysis of GSM Networks in Minna metropolis of Nigeria". *Nigeria Journal of Technology*, 34(2), 359-367, 2015.
- [5] Atenuga, M., & Isabona, J. "On the compromise between network performance and End-User satisfaction over UMTS Radio interface: An Empirical Investigation". *International Journal of Advance Research in Physical Science*, 1(8), 9-18, 2014.
- [6] Isabona, J., & Ekpenyong, M. "End-User satisfaction Assessment Approach for efficient network performance monitoring in wireless communication systems". *Africa Journal of computing & ICT*, 8(1), 1-18, 2015.
- [7] Onyejebu, L. N., & Okafor, N. R. "Congestion control in Asynchronous Transfer mode". *International Journal in Computer Application*, 142(4), 11-15, 2016.
- [8] Kaidioglu, R. "Performance benchmarking of cellular network operators in Turkey". *Graduate School of Natural and Applied Sciences of ATILIM University Turkey*, Electrical and Electronic Engineering, 2010.
- [9] Speech and Multi Transmission Quality (STQ); QoS aspects for popular services in mobile networks; part 1: Assessment of Quality of Service. *ETSI TS 102 250-1, 2.2.1, 2011*.
- [10] Ononiwu, G., Akinwale, B., Agubor, C., & Onojo, J. "Performance Evaluation of Major Mobile network operators in Owerri metropolis of Nigeria". *International Journal of Engineering Technologies in Computational and Applied Sciences*, 18(1), 06-13, 2016.

Conducted and Radiated Interference on Interconnection's Lines

Patricio E. Munhoz-Rojas*

Department of Research, Development and Innovation, Lactec, Brazil

ARTICLE INFO

Article history:

Received: 28 November, 2018

Accepted: 12 January, 2019

Online : 29 January, 2019

Keywords:

Electromagnetic potentials

Electromagnetic interference

Electromagnetic disturbances

ABSTRACT

There exists a profound difficulty of communication between the people that works in the EMC area in circuit terms and the people that works in field terms.

In this paper we show that when the matter is predominantly distributed along a certain direction in space, as for transmission lines, the electromagnetic field can be divided in two modes, each of them with two degrees of freedom, that are practically independent: a longitudinal TM (transverse magnetic) mode and a transversal TE (transverse electric) mode. We also show that two degrees of freedom of the longitudinal mode are the ones that are described by circuit's theory.

This formulation is based on the observation that, when the matter is macroscopically described by constitutive laws, the electromagnetic field within the matter can be fully characterized in terms of the potential fields, in total four degrees of freedom. Using the above formulation, we put forward a generalized formulation of the coupling of an external electromagnetic field to a transmission line, valid in any time scale.

We apply the above concepts to study, in a common theoretical framework, the iconic case of the conducted and radiated interferences on a transmission line, and we show that:

- 1-Differently than what is normally assumed in standard transmission-line theory, the normal operation mode and the internally-produced electromagnetic field are predominantly a longitudinal TM mode;*

- 2-The longitudinal mode is affected by both the conducted disturbances and the radiated disturbances; while*

- 3- The transverse mode is affected only by the radiated disturbances.*

Then, only for systems where the longitudinal mode is predominant, and, the longitudinal and the transversal modes are practically decoupled, EMI can be simulated using circuit simulation software's.

Also, to further illustrate the interpretation power of this formulation, we present some other application examples.

1. Introduction

This paper is an extension of a work originally presented in the 2018 Joint IEEE EMC & APEMC, which took place in Singapore from 14 to 17 May 2018 [1], and its main purpose is to present in a more detailed form both the general theoretical framework and the demonstration that the circuit's theory applies to the two degrees of freedom of the so-called longitudinal mode: the scalar potential " Φ " in the conductor and the magnetic potential " A_{mz} " along the conductor (the current " i " in the conductor is related to A_{mz} through the concept of inductance).

In reference [1], this theory was applied to study conducted and radiated interferences in transmission lines, because there exists confusion among the people that work in the area, in different time scales; which is caused, as mentioned there, by a rather loose definition of conducted and radiated disturbances.

But nor the answer to the question, why do we feel that they are rather loosely defined? was fully explained in the text of reference [1], neither the primary cause of the confusion was identified.

In reference [1] it was recalled that "in IEC, the disturbing electromagnetic fields are divided into: *conducted disturbances* (IEV 161-03-27) and *radiated disturbances* (IEV 161-03-28).

*Patricio E. Munhoz-Rojas, Email: patricio@lactec.org.br

The essential difference between these disturbances, as established in the above definitions, resides in the manner how the energy is transferred to the conductors:

- For conducted disturbances, IEV 161-03-27 says that the energy is transferred via one or more conductors;

- For radiated disturbances, IEV 161-03-28 says that the energy is transferred through space in the form of electromagnetic waves. However, it notes that “The term “radiated disturbance” is sometimes used to cover induction phenomena”.

We feel that the above definitions of conducted and radiated disturbances are rather loosely defined because they appeal to an intuitive concept of energy on the part of the reader, taking advantage of the fact that the electromagnetic energy is very seldom calculated in interference problems. Besides that, the concept of electromagnetic energy of the people that works in circuit terms is very different than the concept of the people that works in field terms.

IEV definition of electromagnetic energy 121-11-64 says that “The energy associated with an electromagnetic field, in a linear medium within a domain V , is given by the volume integral

$$\frac{1}{2} \int_V (\vec{E} \cdot \vec{D} + \vec{H} \cdot \vec{B}) dV$$

where \mathbf{E} , \mathbf{D} , \mathbf{H} and \mathbf{B} are the four vector quantities determining the electromagnetic field”.

This definition in field terms is not easy to understand for the people working in the low-frequency regime, which is used to work in circuit terms. Also, it is not frequently known what the relation between the above defined concept of electromagnetic energy and the concept of energy and power, used in circuit terms, is.

Then, we can trace back the primary cause of the confusion to the fact that some people work in circuit terms, while other people work in field terms.

As mentioned in reference [1], this is the origin of the different approach and the rather different language used by the people working in the low-frequency regime, as the power quality area, and the people working in the high-frequency regime, both in the lightning protection area and in the EMC area, facts that cause a certain degree of confusion and make it difficult the communication among the people working in these different areas.

Also, as very well observed in reference [2], “Within IEC, power quality is treated within the standards on electromagnetic compatibility (EMC)”. This is because EMI problems, both in the low-frequency regime and in the high-frequency regime, are quantified and measured in terms of *currents* in the conductors and *voltages* between conductors, which is partly due to the widely accepted belief that the EMI on transmission lines is a completely known problem, and it can be simulated using various circuit simulation software’s.

But this belief assumes that the problem of determining to what kind of electromagnetic systems the circuit theory is applicable, or which are the limits of the validity of the circuit theory, is a solved problem. Assumption that is by no means valid [3].

In this paper, as in reference [1], we first demonstrate that the characterization of the electromagnetic field in terms of the *electric scalar potential* “ Φ ” and of the *magnetic vector potential* “ \mathbf{A}_m ” [4], which Maxwell called *electromagnetic momentum* [5], when the matter can be macroscopically described by *constitutive laws*, is totally general. This characterization has in total four degrees of freedom.

Second, we demonstrate that the circuit’s theory applies to the two degrees of freedom of the so-called longitudinal mode: the scalar potential “ Φ ” in the conductor and the magnetic potential “ \mathbf{A}_{mz} ” along the conductor (the current “ i ” in the conductor is related to \mathbf{A}_{mz} through the concept of inductance).

This paper is organized as follows:

In part 2, which closely follows the description given in part II of reference [1], we first review the different forms of description of the electromagnetic field within the matter [4, 6-9] showing that all of them have in common the equations involving the vector fields \mathbf{E} (electric field strength) and \mathbf{B}_m (magnetic flux density), which are commonly called “*electric field*” and “*magnetic field*”.

Also, as in reference [1], it is shown that:

- When the matter can be macroscopically described by *constitutive laws*, which are relations between the other fields (\mathbf{H} , \mathbf{D} and \mathbf{J}_{efr}) needed to describe the electromagnetic field within the matter and the fields \mathbf{E} and \mathbf{B}_m ; then, the electromagnetic field within the matter is fully characterized by the fields \mathbf{E} and \mathbf{B}_m (six degrees of freedom); and

- As the common equations for the fields \mathbf{E} and \mathbf{B}_m can be solved in terms of the so-called *magnetic vector potential* “ \mathbf{A}_m ” and *electric scalar potential* “ Φ ”, in the case when the matter is described by constitutive laws, the number of degrees of freedom needed for the characterization of the electromagnetic field within the matter can be reduced from six (\mathbf{E} , \mathbf{B}_m) to four (\mathbf{A}_m , Φ).

Then, as it happens in the vacuum [10], when the matter is macroscopically described by *constitutive laws*, the electromagnetic field within the matter is fully described by these two potential fields \mathbf{A}_m and Φ .

The extension made in this paper mainly refers to alert the reader on the differences, arising from the different formulations of electromagnetic theory and not always acknowledged, between the formulations utilized by the different application software’s.

In part 3, which closely follows the line of reasoning of part III of reference [1], we first show that when the matter is predominantly distributed along a certain direction in space, as for transmission lines, the four degrees of freedom of the potentials can be separated into two independent modes, each of them with two degrees of freedom:

- a *longitudinal mode* constituted by Φ and the component of \mathbf{A}_m along the line, which is a TM (transverse magnetic) mode [11], and

- a *transversal mode* constituted by the two components of \mathbf{A}_m transversal to the line, which is a TE (transverse electric) mode [11].

As in reference [1], it is also shown that the longitudinal mode is the one described by circuit theory and the relation between its two degrees of freedom is given by Kirchhoff's laws, which are deeply related to the existence and predominance of the longitudinal mode.

As an extension made in this paper, using the longitudinal TM mode as the fundamental building block, instead of the TEM (transverse electromagnetic) mode as in traditional transmission-line theory [12], and assuming that the longitudinal and the transversal modes are practically independent, we present the derivation of a generalized theory of the electromagnetic field coupling to a multiconductor line, in time domain, that, as usual, predicts the propagation of the scalar potential and the current along the line.

For the longitudinal TM mode, as in traditional multiconductor transmission-line theory [12-14], the line voltages have a unique value, independent of the integration path from the reference to the conductor.

The independent transversal mode produces additional induced voltages (integration path dependent) between the conductors of the line.

As mentioned in reference [1], this treatment can also be extended to lines with imperfections, discontinuities or with conductors attached perpendicular to the line, such as, the terminations, equipment connections or groundings [15].

Also, as an extension and an application of the concept of longitudinal mode, the derivation of a general theory of the coupling of an external electromagnetic field to a conductor line is presented. This theory can be applied to different problems, such as, electromagnetic neural stimulation [16-20] and the calculation of lightning-induced voltages [13]. In order to not deviate the attention away from the main purpose of the paper and to avoid too many mathematical derivations in the main body of the paper, we present it in Annex 1.

The generalized theory of the electromagnetic field coupling to a multiconductor line, under the proper simplifications, reduces to the standard coupling theories [12-14,21]. In Annex 2, we present a detailed comparison of this generalized theory with the most important classical theories of the electromagnetic field coupling to a multiconductor line.

In part 4, as in part IV of reference [1], we apply the above theory to analyze the interference on a transmission line produced by external disturbances, which are commonly classified into conducted and radiated disturbances.

As in reference [1]:

- We assume that the *normal operation mode*, which is driven by normal lumped external excitation sources, is a longitudinal mode;

- We divide the externally produced disturbances in two classes: *longitudinal mode disturbances* and *transversal mode disturbances*;

- We divide the longitudinal mode disturbances in two classes: the *scattered* and the externally produced; these last ones are also divided in two classes: the remotely produced and that produced

by the *impressed current*, which is injected by lumped external sources.

We show that:

- the *conducted disturbances* are longitudinal mode disturbances that affect only the longitudinal mode; but,

- the *radiated disturbances* are composed of longitudinal mode disturbances and transversal mode disturbances, both of which affect the longitudinal mode; while the transversal mode is only affected by the transversal mode disturbances.

This is the reason why, only when the longitudinal and the transversal modes are practically decoupled, EMI can be simulated using circuit simulation software's.

Also, this explains why current injection and capacitive clamp testing methods represent only the effect of disturbances, both conducted and radiated, on the longitudinal mode.

In part 5, in order to illustrate the interpretation power of this approach, we present the results of other application cases together the important practical and engineering conclusions that has gone unnoticed in other calculations made with previously proposed approaches/software tools. Finally, in part 6 we present our summary and our main conclusions.

2. Description of the Electromagnetic Field

The description of the electromagnetic field given in this paper closely parallels the description given in reference [1] having only being added some complementary explanations, being (1) to (8) the same of reference [1] and our (13) is equal to (9) of reference [1].

The Maxwell-Hertz classical formulation of the electromagnetic theory [4,6], which refers to 5 vector fields and 1 scalar field (\mathbf{E} , \mathbf{H} , \mathbf{D} , \mathbf{B}_m , \mathbf{J}_{efr} and ρ_{efr}), is commonly called *Maxwell equations* (IEV 121-11-62):

$$\begin{aligned} \nabla \cdot \vec{B}_m &= 0; & \nabla \cdot \vec{D} &= \rho_{\text{efr}}; \\ \nabla \times \vec{E} &= -\frac{\partial \vec{B}_m}{\partial t}; & \nabla \times \vec{H} &= \vec{J}_{\text{efr}} + \frac{\partial \vec{D}}{\partial t}. \end{aligned} \quad (1)$$

Where the four quantities determining the *electromagnetic field* (IEV 121-11-61) are: \mathbf{E} the *electric field strength* (IEV 121-11-18), \mathbf{H} the *magnetic field strength* (IEV 121-11-56), \mathbf{B}_m the *magnetic flux density* (IEV 121-11-19) and \mathbf{D} the *displacement* (IEV 121-11-40). The vector field \mathbf{J}_{efr} is the *electric (conduction) current density* (IEV 121-11-11) and the scalar field ρ_{efr} is the *electric charge density or volumic (electric) charge* (IEV 121-11-07), which according to IEV 121-11-61 are needed to characterize **the electric and magnetic conditions of a material medium together the electromagnetic field**.

Using the following definitions:

$$\begin{aligned} \rho_{\text{efr}} &\equiv -\nabla \cdot \vec{M}_{\text{efr}}; \\ \vec{J}_{\text{efr}} &\equiv \frac{\partial \vec{M}_{\text{efr}}}{\partial t} + \nabla \times \frac{\vec{M}_{\text{mfr}}}{\mu_0}; \\ \vec{D} &\equiv \vec{B}_e - \vec{M}_{\text{efr}}; & \vec{H} &\equiv \vec{H}_e + \frac{\vec{M}_{\text{mfr}}}{\mu_0}. \end{aligned} \quad (2)$$

Where: \mathbf{M}_{efr} and \mathbf{M}_{mfr} are the electric and magnetic matter fields produced by the free electric charges; \mathbf{B}_e , which is called the *electric flux density*, is a vector field composed of \mathbf{D} and \mathbf{M}_{efr} ; and \mathbf{H}_e , which is called the *magnetic field intensity*, is a part of the vector field \mathbf{H} that is composed of \mathbf{H}_e and \mathbf{M}_{mfr} .

Then, (1) can be written in a more symmetrical form [7] as:

$$\begin{aligned} \nabla \cdot \vec{B}_m &= 0; & \nabla \cdot \vec{B}_e &= 0; \\ \frac{\partial \vec{B}_m}{\partial t} &= -\nabla \times \vec{E}; & \frac{\partial \vec{B}_e}{\partial t} &= \nabla \times \vec{H}_e. \end{aligned} \quad (3)$$

Where:

$$\begin{aligned} \vec{B}_e &= \vec{D} + \vec{M}_{\text{efr}} \equiv (\epsilon_0 \vec{E} + \vec{M}_{\text{eb}}) + \vec{M}_{\text{efr}} = \epsilon_0 \vec{E} + (\vec{M}_{\text{eb}} + \vec{M}_{\text{efr}}); \\ \mu_0 \vec{H}_e &= \mu_0 \vec{H} - \vec{M}_{\text{mfr}} \equiv (\vec{B}_m - \vec{M}_{\text{mb}}) - \vec{M}_{\text{mfr}} = \vec{B}_m - (\vec{M}_{\text{mb}} + \vec{M}_{\text{mfr}}). \end{aligned} \quad (4)$$

\mathbf{M}_{eb} and \mathbf{M}_{mb} are the electric and magnetic matter fields produced by the bound charges.

It is important to note that our definition of *electric flux density* \mathbf{B}_e is **different** from the definition of IEV 121-11-40, where they call “electric flux density” to the *displacement* \mathbf{D} (saying that this last terminology is obsolete), despite not being divergence-less. Also, our definition of *magnetic field intensity* \mathbf{H}_e is **different** from the definition of \mathbf{H} , the *magnetic field strength* of IEV 121-11-56.

Equations (3) constitute the formulation of the electromagnetic theory that is called “Symmetrical theory of electromagnetism” [7].

The two vector equations in (3) can be interpreted as state equations, with the flux fields \mathbf{B}_e and \mathbf{B}_m representing the *electric state* and the *magnetic state*, respectively; and the intensity fields (\mathbf{E} and \mathbf{H}_e) being the necessary inputs to produce a change in the states.

Using (4), (1) and (3) can be written as:

$$\begin{aligned} \nabla \cdot \vec{B}_m &= 0; & \nabla \cdot \epsilon_0 \vec{E} &= \rho_{\text{et}}; \\ \nabla \times \vec{E} &= -\frac{\partial \vec{B}_m}{\partial t}; & \nabla \times \frac{\vec{B}_m}{\mu_0} &= \vec{J}_{\text{et}} + \epsilon_0 \frac{\partial \vec{E}}{\partial t}. \end{aligned} \quad (5)$$

Where:

$$\begin{aligned} \vec{M}_e &\equiv \vec{M}_{\text{efr}} + \vec{M}_{\text{eb}}; & \vec{M}_m &\equiv \vec{M}_{\text{mfr}} + \vec{M}_{\text{mb}}; \\ \rho_{\text{et}} &\equiv -\nabla \cdot \vec{M}_e; \\ \vec{J}_{\text{et}} &\equiv \frac{\partial \vec{M}_e}{\partial t} + \nabla \times \frac{\vec{M}_m}{\mu_0}. \end{aligned} \quad (6)$$

From (5) and (6) it is clearly seen that the sources of \mathbf{E} and \mathbf{B}_m are \mathbf{M}_e and \mathbf{M}_m . Equations (5) constitute the Feynman’s formulation of the electromagnetic theory [8].

If instead of separating the matter fields \mathbf{M}_e and \mathbf{M}_m into fields produced by free and bound charges, as in (6), we make the separation of the matter fields into fields produced by charges that are internal (\mathbf{M}_{ei} and \mathbf{M}_{mi}) or external (\mathbf{M}_{eex} and \mathbf{M}_{mex}) to the piece of matter considered; then we have:

$$\begin{aligned} \nabla \cdot \vec{B}_m &= 0; & \nabla \cdot \vec{D}' &= \rho_{\text{ex}}; \\ \nabla \times \vec{E} &= -\frac{\partial \vec{B}_m}{\partial t}; & \nabla \times \vec{H}' &= \vec{J}_{\text{ex}} + \frac{\partial \vec{D}'}{\partial t}. \end{aligned} \quad (7)$$

Where:

$$\begin{aligned} \vec{M}_e &\equiv \vec{M}_{\text{eex}} + \vec{M}_{\text{ei}}; & \vec{M}_m &\equiv \vec{M}_{\text{mex}} + \vec{M}_{\text{mi}}; \\ \rho_{\text{ex}} &\equiv -\nabla \cdot \vec{M}_{\text{eex}}; \\ \vec{J}_{\text{ex}} &\equiv \frac{\partial \vec{M}_{\text{eex}}}{\partial t} + \nabla \times \frac{\vec{M}_{\text{mex}}}{\mu_0}; \\ \vec{D}' &\equiv \epsilon_0 \vec{E} + \vec{M}_{\text{ei}}; & \vec{H}' &\equiv \frac{\vec{B}_m - \vec{M}_{\text{mi}}}{\mu_0}. \end{aligned} \quad (8)$$

Equations (7) constitute the Landau & Lifshitz’s formulation of the electromagnetic theory [9].

All the usual formulations of the electromagnetic theory, as (1), (3), (5) or (7), have in common the equations of the left-hand side involving the vector fields \mathbf{E} and \mathbf{B}_m .

Usually, these left-hand side equations are solved in terms of the so-called *magnetic vector potential* “ \mathbf{A}_m ” and *electric scalar potential* “ Φ ”:

$$\begin{aligned} \nabla \cdot \vec{B}_m &= 0 \Rightarrow \vec{B}_m = \nabla \times \vec{A}_m; \\ \frac{\partial \vec{B}_m}{\partial t} &= -\nabla \times \vec{E} \Rightarrow \nabla \times \left(\vec{E} + \frac{\partial \vec{A}_m}{\partial t} \right) = \vec{0} \rightarrow \vec{E} = -\left(\nabla \phi_m + \frac{\partial \vec{A}_m}{\partial t} \right). \end{aligned} \quad (9)$$

In most electromagnetic theories, the potentials have not physical reality, then the divergent of the magnetic potential can be arbitrarily chosen. The most common choices are:

$$\begin{aligned} \nabla \cdot \left(\frac{\vec{A}_m}{\mu_0} \right) + \frac{\partial}{\partial t} (\epsilon_0 \phi_m) &= 0 \quad \text{Lorenz' gauge}; \\ \nabla \cdot \vec{A}_m &= 0 \quad \text{Coulomb' gauge}; \\ \nabla \cdot \left(\frac{\vec{A}_m}{\mu} \right) + \frac{\partial}{\partial t} (\epsilon \phi_m) &= 0 \quad \text{Maxwell' gauge}. \end{aligned} \quad (10)$$

In the “Symmetrical theory of electromagnetism” [7], there exist potentials that have physical reality, which are those related to the Hertz’ potentials, which fulfill the following restriction:

$$\nabla \cdot \left(\frac{\vec{A}_m}{\mu_0} \right) + \frac{\partial}{\partial t} (\epsilon_0 \phi_m) = g_e c \approx 0. \quad (11)$$

Saying that the Lorenz’ gauge has physical reality. However, as in the normal Lorenz’ gauge, there exist a manifold of magnetic potentials that are mathematically equivalent to the real potentials, in the sense that they produce the same magnetic flux density and the same electric field strength, which are the quantities that are universally recognized as having physical reality. These are:

$$\begin{aligned} \vec{A}_m^* &= \vec{A}_m + \nabla \chi; & \phi_m^* &= \phi_m - \frac{\partial \chi}{\partial t}; \\ \vec{B}_m &= \nabla \times \vec{A}_m = \nabla \times \vec{A}_m^*; \\ \vec{E} &= -\nabla \left(\phi_m - \frac{\partial \chi}{\partial t} \right) - \frac{\partial}{\partial t} (\vec{A}_m + \nabla \chi) = -\nabla \phi_m - \frac{\partial \vec{A}_m}{\partial t}; \\ \text{With } \nabla^2 \chi - \mu_0 \epsilon_0 \frac{\partial^2 \chi}{\partial t^2} &= 0. \end{aligned} \quad (12)$$

In the macroscopic formulation of the electromagnetic theory, usually the vector fields appearing in the right-hand side of (1), (3), (5) or (7) are expressed in terms of the fields \mathbf{E} and \mathbf{B}_m , by means of the so-called *constitutive laws*.

For example, for the Maxwell-Hertz's formulation (see (1)), we have:

$$\vec{J}_{efr} = \sigma \vec{E}; \quad \vec{D} = \varepsilon \vec{E}; \quad \vec{H} = \frac{\vec{B}_m}{\mu}. \quad (13)$$

The parameters “ σ ”, “ ε ” and “ μ ”, appearing in (13), are called *electric conductivity*, *electric permittivity* and *magnetic permeability*, respectively.

Then, (1) can be written in terms of the potentials as:

$$\begin{aligned} \vec{B}_m &= \nabla \times \vec{A}_m; \quad \vec{E} = -\left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right); \\ \nabla \cdot \left(\varepsilon \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right)\right) &= -\rho_{efr}; \\ \nabla \times \left(\frac{\nabla \times \vec{A}_m}{\mu}\right) &= -\sigma \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right) - \frac{\partial}{\partial t} \left(\varepsilon \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right)\right). \end{aligned} \quad (14)$$

From the last two equations of (14) we can obtain:

$$\begin{aligned} \nabla \times \left(\frac{\nabla \times \vec{A}_m}{\mu}\right) &= -\sigma \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right) - \frac{\partial}{\partial t} \left(\varepsilon \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right)\right) \Rightarrow \\ \frac{\partial}{\partial t} \left(\nabla \cdot \left(\varepsilon \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right)\right)\right) &= -\nabla \cdot \left(\sigma \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right)\right) = -\frac{\partial \rho_{efr}}{\partial t}. \end{aligned} \quad (15)$$

The last two equations of (14) are four scalar equations to determine the components of the four-vector formed by “ \mathbf{A}_m ” and “ Φ/c ” [10], which we will call “ \mathbf{A}_{em} ”, whose source is the free electric charge density. Equations (15) says that for a conduction dominated medium the real source is the time variation of the free electric charge density.

Then, in the case where the matter is described via *constitutive laws*, the field “ \mathbf{A}_{em} ”, with only four degrees of freedom, fully describes the electromagnetic field.

This is the formulation applied in [22].

When dealing with interference problems is very important to distinguish what belongs to the system being studied and what is considered an externally applied electromagnetic field.

Then, we will write the equations corresponding to (13) and (14) for the Landau & Lifshitz's formulation (see (7)), which makes this separation. In this case we have:

$$\frac{\partial \vec{D}'}{\partial t} = \sigma^* \vec{E} + \frac{\partial}{\partial t}(\varepsilon \vec{E}); \quad \vec{H}' = \frac{\vec{B}_m}{\mu^*}. \quad (16)$$

The parameters “ σ^* ”, “ ε ” and “ μ^* ”, appearing in (16), are also called *electric conductivity*, *electric permittivity* and *magnetic permeability*, respectively.

Then, (7) can be written as:

$$\begin{aligned} \vec{B}_m &= \nabla \times \vec{A}_m; \quad \vec{E} = -\left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right); \\ \nabla \times \left(\frac{\nabla \times \vec{A}_m}{\mu^*}\right) &= \vec{J}_{ex} - \sigma^* \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right) - \frac{\partial}{\partial t} \left(\varepsilon \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right)\right); \\ \nabla \cdot \left(\sigma^* \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right) + \frac{\partial}{\partial t} \left(\varepsilon \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right)\right)\right) &= -\frac{\partial \rho_{ex}}{\partial t} = \nabla \cdot \vec{J}_{ex}. \end{aligned} \quad (17)$$

Again, the last two equations of (17) are four scalar equations to determine the components of the four-vector “ \mathbf{A}_{em} ”, formed by

“ \mathbf{A}_m ” and “ Φ/c ”, whose sources are the external electric charge density and the external current density. Then, also in this case the field “ \mathbf{A}_{em} ”, with only four degrees of freedom, fully describes the electromagnetic field.

We can write the last two equations of (17) in a different manner:

$$\begin{aligned} \mu^* \varepsilon \frac{\partial^2 \vec{A}_m}{\partial t^2} + \mu^* \left(\sigma^* + \frac{\partial \varepsilon}{\partial t}\right) \frac{\partial \vec{A}_m}{\partial t} - \nabla^2 \vec{A}_m - \left(\frac{\nabla \mu^*}{\mu^*}\right) \times (\nabla \times \vec{A}_m) &= \\ = \mu^* \vec{J}_{ex} - \mu^* \left(\sigma^* + \frac{\partial \varepsilon}{\partial t}\right) \nabla \phi + \nabla (\mu^* \varepsilon) \frac{\partial \phi}{\partial t} - \nabla \left(\nabla \cdot \vec{A}_m + \mu^* \varepsilon \frac{\partial \phi}{\partial t}\right); \\ \nabla \cdot \left(\sigma^* \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right) + \frac{\partial}{\partial t} \left(\varepsilon \left(\nabla \phi + \frac{\partial \vec{A}_m}{\partial t}\right)\right)\right) &= -\frac{\partial \rho_{ex}}{\partial t}. \end{aligned} \quad (18)$$

Equations (18) show that, in the case of the electromagnetic field within the matter, even in the Maxwell's gauge, in general, the equations for ϕ and \mathbf{A}_m are not separable, as in the case of the vacuum. Besides that, the simple retarded solutions are not applicable.

For the special case of time-harmonic electromagnetic fields, or in frequency domain, (17) is simplified, by the fact that we can express the external charge density in terms of the divergence of the current density, and for the last two equations of (17) we have:

$$\begin{aligned} \nabla \cdot \left(\varepsilon (\nabla \Phi_\omega + i\omega \vec{A}_{m\omega})\right) &= -i \frac{\nabla \cdot \vec{J}_{ex\omega}}{\omega} - \nabla \cdot \left(\sigma^* \left(\vec{A}_{m\omega} - i \frac{\nabla \Phi_\omega}{\omega}\right)\right); \\ \nabla \times \left(\frac{\nabla \times \vec{A}_{m\omega}}{\mu^*}\right) &= (\vec{J}_{ex\omega} - \sigma^* \nabla \Phi_\omega) - i\omega (\sigma^* \vec{A}_{m\omega} + \varepsilon \nabla \Phi_\omega) + \omega^2 \varepsilon \vec{A}_{m\omega}. \end{aligned} \quad (19)$$

Equations (19) show clearly that the sources of the induced electromagnetic field are the external electric charge density and the external current density.

This is the formulation applied in [23,24].

3. Transmission-Lines

As in reference [1], and our (20) to (24) are the same than (11) to (15) of reference [1], in the case of the transmission lines, matter is predominantly distributed along a certain direction in space, which we will call it “z”, making this a preferential direction.

As the four-dimensional vector field “ \mathbf{A}_{em} ” transforms as a four-vector [10], for a transformation of coordinates between a system moving with a velocity “v”, along the preferential z-axis, relative to another “rest” system, which we denote by the index “0”, we have:

$$\begin{bmatrix} \phi \\ c \\ A_{mx} \\ A_{my} \\ A_{mz} \end{bmatrix} = \begin{bmatrix} \gamma \left(\frac{\phi_0}{c} - \frac{v}{c} A_{m0z}\right) \\ A_{m0x} \\ A_{m0y} \\ \gamma \left(A_{m0z} - \frac{v}{c} \frac{\phi_0}{c}\right) \end{bmatrix}; \quad \gamma \equiv \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}. \quad (20)$$

From (20) we can see that, for horizontal (along the z-axis) ideal transmission lines (having an uniform cross section), we can divide the electromagnetic field into two independent modes, each one with two degrees of freedom: a first one composed of “ Φ ” and “ A_{mz} ”, which will be called *longitudinal mode*; and a second one composed of “ A_{mx} ” and “ A_{my} ”, which will be called *transversal mode*.

Transmission lines constituted of insulated filamentary conductors, disposed along the z-axis, can internally produce only the longitudinal mode. The transversal mode must be externally produced.

The transversal mode can only be internally produced if the transversal extension of the conductive matter is relevant or if the filamentary conductors are not insulated.

For the longitudinal mode ($A_{mx} = A_{my} = 0$), from the first two equations of (11) or (14), we have:

$$\begin{aligned} E_x &= -\frac{\partial\Phi}{\partial x}; & E_y &= -\frac{\partial\Phi}{\partial y}; & E_z &= -\frac{\partial\Phi}{\partial z} - \frac{\partial A_{mz}}{\partial t}; \\ B_{mx} &= \frac{\partial A_{mz}}{\partial y}; & B_{my} &= -\frac{\partial A_{mz}}{\partial x}; & B_{mz} &= 0. \end{aligned} \quad (21)$$

Equations (21) say that the longitudinal mode is a TM mode, and, that the transversal electric field is *conservative* and is equal to the transversal gradient of the scalar potential Φ . This allows the definition of *transverse voltages* that are single valued [11, 12, 14].

For the transversal mode ($\Phi = A_{mz} = 0$), we have:

$$\begin{aligned} E_x &= -\frac{\partial A_{mx}}{\partial t}; & E_y &= -\frac{\partial A_{my}}{\partial t}; & E_z &= 0; \\ B_{mx} &= -\frac{\partial A_{my}}{\partial z}; & B_{my} &= \frac{\partial A_{mx}}{\partial z}; & B_{mz} &= \left(\frac{\partial A_{my}}{\partial x} - \frac{\partial A_{mx}}{\partial y} \right). \end{aligned} \quad (22)$$

Equations (22) say that the transversal mode is a TE mode, with a non-conservative transversal electric field.

As an example of a pure longitudinal mode, we will see first the special case treated by Schelkunoff [11], of an infinite horizontal hollow conductor of arbitrary cross-section, rigid and made of perfectly conducting matter. Here, if we can assume that “ Φ ” and “ A_{mz} ” are separable, we have:

$$\Phi = -T(x, y)V(z, t); \quad \frac{A_{mz}}{\mu_0} = -T(x, y)I(z, t). \quad (23)$$

Where, in equations (23) the T functions are dimensionless functions of x and y.

Using (23) in (21), we have the Schelkunoff’s TM modes [11]:

$$\begin{aligned} E_x &= \frac{\partial T}{\partial x}V; & E_y &= \frac{\partial T}{\partial y}V; & E_z &= T\left(\frac{\partial V}{\partial z} + \mu_0 \frac{\partial I}{\partial t}\right); \\ B_{mx} &= -\mu_0 \frac{\partial T}{\partial y}I; & B_{my} &= \mu_0 \frac{\partial T}{\partial x}I; & B_{mz} &= 0. \end{aligned} \quad (24)$$

Also, from the equations on the right-hand side of (1) and (3), we have:

$$\begin{aligned} (\nabla \times \vec{H})_x &= -\frac{\partial}{\partial z} \left(\frac{\partial T}{\partial x} I \right) = -\frac{\partial T}{\partial x} \frac{\partial I}{\partial z} = \varepsilon \frac{\partial}{\partial t} \left(\frac{\partial T}{\partial x} V \right) = \varepsilon \frac{\partial T}{\partial x} \frac{\partial V}{\partial t}; \\ (\nabla \times \vec{H})_y &= \frac{\partial}{\partial z} \left(-\frac{\partial T}{\partial y} I \right) = -\frac{\partial T}{\partial y} \frac{\partial I}{\partial z} = \varepsilon \frac{\partial}{\partial t} \left(\frac{\partial T}{\partial y} V \right) = \varepsilon \frac{\partial T}{\partial y} \frac{\partial V}{\partial t}; \\ \rightarrow \frac{\partial I}{\partial z} + \varepsilon \frac{\partial V}{\partial t} &= 0. \end{aligned} \quad (25)$$

Equations (24) and (25) display the main characteristic of Schelkunoff’s TM modes, which is the fact that, even in the case of a hollow conductor made of perfectly conducting matter, despite having an electric field strength value equal to zero within the conductor, there can be an electromagnetic field inside the conductor, which propagates along the z axis, following transmission-line equations. The existence of this longitudinal

mode and its predominance, even for hollow conductors that are slowly bent, leads to the theory of waveguides.

The other very important special case, where the longitudinal mode is predominant, is the case of a multi-conductor transmission line having, in general, many (N+1) filamentary conductors.

Here, following the same line of reasoning and adopting the terminology utilized in reference [1], we will describe the interaction of an arbitrary external electromagnetic field with a straight segment of a multiconductor line.

In Annex 1 the case of an elementary single-wire line is treated in detail. There we present the derivation of a general theory of the coupling of an external electromagnetic field to a conductor line and a detailed comparison of this generalized theory with the most important classical theories. This theory can be applied to different problems, such as, electromagnetic neural stimulation [16-20], the calculation of lightning-induced voltages [13].

In the case of a multiconductor line, we can obtain from (9) or (21) the value of the longitudinal electric field strength, **at a point internal to the “j” conductor** (see Figure 1), for $j = 0 \dots N$:

$$-\frac{\partial\Phi^j}{\partial z} - \frac{\partial A_{mz}^j}{\partial t} = E_z(x_j, y_j, z, t). \quad (26)$$

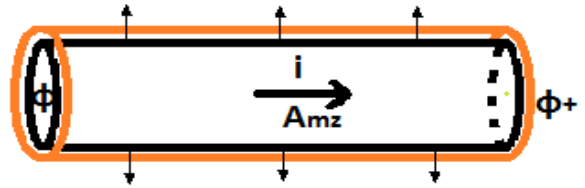


Figure 1 – Segment of a conductor, of length “ Δz ”, between two scalar potential nodes.

Also, integrating the divergence of the last equation of (14) in a closed surface that involves a **segment of length “ Δz ” of the horizontal “j” conductor** (see Figure 2), we have:

$$\frac{\partial i^j}{\partial z} + \frac{\partial \lambda^j}{\partial t} = -I_{LS}^j. \quad (27)$$

Where, “ i ” is the total current flowing through the cross-section of the “j” conductor, “ λ ” is the charge accumulated on the surface of the “j” conductor, per unit length, and “ I_{LS}^j ” is the conduction current flowing out of the “j” conductor through the lateral surface, per unit length.

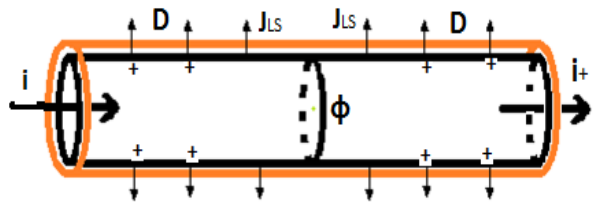


Figure 2 – Segment of a conductor, of length “ Δz ”, around a scalar potential node.

If, as usual, we assume that for this kind of transmission line, for which the distances among the different conductors of the line is much shorter than its length, the Maxwell’s concept of inductance and capacitance coefficients are valid [25,26].

This will allow us to express the magnetic potential in a short segment of a filamentary conductor, at a point "z" along the line, in terms of the current in all the conductors at the same point "z", and, the charge at the surface of a conductor, at a point "z", in terms of the scalar potential at all the conductors at the same point "z". Then, we have:

$$\begin{aligned} A_{mz}^j &= A_{mz}^{j(\Delta)} + A_{mz}^{j(ex)}; & \Phi^j &= \Phi^{j(\Delta)} + \Phi^{j(ex)}; \\ A_{mz}^{j(\Delta)} &= [L_k^j] i^k(z, t); & \lambda^j &= [C_k^j] \Phi^{k(\Delta)}(z, t). \end{aligned} \quad (28)$$

Where: "A_{mz}^{j(Δ)}" and "Φ^{j(Δ)}" are the potentials produced by the matter existing in all the conductors within the segment "Δz", at the point "z"; and "A_{mz}^{j(ex)}" and "Φ^{j(ex)}", are the potentials produced by the matter existing in all the conductors, outside the segment "Δz", plus the potentials "A_{mz}^{j(ex)}" and "Φ^{j(ex)}" representing the externally applied field.

Using Ohm's law for the line conductors, we have:

$$E_z(x_j, y_j, z, t) = R_{(ej)} i^j(z, t). \quad (29)$$

Where "R_(ej)" is the resistance, per unit length, of the conductor "j".

Extracting from the conduction current "I_{LS}^j", flowing out of the conductor through the lateral surface, the part that is due to the linear leakage current, we can write:

$$I_{LS}^j(z, t) = [G_k^j] \Phi^{k(\Delta)} + I_{LS}^{j*}(z, t). \quad (30)$$

Then, (26) and (27) can be written as:

$$\frac{\partial \Phi^j}{\partial z} + \left(R_{(ej)} i^j + [L_k^j] \frac{\partial i^k}{\partial t} \right) = - \frac{\partial A_{mz}^{j(ex)}}{\partial t}; \quad (31)$$

and

$$\frac{\partial i^j}{\partial z} + [G_k^j] \Phi^{k(\Delta)} + [C_k^j] \frac{\partial \Phi^{k(\Delta)}}{\partial t} = -I_{LS}^{j*}. \quad (32)$$

Where: "[L*]" is the matrix of the inductance coefficients, per unit length; "[C*]" is the matrix of the capacitance coefficients, per unit length; and "[G*]" is the matrix of conductance coefficients, per unit length, between the segments of all conductors at the point "z".

To describe a rectilinear transmission line of a uniform cross-section and of a finite length, we can use here again, as usual in transmission-line theory [12], the matrices [L], [C] and [G], which are calculated neglecting the retardation effects and when every conductor of the line is uniformly charged with a charge density, per unit length, which is equal to its value at the point "z", and the current in all segments along the line, in any conductor, is equal to its value at the point "z". The potentials produced in this situation are the *static potentials* "A_{mz}^(st)" and "Φ^(st)". Using for every conductor "j" the following definitions, we have:

$$\begin{aligned} A_{mz}^{j(int)} &\equiv A_{mz}^j - A_{mz}^{j(ex)}; & A_{mz}^{j(c)} &\equiv A_{mz}^{j(int)} - A_{mz}^{j(st)}; \\ \rightarrow A_{mz}^j &= A_{mz}^{j(st)} + \left(A_{mz}^{j(ex)} + A_{mz}^{j(c)} \right); \\ A_{mz}^{j(st)} &\equiv [L_k^j] i^k(z, t); & & \\ \Phi^{j(int)} &\equiv \Phi^j - \Phi^{j(ex)}; & \Phi^{j(c)} &\equiv \Phi^{j(int)} - \Phi^{j(st)}; \\ \rightarrow \Phi^j &= \Phi^{j(st)} + \left(\Phi^{j(ex)} + \Phi^{j(c)} \right); \\ \lambda^j(z, t) &\equiv [C_k^j] \Phi^{k(st)}; & [G_k^j] \Phi^{k(\Delta)} &\equiv [G_k^j] \Phi^{k(st)}. \end{aligned} \quad (33)$$

Where the usual summation rule over repeated index is applied.

Then, we can write (31) and (32) as:

$$\frac{\partial \Phi^j}{\partial z} + \left(R_{(ej)} i^j + [L_k^j] \frac{\partial i^k}{\partial t} \right) = - \frac{\partial \left(A_{mz}^{j(ex)} + A_{mz}^{j(c)} \right)}{\partial t}; \quad (34)$$

and

$$\begin{aligned} \frac{\partial i^j}{\partial z} + [G_k^j] \Phi^k + [C_k^j] \frac{\partial \Phi^k}{\partial t} &= \\ = [G_k^j] \left(\Phi^{k(ex)} + \Phi^{k(c)} \right) + [C_k^j] \frac{\partial \left(\Phi^{k(ex)} + \Phi^{k(c)} \right)}{\partial t} - I_{LS}^{j*}. \end{aligned} \quad (35)$$

As mentioned in reference [1], "the main variables in (34) and (35) are the scalar and the magnetic potentials, which are not uniquely determined (see (12)). Their values, as well as the value of the inductance and capacitance coefficients, are affected by the choice of the reference point, which is the point where the value of the potentials is equal to zero.

But, as the fields **E** and **B_m** are uniquely determined and, for practical purposes, the important voltages are the ones occurring between the conductors of the line, this fact is of little or no importance."

We will choose the reference for the potentials at the infinity, or at a line located very far from the multiconductor transmission line.

With this choice of reference, (34) can be interpreted as the application of Kirchhoff's circuits law applied to a "mesh" formed by a segment of the conductor "j" of the line and the reference line; and (35) as the application of Kirchhoff's circuits law applied to a potential node on the conductor "j" of the line. Then, we can see that the longitudinal TM mode is the one described by circuit theory.

Also, (34) and (35) are completely general and rigorous equations that describe the interaction of an external electromagnetic field with the considered filamentary multiconductor transmission line, being the only assumptions in deriving these equations: the thin-wire approximation for all the conductors and the validity of Ohm's law for all the conductors. Thus, they constitute a generalized formulation of the electromagnetic coupling to a transmission line that, under the proper simplifications, reduces to the standard coupling theories [12-14,21] (see Annex 2).

One of the advantages of the time domain formulation of the longitudinal mode is that (34) and (35) can also be written as state equations:

$$\frac{\partial i^j}{\partial t} = -[L_j^k]^{-1} \left(\frac{\partial \Phi^k}{\partial z} + R_{(ck)} i^k \right) - [L_j^k]^{-1} \left(\frac{\partial \left(A_{mz}^{k(ex)} + A_{mz}^{k(c)} \right)}{\partial t} \right); \quad (34a)$$

and

$$\begin{aligned} \frac{\partial \Phi^j}{\partial t} &= -[C_j^k]^{-1} \left(\frac{\partial i^k}{\partial z} + [G_m^k] \Phi^m \right) + \\ + \left(\frac{\partial \left(\Phi^{j(ex)} + \Phi^{j(c)} \right)}{\partial t} + [C_j^k]^{-1} [G_m^k] \left(\Phi^{m(ex)} + \Phi^{m(c)} \right) - [C_j^k]^{-1} I_{LS}^{k*} \right). \end{aligned} \quad (35a)$$

Then, (34a) and (35a), which are a generalization of (21a) and (22a) of reference [1], as already mentioned in reference [1], “can be interpreted as saying that the scalar potential “ Φ ” and the current “ j ” represent the time evolving state of the system, at a point internal to the “ j ” conductor.” **Showing that the essence of the circuit theory is to assume that the longitudinal mode is predominant.**

Equations (34a) and (35a), after being spatially discretized into N segments (for $n = 1 \dots N$), can be written as:

$$\frac{d i_n^j}{d t} = -[L_j^k]^{-1} (R_{(c k)} i_n^k) - [L_j^k]^{-1} \frac{(\Phi_{n+1}^k - \Phi_n^k)}{\Delta z} + [L_j^k]^{-1} \left(\frac{\partial (A_{mzn}^{k(ex)} + A_{mzn}^{k(c)})}{\partial t} \right); \quad (34b)$$

and

$$\frac{d \Phi_n^j}{d t} = -[C_j^k]^{-1} [G_m^k] \Phi_n^m - [C_j^k]^{-1} \frac{(i_n^k - i_{n-1}^k)}{\Delta z} + \left(\frac{\partial (\Phi_n^{j(ex)} + \Phi_n^{j(c)})}{\partial t} + [C_j^k]^{-1} [G_m^k] (\Phi_n^{m(ex)} + \Phi_n^{m(c)}) + [-C_j^k]^{-1} I_{LSn}^{k*} \right). \quad (35b)$$

Equations (34b) and (35b), as also mentioned in reference [1], “are a system of coupled ordinary differential equations (ODE), for the scalar potential at potential nodes and the current at current nodes.” Also, as already mentioned in reference [15], “this system of coupled ordinary differential equations can then be solved using the powerful ODE solvers now existing” [27].

The main advantage of describing uniform lines by per-unit length parameters, calculated using the static potentials is that these parameters are constant and easy to calculate, particularly, when the reference is taken at the infinity, because it is a reference independent of the direction of the line. For obtaining approximate solutions of non-uniform lines, assuming that the longitudinal mode is predominant, the non-uniform line is usually modeled as a cascade of uniform sections, conductively connected [28]; or in the case of lines with periodically varying cross-section, such as the cables composed of twisted-wire pairs, the line is modeled as an equivalent line, having per-unit length parameters equal to the average over the period [29].

Up to now, we have studied the uniform or slowly non-uniform part of the line.

As mentioned in reference [1], using reference [15], “we will take advantage of this formulation that allows to include the interaction of a segment of conductor with any known arbitrary external electromagnetic field, which can be described by $\Phi^{(ex)}$ and $\mathbf{A}^{(ex)}$, to include the terminations of the line, the discontinuities or even conductors attached perpendicular to the line, such as, groundings or other equipment connected.”

Considering that the longitudinal mode is predominant, we can model them as circuit elements located between two potential nodes, which are displaced in space along a certain direction.

For a vertical conductor (y axis direction) that is relatively small (compared to the minimum wavelength of interest), so that we can neglect the corrections due to the time delay in the

production of the potentials, and neglecting also the non-linear transversal conduction current I_{LS}^* , we have in a manner analogous to (31) and (32):

$$\frac{\partial (\Phi - \Phi^{(ex)})}{\partial y} + R_c i + L \frac{\partial i}{\partial t} = - \left(\frac{\partial \Phi^{(ex)}}{\partial y} + \frac{\partial A_{my}^{(ex)}}{\partial t} \right); \quad (36)$$

and

$$\frac{\partial i}{\partial y} + G (\Phi - \Phi^{(ex)}) + C \frac{\partial (\Phi - \Phi^{(ex)})}{\partial t} = 0. \quad (37)$$

This is the explanation of the important remark quoted in the conclusions of reference [14] “When using the scattered-voltage formulation, it must be remembered that the vertical component of the incident electric field appears as a voltage source in the line terminations.”

If, like in the case of a grounding of the reference conductor, the vertical conductor is connected at the potential node “ $\Phi_n^{(0)}$ ”, the modified equations for this node are:

$$\frac{d i_n^0}{d t} = -[L_0^k]^{-1} (R_{(c k)} i_n^k) - [L_0^k]^{-1} \frac{(\Phi_{n+1}^k - \Phi_n^k)}{\Delta z} + [L_0^k]^{-1} \left(\frac{\partial (A_{mzn}^{k(ex)} + A_{mzn}^{k(c)})}{\partial t} \right); \quad (38)$$

and

$$\frac{d \Phi_n^0}{d t} = -[C_0^k]^{-1} [G_m^k] \Phi_n^m - [C_0^k]^{-1} \frac{(i_n^{k1} + i_n^k - i_{n-1}^k)}{\Delta z} + \left(\frac{\partial (\Phi_n^{0(ex)} + \Phi_n^{0(c)})}{\partial t} + [C_0^k]^{-1} [G_m^k] (\Phi_n^{m(ex)} + \Phi_n^{m(c)}) + [-C_0^k]^{-1} I_{LSn}^{k*} \right). \quad (39)$$

Where, in this case, only “ i_n^{01} ”, which is the current in the first segment of the vertical conductor, is different from zero.

As mentioned in reference [1], “this formulation allows representing the groundings not as a connection to the ground, but as a connection to the grounding electrodes, whose potential is to be considered as one of the states of the system.

This is very convenient for the real cases where the ground is not perfectly conducting, and it is considered as the return conductor of the line.”

Finally, as in reference [1], to include the effect of the transversal mode, we will define the so-called *conductor voltages*.

Usually, in standard transmission-line theory, the *conductor voltages* are defined relative to one of the conductors chosen to be the reference [12, 30]. This was the choice indirectly adopted in reference [1].

Here, we will define the *conductor “j” voltage*, at a point “z” along the conductor, as the integral of the transversal electric field strength “ \mathbf{E}_t ” between the conductor “j” and the reference of the potentials:

$$\begin{aligned} V^j(z, t) &\equiv - \int_{\infty}^{(x_j, y_j)} \vec{E}_t \cdot d\vec{l}_t = \\ &= \Phi^j + \int_{\infty}^{(x_j, y_j)} \frac{\partial \vec{A}_{mt}^{(ex)}}{\partial t} \cdot d\vec{l}_t \equiv \Phi^j + V_{in}^j. \end{aligned} \quad (40)$$

Where we will assume that “ A_{mt} ” is the externally produced field, but, as mentioned in reference [1], “the formulation also allows to include the transversal field produced by transversal conductors.”. Then, we can modify (34) and (35), by means of (40), and obtain:

$$\begin{aligned} \frac{\partial V^j}{\partial z} + R_{(ej)} i^j + [L_k^j] \frac{\partial i^k}{\partial t} = \\ = - \frac{\partial}{\partial t} \left(A_{mz}^{j(ex)} - \int_{\infty}^{(x_j, y_j)} \frac{\partial \vec{A}_{mt}^{(ex)}}{\partial z} \cdot d\vec{l}_i \right) - \frac{\partial A_{mz}^{j(c)}}{\partial t}; \end{aligned} \quad (41)$$

and

$$\begin{aligned} \frac{\partial i^j}{\partial z} + [G_k^j] V^k + [C_k^j] \frac{\partial V^k}{\partial t} = [G_k^j] V^{k(ex)} + [C_k^j] \frac{\partial V^{k(ex)}}{\partial t} \\ + \left([G_k^j] \Phi^{k(c)} + [C_k^j] \frac{\partial \Phi^{k(c)}}{\partial t} - I_{LS}^* \right). \end{aligned} \quad (42)$$

Where, as mentioned in reference [1], “we can see that writing the equations in terms of the *conductor voltages* produce new terms in (41) and (42), due to the existence of induced voltages (integration path dependent) produced by the external transversal mode.”

For the special case of time-harmonic electromagnetic fields, or in frequency domain, (41) and (42) can be written as:

$$\frac{dV_{\omega}^j(z)}{dz} + \left([R_k^j] + i\omega [L_k^j] \right) I_{\omega}^k(z) = S_{E_{\omega}}^j(z); \quad (43)$$

and

$$\frac{dI_{\omega}^j(z)}{dz} + \left([G_k^j] + i\omega [C_k^j] \right) V_{\omega}^k(z) = S_{H_{\omega}}^j(z). \quad (44)$$

Which, without the sources, are the standard transmission-line equations in frequency domain, such as shown in (2) of reference [30] or (6.7) and (6.8) of reference [31].

As already mentioned in reference [1], “we must note that, for the longitudinal mode, the important values of the externally produced field are at points located inside the conductors, and its effect can be represented as a voltage or a current source; while, for the transversal mode, the important values of the externally produced field are at points located in the dielectric between the conductors, and its effect is to produce an induced voltage between the conductors.”.

Also, as mentioned in reference [1], “it must be emphasized that, in order to define the concept of voltage associated with one point of the line, it is necessary to introduce a reference.

When the reference is at the infinity, or in the case of an imperfectly conducting Earth where the “Reference perfect conductor plane” is located at a remote position within the Earth, both the conductor “j” voltage and the scalar potential of the “j” conductor, will have embedded the complexities of the electromagnetic field distribution in the ground.

Calculate the electromagnetic field distribution within the Earth is a very difficult task, but fortunately, the voltages and the potential differences **between aerial conductors**, which are the ones of practical importance, will not depend directly on those complexities.”.

4. Interferences produced by external Disturbances

In this part, as in part IV of reference [1], we will apply the generalized theory of the electromagnetic coupling to a transmission line, developed in part 3, to analyze the interference

on a transmission line produced by external disturbances, which are commonly classified into conducted and radiated disturbances.

To study the interference due to an external disturbance, we must first separate the sources of the externally produced electromagnetic fields in two classes: the normal external sources and the disturbing external sources.

In the interconnections, which are multiconductor transmission lines, the dominant mode is the longitudinal mode, and, the longitudinal and the transversal modes are practically decoupled. Then, as in reference [1], we will assume that the *normal operation mode*, which is driven by normal lumped external excitation sources, is a longitudinal mode.

IEV 161-03-27 says, for conducted disturbances, that the energy is transferred via one or more conductors. So, conducted disturbances are locally produced.

IEV 161-03-28 says, for radiated disturbances, that the energy is transferred through space in the form of electromagnetic waves; and, it notes that “The term “radiated disturbance” is sometimes used to cover induction phenomena”. Then, radiated disturbances are always remotely produced.

Here, as in reference [1], “we will divide the externally produced disturbances in two classes: the first class that produce Φ and A_{mz} , which will be called *longitudinal mode disturbances*; and the second class that produce A_{mx} and A_{my} , which will be called *transversal mode disturbances*.”

Also, as in reference [1], within the internally-produced longitudinal mode, “ $\Phi^{(int)}$ ” and “ $A_{mz}^{(int)}$ ”, we must make a distinction between the scattered part and the part that is produced by the current that is injected by lumped external sources, both normal and disturbing.

Then, we have:

$$\begin{aligned} \Phi^j &= \Phi^{j(int)} + \Phi^{j(ex)} \equiv \left(\Phi^{j(sc)} + \Phi^{j(exinjN)} + \Phi^{j(exinjD)} \right) + \Phi^{j(ex)}; \\ A_{mz}^j &= A_{mz}^{j(int)} + A_{mz}^{j(ex)} \equiv \left(A_{mz}^{j(sc)} + A_{mz}^{k(exinjN)} + A_{mz}^{k(exinjD)} \right) + A_{mz}^{j(ex)}; \\ \Rightarrow i^j &= [L_j^k]^{-1} \left(A_{mz}^{k(sc)} + A_{mz}^{k(exinjN)} + A_{mz}^{k(exinjD)} \right) = i^{j(sc)} + i^{j(exinj)}; \\ A_{mx}^j &= A_{mx}^{j(ex)}; \quad A_{my}^j = A_{my}^{j(ex)}. \end{aligned} \quad (45)$$

Where: the index “N” indicates normal, the index “D” indicates disturbing and the index “exinj” indicates produced by the lumped external sources, conductively connected to the line, that injects current in the line. Our (45) is equal to (33) of reference [1].

In this paper, neglecting the internally-produced transversal magnetic potential, which is produced by the current in the transversal conductors and the transversal leakage currents, we have assumed, as in reference [1], that the transversal mode is produced only by the radiated disturbances.

As mentioned in reference [1], the effect of the transversal mode is:

- to produce an induced voltage between the conductors, and
- to act as a lumped voltage source located at the transversal conductors, such as the line terminations and groundings.

Then:

- The so-called “*conducted disturbances*” (IEV 161-03-27) are characterized by the fields $\Phi^{(exinjD)}$ and $A_{mz}^{(exinjD)}$, which are

produced by a current locally injected into the line, by lumped disturbing external sources conductively connected to it; and
 - The so-called “*radiated disturbances*” (IEV 161-03-28) are characterized by the fields $\Phi^{(ex)}$, $A_{mz}^{(ex)}$, $A_{mx}^{(ex)}$ and $A_{my}^{(ex)}$, which are produced by remotely located sources.

The confusing note existing in IEV 161-03-28 that “The term “*radiated disturbance*” is sometimes used to cover induction phenomena”, can be explained because the electromagnetic field, produced by remotely located sources, includes both the induction field and the radiation field:

$$\begin{aligned} \Phi^{j(ex)} &= \Phi^{j(ind)}; \\ A_{mx}^{j(ex)} &= A_{mx}^{j(ind)} + A_{mx}^{j(rad)}; \\ A_{my}^{j(ex)} &= A_{my}^{j(ind)} + A_{my}^{j(rad)}; \\ A_{mz}^{j(ex)} &= A_{mz}^{j(ind)} + A_{mz}^{j(rad)}. \end{aligned} \tag{46}$$

Where: the index “*ind*” indicates *induction field*, which is negligible in the far field region, and the index “*rad*” indicates *radiation field*.

From equations (45) we can see that:

- 1- The conducted disturbances, “ $\Phi^{(exinjD)}$ ” and “ $A_{mz}^{(exinjD)}$ ”, which produce scalar potentials and inject currents into the line, are longitudinal mode disturbances that affect only the longitudinal mode; while
- 2- The radiated disturbances, which are composed of longitudinal mode disturbances, $\Phi^{(ex)}$ and $A_{mz}^{(ex)}$, and transversal mode disturbances, $A_{mx}^{(ex)}$ and $A_{my}^{(ex)}$, affect:
 - The longitudinal mode, directly by the longitudinal mode disturbances through the scalar potential “ $\Phi^{(ex)}$ ” and the magnetic potential “ $A_{mz}^{(ex)}$ ” along the conductors; and indirectly by the transversal mode disturbances through the lumped voltage sources at the transversal conductors and terminations, which represent the effect of “ $A_{mt}^{(ex)}$ ” along the transversal conductors;
 - The transversal mode, through the magnetic potentials $A_{mx}^{(ex)}$ and $A_{my}^{(ex)}$, which produce an induced voltage between the conductors.

This is the reason why, only when the longitudinal is predominant, and, the longitudinal and the transversal modes are practically decoupled, EMI can be simulated using circuit simulation software’s.

Also, this explains why current injection and capacitive clamp testing methods represent only the effect of disturbances on the longitudinal mode.

Examples of radiated disturbances, where the longitudinal mode disturbances are dominant, can be seen in references [14,21,32], which deal with lightning-induced voltages; and in reference [33], which deal with voltages induced in twisted-wire pairs by a parallel wire excited by a voltage source.

Examples of radiated disturbances, where the transversal mode disturbances are important can be seen in:

- reference [14], where the exciting source is a lumped voltage source, which represents the induced voltage produced by the transversal mode disturbance, along a transversal conductor at input terminal of the line;
- reference [34], where the induced voltage, produced by the transverse mode disturbance, is an important part of the total voltage, and even the dominant part for the first three microseconds.

In reference [2], the conducted disturbances are defined as “Any deviation from the ideal voltage or current waveform”, meaning the presence of a disturbing scalar potential or a longitudinal magnetic potential produced by the current, which is locally injected by lumped disturbing external sources. Examples of common types and new types of power quality disturbances can be seen in reference [2].

5. Application Case Results

To show the usefulness of this formulation, we present, as in reference [1], the results of some application cases to real transmission lines with transversal conductors, which are connected to the earthing electrodes, in the case of a real ground that is not a perfect conductor. Adding here, in order to show the interpretation power of this formulation, some important practical and engineering conclusions that had gone unnoticed in other calculations made with previously proposed approaches/software tools.

Firstly, as in reference [1], we present in Figure 3 that corresponds to Figure 1 of reference [1], the results of the calculation of the phase-to-neutral voltage, induced by a vertical return stroke that strikes close to a line, which have a neutral wire grounded at only one point, with a grounding resistance of 100 Ω .

As mentioned in reference [1], the purpose of this example is to show that, in this case, the impinging electromagnetic pulse, composed of both a longitudinal mode disturbance and a transversal mode disturbance, initially practically produces only common mode on the line, and, the phase-to-neutral voltage is negligible until the instant when the electromagnetic disturbance reaches the grounding conductor of the neutral wire.

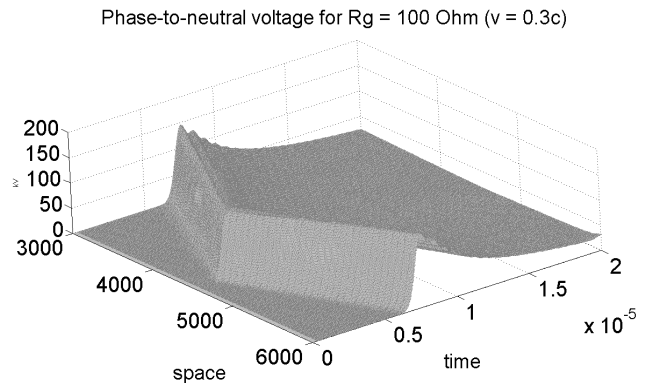


Figure 3 - Phase-to-neutral voltage induced by a 100 kA (2x40 μ s) return stroke, with velocity 0.3c, which strikes at z = 4000 m, at 100 m from a line, with a neutral wire having an isolated grounding at z = 4500 m, with Rg = 100 Ω . Data taken from [15]

Then, due to the presence of the transversal grounding conductor of the neutral wire, when the electromagnetic transversal mode disturbance reaches the grounding conductor, the current produced by the impinging electric field along the conductor generates voltage reduction waves, which are different in size in the different conductors, thus producing a phase-to-neutral voltage pulse, which propagates in both directions along the line, representing the conversion of common mode into differential mode.

This effect that had gone unnoticed in other calculations made with previously proposed approaches/software tools [15,35], has the important engineering conclusion that the main mitigating effect for the induced voltages is not produced by the shielding

wire acting as an extended protective device but by the shielding wire grounding acting as a localized protective device.

As a second example, we present in Figure 4 that corresponds to Figure 2 of reference [1], the results of the calculation of the phase-to-neutral voltage, induced by a vertical return stroke that strikes close to a line, having a neutral wire that is periodically grounded.

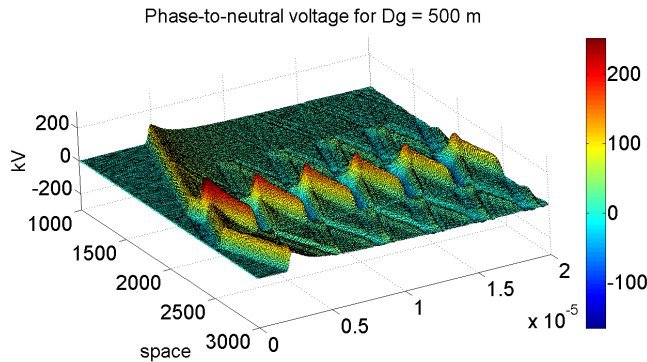


Figure 4 - Phase-to-neutral voltage, in a line periodically grounded each 500 m ($R_g = 10$ Ohms), produced by a 30kA return stroke ($0.3 \times 40 \mu s$), occurring at $z=2375$ m, 50 m from the line. Data taken from [15].

From Figure 4 it can be seen, as already observed in Figure 3, that the impinging electromagnetic pulse produces initially only a common mode perturbation on the line, which propagates in both directions starting from the point of the line which is closest to the stroke location, and the phase-to-neutral voltage is negligible until the instant when the electromagnetic disturbance reaches a transversal grounding conductor of the neutral wire. Then, a phase-to-neutral voltage pulse is produced, due to the injection of current in the transversal wire, which propagates in both directions along the line.

In this case, the effect of the presence of the transversal periodic grounding conductors is to confine the bigger phase-to-neutral overvoltage in the region of the line which is closest to the return stroke location.

Also in this case, the calculations pointed to an effect, which had gone unnoticed in other calculations made with previously proposed approaches/software tools [36], that the presence of the grounding reduces the overvoltage at its location and behind it (seen from the return stroke location), and also it reduces the overvoltage in front of it, but only within a certain distance, fact which defines an “effective distance”.

Then, as mentioned in reference [15], “the shielding wire groundings not only do not protect the line span in front of the return stroke location, but they also produce big positive phase-to-neutral overvoltage”.

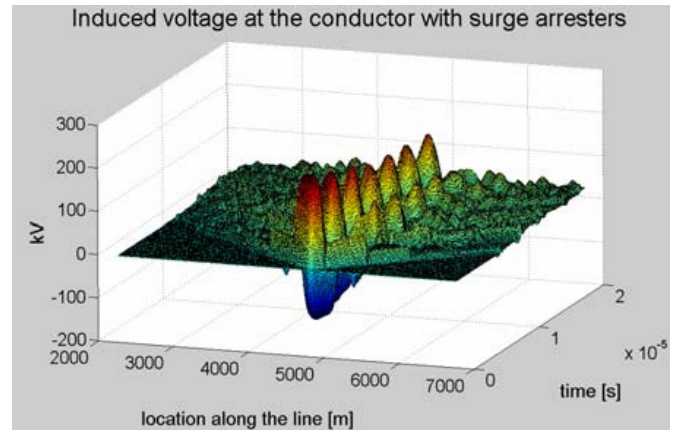
In the next example, we present in Figure 5 that corresponds to Figure 3 of reference [1], the results of the calculation of the voltage induced on the conductors of a line, by a vertical return stroke that strikes close to a line, having surge arresters placed periodically on one of the conductors.

From Figure 5, we can see that the effect of the presence of the transversal surge arresters, with its grounding conductors, is:

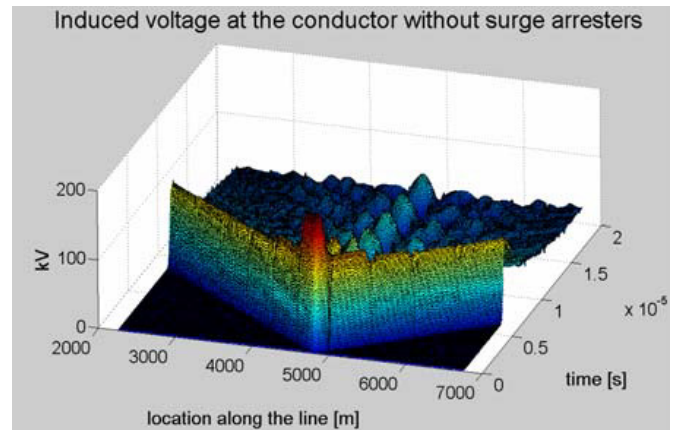
- On the conductor with surge arresters, to confine the phase-to-reference ground overvoltage, which is oscillating, in the region of the line close to the return stroke location, being the bigger

overvoltage, in the region of the line closest to the return stroke location; and,

- On the conductor without surge arresters, to produce just a mild dampening of the overvoltage, produced by the impinging electromagnetic pulse, outside the region of the line closest to the return stroke location.



a) Voltage at the conductor with surge arresters.



b) Voltage at the conductor without surge arresters.

Figure 5– Voltages along a line with surge arresters on one of the conductors, at each 450 m ($R_g = 10$ Ohms), produced by a 45kA lightning discharge ($2 \times 40 \mu s$) propagating with a velocity $v = 0.3 c$, which strikes at $z = 4800$ m, at 70 m from the line. Data taken from [37].

As a result, they produce a conversion of common mode into differential mode overvoltage.

Also, in this case, the calculations pointed to an effect that had gone unnoticed in other calculations made with previously proposed approaches/software tools [36] and which has engineering importance. As mentioned in reference [37], “Multiple surge arresters, placed at periodic intervals along the line, protect the line outside the span in front of the lightning strike location; but the protection afforded to the span in front of the lightning strike location depends on the risetime of the induced voltages. Consequently, any statement on the effectiveness of an interval between arresters should be qualified by a risetime for which it is valid.”

6. Summary and Conclusions

We have shown that, when the matter is macroscopically described by *constitutive laws*, the electromagnetic field within the matter can be fully described by the potentials: the *magnetic vector*

potential “ A_m ” and the electric scalar potential “ Φ ”, with its four degrees of freedom. This conclusion is valid for any time scale and for any frequency, provided the constitutive laws are valid.

We have shown that, when the matter is predominantly distributed along a certain direction in space, as in a transmission line, the electromagnetic field can be divided into two practically independent modes, each one with two degrees of freedom: a longitudinal mode having “ Φ ” and “ A_{mz} ”, and, a transversal mode having “ A_{mx} ” and “ A_{my} ”. Kirchhoff’s laws and circuit theory applies to the two degrees of freedom of the longitudinal mode: the scalar potential “ Φ ” and the current “ i ” in the conductor, which is related to “ A_{mz} ” by the concept of inductance. They represent the time evolving state of the system, at points internal to the conductors.

Using the longitudinal mode as the fundamental building block, and if the longitudinal and the transversal modes are practically independent, we present the derivation of a generalized theory of the electromagnetic field coupling to a multiconductor line, in time domain, that, as usual, predicts the propagation of the scalar potential and the current along the line. We have shown that this generalized coupling theory, under the proper simplifications, reduces to the standard coupling theories and the transmission-line equations there obtained also reduce to the standard transmission-line equations.

We have also shown that, when the longitudinal is predominant, the theory can be extended to include the terminations of the line, the discontinuities or even conductors attached perpendicular to the line.

Also, this formulation allows representing the groundings not as a connection to the reference ground, but as a connection to the grounding electrodes, whose potential is to be considered as one of the states of the system. This is very convenient for the real cases where the ground is not perfectly conducting.

Analyzing the interference produced by external disturbances in these terms, we have assumed that the normal operation mode, which is a differential mode driven by normal lumped external excitation sources, is a longitudinal mode.

We also have divided the externally produced disturbances in two classes: longitudinal mode disturbances and transversal mode disturbances.

We have shown that the conducted disturbances are longitudinal mode disturbances that affect only the longitudinal mode, and the radiated disturbances are composed of longitudinal mode disturbances and transversal mode disturbances, both of which affect the longitudinal mode.

We have shown that the transversal mode disturbances, which are part of the radiated disturbances, affect also the transversal mode. The transversal mode produces induced voltages (integration path dependent) in the insulation between the conductors of the line, and, between any conductor and the reference adopted.

This is the reason why, only when the longitudinal is predominant, and, the longitudinal and the transversal modes are practically decoupled, EMI can be simulated using circuit simulation software’s.

In order to illustrate the usefulness of this interpretation, we have presented the results of the application of this formulation to real lines having perpendicular conductors attached to them, some of them connected to earthing electrodes. We have shown that the

main effect of the presence of these conductors is to produce the conversion of common mode into differential mode.

Acknowledgment

The author would like to thank to LACTEC for supporting this work, and, to thank to the members of CIGRE WG C4.44, to his colleagues at LACTEC and to two anonymous reviewers for their comments and questions.

Annexes

1. Electromagnetic Field Coupling to a Conductor Line

Here, following the same line of reasoning utilized in part 3 and adopting the terminology utilized in reference [1], the very important special case of an infinite horizontal filamentary solid conductor submitted to an externally applied electromagnetic field, which has been studied since Maxwell’s time, will be studied.

Historically, firstly it was studied the electromagnetic field produced by it, when the externally applied field impressed a sinusoidal current in the conductor [38]; but lately, it has been studied the scattered field produced by it, when excited by an externally impinging full-wave electromagnetic field (time-harmonic electromagnetic field) [13].

In this paper, we will study the case of a horizontal filamentary solid conductor, of an infinite extent, submitted to an arbitrary externally applied electromagnetic field.

For the value of the longitudinal electric field strength, at a point internal to the conductor (see Figure I-1), from (9) we have:

$$-\frac{\partial \Phi(x_1, y_1, z, t)}{\partial z} - \frac{\partial A_{mz}(x_1, y_1, z, t)}{\partial t} = E_z(x_1, y_1, z, t). \quad (I-1)$$

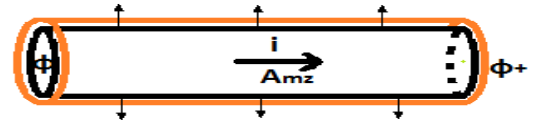


Figure I-1 – Segment of the conductor, of length “ Δz ”, between two scalar potential nodes.

Also, from the equations on the right-hand side of (1) and (3), we have:

$$\nabla \cdot \frac{\partial \vec{B}_e}{\partial t} = \nabla \cdot \left(\vec{J} + \frac{\partial \vec{D}}{\partial t} \right) = \nabla \cdot \vec{J} + \frac{\partial \rho_{eff}}{\partial t} = 0. \quad (I-2)$$

Integrating (I-2) in a closed surface that involves a segment of length “ Δz ” of the horizontal conductor (see Figure I-2), we have:

$$\frac{\partial i(z, t)}{\partial z} + I_{LS}(z, t) = -\frac{\partial \lambda(z, t)}{\partial t}. \quad (I-3)$$

Where: “ i ” is the total current flowing through the cross-section of the conductor, “ I_{LS} ” is the conduction current flowing out of the conductor through the lateral surface, per unit length, and “ λ ” is the free electric charge accumulated on the surface of the conductor, per unit length.

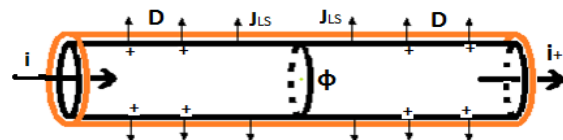


Figure I-2 – Segment of a conductor, of length “ Δz ”, around a scalar potential node.

Equation (1-3) can be interpreted as the application of the Kirchhoff's law to that particular "potential node".

If we separate the potentials produced by the matter existing within the segment " Δz " (" $A_{mz}^{(\Delta)}$ " and " $\Phi^{(\Delta)}$ "), from the potentials produced by the rest of the matter, we have:

$$A_{mz} = A_{mz}^{(\Delta)} + A_{mz}^{(ext)}; \quad \Phi = \Phi^{(\Delta)} + \Phi^{(ext)}; \quad (1-4)$$

$$A_{mz}^{(\Delta)}(x_1, y_1, z, t) = L^* i(z, t); \quad \lambda(z, t) = C^* \Phi^{(\Delta)}(x_1, y_1, z, t).$$

Where: the potentials produced by the rest of the matter, " $A_{mz}^{(ext)}$ " and " $\Phi^{(ext)}$ ", are the potentials produced by the matter existing in the conductor, outside the segment " Δz ", plus the potentials " $A_{mz}^{(ex)}$ " and " $\Phi^{(ex)}$ " representing the externally applied field; and, " L^* " and " C^* " are, respectively, the inductance per unit length and the capacitance per unit length of the segment " Δz " of the conductor.

Using Ohm's law for the conductor, we have:

$$E_z(x_1, y_1, z, t) = R_c i(z, t). \quad (1-5)$$

Where, " R_c " is the resistance, per unit length, of the conductor.

Then, (1-1) and (1-3) can be written as:

$$\frac{\partial \Phi}{\partial z} + R_c i + L^* \frac{\partial i}{\partial t} = - \frac{\partial A_{mz}^{(ext)}(x_1, y_1, z, t)}{\partial t}; \quad (1-6)$$

and

$$\frac{\partial i}{\partial z} + C^* \frac{\partial \Phi}{\partial t} = C^* \frac{\partial \Phi^{(ext)}(x_1, y_1, z, t)}{\partial t} - I_{LS}(z, t). \quad (1-7)$$

As normally, at least a part of the conduction current " I_{LS} ", flowing out of the conductor through the lateral surface, is due to the linear leakage current, we can write:

$$I_{LS}(z, t) = G^* \Phi^{(\Delta)}(x_1, y_1, z, t) + I_{LS}^*(z, t). \quad (1-8)$$

And (1-7) can be written as:

$$\frac{\partial i}{\partial z} + G^* \Phi + C^* \frac{\partial \Phi}{\partial t} = G^* \Phi^{(ext)} + C^* \frac{\partial \Phi^{(ext)}}{\partial t} - I_{LS}^*. \quad (1-9)$$

Equations (1-6) and (1-9) have the form of transmission-line equations, and, they are completely general and rigorous equations that describe the interaction of an external electromagnetic field with a straight finite segment of a single-wire line, being the only assumptions in deriving these equations: the thin-wire approximation and the validity of Ohm's law for the conductor.

They can be applied to short horizontal single-wire lines and to long horizontal single-wire lines; also, they can be applied to bent lines, because, in (1-6) and (1-9), " z " represents the direction of the segment of conductor being described, which can be any direction in real space.

Then, they can be applied to different problems, such as, electromagnetic neural stimulation and the calculation of lightning-induced voltages.

As mentioned in reference [1], the main variables in (1-6) and (1-9) are the potentials, which we have seen are very useful for classification purposes. But they are not uniquely determined.

Their values, as well as the value of the inductance and capacitance coefficients, are affected by the choice of the reference point, which is the point where the value of the potentials is equal to zero. The reference point is normally chosen at the infinity, and with this choice, from (5) we have, in the Lorenz' gauge, the usual retarded potentials [4]:

$$\Phi(\vec{r}, t) = \frac{1}{4\pi\epsilon_0} \iiint_{all\ space} \frac{\rho_{et}(\vec{r}', t')}{|\vec{r} - \vec{r}'|} dV'; \quad (1-10)$$

$$\vec{A}_m(\vec{r}, t) = \frac{\mu_0}{4\pi} \iiint_{all\ space} \frac{J_{et}(\vec{r}', t')}{|\vec{r} - \vec{r}'|} dV'.$$

Another usual choice assumes the existence of a line along the z direction, where:

$$E_z(ref) = A_{mz}(ref) = \Phi_m(ref) = 0. \quad (1-11)$$

This line, which is taken as the reference, could be located at the infinite or very far from the structures of interest, in such a way that the potentials produced by the matter behind the reference are negligible.

Sometimes the reference is chosen on a perfectly conducting conductor. When this perfect conductor has an infinite plane surface, the potentials produced by the matter behind the reference, are represented by the potentials produced by the images of the matter on the perfect conductor plane [13].

With any choice of reference, as also mentioned in reference [1], Φ in (1-6) and (1-9), can also be interpreted as the **potential difference** between the conductor and the reference, at a value of z and t .

In the classical transmission-line theory, which is geared for long horizontal transmission lines, the values utilized, for the capacitance per unit length " C " and for the inductance per unit length " L " of the line, are those calculated, neglecting the retardation effects, for an infinite straight line which is uniformly charged with a charge density, per unit length, equal to " λ " and with a current " i " equal in all segments along the line [12]. In this situation, the potentials produced are the so-called *static potentials* " $A_{mz}^{(st)}$ " and " $\Phi^{(st)}$ ".

We will define the so-called *internally-produced potentials* " $A_{mz}^{(int)}$ " and " $\Phi^{(int)}$ ", which are the potentials really produced by the matter existing in the conductor; and, we will also define the potentials " $A_{mz}^{(c)}$ " and " $\Phi^{(c)}$ ", which are the difference between the internally-produced potentials and the static potentials.

$$A_{mz}^{(int)} \equiv A_{mz} - A_{mz}^{(ex)}; \quad A_{mz}^{(c)} \equiv A_{mz}^{(int)} - A_{mz}^{(st)}; \quad A_{mz}^{(st)} = Li(z, t);$$

$$\rightarrow A_{mz} = A_{mz}^{(st)} + (A_{mz}^{(ex)} + A_{mz}^{(c)}); \quad (1-12)$$

$$\Phi^{(int)} \equiv \Phi - \Phi^{(ex)}; \quad \Phi^{(c)} \equiv \Phi^{(int)} - \Phi^{(st)}; \quad \lambda(z, t) = C\Phi^{(st)};$$

$$\rightarrow \Phi^{(st)} = \Phi - (\Phi^{(ex)} + \Phi^{(c)}); \quad G^* \Phi^{(\Delta)} = G\Phi^{(st)}.$$

Writing (1-6) and (1-9) in terms of the quantities defined in (1-12), we have:

$$\frac{\partial \Phi}{\partial z} + R_c i + L \frac{\partial i}{\partial t} = - \frac{\partial (A_{mz}^{(ex)} + A_{mz}^{(c)})}{\partial t}; \quad (1-13)$$

and

$$\begin{aligned} \frac{\partial i}{\partial z} + G\Phi + C \frac{\partial \Phi}{\partial t} &= \\ &= G\left(\Phi^{(ex)} + \Phi^{(c)}\right) + C \frac{\partial \left(\Phi^{(ex)} + \Phi^{(c)}\right)}{\partial t} - I_{LS}^* \end{aligned} \quad (1-14)$$

To allow an easy comparison with the equations utilized by the people interested in lightning-induced voltages, we will write (1-13) and (1-14) in terms of the internally-produced scalar potential. Then, we have:

$$\frac{\partial \Phi^{(int)}}{\partial z} + R_c i + L \frac{\partial i}{\partial t} = - \left(\frac{\partial \Phi^{(ex)}}{\partial z} + \frac{\partial A_{mz}^{(ex)}}{\partial t} \right) - \frac{\partial A_{mz}^{(c)}}{\partial t}; \quad (1-15)$$

and

$$\frac{\partial i}{\partial z} + G\Phi^{(int)} + C \frac{\partial \Phi^{(int)}}{\partial t} = C \frac{\partial \Phi^{(c)}}{\partial t} + G\Phi^{(c)} - I_{LS}^* \quad (1-16)$$

For the special case of time-harmonic electromagnetic fields, or in frequency domain, (1-15) and (1-16) can be written as:

$$\frac{d\Phi_{\omega}^{(int)}}{dz} + (R_c + i\omega L)I_{\omega} = E_z^{(ex)} - i\omega A_{mz}^{(c)}; \quad (1-17)$$

and

$$\frac{dI_{\omega}}{dz} + (G + i\omega C)\Phi_{\omega}^{(int)} = (G + i\omega C)\Phi_{\omega}^{(c)} - I_{LS\omega}^* \quad (1-18)$$

To show the generality of (1-15) and (1-16), we will compare (1-17) and (1-18) with reference [13]; where the problem of a single-wire line above a perfectly conducting ground, in presence of an impinging electromagnetic field, is considered.

The result there presented is a system of “generalized” or “full-wave” equations, containing electrodynamics corrections (including radiation) to the standard theory of transmission lines. Before making the comparison we must note that:

- In reference [13], the total fields are not divided in: internally produced by the single-wire and externally produced, but, they are divided in what they call “scattered fields”, which are produced by both: the single-wire line and its image in the perfectly conducting ground, and what they call “the exciting electric field “E^{ex}”, which is obtained by the sum of the incident field “Eⁱ” and the ground-reflected field “E^r”, both determined in the absence of the wire”.

- Reference [13] only considers an externally impinging electromagnetic field, neglecting the eventual localized external sources conductively connected to the single-wire. Then, the *internally-produced potentials* reduce to the *scattered potentials*. Reference [13] calculates the potentials produced by the matter really existing in the conductor, outside the segment “Δz”, considering the time delay in the production of the potentials. They in fact calculate the potentials “A_{mz}^(c)” and “Φ^(c)”.

Then, when the perfectly conducting ground is taken as the reference for the potentials, neglecting the conductor resistance, per unit length, “R_c” and the transversal currents “I_{LS}”, (1-17) and (1-18) reduce to (9) and (10) of reference [13].

2. Comparison of our Field Coupling model with standard coupling theories

Here, as in part 3, we will adopt the same terminology utilized in reference [1].

In order to allow an easy comparison with classical transmission-line theories, we will write (34) and (35) in terms of the internally-produced scalar potential. Then, we have:

$$\frac{\partial \Phi^{j(int)}}{\partial z} + \left(R_{(cj)} i^j + [L_k^j] \frac{\partial i^k}{\partial t} \right) = E_z^{j(ex)} - \frac{\partial A_{mz}^{j(c)}}{\partial t}; \quad (2-1)$$

and

$$\begin{aligned} \frac{\partial i^j}{\partial z} + [G_k^j] \Phi^{k(int)} + [C_k^j] \frac{\partial \Phi^{k(int)}}{\partial t} &= \\ &= [G_k^j] \Phi^{k(c)} + [C_k^j] \frac{\partial \Phi^{k(c)}}{\partial t} - I_{LS}^{j*} \end{aligned} \quad (2-2)$$

Also, in order to compare our results with reference [14], which is considered to represent the classical transmission-line theory [13], for this quasi-longitudinal mode (because of the presence of “I_{LS}”), neglecting the transversal magnetic potential “A_{mt}” produced by the transversal current “I_{LS}”, we will define the *internally-produced line voltage with respect to the reference conductor*, as in reference [14]:

$$V^{j(int)}(z,t) = \Phi^{j(int)}(x_j, y_j, z, t) - \Phi^{0(int)}(x_0, y_0, z, t). \quad (2-3)$$

The *internally-produced line voltage*, defined in (2-3), when the external sources conductively connected to the line are not considered, is identical to the so-called *scattered line voltage*, defined in (7) of reference [14], which is usually adopted in the transmission-line theory utilized by the people interested in lightning-induced voltages [13].

Using the definition of (2-3), (2-1) and (2-2) can be written as:

$$\begin{aligned} \frac{\partial V^{j(int)}}{\partial z} + \left(R_{(cj)} i^j + ([L_k^j] - [L_k^0]) \frac{\partial i^k}{\partial t} \right) &= \\ &= E_z^{j(ex)} + \left(R_{(c0)} i^0 - E_z^{0(ex)} \right) - \frac{\partial A_{mz}^{j(c)}}{\partial t} + \frac{\partial A_{mz}^{0(c)}}{\partial t}; \end{aligned} \quad (2-4)$$

and

$$\begin{aligned} \frac{\partial i^j}{\partial z} + [G_k^j] V^{k(int)} + [C_k^j] \frac{\partial V^{k(int)}}{\partial t} &= -[G_k^j] \Phi^{0(int)} + \\ &- [C_k^j] \frac{\partial \Phi^{0(int)}}{\partial t} + \left([G_k^j] \Phi^{k(c)} + [C_k^j] \frac{\partial \Phi^{k(c)}}{\partial t} - I_{LS}^{j*} \right). \end{aligned} \quad (2-5)$$

If we separate, from the current in the reference conductor, the part corresponding to the return current of the other conductors, we can define:

$$i^0 = i^{0*} - \sum_{k \neq 0} i^k. \quad (2-6)$$

Then, (2-4) can be written as:

$$\begin{aligned} \frac{\partial V^{j(int)}}{\partial z} + \left([R_k^j] i^k + ([L_k^j] - [L_k^0]) \frac{\partial i^k}{\partial t} \right) &= \\ &= E_z^{j(ex)} - E_z^{0(ex)} + R_{(c0)} i^{0*} - \frac{\partial}{\partial t} \left(A_{mz}^{j(c)} - A_{mz}^{0(c)} \right); \end{aligned} \quad (2-7)$$

To show the generality of (34) and (35), we will compare (2-7) and (2-5), which are equivalent to (34) and (35), with the corresponding equations of reference [14]:

- Neglecting the differences between the static magnetic potentials and the internally-produced magnetic potentials, the unbalanced current in the reference conductor, and if, from the externally applied electromagnetic field, only the externally impinging or incident electromagnetic field is considered, (2-7) reduce to (16) or (33) of reference [14];

- Neglecting the differences between the static magnetic potentials and the internally-produced magnetic potentials (the electro-dynamics corrections (including radiation) mentioned in reference [13]), the non-linear conduction current I_{LS}^* , flowing out of the “j” conductor through the lateral surface, and assuming that the internally-produced line voltage is equal to the scattered line voltage, (2-5) almost reduce to (32) of reference [14].

The terms remaining in the right-hand side of (2-5) come from an assumption that is made here, which is different from the respective assumption made in reference [14]: The capacitance coefficients are here defined in the Maxwell’s way, in terms of the scalar potentials in the conductors; while in reference [14] they are defined in terms of the scattered line voltages with respect to the reference conductor.

Then, as the scattered line voltages represent only differential modes, the remaining terms represent the conversion of common mode into differential mode.

When the internally-produced line voltage “U^j”, in the conductor “j”, is defined with respect to the reference line for the potentials, we have:

$$U^{j(int)}(z,t) = \Phi^{j(int)}(x_j, y_j, z, t). \quad (2-8)$$

Using the definition of (2-8), (2-1) and (2-2) now read as:

$$\frac{\partial U^{j(int)}}{\partial z} + \left(R_{(cj)} i^j + [L_k^j] \frac{\partial i^k}{\partial t} \right) = E_z^{j(ex)} - \frac{\partial A_{mz}^{j(c)}}{\partial t}; \quad (2-9)$$

and

$$\begin{aligned} \frac{\partial i^j}{\partial z} + [G_k^j] U^{k(int)} + [C_k^j] \frac{\partial U^{k(int)}}{\partial t} &= \\ = [G_k^j] \Phi^{k(c)} + [C_k^j] \frac{\partial \Phi^{k(c)}}{\partial t} - I_{LS}^* &. \end{aligned} \quad (2-10)$$

Equations (2-9) and (2-10), which are equivalent to (34) and (35), are also completely general and rigorous equations that describe the interaction of an external electromagnetic field with the considered filamentary multi-conductor transmission line, being the only assumptions in deriving these equations: the thin-wire approximation for all the conductors and the validity of Ohm’s law for all the conductors.

Now, if we compare (2-9) and (2-10), with the corresponding equations of reference [14]:

- Neglecting the differences between the static magnetic potentials and the internally produced magnetic potentials (the electro-dynamics corrections (including radiation) mentioned in reference [13]), and if, from the externally applied electromagnetic field, only the externally impinging or incident electromagnetic field is considered, (2-9) reduce to (16) or (33) of reference [14];

- Neglecting all the terms of the right-hand side of (2-10) that represent electro-dynamics corrections (including radiation) and the non-linear transversal conduction current I_{LS}^* ; and, assuming that the internally-produced line voltage is equal to the scattered line voltage, (2-10) reduce to (32) of reference [14].

Then, we also obtain the advantage for the numerical solution of the system of equations, claimed in reference [14], of having the source term appearing in only one kind of equations.

In fact, this is the formulation utilized for the numerical examples shown in reference [14], which refer to a two-conductor line over a perfectly conducting ground that is taken as a reference plane, not as a reference conductor.

Equations (34) and (35), which are a generalization of (21c) and (22c) of reference [1], as mentioned in reference [1], also represent a generalization of Rusck’s coupling theory [21], by including, besides the corrections due to the time delay in the production of the potentials, the effect of the externally produced magnetic potential along the direction of the line (a shortcoming of that theory pointed out a long time ago [39]) and also the effect of the conductive imperfections of the matter, both of the conductors and of the dielectrics between the conductors.

To include the effect of the transversal mode, we will define the so-called *conductor voltages*.

Usually, in standard transmission-line theory, the *conductor voltages* are defined relative to one of the conductors chosen to be the reference [12,30]. Then, to include the effect of the transversal mode, here, we will also define the *conductor “j” voltage* as the integral of the transversal electric field strength “E_t” between the conductor “j” and the reference conductor:

$$\begin{aligned} V^{j(int)} + V^{j(ex)} &= V^j(z,t) \equiv - \int_{(x_0,y_0)}^{(x_j,y_j)} \vec{E}_t \cdot d\vec{l}_t = \\ &= \Phi^j - \Phi^0 + \int_{(x_0,y_0)}^{(x_j,y_j)} \frac{\partial \vec{A}_{mt}^{(ex)}}{\partial t} \cdot d\vec{l}_t. \end{aligned} \quad (2-11)$$

Where, as usual, neglecting the transversal magnetic potential produced by the transversal currents, “A_{mt}” is the externally produced field.

Then, we can modify (34) and (35), by means of (2-11), and obtain:

$$\begin{aligned} \frac{\partial V^j}{\partial z} + \left([R_k^j] i^k + ([L_k^j] - [L_k^0]) \frac{\partial i^k}{\partial t} \right) &= \\ = E_z^{j(ex)} - E_z^{0(ex)} + \frac{\partial V^{j(ex)}}{\partial z} + \left(R_{(c0)} i^{0*} - \frac{\partial}{\partial t} (A_{mz}^{j(c)} - A_{mz}^{0(c)}) \right) &. \end{aligned} \quad (2-12)$$

and

$$\begin{aligned} \frac{\partial i^j}{\partial z} + [G_k^j] V^k + [C_k^j] \frac{\partial V^k}{\partial t} &= [G_k^j] V^{k(ex)} + [C_k^j] \frac{\partial V^{k(ex)}}{\partial t} + \\ + \left([G_k^j] (\Phi^{k(c)} - \Phi^{0(int)}) + [C_k^j] \frac{\partial}{\partial t} (\Phi^{k(c)} - \Phi^{0(int)}) - I_{LS}^* \right) &. \end{aligned} \quad (2-13)$$

In equations (2-12) and (2-13), which are a generalization of (28) and (29) of reference [1], we can also see, as mentioned in reference [1], that writing the equations in terms of the *conductor voltages* produce new source terms in (2-12) and (2-13), due to the existence of induced voltages produced by the externally applied electromagnetic field.

Apart from the assumptions of the thin-wire approximation and the validity of Ohm's law for all the conductors, and the assumption that the transversal mode is only externally produced, (2-12) and (2-13) are quite general. Neglecting the terms that appear within the parenthesis in the right-hand side of (2-12) and (2-13), they reduce to (3a) and (3b) of reference [12].

For the special case of time-harmonic electromagnetic fields, or in frequency domain, (2-12) and (2-13) can be written as:

$$\frac{dV_{\omega}^j(z)}{dz} + \left([R_k^j] + i\omega [L_k^j] \right) I_{\omega}^k(z) = S_{E\omega}^j(z); \quad (2-14)$$

and

$$\frac{dI_{\omega}^j(z)}{dz} + \left([G_k^j] + i\omega [C_k^j] \right) V_{\omega}^k(z) = S_{H\omega}^j(z). \quad (2-15)$$

Which, without the sources, are the standard transmission-line equations in frequency domain, such as shown in (2) of reference [30] or (6.7) and (6.8) of reference [31].

References

- [1] P. E. Munhoz-Rojas, "Conducted and Radiated Interference on the interconnections of Renewable Energy Farms" in Proceedings of the 2018 Joint IEEE EMC & APEMC, Singapore, 798-804, 2018.
- [2] CIGRE Technical Brochure 719, "Power Quality and EMC issues with future electricity networks", JWG C4-24/CIREC, www.e-cigre.org, 2018, p.56 .
- [3] C. L. Holloway, E.F. Kuester, A.E. Ruehli and G. Antonini, "Partial and internal inductance: Two of Clayton R. Paul's many passions", IEEE Trans. on EMC 55(4), 600-613, 2013.
- [4] J.D. Jackson, Classical Electrodynamics, 3rd ed., New York: John Wiley & Sons, Inc., 1999.
- [5] J.C. Maxwell, "A dynamical theory of the electromagnetic field", Roy. Soc. Trans. Vol.155, 526-597, 1864.
- [6] A. Sommerfeld, Electrodynamics, New York: Academic Press, 1952.
- [7] P. E. Munhoz-Rojas, "A Symmetrical Theory of Electromagnetism" in PIERS Proceedings - Moscow2012, 1199-1203, 2012.
- [8] R.P. Feynman, R. Leighton and M. Sands, The Feynman Lectures on Physics - The Electromagnetic Field, vol. 2, Reading, Massachusetts: Addison-Wesley, 1965.
- [9] Landau, L.D., and E.M. Lifshitz, Course of Theoretical Physics - Vol. 8: Electrodynamics of continuous media, (Spanish translation) Editorial Reverté, Barcelona, 1975, pp. 285-286.
- [10] Landau, L.D., and E.M. Lifshitz, Course of Theoretical Physics - Vol. 2: Field Theory, 2nd ed., (Portuguese translation) Editora Mir, Moscow, 1980, pp. 31, 69.
- [11] S.A. Schelkunoff, "Forty years ago: Maxwell's theory invades engineering - and grows with it", IEEE Trans. on Education 15(1), 2-14, 1972.
- [12] C. R. Paul, "A brief history of work in transmission lines for EMC applications", IEEE Trans. on EMC 49(2), 237-252, 2007.
- [13] F. Rachidi, "A review of field-to-transmission line coupling models with special emphasis to lightning-induced voltages on overhead lines", IEEE Trans. on EMC 54(4), 898-911, 2012.
- [14] A.K. Agrawal, H.J. Price and S.H. Gurbaxani, "Transient response of multiconductor transmission lines excited by a nonuniform electromagnetic field", IEEE Trans. on EMC 22(2), 119-129, 1980.
- [15] P. E. Munhoz-Rojas, "The effect of discontinuities in a multiconductor line on lightning-induced voltages", IEEE Trans. on EMC 51(1), 53-66, 2009.
- [16] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," J. Physiol., vol. 117, 500-544, 1952.
- [17] D. R. McNeal, "Analysis of a model for excitation of myelinated nerve," IEEE Trans. Biomed. Eng., vol. BME-23, 329-337, 1976.
- [18] J. P. Reilly, V. T. Freeman and W. D. Larkin, "Sensory effects of transient electrical stimulation - Evaluation with a neuroelectric model" IEEE Trans. Biomed. Eng., vol. BME-32, 1001-1011, 1985.
- [19] F. Rattay, "Analysis of models for external stimulation of axons" IEEE Trans. Biomed. Eng., vol. BME-33, 974-977, 1986.
- [20] B. J. Roth and P. J. Basser, "A model of the stimulation of a nerve fiber by electromagnetic induction" IEEE Trans. Biomed. Eng., vol. BME-37, 588-597, 1990.
- [21] S. Rusck, "Induced lightning overvoltages on power transmission lines with special reference to the overvoltage protection of low voltage networks", in *Trans. of the Royal Institute of Technology*, Stockholm, Sweden, vol.120, 1958.
- [22] T.W. Dawson, J. de Moerloose and M.A. Stuchly, "Hybrid finite-difference method for high-resolution modelling of low-frequency electric induction in humans", J. of Comp. Phys. 136, 640-653, 1997.
- [23] X.L. Chen, S. Benkler, C.H. Li, N. Chavannes and N. Kuster, "Low frequency electromagnetic field exposure study with possible human body model" in 2010 IEEE International Symposium on Electromagnetic Compatibility, Fort Lauderdale, FL, USA, 702-705, 2010.
- [24] P. E. Munhoz-Rojas, C.S. Segura-Salas, A.A. Costa, R. Martins and J. Hoffmann, "Fields and current densities induced in the human body by low-frequency electromagnetic fields" in Proceedings of the 2018 Joint IEEE EMC & APEMC, Singapore, 1267-1273, 2018.
- [25] J.C. Maxwell, A Treatise on Electricity and Magnetism, 2nd Edition, Vol. II, Oxford: Clarendon Press, 1881, pp.207-208.
- [26] J.C. Maxwell, A Treatise on Electricity and Magnetism, Vol. I, Oxford: Clarendon Press, 1873, pp.89-90.
- [27] L. F. Shampine, Numerical Solution of Ordinary Differential Equations. New York: Chapman & Hall, 1994.
- [28] C. R. Paul and J. W. McKnight, "Prediction of crosstalk involving twisted pairs of wires-Part I: A Transmission-line model for twisted-wire pairs", IEEE Trans. on EMC 21(2), 92-105, 1979.
- [29] G. Spadacini, F. Grassi, F. Marliani and S. A. Pignari, "Transmission-line model for field-to-wire coupling in bundles of twisted-wire pairs above ground", IEEE Trans. on EMC 56(6), 1682-1690, 2014.
- [30] S. Greedy, C. Smartt, M. Basford and D. Thomas, "Open source cable models for EMI simulations", IEEE EMC Magazine 7(3), 69-81, 2018.
- [31] CIGRE Technical Brochure 707, "EMC in wind energy systems", WG C4-30, www.e-cigre.org, 2017, pp.45-46 .
- [32] C. A. Nucci, F. Rachidi, M. V. Ianoz and C. Mazzetti, "Lightning-induced voltages on overhead lines", IEEE Trans. on EMC 35(1), 75-86, 1993.
- [33] A. Tatematsu, F. Rachidi and M. Rubinstein, "A technique for calculating voltages induced on twisted-wire pairs using the FDTD method", IEEE Trans. on EMC 59(1), 301-304, 2017.
- [34] P. Munhoz-Rojas and C. L. da S. Pinto, "Analysis of lightning-induced voltages on an overhead line" in Proc. of 8th Int. Symp. Lightning Protection, Sao Paulo, Brazil, 57-62, 2005.
- [35] S. Yokoyama, "Calculation of lightning-induced voltages on overhead multiconductor system", IEEE Trans. Power App. Syst., vol. PAS-103, no. 1, pp. 100-108, Jan. 1984.
- [36] M. Paolone, C. A. Nucci, E. Petrace, and F. Rachidi, "Mitigation of lightning-induced overvoltages in medium voltage distribution lines by means of periodical grounding of shielding wires and of surge arresters: Modeling and experimental validation", IEEE Trans. Power Del., 19(1), 423-431, 2004.
- [37] P. E. Munhoz-Rojas and C.L.da S. Pinto, "Calculations of lightning-induced voltages in distribution lines with LSA", Journal of Energy, 60, Special Issue, 95-100, 2011.
- [38] Hertz H., "The forces of electric oscillations, treated according to Maxwell's theory" (in German), *Wiedemann's Annalen*, 36, p.1, 1889. (English translation in Electric Waves, by Heinrich Hertz, translated by D.E. Jones, MacMillan and Co., London and New York, IX,137-159, 1893).
- [39] V. Cooray, "Calculating lightning-induced overvoltages in power lines: A comparison of two coupling models", IEEE Trans. on EMC 36(3), 179-182, 1994.

Can parallelization save the (computing) world?

János Végh^{*1,2} József Vásárhelyi¹, Dániel Drótos²

¹*Kalimános BT, Debrecen, Hungary*

²*Hungarian Academy of Sciences, Institute for Nuclear Research, H-4032 Debrecen, Hungary*

ARTICLE INFO

Article history:

Received: 29 November, 2018

Accepted: 13 January, 2019

Online: 05 February, 2019

Keywords:

Parallelization

Performance stalling

Supercomputer

Single-processor approach

ABSTRACT

As all other laws of the growth in computing, the growth of computing performance also shows a "logistic curve"-like behavior, rather than an unlimited exponential growth. The stalling of the single-processor performance experienced nearly two decades ago forced computer experts to look for alternative methods, mainly for some kind of parallelization. Solving the task needs different parallelization methods, and the wide range of those distributed systems limits the computing performance in very different ways. Some general limitations are shortly discussed, and a (by intention strongly simplified) general model of performance of parallelized systems is introduced. The model enables to highlight bottlenecks of parallelized systems of different kind and with the published performance data enables to predict performance limits of strongly parallelized systems like large scale supercomputers and neural networks. Some alternative solution possibilities of increasing computing performance are also discussed.

1 Introduction

Computing in general shows a growth described by a logistic curve [1] rather than an unlimited exponential curve. Since about 2000, a kind of stalling was experienced in practically all components contributing to the single-processor performance, signaling that drastic changes in the approach to computing is needed [2]. However, the preferred way chosen was to continue the traditions of the Single Processor Approach (SPA) [3]: to assemble systems comprising several segregated processing units, connected in various ways.

The Moore-observation has already been terminated in the sense that no more transistors can be added to a single processor, but persists in the sense that more transistors can be placed on a single chip in the form of several cores, complete systems or networks. In this way *the nominal computing performance* of processors keeps raising, but the task to produce some *payload computing performance* remains mostly for the software. However, *"parallel programs ... are notoriously difficult to write, test, analyze, debug, and verify, much more so than the sequential versions"* [4]. *"The general problem [with parallel programming] is that the basic sentiment seems to go against history. Much of the*

progress attributed to the Industrial Revolution is due to using more power for reducing human effort" [5].

The today's processors comprise a lot of different parallelization solutions [6]. As the result of the development, *"computers have thus far achieved ... tremendous hardware complexity, a complexity that has grown so large as to challenge the industry's ability to deliver ever-higher performance"* [7]. The huge variety of solutions make the efficiency of those systems hard to interpret and measure [8].

In section 2 some issues demonstrate why the present inflexible architectures represent serious issues, like bottleneck for parallelized architectures and tremendous non-payload activity during thread-level parallelization.

A (by intention strongly) simplified operating model of parallel operation is presented in section 3. The basis of the section is Amdahl's law, which is reinterpreted for the modern computing systems for the purposes of the present study. The technical implementation of parallel systems from sequentially working single-threads using clock signals have their performance limitations as derived in section 4. Both the inherent performance bound (stemming out from the paradigm and the clock-driven electronic technology)

* János Végh & Vegh.Janos@gmail.com

and the specific one for supercomputers are discussed.

Some aspects of one of the obvious fields of parallel processing: the limitations of supercomputing is discussed in section 5, where short-term prediction for supercomputer performance is also given and the role of different contributions is demonstrated through some issues of brain simulation. Section 6 discusses some possible ways out of the experienced stalling of performance through parallelization. In section 7 it is concluded that mainly the 70-years old computing paradigm itself (and its consequences: the component and knowledge base) limits the development of computing performance through utilizing parallelized sequential systems.

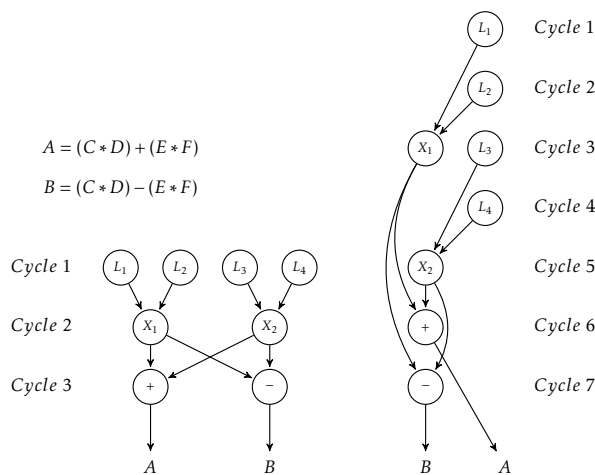


Figure 1: Executing the example calculation on a dual issue processor (borrowed from [6])

2 Some architectural issues

The intention with inventing "computer" was to automate mechanical computing on a single computing resource. Today masses of processors are available for solving a task, and quite typical that the fraction of real computing is much less than the part which imitates other functionality through computations (*if all you have is a hammer, everything looks like a nail*).

2.1 The theoretical parallelism and limits of its technical implementation

When speaking about parallelization, one has to distinguish the theoretically possible and the technically implementable parallelization. Modern processors comprise several (hidden) processing units, but the efficacy of their utilization is rather poor: the theoretically achievable parallelism in most cases cannot even be approached. To find out the achievable maximum parallelisation, the different dependencies between the data (as arguments of some operations) shall be considered. Let us suppose we want to calculate expressions (the example is borrowed from [6])

$$A = (C * D) + (E * F)$$

$$B = (C * D) - (E * F)$$

where we have altogether 4 load operations, 2 multiplications, and 2 additions. To achieve the maximum parallelism enabled by data dependencies the theoretical parallelism assumes the presence of several Processing Unit (PU): (at least) 4 memory access units, 2 multipliers and 2 adders (or equally: an at least 4-issue universal processor). With a such theoretical a processor (see Fig. 1, left side) we could load all the four operands in the first machine cycle, to do the multiplications in the second cycle, and to make the addition/subtraction in the last cycle.

In the practical processors, however, the resources are limited and inflexible. In a real (single issue) processor this task can be solved in 8 cycles, every operation needs a clock cycle.

In a dual-issue processor (see Fig. 1, right side) one issue is able to make calculations, the another one load/store operations. In the real configurations, however, only one memory load operation can be made at a time (mainly due to components manufactured for building SPA computers), so the PUs must wait until all their operands become available. In this case *there is only one single clock tick when the two issues work in parallel*, and the 8 operations are executed in 7 cycles, in contrast with the theoretical 3 cycles. All this at the expense of (nearly) double hardware (HW) cost.

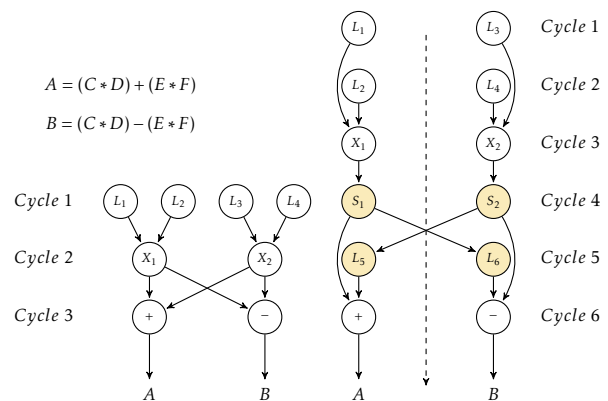


Figure 2: Executing the example calculation on a dual core processor (borrowed from [6])

A different idea for distributing the task could be to utilize a dual-core processor, see Fig. 2, right side. Here the multiplications are performed by independent processors, and also have independent capacity for making summation and subtraction separately. This (provided that the processors can load their arguments independently) requires only 4 cycles. However, the "other half" of the result of multiplications belongs to the other processor. The segregated processors can exchange data only through a memory location, so two extra states (in both processors) are inserted for saving and loading the needed data): i.e. the processors execute 12 operations instead of 8 (including 4 slow memory operations), altogether they solve the task in 2*6 clock cycles, instead of 8 (and assume, that the memory will make synchronization, i.e. can assure that reading the intermediate memory locations takes

place only when the other party surely wrote that in). That is, (at least) double HW cost, double memory bandwidth, 50% more instructions and only 25% performance increase. Here the additional reason (to the inflexible architecture) is the computing paradigm: the single processor-single process principle.

In both cases the performance increase can be implemented at a high price: considerably increased HW cost (and mainly in the latter case: longer, more complex software (SW) development). This latter increase gets exponential when solving a real-life program and using more segregated processors. This somewhat theoretical summary is sufficiently underpinned by experiences of practitioners [9].

2.2 Multitasking

With the growing gap between the speed of the processing and the growing demand for executing complex tasks, the need for sharing the single computing resource between processes appeared. The missing features were noticed early [10].

One of the major improvements for enhancing single-processor performance was introducing the register file. It is practically a different kind of memory, with separate bus, quick access and short address. Although the processors could work with essentially less registers [11], today the processors commonly include dozens of registers. However, the processors have been optimized for single-task regime, so utilizing them for task-level parallelization required serious modifications, both on HW and SW side. What is very advantageous for single-thread computing: to have register access time for computations, is very disadvantageous for multitasking: the register contents (or at least part of them) must be saved and restored when switching to another task. The context switching, however, is very expensive in terms of execution time [12]. Also, the too complex and over-optimized architecture considerably increases the internal latency time of the processor when used in multi-processor environment: the same architecture cannot be optimized for both work regimes,

Especially in the real-time applications the response time is very critical. To respond to some external signal, the running task must be interrupted, task and processor context saved, task and processor context set for the interrupt, and only after that responding to the external signal can begin. This feature (a consequence of SPA) is especially disadvantageous for cyber-physical systems, networking devices, etc. Synchronizing tasks for those purposes by the operating system (OS) requires tremendous offset work.

The major reason of the problem (among others) is the loose connection between the HW and the SW: in the paradigm only one processor exists, and for a segregated processor there is no sense "to return control"; for the only processor a "next instruction" must always be available (leading among others to the idea of the "idle" task is OS, and in a broader sense, to the non energy-proportional operating regime [13]). The HW

interrupts have no information what SW process is actually executed. To mitigate the unwanted effects like priority inversion [14] this instrumental artefact was formalized [15], rather than eliminated by a different HW approach.

3 Amdahl's Law and model for parallelized systems

The huge variety of the available parallelization methods, the extremely different technological solutions do not enable to set up an accurate technological model. With the proper interpretation of the Amdahl's Law, the careful analysis enables to spot essential contributions degrading the efficacy of parallelized systems. Despite its simplicity, the model enables to provide a qualitative (or maybe: semi-quantitative) description of the operational characteristics of the parallelized systems, in quite different fields.

3.1 Amdahl's Law

A half century ago Amdahl [3] called first the attention to that *the parallel systems built from segregated single processors have serious performance limitations*. The "law" he formulated was used successfully on quite different fields [16], sometimes misunderstood, abused and even refuted (for a review see [17]). Despite of that "*Amdahl's Law is one of the few, fundamental laws of computing*" [18], for today it is quite forgotten when analyzing complex computing systems. Although Amdahl was speaking about complex computing systems, due to its initial analysis about which contributions degrading parallelism can be omitted, *today Amdahl's Law is commonly assumed to be valid for SW only*. For today, thanks to the development and huge variety of the technological implementations of parallelized distributed systems, it became obvious that the role of the contributing factors shall be revised.

3.1.1 Amdahl's idea

The successors of Amdahl introduced the misconception that Amdahl's law is valid for software only and that the non-parallelizable fraction contains something like *the ratio of the numbers of the corresponding instructions or maybe the execution time of the instructions*. However, *Amdahl's law is valid for any partly parallelizable activity (including computer unrelated ones) and the non-parallelizable fragment shall be given as the ratio of the time spent with non-parallelizable activity to the total time*. In his famous paper Amdahl [3] speaks about "*the fraction of the computational load*" and explicitly mentions, in the same sentence and same rank, algorithmic contributions like "*computations required may be dependent on the states of the variables at each point*"; architectural aspects like "*may be strongly dependent on sweeping through the array along different axes on succeeding passes*" as well as "*physical problems*" like

"propagation rates of different physical effects may be quite different".

In the present complex parallelized systems his reasoning is still valid: one has to consider *the workload of the complex HW/SW system, rather than some segregated component*, and Amdahl's idea describes parallelization imperfectness of any kind. *Notice that the eligibility of neglecting some component changes with time, technology and conditions. When applied to a particular case, it shall be individually scrutinized which contributions can be neglected.*

3.1.2 Deriving the effective parallelization

Successors of Amdahl expressed Amdahl's law with the formula

$$S^{-1} = (1 - \alpha) + \alpha/k \quad (1)$$

where k is the number of parallelized code fragments, α is the ratio of the parallelizable part within the total code, S is the measurable speedup. The assumption can be visualized that in α fraction of the running time the processors are executing parallelized code, in $(1 - \alpha)$ fraction they are waiting (all but one), or making non-payload activity. That is α describes how much, in average, processors are utilized, or how effective (at the level of the computing system) the parallelization is.

For a system under test, where α is not *a priori* known, one can derive from the measurable speedup S an *effective parallelization factor* [19] as

$$\alpha_{eff} = \frac{k}{k-1} \frac{S-1}{S} \quad (2)$$

Obviously, this is not more than α expressed in terms of S and k from (1). For the classical case, $\alpha = \alpha_{eff}$; which simply means that in the *ideal* case the actually measurable effective parallelization achieves the theoretically possible one. In other words, α describes a system the *architecture* of which is completely known, while α_{eff} characterizes the *performance*, which describes both the complex architecture and the actual conditions. It was also demonstrated [19] that α_{eff} can be successfully utilized to describe parallelized behavior from SW load balancing through measuring efficiency of the on-chip HW communication to characterize performance of clouds.

The value α_{eff} can also be used to refer back to Amdahl's classical assumption even in the realistic case when the parallelized chunks have different lengths and the overhead to organize parallelization is not negligible. The speedup S can be measured and α_{eff} can be utilized to characterize the measurement setup and conditions, how much from the theoretically possible maximum parallelization is realized. Numerically $(1 - \alpha_{eff})$ equals with the f value, established theoretically [20].

The distinguished constituent in Amdahl's classic analysis is the parallelizable fraction α , all the rest (including wait time, non-payload activity, etc.) goes into

the "sequential-only" fraction. When using several processors, one of them makes the sequential calculation, the others are waiting (use the same amount of time). So, when calculating the speedup, one calculates

$$S = \frac{(1 - \alpha) + \alpha}{(1 - \alpha) + \alpha/k} = \frac{k}{k(1 - \alpha) + \alpha} \quad (3)$$

hence the efficiency is

$$E = \frac{S}{k} = \frac{1}{k(1 - \alpha) + \alpha} \quad (4)$$

This explains the behavior of diagram $\frac{S}{k}$ in function of k experienced in practice: the more processors, the lower efficiency. It is not some kind of engineering imperfectness, it is just the consequence of Amdahl's law. At this point one can notice that $\frac{1}{E}$ is a linear function of number of processors, and its slope equals to $(1 - \alpha)$. Equation (4) also underlines the importance of the single-processor performance: the lower is the number of the processors used in the parallel system having the expected performance, the higher can be the efficacy of the system.

Notice also, that through using (4), the efficiency $\frac{S}{k}$ can be equally good for describing the efficiency of parallelization of a setup, provided that the number of processors is also provided. From (4)

$$\alpha_{E,k} = \frac{Ek - 1}{E(k - 1)} \quad (5)$$

From the same relationship α can also be expressed in terms of S and k

$$\alpha_{eff} = \frac{k}{k-1} \frac{S-1}{S} \quad (6)$$

If the parallelization is well-organized (load balanced, small overhead, right number of processors), α_{eff} is very close to unity, so tendencies can be better displayed through using $(1 - \alpha_{eff})$ (i.e. the non-parallelizable fraction) in the diagrams below.

The importance of this practical term α_{eff} is underlined by that the achievable speedup (performance gain) can easily be derived from (1) as

$$G = \frac{1}{(1 - \alpha_{eff})} \quad (7)$$

Correspondingly, the resulting maximum performance is

$$P_{resulting} = G * P_{single} \quad (8)$$

3.1.3 The original assumptions

The classic interpretation implies three¹ essential restrictions, but those restrictions are rarely mentioned in the textbooks on parallelization:

- the parallelized parts are of equal length in terms of execution time

¹An additional essential point which was missed by both [20] and [1], that *the same computing model was used in all computers considered.*

- the housekeeping (controlling parallelization, passing parameters, waiting for termination, exchanging messages, etc.) has no cost in terms of execution time
- the number of parallelizable chunks coincides with the number of available computing resources

Essentially, this is why Amdahl's law represents a theoretical upper limit for parallelization gain. It is important to notice, however, that a 'universal' speedup exists only if the parallelization efficiency α is independent from the number of the processors. As will be discussed below, this assumption is only valid if the number of processors is low, so the usual linear extrapolation of the actual performance on the nominal performance will not be valid any more in the case of high number of processors.

3.1.4 The additional factors considered here

In the spirit of the Single Processor Approach (SPA) the programmer (the person or the compiler) has to organize the job: at some point the initiating processor splits the execution, transmits the necessary parameters to some other processing units, starts their processing, then waits for the termination of started processings; see section 4.1. Real-life programs show sequential-parallel behavior [21], with variable degree of parallelization [22] and even apparently massively parallel algorithms change their behavior during processing [23]. All these make Amdahl's original model non-applicable, and call for extension.

As discussed in [24]

- many parallel computations today are limited by several forms of communication and synchronization
- the parallel and sequential runtime components are only slightly affected by cache operations
- wires get increasingly slower relative to gates

In the followings

- the main focus will be on synchronization and communication; they are kept at their strict absolute minimum; and their effect is scrutinized
- the effect of cache will be neglected, and runtime components not discussed separately
- the role of the wires is considered in an extended sense: both the importance of physical distance and using special connection methods will be discussed

3.2 A simplified model for parallel operation

3.2.1 The performance losses

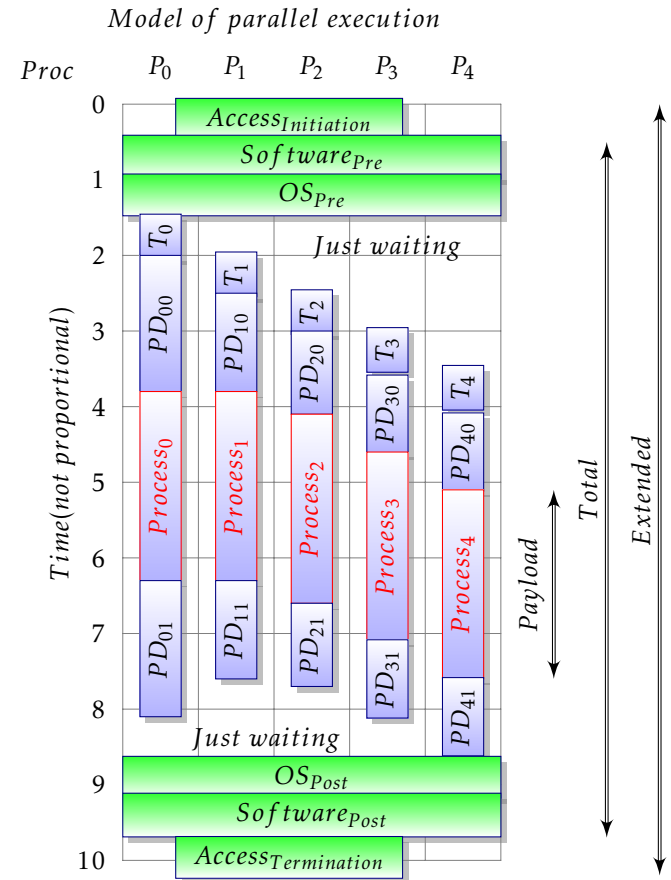


Figure 3: The extended Amdahl's model (strongly simplified)

When speaking about computer performance, a modern computer system is assumed, which comprises many sophisticated components (in most cases embedding complete computers), and their complex interplay results in the final performance of the system. In the course of efforts to enhance processor performance through using some computing resources in parallel, many ideas have been suggested and implemented, both in SW and HW [6]. All these approaches have different usage scenarios, performance and limitations. Because of the complexity of the task and the limited access to the components, empirical methods and strictly controlled special measurement conditions are used to quantize performance [8]. Whether a metric is appropriate for describing parallelism, depends on many factors [20, 25, 27].

As mentioned in section 3, Amdahl listed different reasons why losses in the "computational load" can occur. To understand operation of computing systems working in parallel, one needs to extend Amdahl's original (rather than that of the successors') model in

²This separation cannot be strict. Some features can be implemented in either SW or HW, or shared among them, and also some apparently sequential activities may happen partly parallel with each other.

such a way, that non-parallelizable (i.e. apparently sequential) part comprises contributions from HW, OS, SW and Propagation Delay (PD)², and also some access time is needed for reaching the parallelized system. The technical implementations of different parallelization methods show up infinite variety, so here a (by intention) strongly simplified model is presented. Amdahl's idea enables to put everything that cannot be parallelized into the sequential-only fraction. The model is general enough to discuss qualitatively some examples of parallelly working systems, neglecting different contributions as possible in the different cases. The model can easily be converted to a technical (quantitative) one and the effect of inter-core communication can also easily be considered.

3.2.2 The principle of the measurements

When measuring performance, one faces serious difficulties, see for example [26], chapter 1, both with making measurements and interpreting them. When making a measurement (i.e. running a benchmark) either on a single processor or a system of parallelized processors, an instruction mix is executed many times. The large number of executions averages the rather different execution times [28], with an acceptable standard deviation. In the case when the executed instruction mix is the same, the conditions (like cache and/or memory size, the network bandwidth, Input/Output (I/O) operations, etc) are different and they form the subject of the comparison. In the case when comparing different algorithms (like results of different benchmarks), the instruction mix itself is also different.

Notice that the so called "algorithmic effects" – like dealing with sparse data structures (which affect cache behavior) or communication between the parallelly running threads, like returning results repeatedly to the main thread in an iteration (which greatly increases the non-parallelizable fraction in the main thread) – manifest through the HW/SW architecture, and they can hardly be separated. Also notice that there are fixed-size contributions, like utilizing time measurement facilities or calling system services. Since α_{eff} is a *relative* merit, the *absolute* measurement time shall be long. When utilizing efficiency data from measurements which were dedicated to some other goal, a proper caution must be exercised with the interpretation and accuracy of the data.

3.2.3 The formal introduction of the model

The extended Amdahl's model is shown in Fig. 3. The contributions of the model component XXX to α_{eff} will be denoted by α_{eff}^{XXX} in the followings. Notice the different nature of those contributions. They have only one common feature: *they all consume time*. The vertical scale displays the actual activity for processing units shown on the horizontal scale.

³Although some OS activity was surely included, Amdahl assumed some 20 % SW fraction, so the other contributions could be neglected compared to SW contribution.

Notice that our model assumes no interaction between the processes running on the parallelized systems in addition to the absolutely necessary minimum: starting and terminating the otherwise independent processes, which take parameters at the beginning and return results at the end. It can, however, be trivially extended to the more general case when processes must share some resource (like a database, which shall provide different records for the different processes), either implicitly or explicitly. Concurrent objects have inherent sequentiality [21], and synchronization and communication among those objects considerably increase [22] the non-parallelizable fraction (i.e. contribution $(1 - \alpha_{eff}^{SW})$), so *in the case of extremely large number of processors special attention must be devoted to their role on the efficiency of the application on the parallelized system*.

Let us notice that all contributions have a role during measurement: contributions due to SW, HW, OS and PD cannot be separated, though dedicated measurements can reveal their role, at least approximately. The relative weights of the different contributions are very different for the different parallelized systems, and even within those cases depend on many specific factors, so in every single parallelization case a careful analysis is required.

3.2.4 Access time

Initiating and terminating the parallel processing is usually made from within the same computer, except when one can only access the parallelized computer system from another computer (like in the case of clouds). This latter access time is independent from the parallelized system, and *one must properly correct for the access time when derives timing data for the parallelized system*. Failing to do so leads to experimental artefacts like the one shown in Fig. 7. Amdahl's law is valid only for properly selected computing system. This is a one-time, and usually fixed size time contribution.

3.2.5 Execution time

The execution time *Total* covers all processings on the parallelized system. All applications, running on a parallelized system, must make some non-parallelizable activity at least before beginning and after terminating parallelizable activity. This SW activity represents what was assumed by Amdahl as the total sequential fraction³. As shown in Fig. 3, the *apparent* execution time includes the real payload activity, as well as waiting and OS and SW activity. Recall that the execution times may be different [28], [26], [29] in the individual cases, even if the same processor executes the same instruction, but executing an instruction mix many times results in practically identical execution times, at least at model level. Note that *the standard deviation of the execution times appears as a contribution to the*

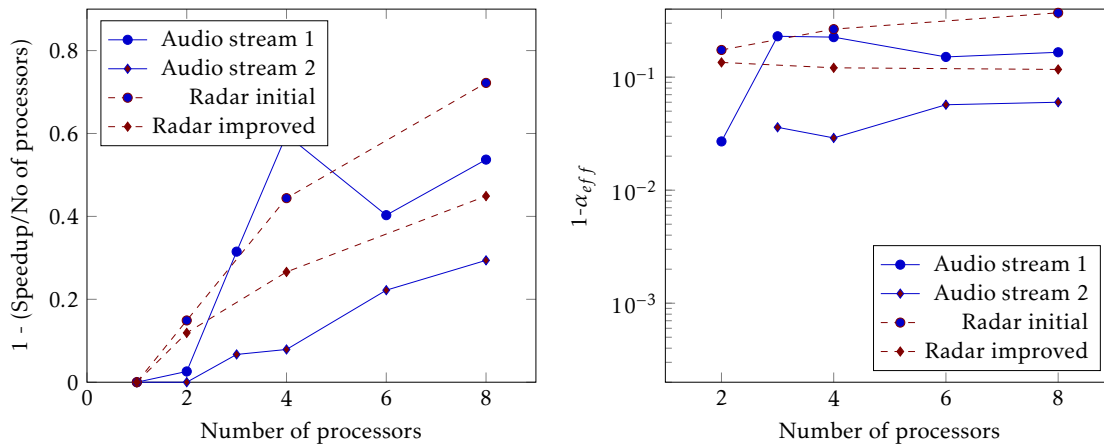


Figure 4: Relative speedup (left side) and $(1 - \alpha_{eff})$ (right side) values, measured running the audio and radar processing on different number of cores. [31]

non-parallelizable fraction, and in this way increases the "imperfectness" of the architecture. This feature of processors deserves serious consideration when utilizing a large number of processors. Over-optimizing a processor for single-thread regime hits back when using it in a parallelized many-processor environment, see also the statistical underpinning in [30].

3.3 Fields of application

The developed formalism can be effectively utilized on quite different fields, for more examples see [19].

3.3.1 Load balancing

The classic field of application is to qualify how effectively a parallelized task uses its resources, as shown in Fig. 4. A compiler making load balancing of an originally sequential code for different number of cores is described and validated in paper [31], by running the executable code on platforms having different number of cores. The authors' first example shows results of implementing parallelized processing of an audio stream manually, with an initial, and a later, more careful implementation. For two different processings of audio streams, using efficiency E as merit enables only to claim a qualitative statement about load balancing, that "The higher number of parallel processes in Audio-2 gives better results", because the Audio-2 diagram decreases less steeply, than Audio-1. In the first implementation, where the programmer had no previous experience with parallelization, the efficiency quickly drops with increasing the number of cores. In the second round, with experiences from the first implementation, the loss is much less, so $1 - E$ rises less speedily.

Their second example is processing radar signals. Without switching the load balancing optimization on, the slope of the curve $1 - E$ is much bigger. It seems to be unavoidable, that as the number of cores increases, the efficiency (according to Eq. (4)) decreases, even at such low number of cores. Both examples leave

the question open whether further improvements are possible or whether the parallelization is uniform in function of the number of cores.

In the right column of the figure (Fig. 10 in [31]) the diagrams show the $(1 - \alpha_{eff})$ values, derived from the same data. In contrast with the left side, these values are nearly constant (at least within the measurement data readback error) which means that the derived parameter is really characteristic to the system. By recalling (6) one can identify this parameter as the resulting non-parallelizable part of the activity, which – even with careful balancing – cannot be distributed among the cores, and cannot be reduced.

In the light of this, one can conclude that both the programmer in the case of audio stream and the compiler in the case of radar signals correctly identified and reduced the amount of non-parallelizable activity: α_{eff} is practically constant in function of number of cores, nearly all optimization possibilities found and they hit the wall due to the unavoidable contribution of non-parallelizable software contributions. Better parallelization leads to lower $(1 - \alpha_{eff})$ values, and less scatter in function of number of cores. The uniformity of the values make also highly probable, that in the case of audio streams further optimization can be done, while processing of radar signals reached its bounds.

3.3.2 SoC communication method

Another excellent example is shown in Fig. 5. In their work [32] the authors have measured the execution time of some tasks, distributed over a few cores in a System on Chip (SoC). It is worth to notice that the method of calculating $(1 - \alpha_{eff})$ is self-consistent and reliable. In the case of utilizing 2 cores only, the same single communication pair is utilized, independently from the chosen communication method. Correspondingly, the diagram lines start at the same $(1 - \alpha_{eff})$ value for all communication methods. It is also demonstrated that "ring" and "nearest neighbor" methods result in the same communication overhead. Notice also that the $(1 - \alpha_{eff})$ values are different for the two

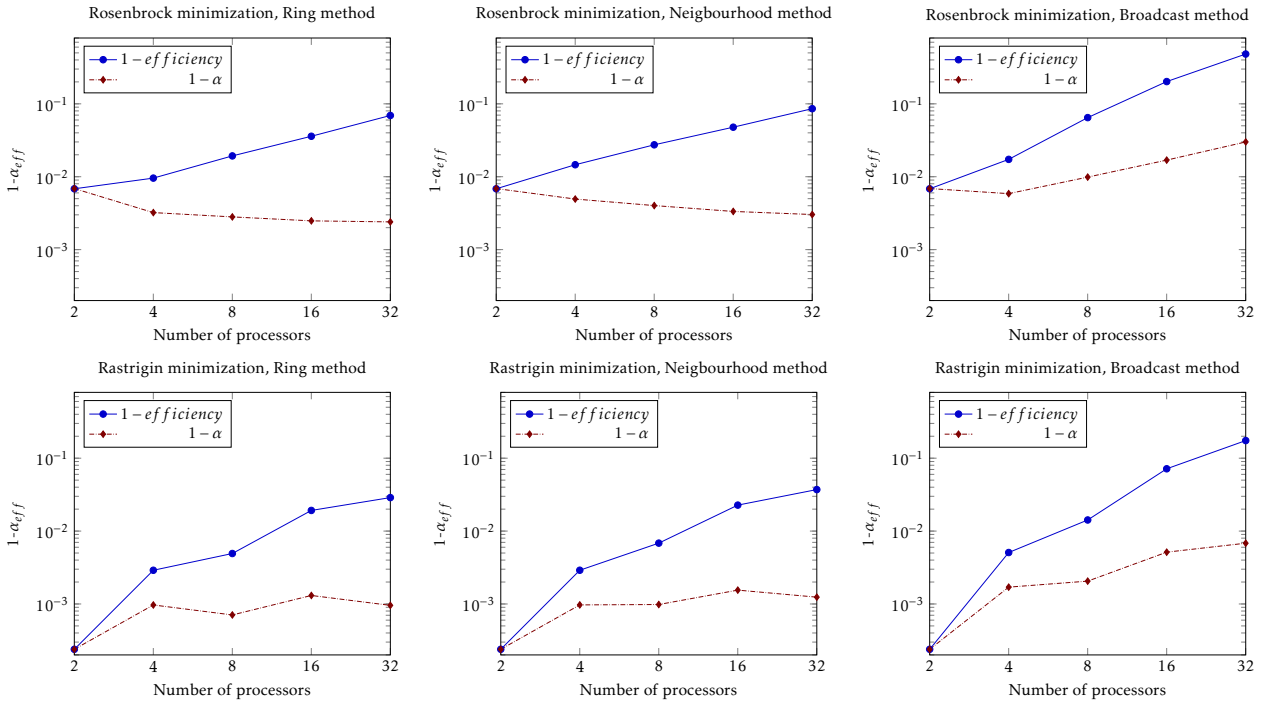


Figure 5: Comparing efficiency and α_{eff} for different communication strategies when running two minimization task on SoC by [32]

mathematical methods, which highlights that *the same architecture behaves differently when executing a different task*.

The commonly used metric ($1-efficiency$) only shows that the efficiency decreases as the number of cores increases. The proposed metric ($1-\alpha_{eff}$) also reveals that while the "ring" and "nearest neighbor" methods scale well with the number of cores, in the case of the "broadcast" method the effective parallelization gets worse as the number of communicating cores increases. The scatter on the diagrams originate from the short measurement time: the authors focused on a different goal, but their otherwise precise measurement can be used for the present goal only with the shown scatter.

3.3.3 Supercomputer performance design

Utilizing the effective parallelization is important in designing supercomputer architectures, too. Since the resulting performance depends on both the number of processors and the effective parallelization, both quantities are correlated with ranking of the supercomputer in Fig. 6. As expected, in TOP50 the higher the ranking position is, the higher is the required number of processors in the configuration, and as outlined above, the more processors, the lower ($1-\alpha_{eff}$) is required (provided that the same efficiency is targeted).

In TOP10, the slope of the regression line sharply changes in the left figure, showing the strong competition for the better ranking position. Maybe this marks the cut line between the "racing supercomput-

ers" and "commodity supercomputers". On the right figure, TOP10 data points provide the same slope as TOP50 data points, demonstrating that to produce a reasonable efficiency, the increasing number of cores must be accompanied with a proper decrease in value of ($1-\alpha_{eff}$), as expected from (4), furthermore, that to achieve a good ranking a good value of ($1-\alpha_{eff}$) must be provided.

3.3.4 Cloud services

In the case of utilizing cloud services (or formerly grids) the parallelized system and the one which interfaces user to its application are physically different. These systems differ from the ones discussed above in two essential points: the access and the inter-node connections are provided through using Internet, and the architecture is not necessarily optimized to offer the best possible parallelization. Since the operation of the Internet is stochastic, the measurements cannot be as accurate as in the cases discussed above. The developed formalism, however, can also be used for this case.

The authors of [33] benchmarked some commercially available cloud services, fortunately using High Performance Linpack (HPL) benchmark. Their results are shown in Fig. 7. On the left side efficiency (i.e. $\frac{R_{Max}}{R_{Peak}}$), on the right side ($1-\alpha_{eff}$) is displayed in function of the processors. They found (as expected) that the efficiency decreases as the number of the processors (nodes) increases. Their one-time measurement

⁴ A long term systematic study [34] derived the results that measured data show dozens of percents of variation in long term run, and also unexpected variation in short term run.

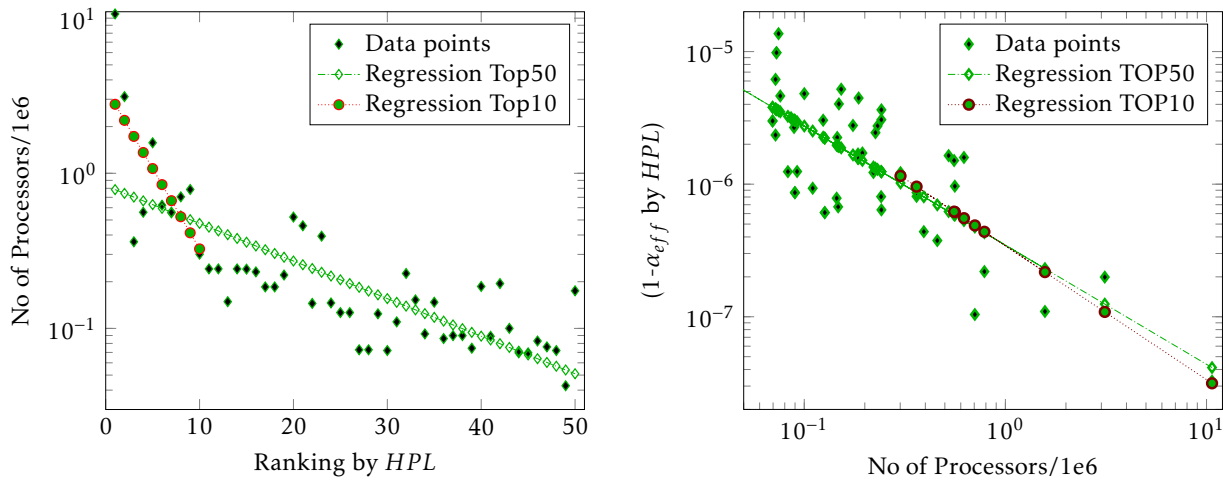


Figure 6: Correlation of number of processors with ranking and effective parallelism with number of processors.

show considerable scatter, as expected⁴.

When compared to Fig. 4, one can immediately notice two essential differences. First, that $\frac{R_{Max}}{R_{Peak}}$ is considerably lower than unity, even for very low number of cores. Second, that $(1 - \alpha_{eff})$ values steeply decrease as number of cores increases, although the model contains only contributions which may only increase as number of cores increases. Both of those deviations are caused by an instrumental artefact made during benchmarking: when acquiring measurement data also the access time must be compensated for, see Fig. 3.

As discussed above, HPL characterizes the setup comprising processors working in parallel, so the benchmark is chosen correctly. If the time is measured on client's computer (and this is what is possible using those services), the time *Extended* is utilized in the calculation in place of *Total*. This artefact is responsible for both mentioned differences. As can be seen from extrapolating the data, the efficiency measured in this way would not achieve 100 % even on a system comprising only one single processor. Since α_{eff} measures the average utilization of processors, this foreign contribution is divided by the number of processors, so with increasing number of processors this foreign contribution decreases, causing to decrease the calculated value of $(1 - \alpha_{eff})$.

At such low number of processors neither of the contributions depending the processor number is considerable, so one can expect that in the case of correct measurement $(1 - \alpha_{eff})$ would be constant. So, extrapolating graphs $(1 - \alpha_{eff})$ to the value corresponding to a one-processor system, one can see that both for Edison supercomputer and Azure A series grid (and maybe also Rackspace) the expected value is approaching unity (but obviously below it). From the slope of the curve (increasing the denominator 1000 times, $(1 - \alpha_{eff})$ reduces to 10^{-3}) one can even find out that $(1 - \alpha_{eff})$ should be around 10^{-3} . Based on these data, one can agree with the conclusion that on a good cloud the benchmark High Performance Conjugate Gradients (HPCG) can run as effectively as on the supercomputer

used in the work. However, $(1 - \alpha_{eff})$ is about 3 orders of magnitude better for TOP500 class supercomputers, but this makes a difference only for HPL class benchmarks and only at large number of processors.

Note that in the case of AWS clouds and Azure F series the α_{eff}^{OS+SW} can be extrapolated to about 10^{-1} , and this is reflected by the fact that their efficiency drops quickly as the number of the cores increases. Interesting to note that ranking based on α_{eff} is just the opposite of ranking based on efficiency (due to the measurement artefact).

Note also that switching hyperthreading on does not change our limited validity conclusions: both efficiency and $(1 - \alpha_{eff})$ remains the same, i.e. the hyper-threaded core seems to be a full value core. The placement group (PG) option did not affect the measured data considerably.

4 Limitations of parallelization

It is known from the ancient times that there is a measure in things; there are certain boundaries (*sunt certi denique fines quos ultra citraque nequit consistere rectum* Horace). The different computing-related technical implementations also have their limitations [24], consequently computing performance itself has its own limitations. In parallelized sequential systems an inherent (stemming out from the paradigm and the clock-driven electronic technology) performance limit is derived. The complex parallelized systems are usually qualified (and ranked) through measuring the execution time of some special programs (benchmarks). Since utilizing different benchmarks results in different quantitative features (and ranking [35]), the role of benchmarks is also discussed in terms of the model. It is explained why certain benchmarks result in apparently controversial results.

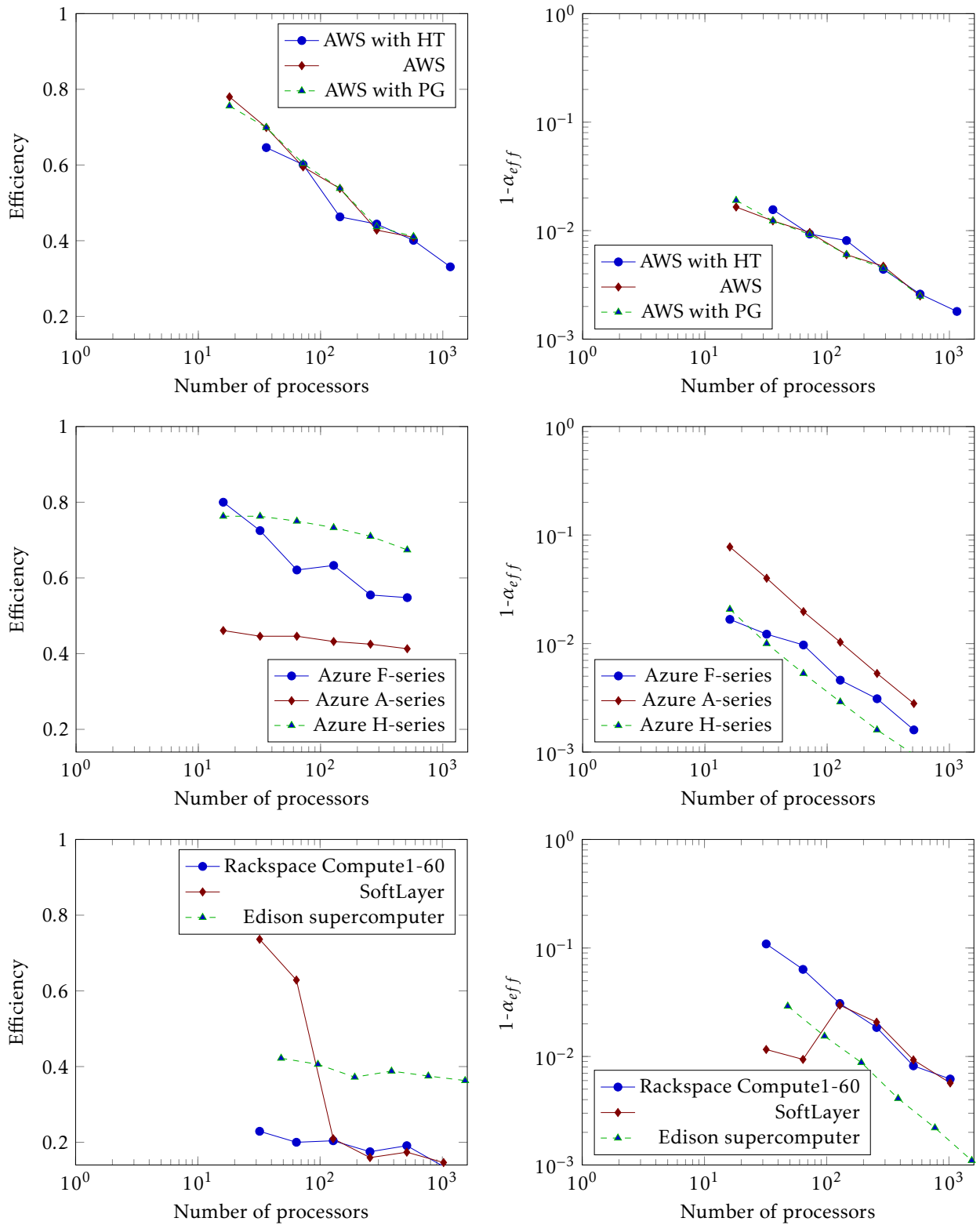


Figure 7: Efficiency (left side) and $(1 - \alpha_{eff})$ (right side) values for some commercial cloud services. Data from [33] are used

4.1 The inherent limit of parallelization

In the SPA initially and finally only one thread exists, i.e. the minimal absolutely necessary non-parallelizable activity is to fork the other threads and join them again. With the present technology, no such actions can be shorter than one clock period. That is, in this simple case the non-parallelizable fraction will be given as the ratio of the time of the two clock periods to the total execution time. The latter time is a free parameter in describing the efficiency, i.e. the value of the effective parallelization α_{eff} also depends on the total benchmarking time (and so does the achievable parallelization gain, too).

This dependence is of course well known for supercomputer scientists: for measuring efficiency with better accuracy (and also for producing better α_{eff} values) hours of execution times are used in practice. For example in the case of benchmarking the supercomputer *Taihulight* [36] 13,298 seconds benchmark runtime was used; on the 1.45 GHz processors it means $2 \cdot 10^{13}$ clock periods. This means that (at such benchmarking time) the inherent limit of $(1 - \alpha_{eff})$ is 10^{-13} (or equivalently the achievable performance gain is 10^{13}). In the followings for simplicity 1.00 GHz processors (i.e. 1 ns clock cycle time) will be assumed.

The supercomputers, however, are distributed systems. In a stadium-sized supercomputer the distance between processors (cable length) about 100 m can be assumed. The net signal round trip time is cca. 10^{-6} seconds, or 10^3 clock periods. The presently available network interfaces have 100...200 ns latency times, and sending a message between processors takes time in the same order of magnitude. This also means that *making better interconnection is not a bottleneck in enhancing performance of large-scale supercomputers*. This statement is underpinned also by statistical considerations [30].

Taking the (maybe optimistic) value $2 \cdot 10^3$ clock periods for the signal propagation time, the value of the effective parallelization $(1 - \alpha_{eff})$ will be at least in the range of 10^{-10} , only because of the physical size of the supercomputer. This also means that the expectations against the absolute performance of supercomputers are excessive: assuming a 10 Gflop/s processor, the achievable absolute *nominal* performance is $10^{10} \cdot 10^{10}$, i.e. 100 EFlops. In the feasibility studies an analysis for whether this inherent performance bound exists is done neither in USA [38, 37] nor in EU[39] nor in Japan [40] nor in China [41].

Another major issue arises from the computing paradigm SPA: only one computer at a time can be addressed by the first one. As a consequence, minimum as many clock cycles are to be used for organizing the parallel work as many addressing steps required. Basically, this number equals to the number of cores in the supercomputer, i.e. the addressing in the TOP10 positions typically needs clock cycles in the order of $5 \cdot 10^5 \dots 10^7$; degrading the value of $(1 - \alpha_{eff})$ into the range $10^{-6} \dots 2 \cdot 10^{-5}$. Two tricks may be used to reduce the number of the addressing steps: either the

cores are organized into *clusters* as many supercomputer builders do, or the processor itself can take over the responsibility of addressing its cores [42]. Depending on the actual construction, the reducing factor can be in the range $10^2 \dots 5 \cdot 10^4$, i.e. the resulting value of $(1 - \alpha_{eff})$ is expected to be in the range of $10^{-8} \dots 2 \cdot 10^{-6}$. Notice that utilizing "cooperative computing" [42] enhances further the value of $(1 - \alpha_{eff})$ and considerably enhances the performance of real-life programs [43], but it means already utilizing a (slightly) different computing paradigm.

An operating system must also be used, for protection and convenience. If one considers the context change with its consumed $2 \cdot 10^4$ cycles [12], the absolute limit is cca. $5 \cdot 10^{-8}$, on a zero-sized supercomputer. This is why *Taihulight* runs the actual computations in kernel mode [42].

It is crucial to understand that the decreasing efficiency (see (4)) is coming from the computing paradigm itself rather than from some kind of engineering imperfections. This inherent limitation cannot be mitigated without changing the computing paradigm.

Although not explicitly dealt with here, notice that the data exchange between the first thread and the other ones also contribute to the non-parallelizable fraction and typically uses system calls, for details see [22, 44] and section 4.2.

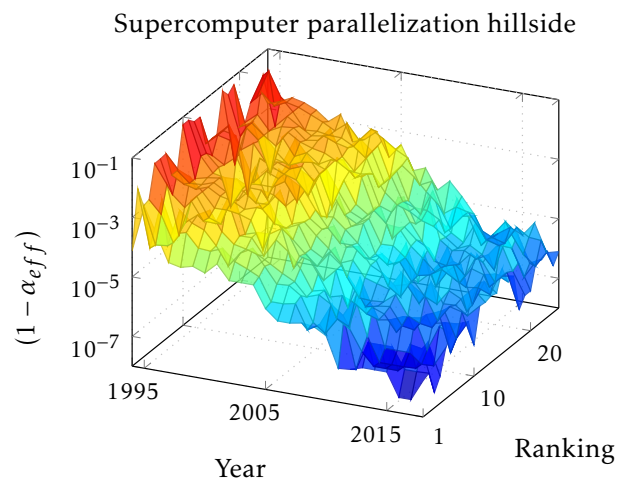


Figure 8: The Top500 supercomputer parallelization efficiency in function of the year of construction and the actual ranking. The $(1 - \alpha)$ parameter for the past 25 years and the (by R_{max}) first 25 computers. Data derived using the HPL benchmark.

4.2 Benchmarking the performance of a complex computing system

As discussed above, measuring the performance of a complex computing system is not trivial at all. Not only finding the proper merit is hard, but also the "measuring device" can basically influence the measurement. The importance of selecting the proper merit and benchmark can be easily understood

through utilizing the model and the well-documented benchmark measurements of the supercomputers.

As experienced in running the benchmarks HPL and HPCG [45] and explained in connection with Fig. 13, the different benchmarks produce different payload performance values and computational efficiencies on the same supercomputer. The model presented in Fig. 3 enables to explain the difference.

The benchmarks, utilized to derive numerical parameters for supercomputers, are specialized *programs*, which run in the HW/OS *environment* provided by the supercomputer under test. Two typical fields of their utilization: to describe the environment supercomputer application runs in, and to guess how quickly an application will run on a given supercomputer.

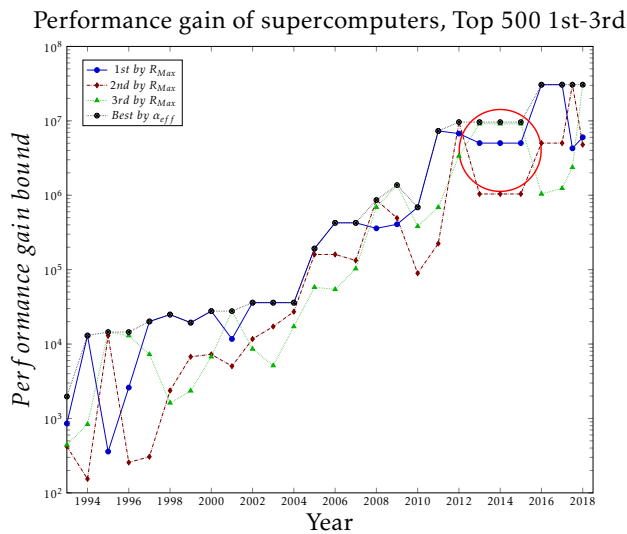


Figure 9: The trend of the development of computing performance gain in the past 25 years, based on the first three (by R_{Max}) and the first (by $(1 - \alpha)$) in the year of construction. Data derived using the HPL benchmark. A saturation effect around 10^7 is expressed.

The (apparently) sequential fraction $(1 - \alpha_{eff})$, as it is obvious from our model, cannot distinguish between the (at least apparently) sequential processing time contributions of different origin, even the SW (including OS) and HW sequential contributions cannot be separated. Similarly, it cannot be taken for sure that those contributions sum up linearly. Different benchmarks provide different SW contributions to the non-parallelizable fraction of the execution time (resulting in different efficiencies and ranking [35]), so comparing results (and especially establishing ranking!) derived using different benchmarks shall be done with maximum care. *Since the efficiency depends heavily on the number of cores, different configurations shall be compared using the same benchmark and the same number of processors (or same R_{Peak}).*

If the goal is to characterize the supercomputer's HW+OS system itself, a benchmark program should

distort HW+OS contribution as little as possible, i.e. the SW contribution must be much lower than the HW+OS contribution. In the case of supercomputers, the benchmark HPL is used for this goal since the beginning of the supercomputer age. The mathematical behavior of HPL enables to minimize SW contribution, i.e. *HPL delivers the possible best estimation for α_{eff}^{HW+OS} .*

If the goal is to estimate the expectable behavior of an application, the benchmark program should imitate the structure and behavior of the application. In the case of supercomputers, a couple of years ago the benchmark HPCG has been introduced for this goal, since *"HPCG is designed to exercise computational and data access patterns that more closely match a different and broad set of important applications, and to give incentive to computer system designers to invest in capabilities that will have impact on the collective performance of these applications"* [45]. However, its utilization can be misleading: *the ranking is only valid for the HPCG application, and only utilizing that number of processors.* HPCG seems really to give better hints for designing supercomputer applications⁵, than HPL does. According to our model, in the case of using the HPCG benchmark, the SW contribution dominates⁶, i.e. HPCG delivers the best estimation for α_{eff}^{SW} for this class of supercomputer applications.

Supercomputer community has extensively tested the efficiency of TOP500 supercomputers when benchmarked with HPL and HPCG [45]. It was found that the efficiency (and R_{Max}) is typically 2 orders of magnitude lower when benchmarked with HPCG rather than with HPL, even at relatively low number of processors.

5 Supercomputing

In supercomputing the resulting payload computing performance is crucial and the number of the processors, their single-processor performance and the speed of their interconnection are critical resources. As today a "gold rush" is experienced with the goal to achieve the dream limit of 1 Eflop/s (10^{18} flop/s) [46], the section scrutinizes the feasibility of achieving that goal. Through applying the model to different kinds of large-scale sequential-parallel computing systems it is shown that such systems require to understand the role and dominance of the contributions of quite different kinds to the performance loss and that the improperly designed HW/SW cooperation provides direct evidence about the existence of the performance bound. The well-documented, strictly controlled measurement database [48] enables to draw both retrospective statistical conclusions on the logic of development behind the performance data, as well as to make predictions for the near future about the performance of supercomputers and its limitations.

⁵This is why for example [47] considers HPCG as "practical performance".

⁶ Returning calculated gradients requires much more sequential communication (unintended blocking).

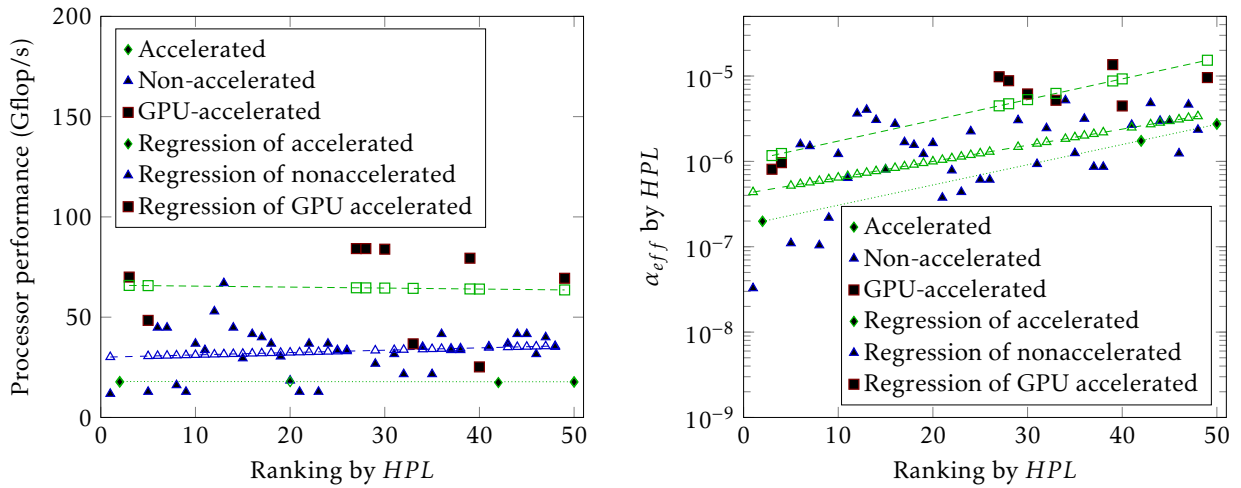


Figure 10: Correlation of performance of processors using accelerator and effective parallelism with ranking, in 2017. The left figure shows that utilizing GPU accelerators increases single-processor performance by a factor of 2...3, but as the right side demonstrates, at the price of increasing the non-parallelizable fraction.

Prediction of R_{Max}^{HPL} of Top10 Supercomputers

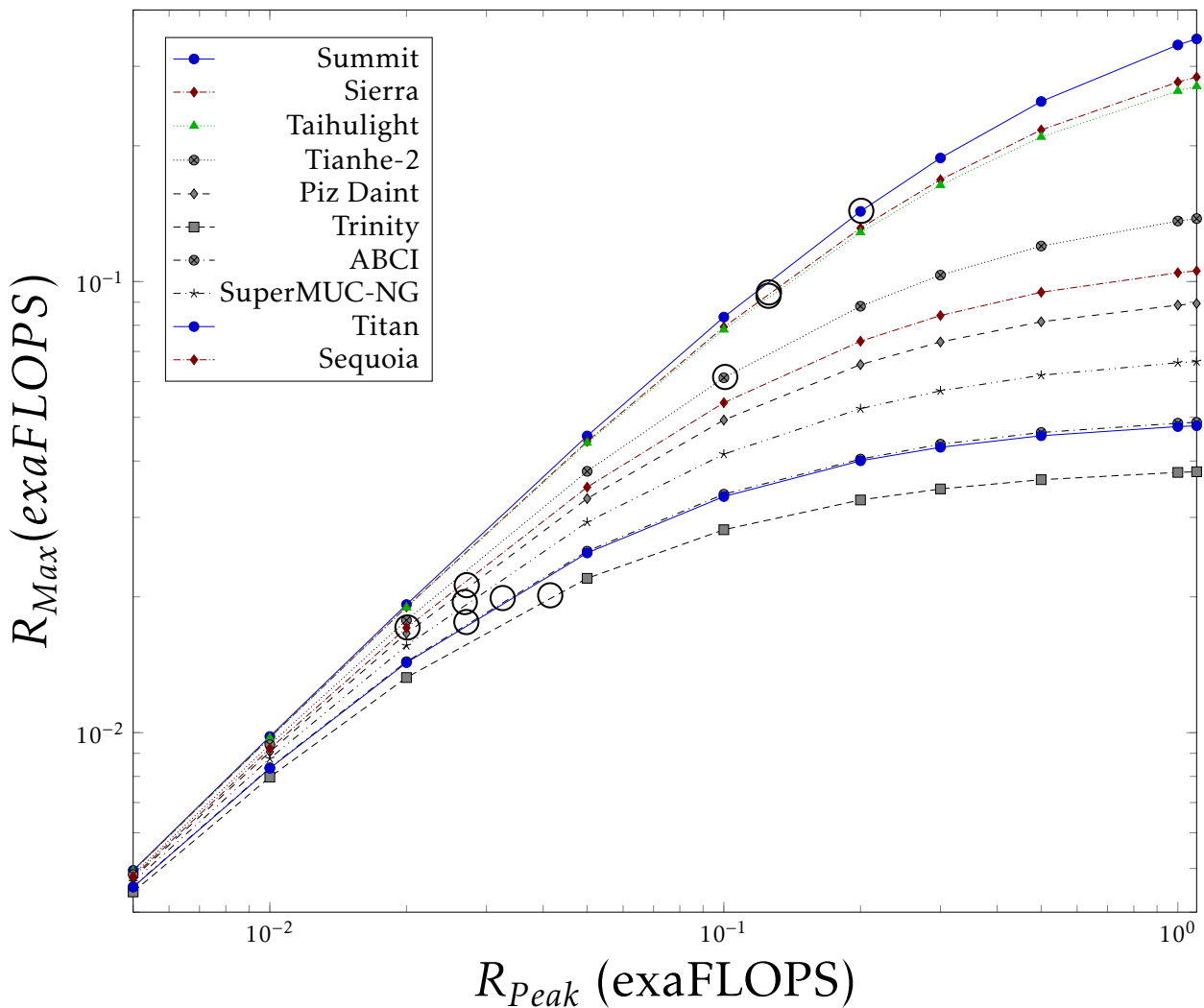


Figure 11: Dependence of payload supercomputer performance on the nominal performance for the TOP10 supercomputers (as of November 2018) in case of utilizing the HPL benchmark. The actual positions are marked by a bubble on the diagram lines.

5.1 Describing the development of super-computing

One can assume that all configurations are built with the actually available best technology and components, and only the really best configurations can be found in circles of TOP lists. The large number of configurations in the rigorously controlled database TOP500 [48] enables to draw reliable conclusions, although with considerable scattering. Fig. 8 depicts the calculated $(1 - \alpha_{eff})$ for the past 26 years and the first 25 computers in the TOP500 list.

Fig. 8 explicitly shows signs of reaching a "flat surface" in values of $(1 - \alpha_{eff})$ in the past few years. This effect can be more clearly studied if considering the TOP3 supercomputers only. Fig. 9 depicts the performance gain (see (7)) of supercomputers for their 26-years history. As seen, symptoms of stalling of the parallelized performance appeared in 2012: in three adjacent years the payload performance gain of the TOP3 supercomputers remained at the same value. With the appearance of *Taihulight* in 2016 apparently a by an order of magnitude higher performance gain has shined up. However, it is just the consequence of using the "cooperative computing" [42], the rest of world remains under that limiting stalling value. The new world champion in 2018 could conquer the slot #1 only due to its higher (accelerator-based) single-processor performance, rather than the enhanced effective parallelization (although its clustering played an important role in keeping it in good condition).

5.2 Predictions of supercomputer performance for the near future

As Eq. (8) shows, the resulting performance can be increased by increasing either the single processor performance or the performance gain. As the computational density cannot be increased any more [49, 50], some kind of accelerator is used to (apparently) increase the single processor performance. The acceleration, however, can also contribute to the non-parallelizable fraction of the computing task.

Fig. 10 shows how utilizing Graphic Processing Unit (GPU) acceleration is reflected in parameters of the TOP50 supercomputers. As the left side of the figure displays, the GPU really increases the single-processor performance by a factor of 2...3 (in accordance with [51]), but at the same time increases the value of $(1 - \alpha_{eff})$. The first one is a linear multiplier, the second one is an exponential divisor. Consequently, at low number of cores it is advantageous to use that kind of acceleration, while at high number of cores it is definitely disadvantageous.

Having the semi-technical model ready, one can estimate how the computing performance will develop in the coming few years. Assuming that the designers can keep the architecture at the achieved performance gain, and virtually changing the number of processors one can estimate how the payload performance changes when adding more cores to the existing TOP10

constructions. Fig. 11 shows the virtual payload performance in the function of the nominal performance for the TOP105 supercomputers (as of November 2018). Notice that the predictions are optimistic, because the performance breakdown shown in Fig. 13 is not included. Even with that optimistic assumption, the 1 Eflo/s payload performance cannot be achieved with the present technology and paradigm.

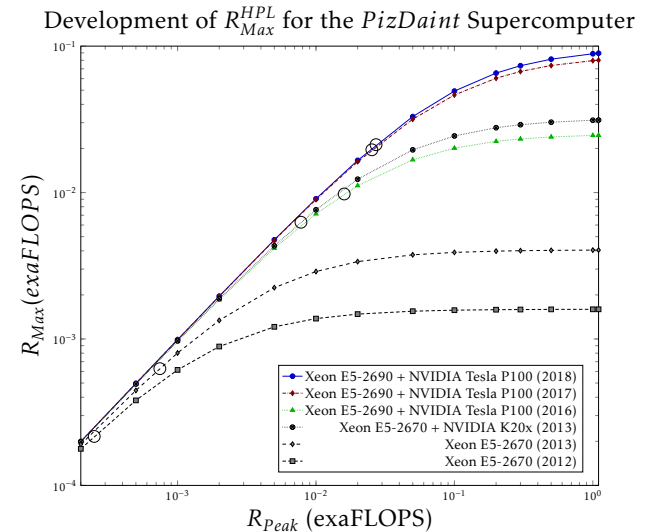


Figure 12: Dependence of the predicted payload performance of supercomputer *Piz Daint* in different phases of the development, with the prediction based on the actual measured efficiency values. The actual positions are marked by bubbles on the diagram lines.

The accuracy of the short-term prediction can be estimated from Fig. 12. Fortunately, supercomputer *Piz Daint* has a relatively long documented history in the TOP500 database [48]. The publishing of the results of the development has started in year 2012. In the next year the number of processor has changed by a factor of three, and the parallelization efficiency considerably improved by the same factor, although presumably some hardware efficacy improvements have also occurred. Despite this, the predicted performance improvement is quite accurate. (unfortunately, two parameters have been changed between two states reported to the list, so their effect cannot be qualified separately.) In year 2013 GPU acceleration with K20 have been introduced, and the number of processors have been increased by a factor of four. The resulting effect is a factor of 10 increase in both the payload performance and the nominal performance. Probable a factor of 2.5 can be attributed to the GPU, which value is in good agreement with the values received in[51, 30], and a factor of 4 to the increased number of cores. The designers were not satisfied with the result, so they changed to TESLA P100. Notice that the change to a more advanced type of GPU results in a slight fallback relative to the prediction in the expected value of R_{Max} : copying between a bigger GPU memory and the main memory increases the non parallelizable

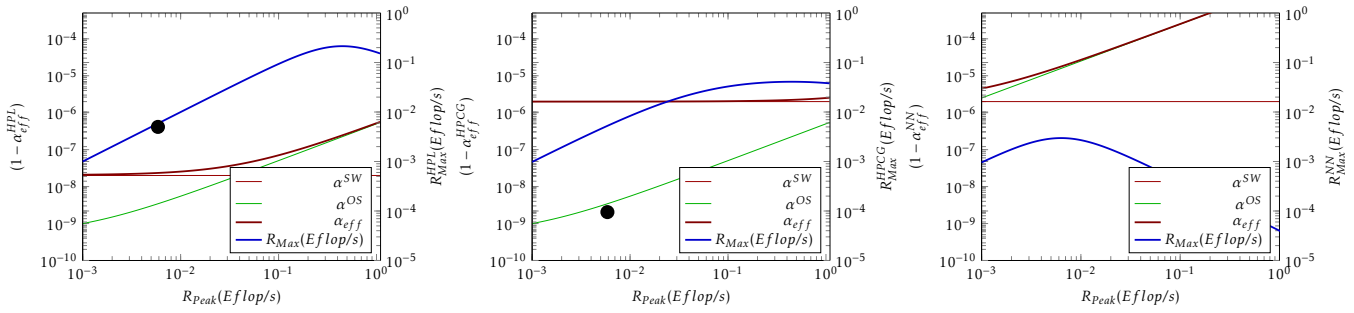


Figure 13: Contributions $(1 - \alpha_{eff}^X)$ to $(1 - \alpha_{eff}^{total})$ and max payload performance R_{Max} of a fictive supercomputer ($P = 1 Gflop/s @ 1GHz$) in function of the nominal performance. The blue diagram line refers to the right hand scale (R_{Max} values), all others ($(1 - \alpha_{eff}^X)$ contributions) to the left scale. The leftmost figure illustrates the behavior measured with benchmark HPL. The looping contribution becomes remarkable around 0.1 Eflops, and breaks down payload performance when approaching 1 Eflops. The black dot marks the HPL performance of the computer used in works [53, 55]. In the middle figure the behavior measured with benchmark HPCG is displayed. In this case the contribution of the application (thin brown line) is much higher, the looping contribution (thin green line) is the same as above. As a consequence, the achievable payload performance is lower and also the breakdown of the performance is softer. The black dot marks the HPCG performance of the same computer. The rightmost figure demonstrates what happens if the clock cycle is 5000 times longer: it causes a drastic decrease in the achievable performance and strongly shifts the performance breakdown toward lower nominal performance values. The figure is purely illustrating the concepts; the displayed numbers are somewhat similar to the real ones.

contribution, which cannot be counterbalanced even by the two times more cores. Finally, the constructors upgraded also the processor (again changing two parameters at a time), but the twice more processors and the new accelerator produced only twice more performance. The upgrade in 2018 accurately predicted by the diagram. As the diagram lines in Fig. 12 accurately display the resulting values of the later implementation if no changes in the architecture (the efficiency) happens, one can rely to the estimations for predicting future values display in Fig. 11, provided that no architectural changes (or changes in the paradigm) occur during the development.

5.3 What factors dominate the performance gain

As discussed above, different goals can be targeted with the parallelized sequential HW/SW systems. Accordingly, different factors dominate in limiting the computational performance. The case of simulating large neural networks is a very good example how the dominance of the different factors changes with the actual conditions.

The idea of simulating neural activity in a way that processors are used to provide the computing facility for solving partial differential equations, seems to be a good idea, as well as to utilize parallelly running SW threads to imitate neural activity of large networks. However, utilizing parallelized sequential systems for that goal implies all limitations discussed above.

As discussed in details in [52], two of the mentioned effects can become dominating in those applications. Since the operating time scale of the biological

networks lies in the *msec* range, brain simulation applications commonly use integration time about 1 *ms* [53]. Since the threads running in parallel need to be synchronized (i.e. they must be put back on the same biological time scale, in order to avoid working with "signals from the future"), quietly and implicitly a 1 *ms* "clock signal" is introduced. Since this clock period is a million times longer than the typical clock period in digital computers, its dominance is seriously promoted in producing non-payload overhead. This implicit clock signal has the same limiting effect for both the many-thread simulation and the special-purpose HW brain simulator [54]. This is why the special HW brain simulator cannot outperform the many-thread version SW running on a general-purpose supercomputer.

The other field-specific effect is that the processing speed is too quick and storage capacity too large for simulating biological systems in an economic way, so the same core is utilized to simulate several neurons (represented by several threads). This method of solution however needs several context changes within the same core, and since the context changes are rather expensive in terms of execution time [12], the overhead part of the process amounts to 10% [53], and the 0.1 value of $(1 - \alpha_{eff})$ competes for dominance with the improperly selected clock period. This is the reason why the many-thread based simulation cannot scale above a few dozens of thousands of neurons [55]. (At the same time, the analog simulators achieve thousands time better performance and scaling.)

It is worth to have look at Fig. 13. The thick blue line shows the dependence of the payload performance on the nominal performance under different condi-

tions and refers to the right side scale. The leftmost figure shows the case of running the benchmark HPL in some fictive (but somewhat similar to *Taihulight*) supercomputer. The middle figure shows how a typical computation (modelled by the benchmark HPCG, utilizing intensively communication between threads) leads to strong degradation of payload performance at higher nominal performances. HPCG is in its behavior greatly similar to AI applications. These real-life applications show also smeared maximum behavior (compared to those of the benchmark HPL).

On the rightmost figure the clock signal is 5,000 times higher, modeling the "hidden clock signal". In this case the performance degradation is much more expressed, and also the breakdown of the performance has been measured [55].

6 How to proceed after reaching limits of parallelization

Today the majority of leading researchers agree that computing (and especially: the computing paradigm) needs renewal (for a review see [56]), although there is no commonly accepted idea for a new paradigm. After the failure of supercomputer *Aurora* (A18) project it became obvious that a processor optimized for SPA regime cannot be optimal at the same time for parallelized sequential regime.

Intel learned the lesson and realized that exa-scale computing needs a different architecture. Although Intel is very cautious with discovering its future plans, especially the exa-scale related ones [57], they already dropped the X86 line [58]. Their new patent [59] attempts to replace the conventional instruction-driven architecture by a data-driven one. However, a serious challenge will be to eliminate the overhead needed to frequently reconfigure the internals of the processor for the new task fraction and it is also questionable that emulating the former X86 architecture (obviously from code compatibility reasons) enables to reduce the inherent overhead coming from that over-complex single-processor oriented architecture. The new and ambitious player in Europe [39, 60], however, thinks that some powerful architecture (although even the processor type is not selected) can overcome the theoretical and technological limits without changing the paradigm. That is, there are not much "drastically new" ideas on board.

As discussed above, from the point of view of parallelism the inherently sequential parts of any HW/SW system form a kind of overhead. The happenings around parallelised sequential processing systems validate the prophecy of Amdahl: *The nature of this overhead [in parallelism] appears to be sequential so that it is unlikely to be amenable to parallel processing techniques. Overhead alone would then place an upper limit on throughput . . . , even if the housekeeping were done in a separate processor* [3]

As Amdahl in 1967(!) warned: *the organization of a single computer has reached its limits and that truly sig-*

nificant advances can be made only by interconnection of a multiplicity of computers in such a manner as to permit cooperative solution [3]. Despite that warning, even today, many-processor systems, including supercomputers, distributed systems and manycore processors as well, comprise many single-processor systems, (rather than "cooperating processors" as envisioned by Amdahl). The suggested solutions typically consider segregated cores, based on SPA components and use programmed communication, like [61].

There are some exceptions, however: solutions like [62, 63] transfer control between (at Instruction Set Architecture (ISA) level) equivalent cores, but under SW control. Direct connection between cores exists only in [42], also under SW control. This processor has kept the supercomputer *Taihulight* in the first slot of the TOP500 list [48] for two years. Optimizing this direct (non-memory related) data transfer resulted in drastic changes [43] in the executing of the HPCG benchmark (mitigating the need to access memory decreases the sequential-only part). Unfortunately, the architectures with programmed core-to-core connection are not flexible enough: the inter-core register-to-register data transfer must be programmed in advance by the programmer. The need for solving architectural inflexibility appeared, in the form of "switchable topology" [64]. The idea of "outsourcing" (sending compute message to neighboring core) [65] also shined up. Both suggestions without theoretical and programming support.

A different approach can be to use a radically new (or at least a considerable generalization of the old) computing paradigm. That radically new paradigm at the same time must provide a smooth transition from the present age to the new one, as well as co-existence for the transition period. The idea of utilizing Explicitly Many-Processor Approach (EMPA) [66] seems to fulfill those requirements, although it is in a very early stage of development.

The technology of manufacturing processors (mainly the electronic component density) as well as the requirements against its utilization have considerably changed since the invention of the first computer. Some of the processing units (the cores within a processor) are in close proximity, but this alone, without cooperation is not sufficient. Even utilizing synergistic cell processors [67] did not result in breakthrough success. Utilizing cooperation of processors in close proximity [42], however, resulted in the best effective parallelism up to now. Establishing further ways of cooperation (like distributing execution of task fragments to cooperating cores in close proximity [66, 68]), however, can considerably enhance the apparent performance through decreasing the losses of parallelization.

7 Summary

The parallelization of otherwise sequential processing has its natural and inherent limitations, does not meet

the requirements of modern computing and does not provide the required flexibility. Although with decreasing efficiency, for part of applications sufficiently performable systems can be assembled and with some tolerance and utilizing special SW constraints, the real-time needs can be satisfied. In the case of extremely large processing capacity, however, bounds of the parallelized sequential systems are faced. *For developing the computing performance further, the 50-years old idea about making systems comprising cooperating processors must be renewed.* The need for cooperative computing is evident and its feasibility was convincingly demonstrated by the success of the world's first really cooperative processor [42]. An extension [66] to the computing paradigm, that considers both the technological state-of-the-art and the expectations against computing, was also presented and some examples of its advantageous features were demonstrated.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment Project no. 125547 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the K funding scheme.

References

- [1] P. J. Denning, T. Lewis, Exponential Laws of Computing Growth, *COMMUN ACM* (2017) 54–65doi:DOI: 10.1145/2976758.
- [2] S. H. Fuller and L. I. Millett, Computing Performance: Game Over or Next Level?, *Computer* 44/1 (2011) 31–38.
- [3] G. M. Amdahl, Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities, in: *AFIPS Conference Proceedings*, Vol. 30, 1967, pp. 483–485. doi: 10.1145/1465482.1465560.
- [4] J. Yang et al., Making Parallel Programs Reliable with Stable Multithreading *COMMUN ACM* 57/3(2014)58–69
- [5] U. Vishkin, Is Multicore Hardware for General-Purpose Parallel Processing Broken?, *COMMUN ACM*, 57/4(2014)p35
- [6] K. Hwang, N. Jotwani, *Advanced Computer Architecture: Parallelism, Scalability, Programmability*, 3rd Edition, Mc Graw Hill, 2016.
- [7] Schlansker, M.S. and Rau, B.R. EPIC: Explicitly Parallel Instruction Computing, *Computer* 33(2000)37–45
- [8] D. J. Lilja, *Measuring Computer Performance: A practitioner's guide*, Cambridge University Press, 2004.
- [9] Singh, J. P. et al, Scaling Parallel Programs for Multiprocessors: Methodology and Examples, *Computer* 27/7(1993),42–50
- [10] Arvind and Iannucci, Robert A., Two Fundamental Issues in Multiprocessing 4th International DFVLR Seminar on Foundations of Engineering Sciences on Parallel Computing in Science and Engineering (1988)61–88
- [11] Mahlke, S.A. and Chen, W.Y. and Chang, P.P. and Hwu, W.-M.W., Scalar program performance on multiple-instruction-issue processors with a limited number of registers *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences* (1992)34 - 44
- [12] D. Tsafirir, The context-switch overhead inflicted by hardware interrupts (and the enigma of do-nothing loops), in: *Proceedings of the 2007 Workshop on Experimental Computer Science, ExpCS '07*, ACM, New York, NY, USA, 2007, pp. 3–3. URL <http://doi.acm.org/10.1145/1281700.1281704>
- [13] Luiz Andr Barroso and Urs Hlzl, The Case for Energy-Proportional Computing *Computer* 40(2007)33–37
- [14] "Babaolu, zalp and Marzullo, Keith and Schneider, Fred B.", "A formalization of priority inversion" *Real-Time Systems*, 5/4(1993)285303
- [15] "L. Sha and R. Rajkumar and J.P. Lehoczky", "Priority inheritance protocols: an approach to real-time synchronization" *IEEE Transactions on Computers*, 39/4(1990)1175–1185
- [16] S. Krishnaprasad, Uses and Abuses of Amdahl's Law, *J. Comput. Sci. Coll.* 17 (2) (2001) 288–293. URL <http://dl.acm.org/citation.cfm?id=775339.775386>
- [17] F. Dévai, The Refutation of Amdahl's Law and Its Variants, in: O. Gervasi, B. Murgante, S. Misra, G. Borruso, C. M. Torre, A. M. A. Rocha, D. Taniar, B. O. Apduhan, E. Stankova, A. Cuzzocrea (Eds.), *Computational Science and Its Applications – ICCSA 2017*, Springer International Publishing, Cham, 2017, pp. 480–493.
- [18] J. M. Paul, B. H. Meyer, Amdahl's Law Revisited for Single Chip Systems, *INT J of Parallel Programming* 35 (2) (2007) 101–123.
- [19] J. Végħ, P. Molnár, How to measure perfectness of parallelization in hardware/software systems, in: *18th Internat. Carpathian Control Conf. ICC*, 2017, pp. 394–399.
- [20] A. H. Karp, H. P. Platt, Measuring Parallel Processor Performance, *COMMUN ACM* 33 (5) (1990)Inherent Sequentiality:2012 539–543. doi: 10.1145/78607.78614.
- [21] F. Ellen, D. Hendler, N. Shavit, On the Inherent Sequentiality of Concurrent Objects, *SIAM J. Comput.* 43 (3) (2012) 519536.
- [22] L. Yavits, A. Morad, R. Ginosar, The effect of communication and synchronization on Amdahl's law in multicore systems, *Parallel Computing* 40 (1) (2014) 1–16.
- [23] K. Pingali, D. Nguyen, M. Kulkarni, M. Burtscher, M. A. Hassaan, R. Kaleem, T.-H. Lee, A. Lenharth, R. Manevich, M. Méndez-Lojo, D. Prountzos, X. SuiThe Tao of Parallelism in Algorithms, *SIGPLAN Not.* 46 (6) (2011) 12–25.
- [24] I. Markov, Limits on fundamental limits to computation, *Nature* 512(7513) (2014) 147–154.
- [25] Sun, Xian-He and Gustafson, John L., Toward a Better Parallel Performance Metric *Parallel Comput.*, 17/10-11(1991)1093–1109
- [26] D.A. Patterson and J.L. Hennessy, *Computer Organization and design. RISC-V Edition*, (2017) Morgan Kaufmann
- [27] S. Orii, Metrics for evaluation of parallel efficiency toward highly parallel processing "Parallel Computing " 36/1(2010)16–25
- [28] P. Molnár and J. Végħ, Measuring Performance of Processor Instructions and Operating System Services in Soft Processor Based Systems. *18th Internat. Carpathian Control Conf. ICC* (2017)381–387
- [29] Randal E. Bryant and David R. O'Hallaron, *Computer Systems: A Programmer's Perspective* (2014) Pearson
- [30] J. Végħ, Statistical considerations on limitations of supercomputers, *CoRR* abs/1710.08951.
- [31] W. Sheng et al., A compiler infrastructure for embedded heterogeneous MPSoCs *Parallel Computing* volume = 40/2(2014)51-68

- [32] L. de Macedo Mourelle, N. Nedjah, and F. G. Pessanha, *Reconfigurable and Adaptive Computing: Theory and Applications*. CRC press, 2016, ch. 5: Interprocess Communication via Crossbar for Shared Memory Systems-on-chip.
- [33] Mohammadi, M. and Bazhiro, T., Comparative Benchmarking of Cloud Computing Vendors with High Performance Linpack Proceedings of the 2nd International Conference on High Performance Compilation, Computing and Communications (2018)1–5
- [34] E. Wustenhoff and T. S. E. Ng, Cloud Computing Benchmark(2017) <https://www.burstorm.com/price-performance-benchmark/1st-Continuous-Cloud-Price-Performance-Benchmarking.pdf>
- [35] IEEE Spectrum, Two Different Top500 Supercomputing Benchmarks Show Two Different Top Supercomputers, <https://spectrum.ieee.org/tech-talk/computing/hardware/two-different-top500-supercomputing-benchmarks-show-two-different-top-supercomputers> (2017).
- [36] J. Dongarra, Report on the Sunway TaihuLight System, Tech. Rep. Tech Report UT-EECS-16-742, University of Tennessee Department of Electrical Engineering and Computer Science (June 2016).
- [37] Robert F. Service, Design for U.S. exascale computer takes shape, *Science*, 359/6376(2018)617–618
- [38] US DOE, The Opportunities and Challenges of Exascale Computing, https://science.energy.gov/-/media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf (2010).
- [39] European Commission, Implementation of the Action Plan for the European High-Performance Computing strategy, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15269 (2016).
- [40] Extremtech Japan Tests Silicon for Exascale Computing in 2021. <https://www.extremtech.com/computing/272558-japan-tests-silicon-for-exascale-computing-in-2021>
- [41] X Liao, Kaiet al Moving from exascale to zettascale computing: challenges and techniques. *Frontiers of Information Technology & Electronic Engineering* 567 19(10) pp: 1236–1244 (2018)
- [42] F. Zheng, H.-L. Li, H. Lv, F. Guo, X.-H. Xu, X.-H. Xie, Cooperative computing techniques for a deeply fused and heterogeneous many-core processor architecture, *Journal of Computer Science and Technology* 30 (1) (2015) 145–162.
- [43] Ao, Yulong et al Performance Optimization of the HPCG Benchmark on the Sunway TaihuLight Supercomputer *ACM Trans. Archit. Code Optim.* 15/1(2018)1–20
- [44] S. Eyerma, L. Eeckhout, Modeling Critical Sections in Amdahl's Law and Its Implications for Multicore Design, *SIGARCH Comput. Archit. News* 38 (3) (2010) 362–370.
- [45] HPCG Benchmark, <http://www.hpcg-benchmark.org/> (2016).
- [46] J. Dongarra, The Global Race for Exascale High Performance Computing(2017) http://ec.europa.eu/newsroom/document.cfm?doc_id=45647
- [47] Tim Dettmers, The Brain vs Deep Learning Part I: Computational Complexity Or Why the Singularity Is Nowhere Near (2015) <http://timdettmers.com/2015/07/27/brain-vs-deep-learning-singularity/>
- [48] TOP500.org, TOP500 Supercomputer Sites. URL <https://www.top500.org/>
- [49] J. Williams et al, Computational density of fixed and reconfigurable multi-core devices for application acceleration Proceedings of Reconfigurable Systems Summer Institute, Urbana, IL, (2008)
- [50] J. Williams et al, Characterization of Fixed and Reconfigurable Multi-Core Devices for Application Acceleration *ACM Trans. Reconfigurable Technol. Syst.* volume = 3/4(2010) 19:1–19:29
- [51] Lee, Victor W. et al, Debunking the 100X GPU vs. CPU Myth: An Evaluation of Throughput Computing on CPU and GPU Proceedings of the 37th Annual International Symposium on Computer Architecture ISCA '10(2010)451–460,
- [52] J. Végth How Amdahl's Law limits the performance of large neural networks *Brain Informatics*, in review (2019).
- [53] van Albada, Sacha J. and Rowley, Andrew G. and Senk, Johanna and Hopkins, Michael and Schmidt, Maximilian and Stokes, Alan B. and Lester, David R. and Diesmann, Markus and Furber, Steve B., Performance Comparison of the Digital Neuromorphic Hardware SpiNNaker and the Neural Network Simulation Software NEST for a Full-Scale Cortical Microcircuit Model, *Frontiers in Neuroscience* 12(2018)291
- [54] S. B. Furber et al, Overview of the SpiNNaker System Architecture *IEEE Transactions on Computers* 62/12(2013)2454-2467
- [55] Ippen, Tammo and Eppler, Jochen M. and Plesser, Hans E. and Diesmann, Markus, Constructing Neuronal Network Models in Massively Parallel Environments, *Frontiers in Neuroinformatics* 11 (2017) 30-41.
- [56] J. Végth, Renewing computing paradigms for more efficient parallelization of single-threads, Vol. 29 of *Advances in Parallel Computing*, IOS Press, 2018, Ch. 13, pp. 305–330.
- [57] Intel, Looking Ahead to Intels Secret Exascale Architecture (2017), <https://www.nextplatform.com/2017/11/14/looking-ahead-intels-secret-exascale-architecture/>
- [58] Intel Intels Exascale Dataflow Engine Drops X86 And Von Neumann (2018) <http://www.freepatentsonline.com/y2018/0189231.html>
- [59] Intel, Processors, methods and systems with a configurable spatial accelerator (2018) <http://www.freepatentsonline.com/y2018/0189231.html>
- [60] European Community, The European Processor Initiative (EPI) to develop the processor that will be at the heart of the European exascale supercomputer effort (2018) <http://e-irg.eu/news-blog/-/blogs/the-european-processor-initiative-epi-to-develop-the-processor-that-will-be-at-the-heart-of-the-european-exascale-supercomputer-effort>
- [61] GlobalFoundries Inc, Administering inter-core communication via shared memory (2013) <https://patents.google.com/patent/US9223505>
- [62] J. Congy, et al, Accelerating Sequential Applications on CMPs Using Core Spilling, *Parallel and Distributed Systems* 18 (2007) 1094–1107.
- [63] ARM, big.LITTLE technology (2011). URL <https://developer.arm.com/technologies/big-little>
- [64] Intel, Switchable topology machine (2018) <https://patents.google.com/patent/US20180113838A1>
- [65] Nokia, Method, apparatus, and computer program product for parallel functional units in multicore processors, (2013) <https://patents.google.com/patent/US20130151817A1/>
- [66] J. Végth, Introducing the explicitly many-processor approach, *Parallel Computing* 75 (2018) 28 – 40.
- [67] M. Kistler et al, Cell multiprocessor communication network: Built for speed, *IEEE Micro* 26/3 (2006) 10–23.
- [68] J. Végth, EMPATHY86: A cycle accurate simulator for Explicitly Many-Processor Approach (EMPA) computer. doi: 10.5281/zenodo.580633. URL <https://github.com/jvegh/EMPATHY86>

Low Contrast Image Enhancement Using Convolutional Neural Network with Simple Reflection Model

Young Shik Moon*, Bok Gyu Han, Hyeon Seok Yang, Ho Gyeong Lee

Hanyang University, Computer Science & Engineering, 15588, Republic of Korea

ARTICLE INFO

Article history:

Received: 20 December, 2018

Accepted: 23 January, 2019

Online : 5 February, 2019

Keywords:

Image Enhancement

Convolutional Neural Network

Reflection Model

Machine Learning

ABSTRACT

Low contrast images degrade the performance of image processing system. To solve the issue, plenty of image enhancement methods have been proposed. But the methods work properly on the fixed environment or specific images. The methods dependent on fixed image conditions cannot perform image enhancement properly and perspective of smart device users, algorithms including iterative calculations are inconvenient for users. To avoid these issues, we propose a locally adaptive contrast enhancement method using CNN and simple reflection model. The experimental results show that the proposed method reduces over-enhancement, while recovering the details of the low contrast regions.

1. Introduction

Image enhancement is an important and typical topic of computer vision. Although the performance of digital systems has improved greatly, the low quality of some images due to external factors such as environment and backlight, which may degrade the performance of image processing systems such as intelligent traffic systems, visual surveillance, and consumer electronics [1]. Especially, low contrast images reduce visibility. Therefore, many smart devices help users take pictures and improve their results with internal image processing methods, but require some input, such as a low contrast area input by the device user, or an average brightness and hue. In specific condition, it is a good solution for the device user who does not consider about the visibility of the image such as counter light images on purpose. However, only chooses the color tone and increases overall brightness is not fit for image enhancement in view of the simple device users.

While maintaining the image quality, to improve the low contrast region of input image is not easy. To solve the problem in this conditional situation, a plenty of methods have been proposed based on mathematical knowledge and traditional image processing method but the methods only perform well in fixed circumstance or make side effect such as over-enhancement or halo effect. To reduce the undesired effect, some enhancement methods use optimization techniques with image decomposition mechanism. However, the computation time is increased and the methods may produce unintended adverse effects depending on the

objective function with decomposed factors. Because of these disadvantage, the methods are not proper to simple device users.

In the past, artificial neural network methods have been one of the most difficult methods to obtain successful results due to small amounts of computation resource and data. However, artificial neural networks are widely used in many fields due to the development of devices and explosively increased amount of data. In addition, as many learning methods are studied, many effective approaches based on input data conditions have been proposed. Recently, the deep learning based methods show the best performance in computer vision sections. The methods are widely being employed on many computer vision tasks that are not easily solved with traditional methods such as image classification, image segmentation, etc. The deep learning based methods with well-labeled training data automatically abstract features of the image differently from the traditional methods and reconstruct the final result with the extracted features. However, the performance of the methods is built on well-labelled data. To overcome the conditional limits, several training techniques are proposed such as semi-supervised learning and weakly-supervised learning. From the perspective of performance, using these skills are able to lead to better results and it can be used widely in general purpose. The performance of the deep learning models trained using non-labelled or rough labelled data shows the power of representation ability.

Focus on the advantage of deep learning, we propose a locally adaptive contrast enhancement method. The method follows previous image enhancement mechanism basically, detecting low contrast region and apply enhancement algorithm. Each person

*Young Shik Moon, 55, Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggi-do, Republic of Korea, ysmoon@hanyang.ac.kr

evaluations for the low contrast region are different. Therefore, we use deep learning mechanism to employ the advantage which extracting feature factor from dataset automatically. Even though our method is based on training with confusing data, the power of representation ability of deep learning shows well results on contrast enhancement purpose. Also, we use a simple reflection model to reduce computational burden.

The remainder of this paper is structured as follows: Section 2 provides background information on image enhancement and briefly introduces previous methods. Section 3 gives a detail of proposed method and shows the detail of our network. While Section 4 provides the experiment results of image contrast enhancement with the proposed method and the previous methods. Finally, Section 5 concludes this paper.

2. Related works

For image enhancement, a lot of image processing methods were proposed previously. Histogram based methods are the most popular due to its simplicity and effectiveness. Basically, histogram equalization (HE) is a well-known traditional contrast enhancement method. The main idea of HE contains automatic calculation of the uniform histogram distribution in dynamic range. But HE method does not preserve to mean intensity of input images. The mean intensity of an input image and result of HE are different and it cause undesired effect such an over-enhancement. To solve the issue, brightness preserving bi-histogram equalization (BBHE) [2] was proposed. It divides histogram of an input image into sub histograms and enhances contrast separately. Adaptive histogram equalization (AHE) is used for enhancing contrast in images. It differs from HE by adaptive method that computes several histograms and each histogram corresponding to a distinct section of an image. AHE divide an input image into small sub images and apply HE to each sub image. But because of the size of the sub image, it is too sensitive to noise. Contrast-limited adaptive histogram equalization (CLAHE) [3] is improved version of AHE. The processes of CLAHE is basically similar to AHE. CLAHE controls the sensitive noise with contrast limited distribution parameters. The neighboring tiles, sub image, are combined by bilinear interpolation. Especially in homogeneous areas, the contrast can be limited to avoid noises by distribution parameter.

Retinex-based methods assume that recognizes the relative brightness of the scene, rather than recognizing the brightness of the scene at a certain position when the human visual system recognizes the scene. Retinex based methods have been proposed widely and single-scale retinex (SSR) [4], multi-scale retinex (MSR) [5] and multi-scale retinex with color restoration (MSRCR) [6] are most well-known algorithms. SSR is a method of using the difference between the center pixel value and the convolution result around the center pixel on the log scale. Several convolution masks were proposed [7, 8] but to select the variance of convolution mask effects largely and inadequate variance value can be lead to deterioration of image quality. The MSR is proposed as a method to mitigate this problem of SSR and the retinex result is calculated as the weighted average value of SSR with multiple variance values. The MSRCR proposed that color restoration function (CRF) added to MSR result to solve a problem that come out the gray scale result if specific color is dominant in the input image.

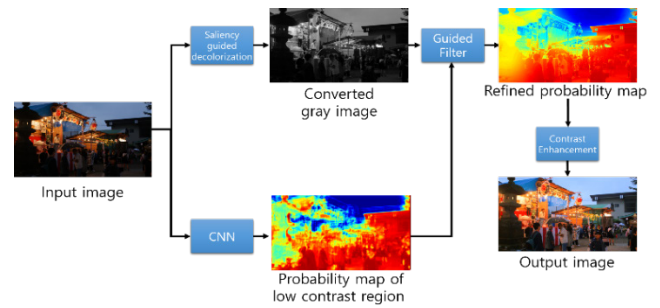


Figure 1. Overall flow of the proposed method.

In recent years, some methods are proposed to enhance low contrast images with adjustment to decomposed illumination image which obtained from various feature models. Different feature models lead to different illumination images and affect the enhanced images. R. Chouhan [9] proposed image enhancement method based on image decomposition technique which follow the discrete wavelet transform (DWT). X. Fu [10] proposed a probabilistic method based on image decomposition. The method decomposed input image into illumination and reflection model then applied the maximum a posteriori (MAP) for estimating enhanced illumination and reflection effectively. Y. ZhenQuiang [11] proposed low-light image enhancement using the camera response model. The camera response model consist of two sub models, camera response function (CRF) and brightness transform function (BTF). Based on these models and estimated exposure ratio map, they enhance image each pixel of the low light image.

The convolutional neural network (CNN) [12] proposed by Y. Lecun has been used widely in computer vision. CNN is being employed on many hard tasks and it showed the best results in many competition in the area. Basically, neural network was proposed for estimating test samples based on training data with hidden layers but calculation with just hidden layer is not suitable for some tasks which related to 2D and 3D data such as images and videos. To preserve locality and flexibility, the Y. Lecun proposed

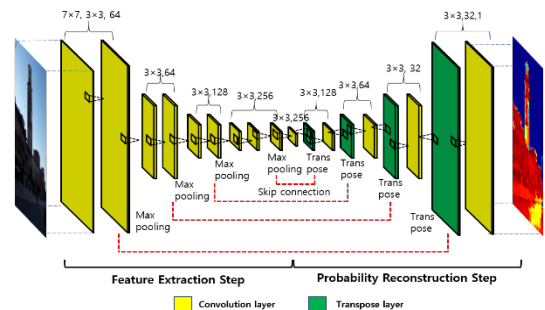


Figure 2. Structure of the convolutional neural network to detect low contrast regions.



Figure 3. Example of training data.

(a) Original image. (b) Labelled image.

basic mechanism of CNN with simple convolutional operation.

One of the most improved tasks based on deep learning in computer vision task is semantic segmentation. Traditional segmentation algorithm was based on mathematical knowledge such as specific feature extraction or the relationship of each pixel. Because of these factors it required in deep research each object which to be segmented. The features that can segment each of object in image are used in restrictive environment. But the methods based on deep learning, It works well without the need to handcraft features of the image based on network characteristics and training data. The most well-known model in semantic segmentation based on deep learning is fully convolutional networks (FCN) proposed by E. Shelhamer [13]. This method shows a basic segmentation network architecture with convolution, pooling and deconvolution. In an input image, important feature information is extracted and compressed by the trained mask through convolution layer and the pooling layer. The compressed information is upsampled in final step with deconvolution layer for matching the size of the input image and the result image. This network not only contains a basic idea for segmentation but also skip-connection mechanism through experimental results. H. Noh [14] proposed a segmentation method with deconvolution technique. The basic model is encoder-decoder network and each lost spatial data by pooling is compensated by deconvolution and unpooling layer.

3. Proposed method

The proposed method consists of four steps in total. The first step is low contrast estimation step that generate low contrast probability map. The second step is a create contrast gray scale step. The third step is a refining probability map step with obtained probability map and converted gray scale image. The final step is an enhancement step that enhance low contrast image with refined probability map and converted gray scale. Figure 1 shows the flow chart of the proposed method.

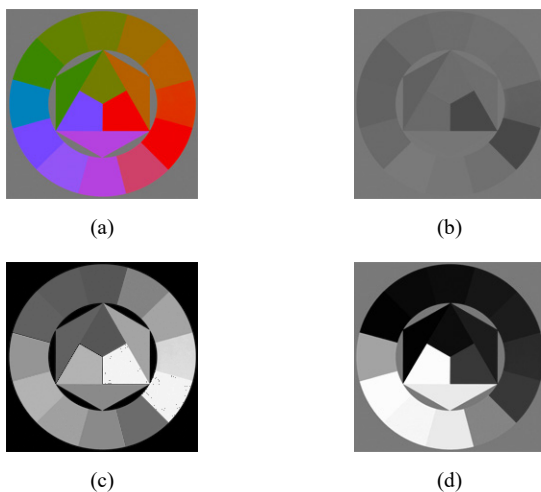


Figure 4. Comparison of gray scale images.

- (a) Input image. (b) Standard gray image.
- (c) Contrast chromatic weight. (d) Contrast gray image.

3.1. Low contrast region probability map

The key idea of the proposed method is to generate probability map of low contrast region. For generation of the probability map quickly and working on small devices such as mobile and tablet PC without computation burden, we construct the network as small

as possible. The Figure 2 shows structure of the proposed method. As shown Figure 2 the network consist of 8 convolution blocks. Inspired by semantic segmentation networks, we design the structure similar to these networks, encoder-decoder networks.

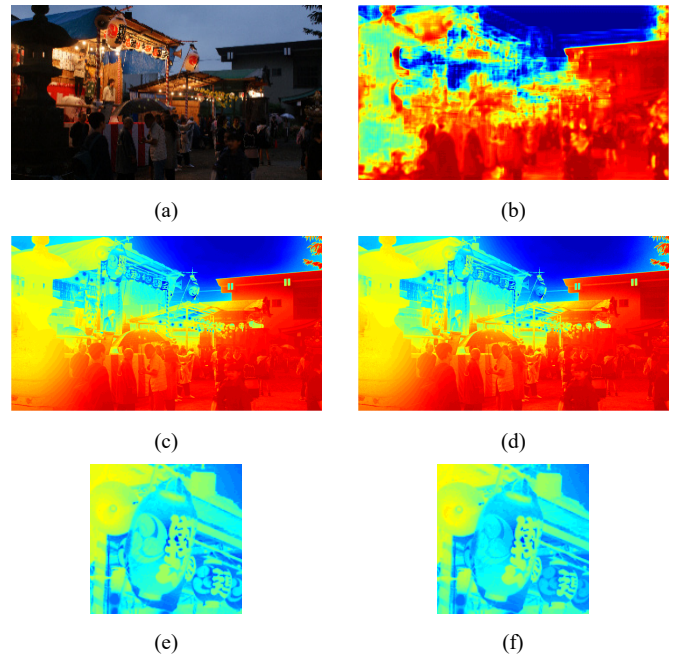


Figure 5. Comparison of refined probability maps.

- (a) Input image. (b) Probability map.
- (c) Refined the (b) with standard gray image.
- (d) Refined the (b) with contrast gray image.
- (e) Part of (c). (f) Part of (d).

From first convolution block to transpose layer, it is called feature extraction step which abstracts low contrast region's feature. All the remaining layers belong to image reconstruction step which rebuild probability map of low contrast regions. To compensate for the lost spatial information by using pooling layer, we use skip connection mechanism and transpose layers. The skip connection was proposed to improve performance of semantic segmentation network that it added the extracted spatial information from the extraction step to the reconstruction step.

We have train the network to generate probability map of low contrast region with low contrast image dataset which collected from on the internet and we used published dataset [15]. We have made ground truth manually and each label consisted of 3 label colors(red, yellow, none). The colors is converted to 1, 0.5 and 0



Figure 6. Result of enhancement step.
(a) Input image. (b) Final result of enhancement step.

for training step. To get enough training data, we have applied data augmentation including rotation (-10, 0, and +10 degree) and mirroring (vertical, horizontal, and diagonal). As result, the size of training data is 6924. Figure 3 shows example of training data.

3.2. Contrast gray image

In the previous works, [1, 16] used Chromatic contrast weights [17] and standard gray scale image for refining the probability map. But only use standard gray scale was not proper because of the image contrast. The standard gray scale image is built with

constant ratio of each channel therefore the image was not be able to contain the contrast of each object's shape. To solve the issue, using the chromatic contrast weight was proposed with handled constant values such as angle of HSV offset and color opponent. The chromatic contrast weight was calculated based on these values, they showed similar values even difference color. To convert an input color image to the gray image appropriately, we use saliency-guided decolorization method [18]. The method based on sparse model and it is able to express contrast with gray value properly. Figure 4 shows gray convert result of standard, chromatic weight and saliency-guided decolorization method.

3.3. Refining outline of probability map

In this step, we refine the probability map with guided filter [19] and converted contrast gray image. The guided filter performs as edge-preserving smooth filter which is well known filter in the computer vision area. The guided filter needs the guided image and filtering image, we chose the probability map as a filtering image and the contrast gray image as guided image. It has performed that the boundary region of the probability map refined similar to contrast gray image. Refined the boundary region and shape prevent halo effect and undesired over-enhancement. Figure 5 shows the result of this step. In (e) and (f) of Figure 5, the refined probability map based on standard gray image shows lost detail shape but (f) shows the detail of tomoe pattern.

3.4. Enhancement

The final step is enhancement step. In this step we follow the reflection model :

$$I(x) = L(x)J(x)+w \quad (1)$$

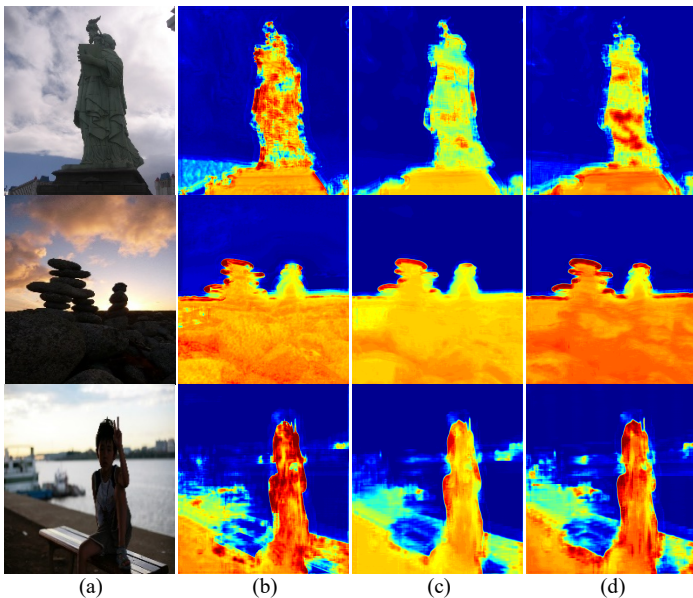


Figure 7. Comparison of the probability maps.
 (a) Input image. (b) Probability map of proposed method.
 (c) Without skip connection. (d) No pooling.

where $I(x)$ is an observed image, $J(x)$ is a desired (enhanced) image, $L(x)$ is illuminance map which is a degraded operator, and w is additive noise. We assume that the image has no noise. So, we can get the enhanced image following (2)

$$J(x) = \frac{I(x)}{F(L(x))+\epsilon} \quad (2)$$

where F is the guided filter and ϵ is small value which is used to avoid dividing by zero. For implementation of (2), we used (2) as

$$\text{Output}_{enhanced}^{ch} = \frac{I^{ch}(x,y)}{(1-\text{Refined map}(x,y))+\epsilon} \quad (3)$$

where I^{ch} is each color channel of the input image, and the refined map is the result computed in the previous step. Figure 6 shows result of the enhancement step.

4. Experiment and results

In the following section we evaluate performance of the proposed method. In Section 4.1, we compare the network performance of the proposed method by comparing two different networks. The performance of the proposed method is then evaluated by comparing the results of the proposed method and the previous methods. The experiments were performed on a PC with a 3.20 GHz Intel Pentium Quad Core Processor and GTX 1080ti graphics card.

4.1. Network structure experimentation

For the comparison of effect of the networks, we made two kind of networks. First network was built without skip-connection and the second network was constructed without pooling layers. Each networks were trained same setting as proposed method such as hyper parameters, size of kernel, and number of channels. The Figure 7 shows the generated probability map of each network. The probability maps in Figure 7, the proposed method can get better shape than the others because of reconstruction power in decoder section. Also, the results of no-skip network and no-pooling networks can be seen quite similar. However, from a detailed point of view, the result of no pooling network shows better shape than no-skip network because of the lost the spatial information by pooling layers. In the previous work [13], they proved the power of skip connection and pooling layer but they used precise labeled data for solving each task. Basically, the neural network is used for estimation of some test data based on training data. Even though the training data we used were not precise, we reconfirmed that the network architecture and the inner method could drive quite successful result.

4.2. Comparison with other methods

In order to evaluate the performance of the proposed method, we compared the performance of the proposed method and the previous methods, CLAHE [3], MSRCR [6], Fu's method [10], and Ying's method [11]. Figure 8 shows the results of each

methods. The first row of Figure 8, the test image was taken from outside which contains the light source and low contrast region such as the counter light region. In the result of CLAHE, the contrast was all most same as input image. Because of the input image contains light source and contrast region together, the size of sub image is sensitive and the distribution of the sub image is controlled by the distribution parameter. The result by MSRCR

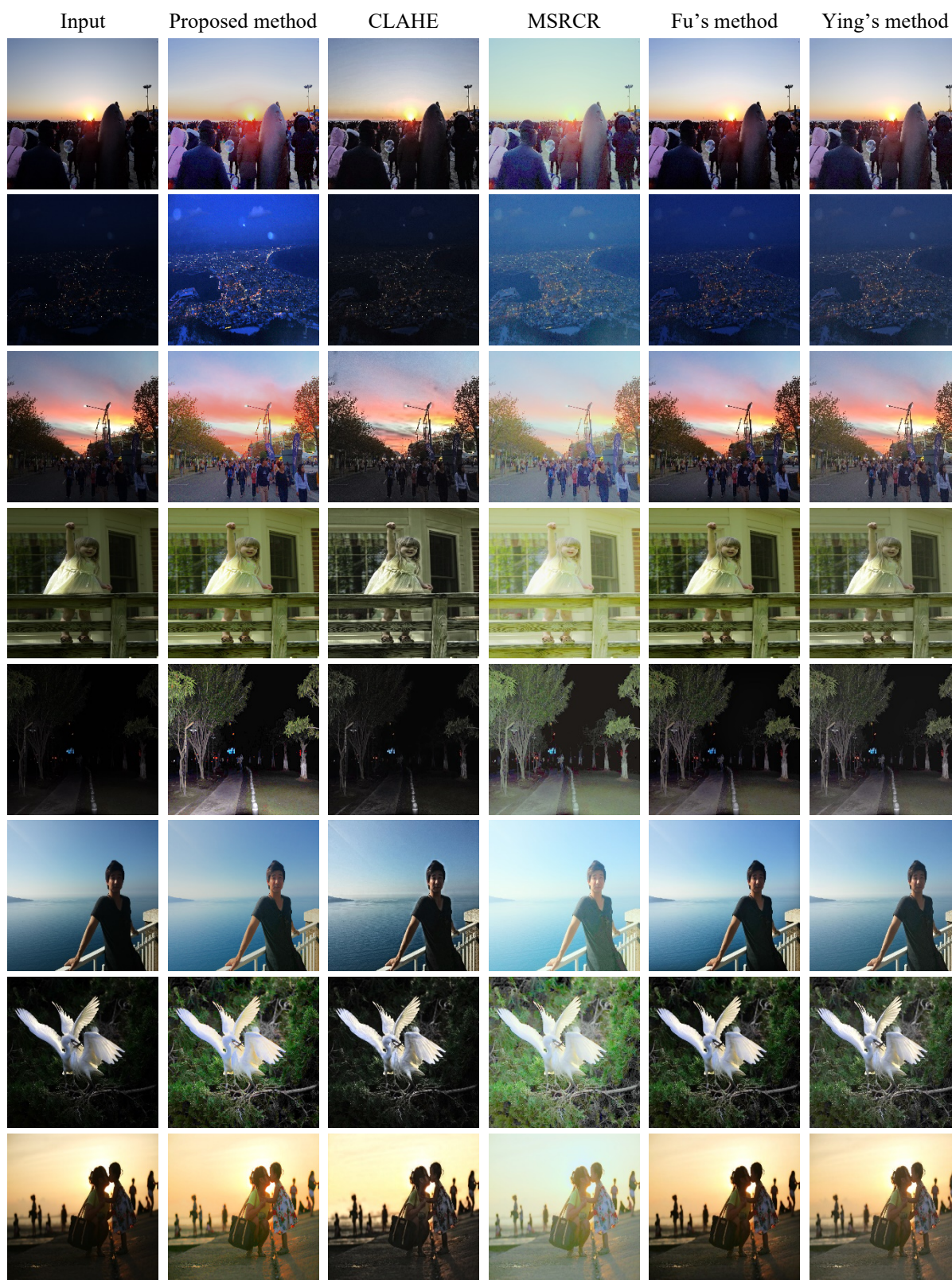


Figure 8. Results of proposed method and previous methods.



(a) (b) (c)
Figure 9. Zoom in the result of Fu's method, Ying's method and proposed method.
(a) Proposed method. (b) Fu's method.
(c) Ying's method.

that not only the contrast is better than CLAHE but also the shape of the statue is seen clearly. But overall brightness is increased too much and the color of the sky looks the same color. The results of Fu's method and Ying's method are improved properly in perspective of contrast and shape. But the overall brightness results are inadequate. In the second row of Figure 8, the results of overall brightness are quite dark. Therefore, the results cannot show the detail of the sky view. For more accurate qualitative comparison, Figure 9 shows zoomed in the results of Fu's method, Ying's method, and proposed method. As you can see images in Figure 9, the shade of the man's face is enhanced properly and we can distinguish the boundary between the nose and the cheek.

5. Conclusion

In this paper, we proposed the image enhancement algorithm with simple reflection model and CNN. In order to enhance low contrast image, plenty of methods have been proposed and result of these methods show good quality from perspective of researcher. But previous methods showed undesired effects such as over- enhancement or unnatural results in unfit circumstance such as illumination setting and image sizes. To solve these problems, CNN was employed to detect low contrast regions in order to preserve the shape of low contrast region and prevent over-enhancement. For the purpose of expression the shape of input images, we employed the saliency-guided decolorization method and the guided filter. We showed our network was suitable for image enhancement through experimental network changes and results, and proved that the results of the proposed method were superior to those of the previous methods through comparison of results with the methods. As a result, the proposed method showed successful enhancement without side effects.

References

- [1] B. K. Han, H. S. Yang, and Y. S. Moon, "Locally Adaptive Contrast Enhancement Using Convolutional Neural Network," IEEE International Conference on Consumer Electronics, Las Vegas, USA, 2018. <https://doi.org/10.1109/ICCE.2018.8326096>
- [2] Y. T. Kim, "Contrast Enhancement Using Brightness Preserving Bi-Histogram Equalization," IEEE Trans., 1997. <http://doi.org/10.1109/30.580378>
- [3] K. Zuiderveld, *Graphics Gems IV*, Academic Press, 1994.
- [4] D. Marini and A. Rizzi, "Computational approach to color adaptation effects," *Image and Vision Computing*, 18(13), pp. 1005-1014. Oct. 2000. [https://doi.org/10.1016/S0262-8856\(00\)00037-8](https://doi.org/10.1016/S0262-8856(00)00037-8)
- [5] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A Multi-Scale Retinex For Bridging the Gap Between Color Images and the Human Observation of Scenes," IEEE Trans. on Image Processing: Special Issue on Color Processing, 6, pp. 965-976, 1997. <http://doi.org/10.1109/83.597272>
- [6] Z. Rahman, G.A. Woodell, and D.J. Jobson, "Multiscale Retinex for Color Image Enhancement," *Proceedings of 3rd IEEE International Conference on Image Processing*, Lausanne, Switzerland, 1996.

- [7] E. Provenzi, M. Fierro, A. Rizzi, L. De Carli, D. Gadia, and D. Marini, "Random spray retinex: A new retinex implementation to investigate the local properties of the model," *IEEE Transactions on Image Processing*, 16(1), pp. 162-171, Jan. 2007. <https://doi.org/10.1109/TIP.2006.884946>
- [8] D.J. Jobson, Z. Rahman, and G.A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Transactions on Image Processing*, 6(3), Mar., 1997. <https://doi.org/10.1109/83.557356>
- [9] R. Chouhan, C. Pradeep Kumar, R. Kumar, and R. K. Jha, "Contrast Enhancement of Dark Images using Stochastic Resonance in Wavelet Domain," *International Journal of Machine Learning and Computing*, 2(5), 2012. <http://doi.org/10.7763/IJMLC.2012.V2.220>
- [10] X. Fu, Y. Liao, D. Zeng, Y. Huang, X. Zhang, and X. Ding, "A Probabilistic Method for Image Enhancement With Simultaneous Illumination and Reflectance Estimation," in *IEEE Transactions on Image Processing*, 24(12), pp. 4965-4977, Dec., 2015. <https://doi.org/10.1109/TIP.2015.2474701>
- [11] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A New Low-Light Image Enhancement Algorithm Using Camera Response Model," 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, pp. 3015-3022, 2017. <http://doi.org/10.1109/ICCVW.2017.356>
- [12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 86(11), pp. 2278-2324, Nov. 1998. <http://doi.org/10.1109/5.726791>
- [13] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), pp. 640-651, 1 April 2017. <http://10.1109/TPAMI.2016.2572683>
- [14] H. Noh, S. Hong and, B. Han, "Learning Deconvolution Network for Semantic Segmentation," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, pp. 1520-1528, 2015. <http://doi.org/10.1109/ICCV.2015.178>
- [15] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the Exclusively Dark dataset," *Computer Vision and Image Understanding*, 178, pp 30-42, 2019. <https://doi.org/10.1016/j.cviu.2018.10.010>
- [16] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and John Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal Processing*, 129, 2016. <https://doi.org/10.1016/j.sigpro.2016.05.031>
- [17] C. O. Ancuti, C. Ancuti, and P. Bekaert, "Enhancing by saliency-guided decolorization," *CVPR 2011*, Colorado Springs, CO, USA, pp. 257-264, 2011. <http://doi.org/10.1109/CVPR.2011.5995414>
- [18] C. Liu and T. Liu, "A sparse linear model for saliency-guided decolorization," 2013 IEEE International Conference on Image Processing, Melbourne, VIC, pp. 1105-1109, 2013. <http://doi.org/10.1109/ICIP.2013.6738228>
- [19] K. He, J. Sun, and X. Tang, "Guided Image Filtering," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6), pp. 1397-1409, June 2013. <http://doi.org/10.1109/TPAMI.2012.213>

Photodecoloration of Methyl Orange Solution Assisted by ZrS₃ Powders

Sofya Artemkina*^{1,2}, Anastassiia Poltarak^{1,2}, Pavel Poltarak^{1,2}, Igor Asanov¹, Vladimir Fedorov^{1,2}

¹*Nikolaev Institute of Inorganic Chemistry SB RAS, Novosibirsk 630090, Russian Federation*

²*Novosibirsk State University, Department of Natural Sciences, Novosibirsk 630090, Russian Federation*

ARTICLE INFO

Article history:

Received: 20 November, 2018

Accepted: 01 February, 2019

Online : 07 February, 2019

Keywords:

Methyl orange

Decoloration

Zirconium trisulfide

Synthesis

Morphology

Catalysis

ABSTRACT

Zirconium trisulfide ZrS₃ is a representative of transition metal polysulfides containing sulfur as S⁻¹ in polysulfide, usually disulfide S₂²⁻ groups. Semiconductive zirconium trisulfide which absorbs visible light near UV edge was considered as a possible photocatalyst. We experimentally studied photodecoloration of methyl orange in presence of ZrS₃. It was shown for the first time that crystalline ZrS₃ strongly deepens photodegradation reaction, and in one case the methylene orange conversion reached almost 100%. The rates of degradation curves were associated with the ZrS₃ samples morphology; the best result revealed for microribbons ZrS₃ synthesized at 650°C.

1. Introduction

Transition metal sulfides are well-known class of inorganic compounds. The very popular representatives CdS, MoS₂ and some others are considered as materials for luminescence [1], photocatalysis [2], piezoelectric materials [3], anodic components for lithium batteries [4], etc. [5].

The distinctive feature of transition metal polysulfides (TMP) is a polysulfide group which is usually a disulfide group (S₂)²⁻ coordinated to metal atoms. For instance, crystalline VS₄ [6], NbS₃ [7], TiS₃ [8], as well as amorphous MoS₃ [9], MoS₄ and WS₅ [10] include disulfide (S₂)²⁻ groups along with sulfide anions S²⁻ coordinated to metal atoms. Like layered transition metal disulfides, TMP also possess low-dimensional structures (layered – crystalline trisulfides, chained – crystalline VS₄, amorphous CrS₃, MoS₃, etc. – presumably chained).

Zirconium trisulfide ZrS₃ is a typical representative of layered transition metal trisulfides. The crystal structure of ZrS₃ (attributed to ZrSe₃ structural type [8]) (Fig. 1) comprise double layers whose surfaces are produced by disulfide groups (S₂)²⁻. Particularly, a basic structural fragment prism {ZrS_{6/2}}, is constructed from two disulfide (S–S)²⁻ groups as well as two sulfide groups S²⁻, the Zr atom is situated near the center of the prism (Fig. 1). Such prisms

{ZrS_{6/2}} with metal atoms situated close to the prism centers are connected to each other via common triangle bases to form infinite columns which are oriented along *b* axis. Additionally, these columns are connected to the neighboring ones forming layers two prisms thick. The infinite layers are bind to the neighboring *via* van der Waals S...S contacts. Hence the layered zirconium trisulfide may be described by ionic model Zr⁴⁺(S₂)²⁻S²⁻.

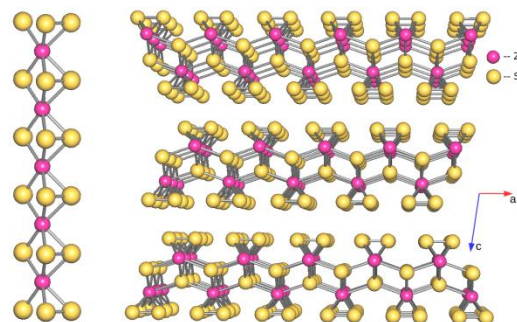


Figure 1: Structure of the wedge-shaped columns (left) within one polymeric layer and arrangement of the layers (right) in the crystal structure of triclinic ZrS₃

Presence of the disulfide group introduces new chemical properties which could be useful in known physico-chemical processes required in modern industry. Sulfur in the disulfide group is charged as –1, so it can be oxidized or reduced [11] in 1-electron ox-red processes, so these systems can be considered as

*Artemkina Sofya, NIIC SB RAS, 3, Akad. Lavrentiev Ave., 630090 Novosibirsk, Russia, +7 383 330 92 53, artem@niic.nsc.ru

electron reservoirs [12]. This property can meet an application in high capacity alkaline batteries [4], [13]. In the ionic model approximation, trisulfides MS_3 ($M = Ti, Zr, Hf$) are d^0 -complexes, a diamagnetic semiconductors. The semiconductor band gaps of the trisulfides MS_3 ($M = Ti, Zr, Hf$) were optically defined (1.8, 2.0, and 2.2 eV, respectively), and were recently examined as cathodes in hydrogen evaluation reaction [14]. In addition, ZrS_3 nanomaterial was used as efficient anode for catalytic oxygen evaluation reaction in both alkaline and neutral aqueous solutions [15].

Photocatalytic degradation of organic molecules is an actual theme now, and it becomes more popular once renewable energy source is used (solar irradiation); the nature of substrates is dictated by polluting agents in natural waters and soils. Semiconductor photocatalysis is an important option for solar energy conversion and pollutant degradation according to the idea of green chemistry. For purposes of degradation of the pollutants, a known set of catalysts is already designed [16]: double layered hydroxides [17], titanium dioxide [18], composites based on TiO_2 [19], other oxides (ZnO [20], WO_3 [21]), composites with oxides [22-24, 25, 26], as well as composites with MoS_2 [27-29].

Many transition metal di- and polysulfides are dark or metal black in color. A pair of "colored exceptions" can be distinguished here, ZrS_3 and HfS_3 , which are orange and dark red, respectively. According to study [30], these compounds absorb visible light in near UV region, so we decided to examine ZrS_3 compound as photocatalyst in a model reaction of methyl orange degradation. Shifting from transition metal disulfides (MoS_2 , WS_2) to transition metal trisulfides as photocatalyst is something new in the present work.

2. Experimental

2.1. Materials

Zirconium powder 99% and crystalline sulfur 99.9% (Acros), were used for synthesis of ZrS_3 samples; TiO_2 Degussa P25 (Degussa AG, Germany), methyl orange (Reakhim, Russia) and distilled water were used for carrying out photodecolorization experiments.

2.2. Syntheses of zirconium trisulfide at different temperatures

ZrS_3 was obtained by heating a stoichiometric mixture of Zr and S. Mixtures of powder zirconium (4.868 g 0.05336 mol) and crystalline sulfur (5.132 g 0.1601 mol) were evacuated in three quartz ampoules (about 20 ml volume) and sealed. The syntheses were carried out at temperatures 350, 500, and 650°C for 150 hours (up to these temperatures, the ampoules were heated for 15 h). The ampoules were cooled down with the furnace. The products were thin powders different shades of red-orange. The products were heated in dynamic vacuum at 200°C during 1 h in order to remove elementary sulfur after syntheses.

2.3. Methods and apparatus

X-ray powder diffraction patterns for solids including exfoliated samples were collected with a Philips PW 1830/1710 automated diffractometer (Cu K_α radiation, graphite monochromator, silicon plate as an external standard). Raman spectra were recorded with a LabRam HR Revolution (Horiba www.astesj.com

Scientific) instrument at wavelength 488nm. UV spectra were recorded with an Agilent Cary 60 Spectrophotometer in the range of 200 – 800 nm. Eppendorf Centrifuge 5430 equipped with container for 15 and 50 ml tubes used for centrifugation of the reactive mixtures. Scanning electron microscopy (SEM) images were collected with Hitachi S3400N instrument.

X-ray photoelectron spectroscopy (XPS) measurements were carried out on a Specs Phoibos-150 spectrometer with an Al K_α monochromatic excitation. The pass energy of an electron analyzer was set at 20 eV. For the compensation of a charging effect from non-conductive samples, a low energy electron beam was applied. Binding energies were measured from the C1s level (285.0 eV) of surface hydrocarbon contaminations. Relative atomic concentrations are calculated from the measured areas of spectra taking into account photoionization cross-sections, inelastic mean free paths and the transmission function of the spectrometer.

The Brunauer–Emmett–Teller (BET) surface areas were determined by N_2 adsorption 5-point measurements in pressure range $p/p_0 = 0.05-0.25$ using SORBTOMETER-M surface area analyzer. The ZrS_3 powder samples were preliminary heated at 200°C during 1 hour in nitrogen flow.

2.4. Photodecoloration experiment

The photocatalytic activity of ZrS_3 powders was evaluated by standard decoloration reaction of dye methyl orange (MO). The reactor for the decoloration experiments consisted of a cylindrical beaker on a magnetic stirrer, with a common LH9-U/BLB/G23 black light lamp (Camelion). Experimental conditions for methyl orange (MO) decoloration on powders were: aqueous solution of MO with initial concentration 0.022 mM, amount of ZrS_3 powder 0.175 g/L (14 mg per each experiment), and UV irradiation time was kept constant for 2-3 hours. The aqueous suspensions were stirred throughout the experiment. In the experiment, to 80 ml of aqueous solution of MO ($C = 0.022$ mM) 14 mg of ZrS_3 powder was added, and mixed during 3 hours excluding irradiation of any light (for dark sorption). At the stage of dark sorption several points were documented in the following way: 1 ml of suspension was collected from the solution with Pasteur pipette, centrifuged with acceleration 3000 g for 10 min for removal of ZrS_3 solid, the transparent solution decanted, and the UV-Vis spectrum in region 200-800 nm written for the solution. After 3 hours of dark sorption a lamp was turned on above the tested mixture and every 15 min points were examined in the same way. Three series of experiments were carried out, each series three times: using powder samples of ZrS_3 synthesized at 350 (ZrS_3 -350), 500 (ZrS_3 -500), and 650°C (ZrS_3 -650). Graphics of MO absorbance vs time of irradiation were built for comparison.

2.5. Photodecoloration experiment with Degussa P25

The experiment with Degussa P25[®] was carried out in similar way, using the same experimental setup. The portion of powder Degussa P25 was 14 mg per every degradation experiment.

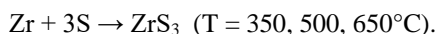
After the experiments the photocatalysts were washed with 20 ml of water, centrifuged and decanted, and then washed with 20 ml of ethanol. After these operations the solids were dried in

air. They were examined by XRD and Raman spectroscopy, as well as XPS for ZrS₃-650.

3. Results and discussion

3.1. Initial powder ZrS₃ samples

Zirconium trisulfide samples were synthesized from the elements in evacuated quartz ampules at different temperatures according to reaction which is rather common for synthesis of transition metal trichalcogenides [31, 32]:



The yields were quantitative or near-quantitative. Preliminary EDS tests revealed a quantity of elementary sulfur in the products, so we heated the ZrS₃ samples in dynamic vacuum in order to remove elementary sulfur. The products were thin powders, red-orange in color (red for ZrS₃-350, yellow-orange for ZrS₃-500 and ZrS₃-650). Powder diffraction patterns (XRD) revealed only monoclinic ZrS₃ phase (ICSD 42-073) as crystalline product of reaction in the cases of temperatures 500 and 650°C (Fig. 2). Positions of some observable reflections were as follows: 2θ 9.97° (001), 20.01° (002), 35.45° (200), 40.55° (004), 43.59° (211), 50.42° (020), 62.99° (006). Small quantity of unreacted metal zirconium was found in ZrS₃-350 (2θ 36.6°). Coherent scattering domain size (CSDS) calculated by Scherrer formula [33] (applied to 001 reflection) from the powder patterns gave 14, 29, and 77 Å for ZrS₃-350, ZrS₃-500, and ZrS₃-650, respectively. So, we noticed here that rise in temperature of synthesis enhances the crystalline ordering of the samples ZrS₃. The Raman spectra of the ZrS₃ samples were very close to each other, and the vibration bands A_g 148 cm⁻¹ (150 cm⁻¹ from [34]), 276 (280), 317 (322), 525 (530) cm⁻¹ were agree with the literature data, the 525 (530) cm⁻¹ band testifies presence of (S-S)²⁻ groups in the samples.

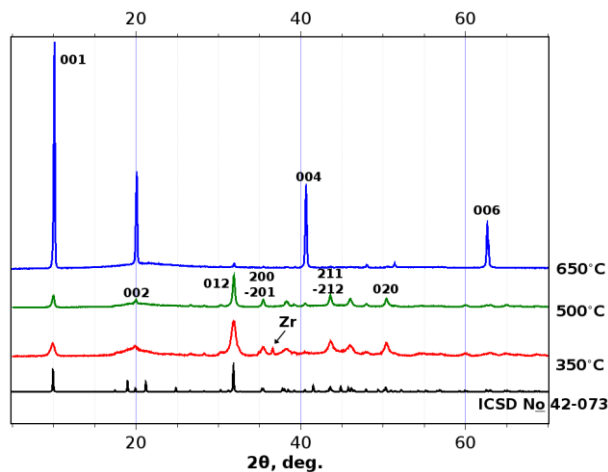
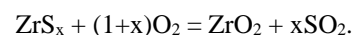


Figure 2: X-ray powder diffraction patterns for crystalline samples ZrS₃-350 (red line), ZrS₃-500 (green line), and ZrS₃-650°C (blue line), and theory calculated XRD pattern for ZrS₃ (black line)

EDS data for the prepared ZrS₃ samples showed rather low content of sulfur (Zr : S ratios were from 2.2 to 2.4), so we considered that the materials demonstrated mainly surface stoichiometry which could include some deal of oxides also. These data contradicted with the XRD data where clear crystalline monoclinic ZrS₃ phase occurred. In order to define the exact Zr : S ratios in these samples, we carried out gravimetric experiments:

www.astesj.com

carefully weighed ZrS₃ samples were put into pre-weighed crucibles and heated in a muffle furnace at 600°C during 20 hours. The products of calcination in the crucibles were weighed, and the quantity of sulfur was calculated using mass difference according to the equation:



After weighing the calcination products, they were examined by X-ray powder diffraction in order to make sure that ZrO₂ (gravimetric form) were the only calcination product. It was found that the bulk material ZrS₃-350 was ZrS_{2.56}, ZrS₃-500 was ZrS_{2.76}, and ZrS₃-650 was ZrS_{2.76}. So we concluded from this data that the ZrS₃ samples synthesized from stoichiometric mixtures of Zr and S are monoclinic ZrS₃ albeit possess remarkable deficiency in sulfur.

Probably our ZrS₃ samples have such a sulfur deficiency due to synthetic procedure: only 3 equivalents of S were taken for reaction per one equivalent of Zr. Crystals of zirconium trisulfide with formulae very close to ZrS₃ are synthesized in gas-transport reactions with some excess of sulfur [35], longer heating, or the stoichiometric mixture which was heated at higher temperature (1000°C) [35], or applied transport species (for example, bromine) [36].

SEM images of the powder ZrS₃ samples revealed different morphologies (Fig. 3). ZrS₃-350 sample is aggregates of lamellar slabs about 5µm size; ZrS₃-500 sample is a mixture of particles of two types: aggregates of lamellar slabs ca. 5µm size and fibers up to 20 µm long; ZrS₃-650 sample is ribbons 0.5-1 µm width and ca. 20µm long. Visible texturing is observed for ZrS₃-650 sample (Fig. 2): apparently, the preferred orientation of the ribbons is layering in parallel to their wide side.

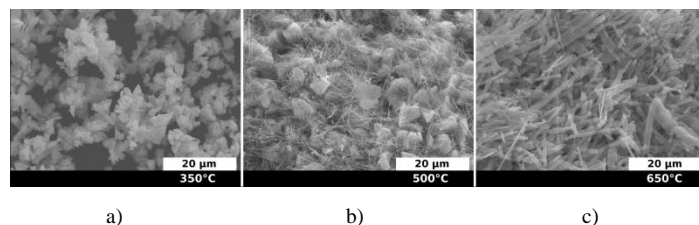


Figure 3: SEM images of powder samples synthesized at ZrS₃-350 (a), ZrS₃-500 (b), and ZrS₃-650 (c)

Strong dependence of the BET surface areas on the ZrS₃ morphology (min BET 9 m²/g for ZrS₃-650, Table 1) indicates that the lamellar slab aggregates make the main contribution to the surface area: BET of the samples ZrS₃-350 and ZrS₃-500 having components of such morphology, are noticeably higher (49 and 24 m²/g, respectively). Degree of dark sorption correlated with the surface area values: the percentage of absorbed MO on ZrS₃ powders increased while increasing their BET in series ZrS₃-650 < ZrS₃-500 < ZrS₃-350. The minimal dark adsorption of ZrS₃-650 is in agreement with the maximal CSDS value, 77 Å, and the minimal BET surface.

3.2. Photodecoloration of MO in presence of ZrS₃ powders

ZrS₃ powders were used as photocatalysts for decomposing of dye molecule MO for the first time (Fig. 4). ZrS₃ powders were used as synthesized. Photocatalytic activity of ZrS₃ was evaluated

in terms of the degradation of MO aqueous solution under UV irradiation. The efficiency of the dye degradation was calculated based on the changes in absorption of MO in UV-Vis spectra. At the lamp irradiation, MO didn't decolorate in absence of ZrS₃.

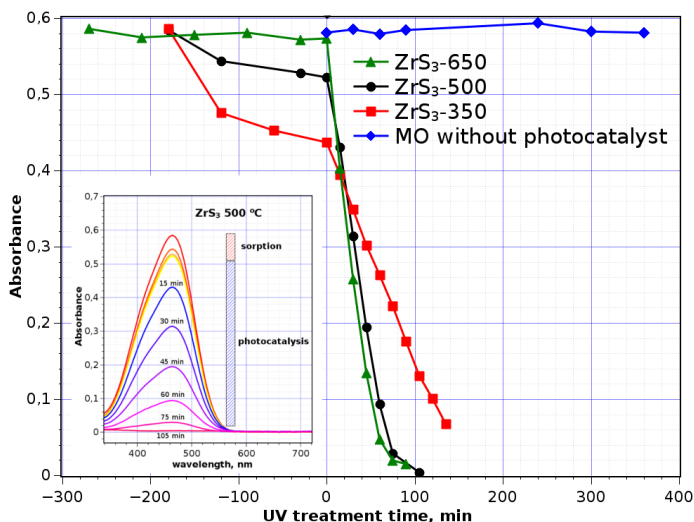


Figure 4: Dependence of MO absorbance in the presence of ZrS₃ vs time of UV irradiation exposure. Systematic errors of measurement are from minimum three experiments for each ZrS₃ sample

From the UV-Vis spectroscopy data (Fig. 4, inset) one can see that MO absorbs on ZrS₃ particles in the cases ZrS₃-350 and ZrS₃-500 in the dark phase whereas presence of ZrS₃-650 decreases MO concentration to a lesser degree. After start of the irradiation, the concentration of MO in the reaction mixture was decreased noticeably, that indicated start of MO photodegradation.

The decoloration dependencies vs time of irradiation are averaged from at least three experiments, and the standard deviations of the data points not exceed 5% from absorbance values at time starting UV treatment (t = 0 min).

Table 1: Some parameters of the decoloration process in presence of different ZrS₃ crystalline samples

Crystalline ZrS ₃ sample	ZrS ₃ -350°C	ZrS ₃ -500°C	ZrS ₃ -650°C
Coherent scattering domain size (CSDS), Å	14	29	77
BET surface area, m ² /g	49	24	9
Adsorption of MO in "dark" time, %	26	11	2.2
Time of irradiation at 50% degradation of MO (0.022mM), min	48	36	27
Time of irradiation at 90% degradation of MO (0.022mM), min	144	71	59
Time of irradiation at 50% degradation of MO (0.022mM) in presence of TiO ₂ Degussa P25, min	138		

At the initial concentration of MO in aqueous solution 0.022 M, photodegradation time was confined by two hours. The longest MO degradation process concerns to ZrS₃-350 sample, 140 min with 90% of the initial MO amount. The photodegradation dependences have linear character at least for the first 60 min, for ZrS₃-500 and ZrS₃-650 they are very close while the line for ZrS₃-350 has gentler slope, and doesn't demonstrate complete degradation of MO in 140 min of the experiment.

One can see ZrS₃-650 is the most efficient sample from the three. The reason of this may be the minimal dark sorption, so the sample surface is better accessible for dissolved oxygen; ZrS₃-650 also possesses larger size of coherent scattering domain (77Å) that allows hoping to have more regular surface of the particles. The MO degradation results obtained are comparable with MO degradation results in the presence of MoS₂ of two morphologies: nanorods and nanoflowers [37].

3.3. Characterization of ZrS₃ powders after the photodecoloration experiments

After the photodegradation experiments, ZrS₃ powder samples retained their structural identity (XRD for ZrS₃-650 is presented in Fig. 5). None of crystalline ZrO₂ phases were detected in the XRD patterns. Raman spectra of ZrS₃ after the decoloration experiments revealed ZrS₃ spectra which were very similar to each other: 105, 119, 146, 239, 275, 316, 357, 523 cm⁻¹. These bands are in good accordance with the literature data: 107, 120, 148, 241, 277, 317, 355, 525 cm⁻¹ [38]. We didn't find in the Raman spectra band respected to amorphous ZrO₂ (148, 263, 476 cm⁻¹) as in [39].

The analysis of XPS data before and after the photocatalytic experiments for ZrS₃-650 as critical sample, revealed certain differences. In the initial powder sample ZrS₃-650 (Fig. 6), the atomic concentration ratio S/Zr was 2.78. S2p spectrum contained two doublets from S₂²⁻ and S²⁻ with positions of S2p_{3/2} components at 162.6 and 161.6 eV, respectively; the ratio of their intensities was 1.78. The values of bond energies were well agreed with the literature data [40]. Spectrum Zr3d_{5/2, 3/2} also contains two doublets with spin-orbital splitting 2.4 eV. The Zr3d_{5/2} components were situated at 181.2 and 182.9 eV with intensity ratio 9:1. The first component belonged to zirconium in ZrS₃ while the second might be associated with an oxide phase on surface of the ZrS₃ particles [41]. The composition of the surface sulfide phase may be thereby formulated as Zr_{0.9}(S₂)_{0.89}S.

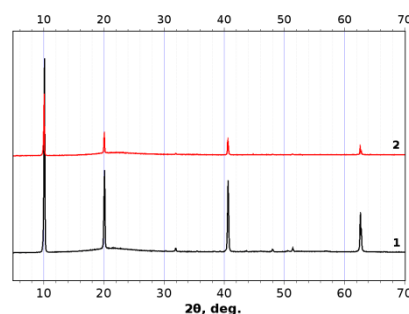


Figure 5: XRD patterns of ZrS₃-650 before (line 1) and after (line 2) photodegradation MO experiment

In the sample ZrS₃-650 after the photodegradation experiment, weak peaks in the region 168.0 eV appeared in S2p_{3/2} spectrum that corresponds to sulfate groups on the surface. Component

corresponding to an oxide phase, strongly arose in the $Zr3d_{5/2, 3/2}$ spectrum. These data indicate the oxidation of the ZrS_3 particles surface after photocatalysis experiment.

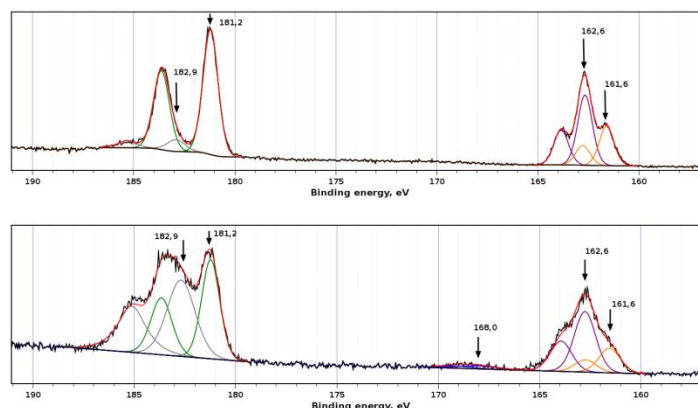


Figure 6: XPS data for ZrS_3 -650 powder sample, as-synthesized (upper graphic) and after photocatalysis experiment (lower graphic)

3.4. Comparison with Degussa P25

We carried out photocatalytic experiment with Degussa P25 using the same experimental setup in order to compare photocatalytic ability of ZrS_3 with the international reference TiO_2 (TiO_2 , 80% anatase and 20% rutile). 14 mg of Degussa P25 in 80 ml of 0.022 mM MO solution was used for the experiment. As a result of experiment in our setup, MO decolorated in presence of Degussa P25 at 50% in 138 min. Although kinetics of MO decoloration was faster for ZrS_3 -650 than for Degussa P25 here, we discussed here only the first cycle of decoloration. Additional studies are required to test ZrS_3 powders in series of decoloration cycles, and examination of the catalysts after the experiments.

The preliminary experiment on photodegradation of MO in aqueous solution in presence of powder ZrS_3 opens way to design of new light-sensitive materials for photocatalysis and other applications. Some additional studies are to carry out in order to realize the mechanism and other characteristics of the photodegradation observed. Authors believe that photosensitive properties of ZrS_3 (and probably some another transition metal trisulfides) may be enhanced by different ways (crystal or surface modification, assembly in composites) to obtain new materials for photocatalysis.

4. Conclusion

Zirconium trisulfide revealed appreciable photocatalytic ability in dye molecules decoloration, on the example of methyl orange decoloration for the first time. The surface of the photocatalyst increases the content of oxide phase after the degradation reaction. It was ascertained that the photocatalytic behavior of ZrS_3 is dependent on its morphology, which in turn is dependent on temperature of synthesis of ZrS_3 . Based on qualitative level results, ZrS_3 -containing species are considered as perspective materials for photocatalysis.

Conflict of Interest

The authors declare no conflict of interest.

www.astesj.com

Acknowledgment

This work was financially supported by Russian Science Foundation (grant RSCF 13-14-00674). Authors are grateful to Dr. Eugeny Maksimovsky for collecting of SEM images and Dr. Konstantin Kovalenko for BET analysis.

References

- [1] P.F. Smet, I. Moreels, Z. Hens, and D. Poelman, "Luminescence in Sulfides: A Rich History and a Bright Future" *Materials*, 3 (4), 2834, 2010. <https://doi.org/10.3390/ma3042834>
- [2] J. Mao, Y. Wang, Z. Zheng, and D. Deng, "The rise of two-dimensional MoS_2 for catalysis" *Frontiers of Physics*, 13 (4), 138118, 2018. <https://doi.org/10.1007/s11467-018-0812-0>
- [3] K. Jinzong, L. Yudong, Z. Yaxing, and Z. Junyi, "Progress in piezotronics of transition-metal dichalcogenides" *Journal of Physics D: Applied Physics*, 51 (49), 493002, 2018. <https://doi.org/10.1088/1361-6463/aadb15>
- [4] E.D. Grayfer, E.M. Pazhetnov, M.N. Kozlova, S.B. Artemkina, and V.E. Fedorov, "Anionic redox chemistry in polysulfide electrode materials for rechargeable batteries" *ChemSusChem*, 10 (24), 4805–4811, 2017. <https://doi.org/10.1002/cssc.201701709>
- [5] C. Yan, C. Gong, P. Wangyang, J. Chu, K. Hu, C. Li, X. Wang, X. Du, T. Zhai, Y. Li, and J. Xiong, "2D Group IVB Transition Metal Dichalcogenides" *Advanced Functional Materials*, 28 (39), 1803305, 2018. <https://doi.org/10.1002/adfm.201803305>
- [6] M.N. Kozlova, Y.V. Mironov, E.D. Grayfer, A.I. Smolentsev, V.I. Zaikovskii, N.A. Nebogatikova, T.Y. Podlipskaya, and V.E. Fedorov, "Synthesis, Crystal Structure, and Colloidal Dispersions of Vanadium Tetrasulfide (VS_4)" *Chemistry – A European Journal*, 21 (12), 4639-4645, 2015. <https://doi.org/10.1002/chem.201406428>
- [7] J. Rijnsdorp, and F. Jellinek, "The crystal structure of niobium trisulfide, NbS_3 " *Journal of Solid State Chemistry*, 25 325-328, 1978.
- [8] S. Furuseth, L. Brattas, and A. Kjekshus, "On the crystal structures of TiS_3 , ZrS_3 , $ZrSe_3$, $ZrTe_3$, HfS_3 , and $HfSe_3$ " *Acta Chemica Scandinavica A*, 29 623-631, 1975.
- [9] S.J. Hibble, and G.B. Wood, "Modeling the Structure of Amorphous MoS_3 : A Neutron Diffraction and Reverse Monte Carlo Study" *Journal of the American Chemical Society*, 126 (3), 959-965, 2004. <https://doi.org/10.1021/ja037666o>
- [10] D.A. Rice, S.J. Hibble, M.J. Almond, K.A.H. Mohammad, and S.P. Pearce, "Novel low-temperature route to known (MnS and FeS_2) and new (CrS_3 , MoS_3 and WS_3) transition-metal sulfides" *Journal of Materials Chemistry*, 2 (8), 895-896, 1992. <https://doi.org/10.1039/JM9920200895>
- [11] M.N. Kozlova, E.D. Grayfer, P.A. Poltarak, S.B. Artemkina, A.G. Cherkov, L.S. Kibis, A.I. Boronin, and V.E. Fedorov, "Oxidizing Properties of the Polysulfide Surfaces of Patronite VS_4 and NbS_3 Induced by (S2)2- Groups: Unusual Formation of Ag_2S Nanoparticles" *Advanced Materials Interfaces*, 1700999, 2017. DOI: 10.1002/admi.201700999
- [12] J. Rouxel, "Anion–Cation Redox Competition and the Formation of New Compounds in Highly Covalent Systems" *Chemistry – A European Journal*, 2 (9), 1053-1059, 1996. <https://doi.org/10.1002/chem.19960020904>
- [13] V.V.T. Doan-Nguyen, K.S. Subrahmanyam, M.M. Butala, J.A. Gerbec, S.M. Islam, K.N. Kanipe, C.E. Wilson, M. Balasubramanian, K.M. Wiaderek, O.J. Borkiewicz, K.W. Chapman, P.J. Chupas, M. Moskovits, B.S. Dunn, M.G. Kanatzidis, and R. Seshadri, "Molybdenum Polysulfide Chalcogenides as High-Capacity, Anion-Redox-Driven Electrode Materials for Li-Ion Batteries" *Chemistry of Materials*, 28 (22), 8357-8365, 2016. <https://doi.org/10.1021/acs.chemmater.6b03656>
- [14] E. Flores, J.R. Ares, I.J. Ferrer, and C. Sanchez, "Synthesis and characterization of a family of layered trichalcogenides for assisted hydrogen photogeneration" *Physica Status Solidi*, 10 (11), 802-806, 2016. <https://doi.org/10.1002/pssr.20160016>
- [15] J. Xie, R. Wang, J. Bao, X. Zhang, H. Zhang, S. Li, and Y. Xie, "Zirconium Trisulfide Ultrathin Nanosheets as Efficient Catalyst for Water Oxidation in Both Alkaline and Neutral Solutions" *Inorganic Chemistry Frontiers*, 1 751-756, 2014. <https://doi.org/10.1039/C4QI00127C>
- [16] C. Byrne, G. Subramanian, and S.C. Pillai, "Recent advances in photocatalysis for environmental applications" *Journal of Environmental Chemical Engineering*, 6 (3), 3531-3555, 2018. <https://doi.org/10.1016/j.jece.2017.07.080>

- [17] G. Starukh, "Photocatalytically enhanced cationic dye removal with Zn-Al layered double hydroxides" *Nanoscale Research Letters*, 12 (1), 391, 2017.
- [18] C. Chen, W. Ma, and J. Zhao, "Semiconductor-mediated photodegradation of pollutants under visible-light irradiation" *Chemical Society Reviews*, 39 (11), 4206-4219, 2010. https://doi.org/**
- [19] L. Zhang, Y. He, Y. Wu, and T. Wu, "Photocatalytic degradation of RhB over MgFe₂O₄/TiO₂ composite materials" *Materials Science and Engineering: B*, 176 (18), 1497-1504, 2011. <https://doi.org/10.1016/j.mseb.2011.09.022>
- [20] R. Kumar, G. Kumar, and A. Umar, "Zinc oxide nanomaterials for photocatalytic degradation of methyl orange: a review" *Nanoscience and Nanotechnology Letters*, 6 (8), 631-650, 2014. <https://doi.org/10.1166/nnl.2014.1879>
- [21] P. Dong, G. Hou, X. Xi, R. Shao, and F. Dong, "WO₃-based photocatalysts: morphology control, activity enhancement and multifunctional applications" *Environmental Science: Nano*, 4 (3), 539-557, 2017. <https://doi.org/10.1039/C6EN00478D>
- [22] X. Yang, H. Fu, K. Wong, X. Jiang, and A. Yu, "Hybrid Ag@TiO₂ core-shell nanostructures with highly enhanced photocatalytic performance" *Nanotechnology*, 24 (41), 415601, 2013. <https://doi.org/10.1088/0957-4484/24/41/415601>
- [23] J. Luo, X. Zhou, L. Ma, and X. Xu, "Enhanced visible-light-driven photocatalytic activity of WO₃/BiOI heterojunction photocatalysts" *Journal of Molecular Catalysis A: Chemical*, 410 168-176, 2015. <https://doi.org/10.1016/j.molcata.2015.09.019>
- [24] S. Chen, Y. Hu, X. Jiang, S. Meng, and X. Fu, "Fabrication and characterization of novel Z-scheme photocatalyst WO₃/g-C₃N₄ with high efficient visible light photocatalytic activity" *Materials Chemistry and Physics*, 149-150 512-521, 2015. <https://doi.org/10.1016/j.matchemphys.2014.11.001>
- [25] S. Bai, K. Zhang, J. Sun, R. Luo, D. Li, and A. Chen, "Surface decoration of WO₃ architectures with Fe₂O₃ nanoparticles for visible-light-driven photocatalysis" *CrystEngComm*, 16 3289-3295, 2014. <https://doi.org/10.1039/C3CE42410C>
- [26] S.-M. Lam, J.-C. Sin, A.Z. Abdullah, and A.R. Mohamed, "Sunlight responsive WO₃/ZnO nanorods for photocatalytic degradation and mineralization of chlorinated phenoxyacetic acid herbicides in water" *Journal of Colloid and Interface Science*, 450 34-44, 2015. <https://doi.org/10.1016/j.jcis.2015.02.075>
- [27] L. Shi, Z. He, and S. Liu, "MoS₂ quantum dots embedded in g-C₃N₄ frameworks: A hybrid 0D-2D heterojunction as an efficient visible-light driven photocatalyst" *Applied Surface Science*, 457 30-40, 2018. <https://doi.org/10.1016/j.apsusc.2018.06.132>
- [28] Q. Qin, Q. Shi, W. Ding, J. Wan, and Z. Hu, "Efficient hydrogen evolution and rapid degradation of organic pollutants by robust catalysts of MoS₂/TNT@CNTs" *International Journal of Hydrogen Energy*, 43 (33), 16024-16037, 2018. <https://doi.org/10.1016/j.ijhydene.2018.07.051>
- [29] H. Li, H. Shen, L. Duan, R. Liu, Q. Li, Q. Zhang, and X. Zhao, "Enhanced photocatalytic activity and synthesis of ZnO nanorods/MoS₂ composites" *Superlattices and Microstructures*, 117 336-341, 2018. <https://doi.org/10.1016/j.spmi.2018.03.028>
- [30] S. Kurita, M. Tanaka, and F. Levy, "Optical spectra near band edge of ZrS₃ and ZrSe₃" *Physical Review B*, 48 (3), 1356-1360, 1993. <https://doi.org/10.1103/physrevb.48.1356>
- [31] A.A. Stabile, L. Whittaker, T.L. Wu, P.M. Marley, S. Banerjee, and G. Sambandamurthy, "Synthesis, characterization, and finite size effects on electrical transport of nanoribbons of the charge density wave conductor NbSe₃" *Nanotechnology*, 22 485201, 2011. <https://doi.org/10.1088/0957-4484/22/48/485201>
- [32] K. Endo, H. Ihara, K. Watanabe, and S.I. Gonda, "XPS study on valence band structures of transition-metal trisulfides, TiS₃, NbS₃, TaS₃" *Journal of solid state chemistry*, 39 215-218, 1981. [https://doi.org/10.1016/0022-4596\(81\)90334-0](https://doi.org/10.1016/0022-4596(81)90334-0)
- [33] A.L. Patterson, "The Scherrer formula for X-ray particle size determination" *Physical Review*, 56 978-982, 1939.
- [34] H. Jin, D. Cheng, J. Li, X. Cao, B. Li, X. Wang, X. Liu, and X. Zhao, "Facile synthesis of zirconium trisulfide and hafnium trisulfide nanobelts: Growth mechanism and Raman spectroscopy" *Solid State Sciences*, 13 (5), 1166-1171, 2011. <https://doi.org/10.1016/j.solidstatesciences.2010.12.017>
- [35] K. Giagloglou, J.L. Payne, C. Crouch, R.K. Gover, P.A. Connor, and J.T. Irvine, "Zirconium trisulfide as a promising cathode material for Li primary thermal batteries" *Journal of the Electrochemical Society*, 163 (14), A3126-A3130, 2016. <https://doi.org/10.1149/2.1351614jes>
- [36] P. Gard, F. Cruege, C. Sourisseau, and O. Gorochoy, "Single-crystal micro-Raman studies of ZrS₃, TiS₃, and several Zr_{1-x}Ti_xS₃ compounds (0<x<=0.33)" *Journal of Raman Spectroscopy*, 17 283-288, 1986. <https://doi.org/10.1002/jrs.1250170310>
- [37] C. Wang, Y. Zhan, and Z. Wang, "TiO₂, MoS₂, and TiO₂/MoS₂ Heterostructures for Use in Organic Dyes Degradation" *ChemistrySelect*, 3 (6), 1713-1718, 2018. <https://doi.org/10.1002/slct.201800054>
- [38] H. Jin, D. Cheng, J. Li, X. Cao, B. Li, X. Wang, X. Liu, and X. Zhao, "Facile synthesis of zirconium trisulfide and hafnium trisulfide nanobelts: Growth mechanism and Raman spectroscopy" *Solid State Sciences*, 13 1166-1171, 2011. <https://doi.org/10.1016/j.solidstatesciences.2010.12.017>
- [39] V.G. Keramidas, and W.B. White, "Raman Scattering Study of the Crystallization and Phase Transformations of ZrO₂" *J. Am. Ceram. Soc.*, 57 (1), 22-24, 1974. <https://doi.org/10.1111/j.1151-2916.1974.tb11355.x>
- [40] F. Jellinek, R.A. Pollak, and M.W. Shafer, "X-ray photoelectron spectra and electronic structure of zirconium trisulfide and triselenide" *Materials Research Bulletin*, 9 (6), 845-856, 1974. [http://dx.doi.org/10.1016/0025-5408\(74\)90121-4](http://dx.doi.org/10.1016/0025-5408(74)90121-4)
- [41] C. Morant, J.M. Sanz, L. Galán, L. Soriano, and F. Rueda, "An XPS study of the interaction of oxygen with zirconium" *Surface Science*, 218 (2), 331-345, 1989. [https://doi.org/10.1016/0039-6028\(89\)90156-8](https://doi.org/10.1016/0039-6028(89)90156-8)

A Novel Pulse Position Modulator for Compressive Data Acquisition

Constantine A. Pappas*

ECE, Stevens Institute of Technology, 07030, USA

ARTICLE INFO

Article history:

Received: 05 December, 2018

Accepted: 28 January, 2019

Online: 11 February, 2019

Keywords:

Nonuniform Sampling

Compressed Sensing

Data Converter

ABSTRACT

This work extends the development of the nonuniform Parallel Digital Ramp Pulse Position Modulation Analog-to-Digital Converter (PDR-ADC) architecture. The continuous to discrete transform of the PDR-ADC is achieved by partitioning the signal amplitude axis into P nonoverlapping partitions that sample the analog input at input signal driven instances. Each partition contains L uniform levels with different quantization step sizes such that the dynamic range of the partitions are related as a geometric series. It is shown that this new architecture satisfies the Nyquist requirement on average (Beutler's condition) and results in a random additive sampling architecture that is alias free (Shapiro-Silverman condition). Additionally, it is shown that the geometric partitioning causes the signal-to-quantization noise ratio (SQNR) to remain approximately constant. A comprehensive design paradigm is presented, including circuits to affect the desired response, the format of the encoded digital samples and the corresponding transformation to determine the equivalent analog voltage. Lastly, although the thrust of this paper is not reconstruction techniques, reconstruction is, nevertheless, compulsory, and recovery and reconstruction is demonstrated through simulations.

1 Introduction

This communication is an extension of work originally presented at a conference on Electrical and Computer Engineering [1] and this current paper significantly expands upon the initial concept proposed in the original material. Although the most common form of sampling is uniform sampling, there are many cases where nonuniform sampling arises and is intentional [2]. Compressed Sensing (CS) [3, 4], is based upon, and operates on nonuniform samples. In CS, the goal is to compress at the time of sampling [5]. To combine acquisition and compression into one step necessitates new hardware innovations. The data converter proposed is a novel approach to nonuniform data acquisition.

The mathematical theory of nonuniform sampling and reconstruction has been well studied [6], and several hardware realizations have been described. In [7], the concept of the Level Crossing (LC) detector for nonuniform sampling was described and developed, in [8], the LC concept was extended and [9], an LC hardware design in 120nm CMOS was fabricated. In

[10], the LC concept was extended to include a triangular dither signal on the input. In [11], the LC sampling scheme was used in an event driven ADC application for electrocardiogram (ECG) signal acquisition. In [12], an adaptive LC sampling scheme was developed, whereby the levels are no longer static but rather adapt to the required signal dynamic range. In [13], a time based ADC was proposed using pulse position modulation (PPM). In [14], a nonuniform sampling system based upon PPM using a reference ramp rest was proposed. In [15], a wideband nonuniform sampling system using a random modulator pre-integrator, similar to direct sequence spread spectrum, was described, and in [16] a nonuniform sampling system based upon pseudo-randomly (PN) triggering the sample-and-hold circuit in an otherwise standard ADC was proposed.

The new nonuniform architecture presented here is based on a parallel implementation of the standard Digital Ramp Pulse Position Modulators (PDR-ADC). The architecture partitions the amplitude range into P nonoverlapping partitions, each governed by its own digital ramp. The ensemble of digital ramps operate from a single counter N bit counter and a single Digital-

*Corresponding Author: Constantine Pappas, cpappas@stevens.edu

to-Analog (DAC) converter, greatly simplifying the synchronization and calibration process. It is shown that our approach exhibits effective compressed sensing performance (compression and acquisition at the same time) at greatly reduced complexity.

2 Conventional PPM

Conventional Pulse Position Modulation (PPM) generates one sample per period of a reference ramp [1]. In the PPM waveform generator circuit shown in Figure 1, $f(t)$ is the input analog signal to be sampled and $r(t)$ is a saw tooth reference waveform with period, T_{ramp} . The comparator is assumed to be referenced from a positive voltage supply, V_{CC} , and a negative supply voltage supply V_{EE} . Pulse Width Modulation (PWM) is an output signal and PPM is the output Pulse Position Modulation signal generated by the monostable multivibrator (one shot) circuit. When triggered, a one shot produces a single pulse of fixed, finite duration.

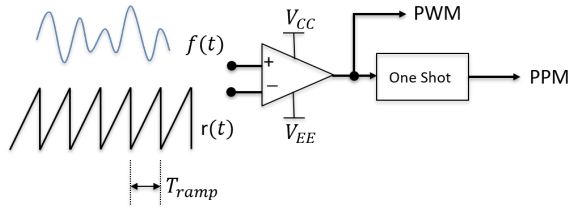


Figure 1: PPM Generator

A comparator, in general, will produce an output at the positive supply rail, V_{CC} , whenever the signal input to the noninverting amplifier terminal, V_{\oplus} , is greater than the signal input to the inverting amplifier terminal, V_{\ominus} . Similarly, the comparator output will be at the negative supply rail, V_{EE} , whenever $V_{\ominus} > V_{\oplus}$. For the circuit in Figure 1, the PWM output signal as a function of time may be expressed as:

$$PWM(t) = \begin{cases} V_{CC} & \text{if } f(t) > r(t) \\ V_{EE} & \text{otherwise} \end{cases} \quad (1)$$

Under the conditions specified, when $r(t)$ exceeds $f(t)$, the one shot will trigger and a sample generated. Assuming the bandwidth of the analog input, $f(t)$, is less than $\frac{1}{T_{ramp}}$, one sample is generated per period of the reference ramp and we may define the sample rate as, $FS = \frac{1}{T_{ramp}} \equiv \frac{1}{T_s}$.

Figure 2 is an illustration of several cycles of the reference ramp signal. In Figure 2, r_n is the n^{th} period of the reference signal. If we define the full scale voltage as, $V_{FS} = V_{CC} - V_{EE}$ and let the negative supply rail be such that, $V_{EE} = -V_{CC}$, as is typically the case, then we may write, $V_{FS} = 2V_{CC}$. We can then define the slope of the reference as, $\beta = \frac{V_{FS}}{T_s}$. By direct enumeration, $r_n(t)$ in Figure 2 is:

$$\begin{aligned} r_0(t) &= \beta t + V_{EE} & 0 \leq t < t_1 \\ r_1(t) &= \beta t + V_{EE} - \beta t_1 & t_1 \leq t < t_2 \\ r_2(t) &= \beta t + V_{EE} - \beta t_2 & t_2 \leq t < t_3 \\ &\vdots \\ r_n(t) &= \beta t + V_{EE} - \beta t_n & nT_s \leq t < (n+1)T_s \end{aligned} \quad (2)$$

Substituting $V_{EE} = -\frac{V_{FS}}{2}$, $V_{FS} = \beta T_s$ and $t_n = nT_s$, (2) becomes:

$$\begin{aligned} r_n(t) &= \beta t - \frac{\beta T_s}{2} - \beta nT_s \\ &= \beta t - \beta T_s \left(n + \frac{1}{2} \right) \quad n = 0, 1, 2, \dots \end{aligned} \quad (3)$$

The comparator triggers when $r(t) = f(t)$, from (3), the discrete sample time, τ_n , of the n^{th} pulse is:

$$\tau_n = nT_s + \frac{f_n}{\beta} + t_d \quad n = 0, 1, 2, \dots \quad (4)$$

where $\beta = \frac{V_{FS}}{T_s}$ and $t_d = \frac{T_s}{2}$.

Equation (4) shows, in conventional PPM not only is the timing of the n^{th} sample proportional to the sample number, n , the timing is also a function of the signal being sampled. Consequently, it is seen, conventional PPM generates nonuniform sampling.

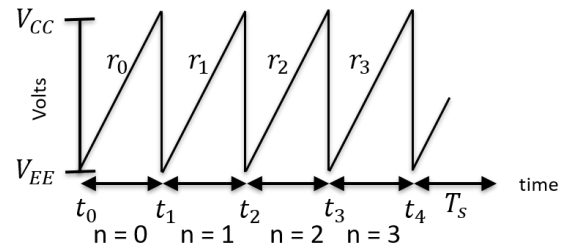


Figure 2: PPM Reference Ramp

3 Δ-PPM

It is possible to generate a PPM signal directly, without explicitly having the PWM output trigger a one-shot. Such a direct generation of the PPM signal is accomplished by resetting the reference ramp after the comparator triggers. Generating a PPM signal by resetting the reference ramp will be called, Δ -PPM, to distinguish it from conventional PPM. Δ -PPM may be generated with the circuit of Figure 3. In Figure 3, V_{CC} and V_{EE} are the power supplies for the comparator, $f(t)$ is the signal to be sampled and $r(t)$ is the reference ramp. Also shown in Figure 3 is a RESET circuit that asynchronously resets the reference ramp at each PPM pulse.

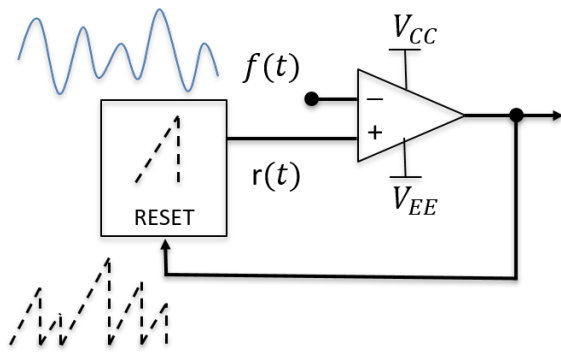


Figure 3: Reset PPM Generator

The previous periodicity of the reference signal is annihilated when the reference ramp is allowed to reset after the comparator triggers. An arbitrary response to the asynchronous triggering is shown in Figure 4. In Figure 4, we call r_n is the n^{th} response of the reference signal, rather than the n^{th} period of the reference, and again define the full scale voltage as, $V_{FS} = V_{CC} - V_{EE}$ and the slope of the reference as, $\beta = \frac{V_{FS}}{T_s}$.

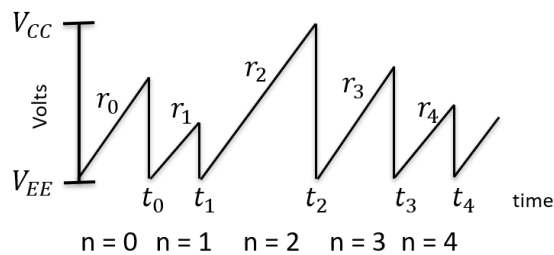


Figure 4: Reset PPM Reference Ramp

By direct enumeration, $r_n(t)$ in Figure 4 is:

$$\begin{aligned} r_0(t) &= \beta t + V_{EE} & 0 \leq t < t_0 \\ r_1(t) &= \beta t + V_{EE} - \beta t_0 & t_0 \leq t < t_1 \\ r_2(t) &= \beta t + V_{EE} - \beta t_1 & t_1 \leq t < t_2 \\ &\vdots \\ r_n(t) &= \beta t + V_{EE} - \beta t_{n-1} & t_{n-1} \leq t < t_n \end{aligned} \quad (5)$$

We note that the interval endpoints cannot be specified as constants, as was the case in conventional PPM, because they evolve dynamically. Again, substituting $V_{EE} = -\frac{V_{FS}}{2}$ with $V_{FS} = \beta T_s$, (5) becomes:

$$r_n(t) = \beta t - \beta t_{n-1} - \frac{\beta T_s}{2} \quad t_{n-1} \leq t < t_n \quad (6)$$

The comparator triggers when $r(t) = f(t)$, from (6), the discrete sample time, τ_n , of the n^{th} pulse is:

$$\tau_n = \tau_{n-1} + \frac{f_n}{\beta} + t_d \quad n = 1, 2, 3, \dots \quad (7)$$

where $\beta = \frac{V_{FS}}{T_s}$ and $t_d = \frac{T_s}{2}$.

Equation (7) shows, in Δ -PPM the timing of the n^{th} sample is a function of the signal being sampled and thus Δ -PPM generates nonuniform sampling, as in conventional PPM. Additionally, from (7), Δ -PPM samples times are a function of the previous sample time.

Two important consequence result if the sample times are a function of the previous sample time. First, Shapiro and Silverman [20] showed that sampling can be made alias free if each sample time is derived from the previous one by the addition of an independent random variable, from (7) it is seen that Δ -PPM provides such a sampling scheme. Second, Beutler [21] showed that if the sampling rate exceeds the Nyquist rate *on average* then the nonuniform sampling set will be stable and can be used to reconstruct a band-limited signal. We now show that Δ -PPM produces a sampling set with an average sampling rate, R_{avg} , that approaches,

$$R_{avg} = \frac{2}{T_s} \approx 2FS.$$

Using (7) iteratively we may write:

$$\tau_n = \tau_o + n \frac{T_s}{2} + \frac{1}{\beta} \sum_{k=1}^n f_k \quad (8)$$

where τ_o is the first sampling instant.

Let $n \rightarrow N$, be the total number of samples such that $\{N : N \in 1, 2, 3, \dots, \infty\}$ and let τ_N be the maximum sample time, then, dividing by N :

$$\frac{\tau_N}{N} = \frac{\tau_o}{N} + \frac{T_s}{2} + \frac{1}{\beta} \left(\frac{1}{N} \sum_{k=1}^N f_k \right) \quad (9)$$

The last term in (9) is the average value of $f(t)$. If the average value equals zero, then the average sample rate is given by:

$$R_{avg} = \frac{2}{T_s} \left(\frac{N}{N + 2\tau_o/T_s} \right) \quad (10)$$

The maximum value of the first sampling instant is, $\tau_o = T_s$, which corresponds to the minimum value of R_{avg} , thus:

$$R_{avg} \geq \frac{2}{T_s} \left(\frac{N}{N + 2} \right) \quad (11)$$

Less formally, for reasonable values of N experienced in practice, we may regard the average sample rate in Δ -PPM as, $R_{avg} = \frac{2}{T_s} \approx 2FS$.

The significance of (7) and (11) are that Δ -PPM is *self-regulating*. Δ -PPM automatically satisfies the Nyquist requirement on average (Beutler's condition [21]) and produces alias free random sampling (Shapiro-Silverman condition [20]). Additionally, Δ -PPM achieves self-regulation with no a priori knowledge of the signal support and does not utilize any particular code or special sequence to generate random samples.

4 Geometric Partitioning

The stylized, equal partitioning, presented [1], was intended to provide an introduction for the essence of the PDR-ADC. We now develop the elaboration of the partitioning scheme, where significant benefits will be obtained.

All data converters need an analog reference voltage to accomplish the continuous to discrete transform and produce a digital word [22]. In general, the converter divides an analog reference voltage, V_{REF} , into a fixed number of analog voltage levels, L . These analog voltage levels are then mapped to a digital number, typically referred to as counts. By convention, the smallest analog voltage level is assigned digital level $[0]$. The remaining analog voltage levels are mapped to digital levels by incrementing the ADC count. In this way, it is not possible to map the analog reference voltage, V_{REF} , to a number that can be reached and assigned a digital count.

All of the digital levels taken together define the scale. In general, there will be L levels and $L - 1$ steps. The reference voltage divided by the number of levels defines the size of the analog step to reach an adjacent level. The step size, Δ , is sometimes referred to as the quantization step size or the least significant bit (LSB). The maximum analog voltage that can be mapped to a digital level is called the *full scale* voltage, V_{FS} . The relationship between the reference voltage, the full scale voltage and the quantization step size are shown in Figure 5 for a converter with $L = 16$ levels.

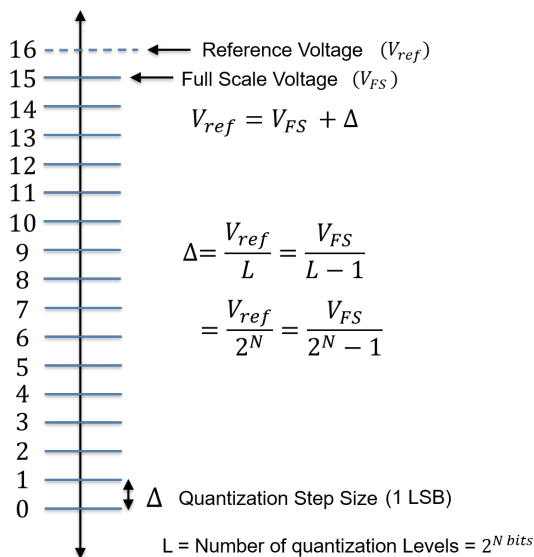


Figure 5: Quantization Step Size

The number of levels, L , is typically designed to be a function of the number of bits as, $L = 2^N$, where N is the number of bits. For the 4 bit converter shown in Figure 5, the reference voltage is divided by $16 = 2^4$. The first digital level is assigned digital count $[0]$, and corresponds to the digital number, $[0\ 0\ 0\ 0]$. The maximum digital count is $[15]$, and corresponds to

the digital number, $[1\ 1\ 1\ 1]$. It is not possible to encode digital count $[16]$ with a 4 bit counter, and consequently, the reference voltage, V_{REF} in Figure 5, is not mapped to a digital number.

The principle of the geometric partitioning is shown in Figure 6 for a system with $\mathcal{P} = 8$ partitions and $L = 4$ levels. In the PDR data converter, the signal amplitude axis is partitioned into \mathcal{P} partitions such that each partition contains L levels, the partitions do not overlap and each partition has a different quantization step size. The geometric partitioning is achieved by relating the spans (the dynamic range) of the partitions as a geometric series.

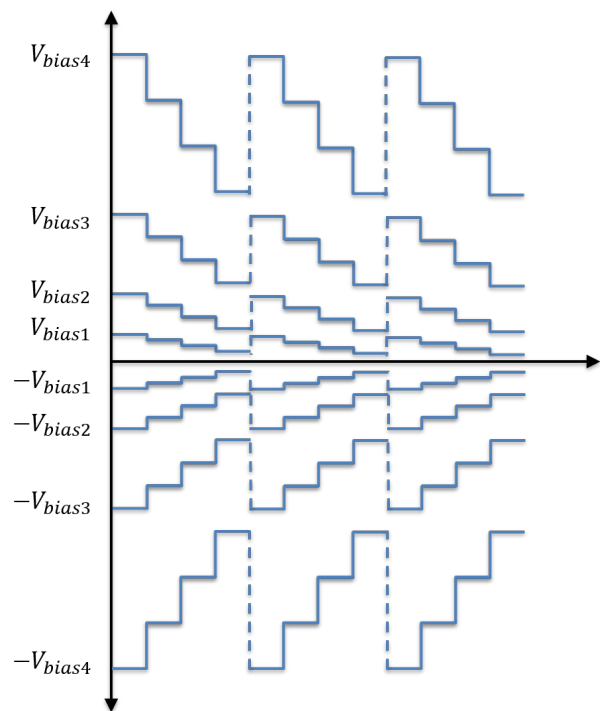


Figure 6: Geometric Partitioning: Conceptual

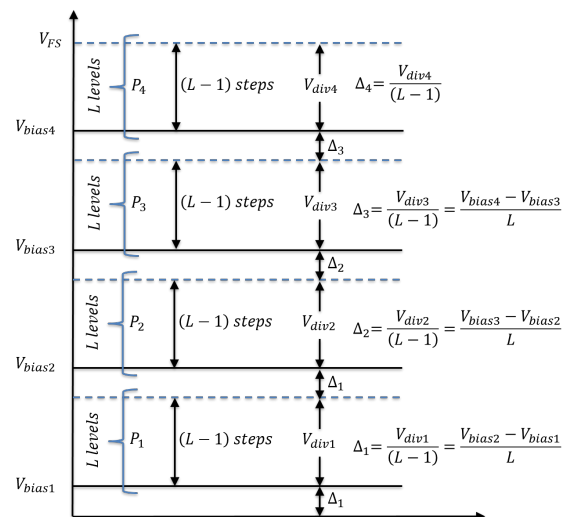


Figure 7: Geometric Partitioning: Detail

The relationship between partition bias voltages, V_{bias_m} , and the quantization step sizes needed to realize the behavior shown in Figure 6, can be better understood and visualized with the aid of Figure 7. Due to the symmetry of the bias voltages, only the positive partitions, P_1, P_2, P_3 and P_4 , as shown in Figure 7, are needed.

Let the total number of partitions, \mathcal{P} , be even, and define the maximum partition number to be, $\mathcal{M} = \frac{\mathcal{P}}{2}$, and let m denote the m^{th} partition. In Figure 7, we denote by V_{div_m} the span of the m^{th} partition. By design, the spans of the partitions are geometrically related, thus:

$$\begin{aligned} V_{div2} &= 2V_{div1} = 2^{(2-1)}V_{div1} \\ V_{div3} &= 4V_{div1} = 2^{(3-1)}V_{div1} \\ V_{div4} &= 8V_{div1} = 2^{(4-1)}V_{div1} \\ &\vdots \\ V_{div_m} &= 2^{(m-1)}V_{div1} \quad m = 1, 2, 3, \dots, \mathcal{M} \end{aligned} \quad (12)$$

To realize (12), a circuit that takes the output from the DAC and applies the appropriate gain to compress the DAC steps is required. The required step compression response is shown in Figure 8 and the circuit to realize the response is shown in Figure 9.

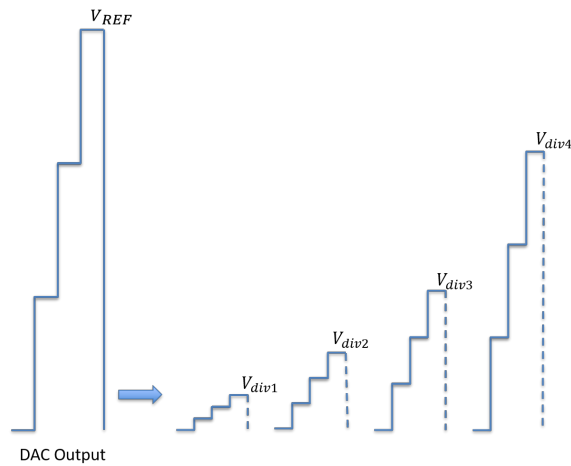


Figure 8: Step Compression Voltages

The step compression circuit in Figure 9 is a non-inverting voltage divider. To obtain design equations for the bias voltages in terms of the reference voltage, V_{REF} , the maximum number of partitions, \mathcal{M} and the number of levels, L , we design the step compression circuit such that the voltage drop across resistor, $V_{R_x} = (2^{(\mathcal{M}-1)} - 1)V_{div1}$. Then, due to the geometric design and Kirchoff's voltage law, it must be the case:

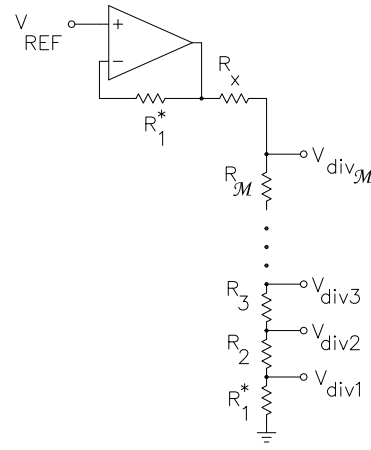


Figure 9: Step Compression Circuit

$$\begin{aligned} V_{REF} &= V_{div1} + V_{div2} + V_{div3} + \dots + V_{div_M} \\ V_{REF} &= V_{div1} (1 + 2 + 4 + 8 + \dots + 2^{(\mathcal{M}-1)}) \\ V_{REF} &= V_{div1} \sum_{k=0}^{\mathcal{M}-1} 2^k \\ V_{REF} &= V_{div1} (2^{\mathcal{M}} - 1) \\ &\vdots \\ V_{div1} &= \frac{V_{REF}}{2^{\mathcal{M}} - 1} \end{aligned} \quad (13)$$

From (12) and (13), the m^{th} voltage divider voltage is given by:

$$V_{div_m} = \left(\frac{2^{(m-1)}}{2^{\mathcal{M}} - 1} \right) V_{REF} \quad (14)$$

Dividing (14) by $L - 1$, the quantization step size, Δ , is:

$$\Delta_m = \left(\frac{2^{(m-1)}}{2^{\mathcal{M}} - 1} \right) \left(\frac{V_{REF}}{L - 1} \right) \quad (15)$$

Using Figure 7 and (14) and (15), we may write the bias voltages, V_{bias_m} , as:

$$\begin{aligned} V_{bias_m} &= \Delta_1 + \sum_{k=1}^{m-1} (V_{div_m} + \Delta_m) \\ V_{bias_m} &= \Delta_1 + \sum_{k=1}^{m-1} \left(\frac{2^{(m-1)}}{2^{\mathcal{M}} - 1} V_{REF} + \frac{2^{(m-1)}}{2^{\mathcal{M}} - 1} \frac{V_{REF}}{L - 1} \right) \end{aligned} \quad (16)$$

$$V_{bias_m} = \Delta_1 + \frac{V_{REF}}{2^{\mathcal{M}} - 1} \left(\frac{L}{L - 1} \right) \sum_{k=1}^{m-1} 2^{k-1}$$

$$V_{bias_m} = \Delta_1 + \frac{V_{REF}}{2^{\mathcal{M}} - 1} \left(\frac{L}{L - 1} \right) [2^{(m-1)} - 1]$$

$$V_{bias_m} = \frac{V_{REF}}{2^{\mathcal{M}} - 1} \left(\frac{L}{L - 1} \right) \left(2^{(m-1)} - 1 + \frac{1}{L} \right)$$

Lastly, the Full Scale voltage shown in Figure 7 may be obtained by adding the results of (16) and (14) evaluating at $m = \mathcal{M}$:

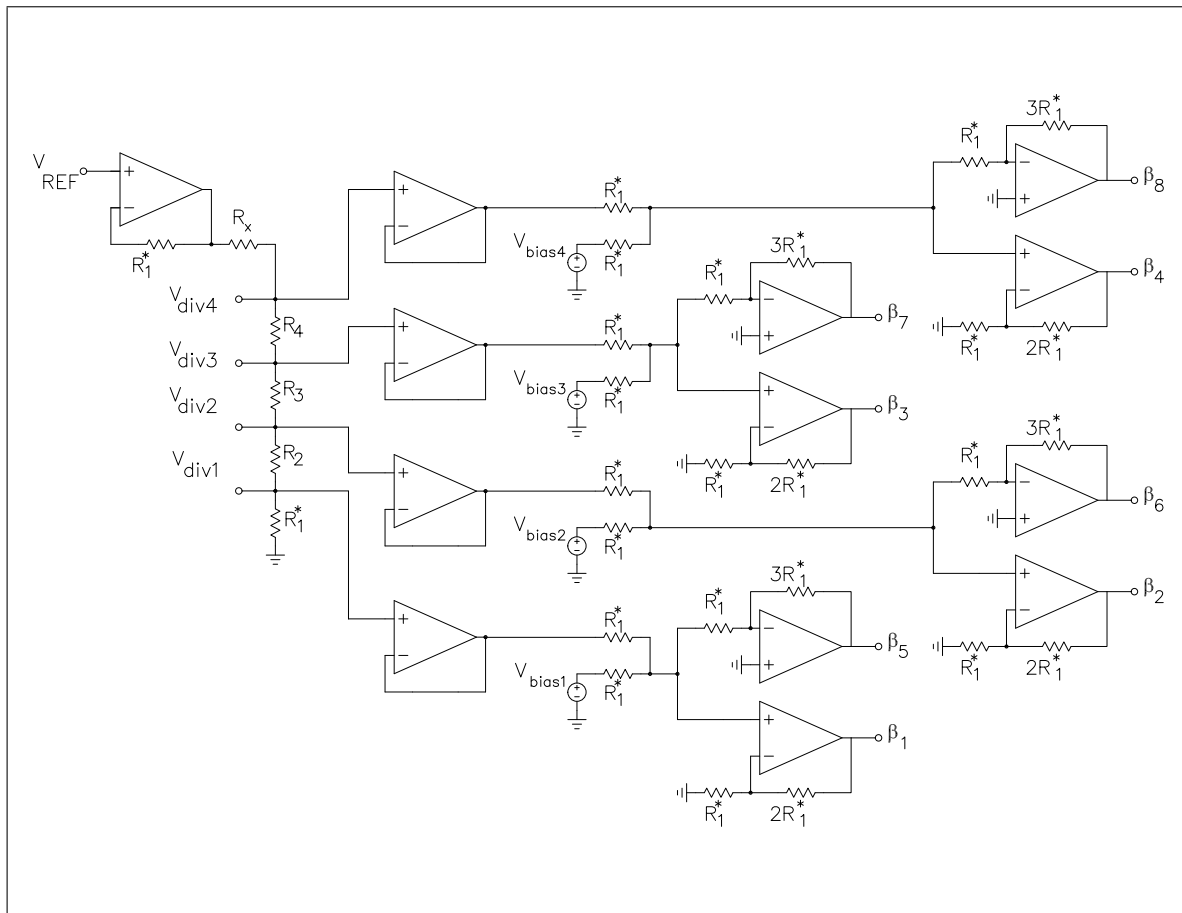


Figure 10: Step Compression & Level Shifting Circuit

$$V_{FS} = \frac{V_{REF}}{2^M - 1} \left(\frac{L}{L-1} \right) \left[2^M - 1 + \frac{1 - 2^{(M-1)}}{L} \right] \quad (17)$$

In terms of the external design parameters, V_{REF} , M and L , the number of partitions, M will have the most influence on the design equations, (16) and (17), since the bias is proportional to $\frac{1}{2^M - 1}$.

5 Step Compression and Level Shifting

We now develop the circuit to generate the parallel digital ramps shown in Figure 6, and we call this function, step compression and level shifting (SCLS). The step compression and level shifting circuit is shown in Figure 10. Step compression is governed by equations (14) and (15) and we seek now to determine design equations for the resistance for the circuit in Figure 10.

5.1 Step Compression

The step compression gains can be realized from circuit analysis by solving for the resistor values, in Figure 8, required to establish the voltage divider voltages, V_{div_m} as given by (14).

Let $\sum R_m = R_1^* + R_2 + R_3 + \dots + R_M$, then from circuit analysis:

$$V_{div1} = V_{REF} \left(\frac{R_1^*}{R_x + \sum R_m} \right) \quad (18a)$$

and in general:

$$V_{div_m} = V_{REF} \left(\frac{R_1^* + R_2 + R_3 + \dots + R_m}{R_x + \sum R_m} \right) \quad (18b)$$

Divide (18b) by (18a):

$$\frac{V_{div_m}}{V_{div1}} = \frac{R_1^* + R_2 + R_3 + \dots + R_m}{R_1^*} \quad (19)$$

Substituting the geometric relation given in (12) for V_{div_m} :

$$2^{(m-1)} R_1^* = R_1^* + R_2 + R_3 + \dots + R_m = \sum R_m \quad (20a)$$

from which it is seen:

$$\begin{aligned} m = 2 &\longrightarrow R_2 = R_1^* \\ m = 3 &\longrightarrow R_3 = 2R_1^* \\ m = 4 &\longrightarrow R_4 = 4R_1^* \end{aligned} \quad (20b)$$

⋮

and in general:

$$R_m = 2^{(m-2)} R_1^* \quad \text{for } m > 1 \quad (20c)$$

Lastly, from Equation (13) and (18a):

$$\frac{1}{2^M - 1} = \frac{R_1^*}{R_x + \sum R_M} \quad (21a)$$

$$R_x = (2^M - 1)R_1^* - \sum R_M$$

and from (20a):

$$R_x = (2^M - 1)R_1^* - 2^{(M-1)}R_1^* \quad (21b)$$

$$R_x = [2^{(M-1)} - 1]R_1^*$$

To complete the design of the Step Compression circuit, we must specify a value for resistor R_1^* . To do so, we seek a design equation that relates the RMS noise voltage of the resistor to the voltage of the smallest quantization step size, Δ_1 in Figure 7. We then design the resistor value be less than this value so that the system is limited by the quantization noise of the LSB rather than the thermal noise of the resistor.

From (15):

$$\Delta_1 = \frac{V_{REF}}{(2^M - 1)(L - 1)} \quad (22)$$

Assuming the noise is Gaussian distributed, then approximately all of the noise is contained within 6.6 standard deviations¹. The peak-to-peak thermal noise voltage of the resistor is given by:

$$V_{ppN_{th}|R_1^*} = 6.6 \sqrt{4RkTB} \quad (23)$$

where R is the resistance in Ohms, k is Boltzmann's constant ($k \approx 1.30865 \times 10^{-23} J/K$), T is the temperature in Kelvin and B is the bandwidth in Hertz.

To design for the thermal noise voltage of R_1^* to be less than Δ_1 , we should select R_1^* , such that:

$$R_1^* \leq \left(\frac{1}{6.6}\right)^2 \left(\frac{V_{REF}}{(L-1)(2^M-1)}\right)^2 \left(\frac{1}{4kTB}\right) \quad (24)$$

5.2 Level Shifting

Level shifting is governed by the bias voltages as given by (16). The aim of the level shifting circuit is to shift the step compressed voltages, shown in Figure 8, to the required bias level, shown in Figure 6. The β_m signals, in Figure 10, are the final step compressed and level shifted signals that are feed back to the input comparators [1]. From the symmetry of the design, the positive shifted β signals are given by:

$$\beta_m = V_{div_m} + V_{bias_m} \quad \text{for } 1 \leq m \leq M \quad (25)$$

and the negative shifted signals are given by:

$$\beta_{M+m} = -(V_{div_m} + V_{bias_m}) \quad \text{for } 1 \leq m \leq M \quad (26)$$

Specifically, for the β signals shown in Figure 10,;

$$\begin{aligned} \beta_1 &= V_{div1} + V_{bias1} & \beta_5 &= -(V_{div1} + V_{bias1}) \\ \beta_2 &= V_{div2} + V_{bias2} & \beta_6 &= -(V_{div2} + V_{bias2}) \\ \beta_3 &= V_{div3} + V_{bias3} & \beta_7 &= -(V_{div3} + V_{bias3}) \\ \beta_4 &= V_{div4} + V_{bias4} & \beta_8 &= -(V_{div4} + V_{bias4}) \end{aligned} \quad (27)$$

¹ 6σ is often used, however, 6.6σ is becoming an industry standard.

The design is readily obtained if the inverting and noninverting amplifiers in 10 are solved with the feedback resistors as arbitrary unknown resistors (2 degrees of freedom) and the remaining resistors fixed, equal to resistor R_1^* . A pair of inverting and noninverting amplifiers are shown in Figure 11 where R_{\ominus} , is the resistor in the inverting amplifier and R_{\oplus} is the resistor in the non-inverting amplifier.

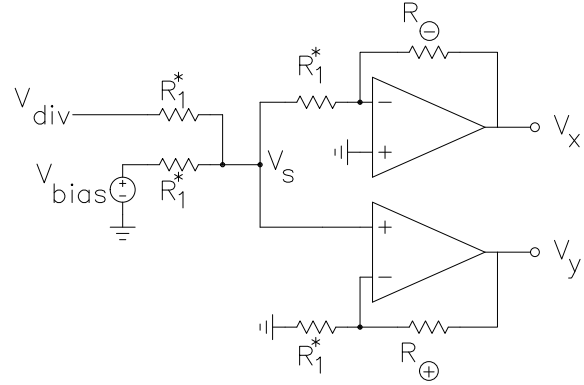


Figure 11: Level Shifting Feedback Scaling

From circuit analysis, the node voltage, V_s is:

$$\frac{V_s - V_{div}}{R_1^*} + \frac{V_s - V_{bias}}{R_1^*} + \frac{V_s}{R_1^*} = 0 \quad (28)$$

$$V_s = \frac{V_{div} + V_{bias}}{3}$$

For the inverting amplifier we have, $\frac{V_s}{R_1^*} = -\frac{V_x}{R_{\ominus}}$, from which:

$$V_x = -\frac{R_{\ominus}}{R_1^*} \left(\frac{V_{div} + V_{bias}}{3} \right) \quad (29)$$

The required inverting sum, as given in (27), is obtained when:

$$R_{\ominus} = 3R_1^* \quad (30)$$

Similarly, for the non-inverting amplifier we have:

$$V_y = \left(1 + \frac{R_{\oplus}}{R_1^*} \right) V_s, \text{ from which:}$$

$$V_y = \left(1 + \frac{R_{\oplus}}{R_1^*} \right) \left(\frac{V_{div} + V_{bias}}{3} \right) \quad (31)$$

The required non-inverting sum, as given in (27), is obtained when:

$$R_{\oplus} = 2R_1^* \quad (32)$$

6 Maximum Sample Rate

Samples are generated by the precision windowed dual slope edge detector shown in Figure 12. The edge detector triggers an asynchronous LOAD flip-flop that stores the instantaneous value of the counter to memory. The LOAD signal, additionally, “sets” an RS flip-flop that drives the synchronous RESET flip-flop. On the next CLOCK, T_{clk} , the RESET flip-flop resets the counter. The counter generates a HOLD OFF signal that inhibits additional LOAD signals until the counter has settled, where the safety time to settle is 1 CLOCK pulse. The sampling circuit is shown in Figure 12.

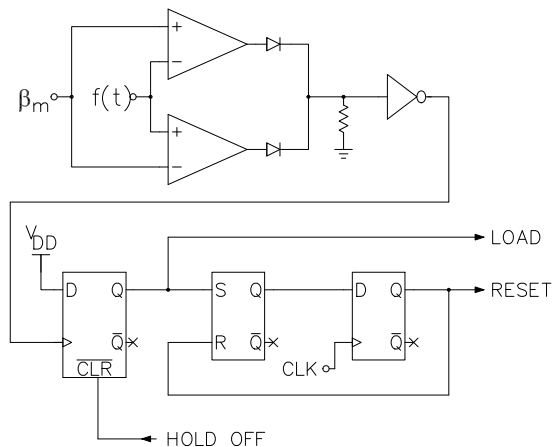


Figure 12: Windowed Synchronous One-Shot

In the PDR data converter, we must ensure that the counter has settled before the next LOAD/RESET cycle. The worst case signaling (fastest signal) corresponds if the LOAD signal aligns with a CLOCK edge. In such a case, a minimum of 3 CLOCKS is required to guarantee correct conversion before the next LOAD signal is allowed to register a new sample, this worst case timing is shown in Figure 13.

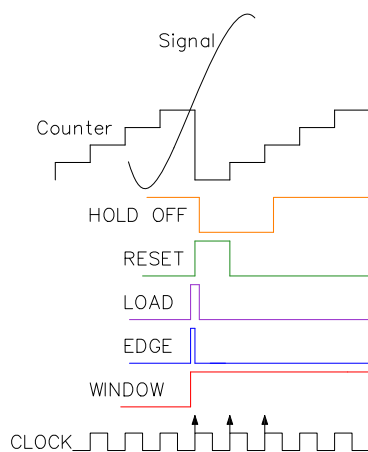


Figure 13: LOAD/RESET Timing

To determine the maximum signal frequency that the PDR-ADC can accommodate, we equate the maximum rate of change of the input signal, to the maximum rate of change allowable by the PDR. With an

input sinusoid of the form, $f(t) = V_{FS} \sin(2\pi f_{sig} t)$, the maximum rate of change is: $\frac{\Delta V}{\Delta t} = 2\pi f_{sig} V_{FS}$. The fastest rate of change the PDR can respond to is: $\frac{\Delta V}{\Delta t} = \frac{2V_{FS}}{3T_{clk}}$. The maximum signal frequency is:

$$f_{sigmax} = \frac{1}{3\pi T_{clk}} \quad (33)$$

7 The Counting Vector

In this section, we establish some important properties of the counting vector. Each intersection of $f(t)$ with a reference counter step contributes to the counting vector, α . The counting vector is responsible for determining the time of each sample, the amplitude of each sample, and the number of samples acquired. The counting vector also contains the information about the number of missing samples and where these missing samples are located. Information regarding the missing samples is critically important for reconstruction as the location and number of the missing samples (the data to be interpolated) must be known.

We denote by α , the number of counts accumulated by the counter. If the counter is strictly counting up, then the time to accumulate α counts, $t_\alpha = (\alpha + 1)T_{clk}$, where T_{clk} is the clock period. The time of the n^{th} sample, $t_{[n]}$, is the cumulative sum of the t_α 's:

$$t_{[n]} = T_{clk} \sum_{i=1}^{i=n} (\alpha_i + 1) \quad (34)$$

If, instead, the counter is strictly counting down, we denote by α_c (the complement of α) the number of counts accumulated by the counter. In this way, $\alpha_c = (L - 1) - \alpha$, where $L - 1$ is the maximum value of the counter. In this case, the time of the n^{th} sample is:

$$t_{[n]} = T_{clk} \sum_{i=1}^{i=n} (L - \alpha_i) \quad (35)$$

The direction of the counter can be changed, by using a toggle flip-flop clocked on each RESET signal, and the system will continue to maintain the correct timing of each sample.

To recover the signal requires a method to resolve the counter slope, the partition number and the count value. We append, to the counting vector, additional bits that correspond to the counter slope and the partition number. As a specific example of accommodating partition encoding, suppose a PDR-ADC is designed with $\mathcal{P} = 8$ partitions and $L = 256$ levels per partition. The partition encoder requires 3 bits, the count slope requires 1 bit and the count value requires 8 bits, the data word stored in memory will be of the form:

$$rawData = \overbrace{\begin{matrix} \text{Slope} & \text{Partition number} \\ \boxed{B12} & \boxed{B11} & \boxed{B10} & \boxed{B9} \end{matrix}}^{\text{Count value}} \overbrace{\begin{matrix} \boxed{B8} & \boxed{B7} & \boxed{B6} & \boxed{B5} & \boxed{B4} & \boxed{B3} & \boxed{B2} & \boxed{B1} \end{matrix}}$$

In this example, the partition number in each data word is determined by:

$$m_i = rawData_i [B_{11} : B_9][4 \ 2 \ 1]^T \quad (36)$$

and each α_i is determined by:

$$\begin{aligned} \alpha_i &= rawData_i [B_8 : B_1] \mathbf{W}^T \\ &\equiv \mathbf{B}_i \mathbf{W}^T \end{aligned} \quad (37)$$

where, $\mathbf{W} = [128 \ 64 \ 32 \ 16 \ 8 \ 4 \ 2 \ 1]$.

In general, once the i^{th} partition number, m_i , and the i^{th} count, α_i , have been determined, the sampled voltage value is given by:

$$V_i = \alpha_i \Delta_{m_i} + V_{bias_{m_i}} \quad (38)$$

where Δ_m is given by (15) and V_{bias_m} is given by (16).

Lastly, since $\alpha_i + 1$ is the number of clocks to obtain α_i counts, then α_i is equal to the number of missing samples if the signal had been sampled at a rate equal to T_{clk} .

8 Signal to Quantization Noise Ratio (SQNR)

The signal to noise ratio (SNR) is always equal to the ratio of the signal power, P_{sig} , to the noise power, P_{N_Q} . In data converters, the “noise” added by the system due to the act of approximating (truncating) a continuous function to finite precision is quantization noise, and the ratio of interest is the signal to quantization noise, $SQNR = P_{sig}/P_{N_Q}$. In uniform quantizers, the quantization noise power is well approximated by, [17], [23]:

$$P_{N_Q} = \frac{\Delta^2}{12} \quad (39)$$

In the PDR data converter, from (15), for any adjacent partitions:

$$\frac{\Delta_m}{\Delta_{m+1}} = \frac{1}{2} \quad (40)$$

from (39) and (40):

$$P_{N_{Q_m}} = \frac{P_{N_{Q_{m+1}}}}{4} \quad (41)$$

For the PDR, (41) states, the quantization noise power decreases by a factor of 4 when transitioning from a higher partition to a lower level partition. This effect is shown in the bottom panel in Figure 14, where, for comparison, we have also plotted the quantization error of a uniform quantizer in the top panel.

Let a signal of the form, $y = A_o \sin(\omega_o t)$, be input to a uniform quantizer with quantization noise power, $P_{N_{Q_u}}$, the SQNR is given by:

$$SQNR_u = \frac{P_{sig}}{P_{N_{Q_u}}} = \frac{A_o^2}{2} \frac{1}{P_{N_{Q_u}}} \quad (42)$$

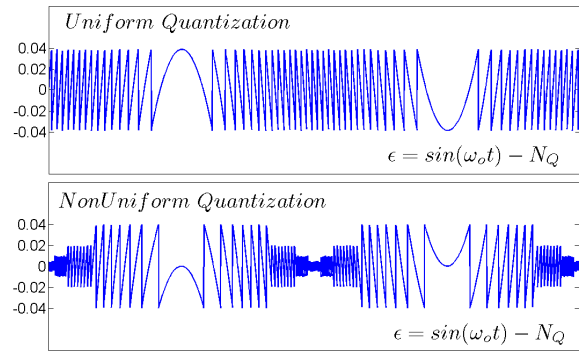


Figure 14: Quantization Error Comparison

Now suppose a signal of the form, $y = \frac{A_o}{2} \sin(\omega_o t)$ is input to this uniform quantizer, the SQNR becomes:

$$SQNR_u = \left(\frac{1}{4}\right) \frac{A_o^2}{2} \frac{1}{P_{N_{Q_u}}} \quad (43)$$

From (42) and (43), in a uniform quantizer, when the input signal amplitude decreases by a factor of 2, the SQNR degrades by a factor of 4.

Now consider a signal of the form, $y = A_o \sin(\omega_o t)$, input to the nonuniform PDR quantizer with quantization noise power, $P_{N_{Q_{nu}}}$, the SQNR is given by:

$$SQNR_{nu} = \frac{P_{sig}}{P_{N_{Q_{nu}}}} = \frac{A_o^2}{2} \frac{1}{P_{N_{Q_{nu}}}} \quad (44)$$

Again, suppose the input signal amplitude decreases by a factor of 2 and let $y = \frac{A_o}{2} \sin(\omega_o t)$ be input to the nonuniform PDR quantizer, then, by (41), the SQNR becomes:

$$\begin{aligned} SQNR_{nu} &= \left(\frac{1}{4}\right) \frac{A_o^2}{2} \frac{1}{P_{N_{Q_{nu}}}/4} \\ &= \frac{A_o^2}{2} \frac{1}{P_{N_{Q_{nu}}}} \end{aligned} \quad (45)$$

From (44) and (45), it is seen, the geometric partitioning of the PDR-ADC attempts to maintain the signal-to-quantization noise ratio constant, this is a significant improvement compared to uniform quantization data converters.

We shall now be concerned with determining this constant. In an N bit uniform quantizer, the quantization step size, Δ_u , is:

$$\Delta_u = \frac{2V_{FS}}{2^N - 1} \quad (46)$$

Suppose we have a PDR data converter, with maximum partition number, \mathcal{M} , and $L = 2^N$ levels per partition, where N is the same as the uniform quantizer in (46). In the PDR, the largest quantization step size is given by (15), evaluated at $m = \mathcal{M}$.

$$\begin{aligned} \Delta_{\mathcal{M}} &= \left(\frac{2^{(\mathcal{M}-1)}}{2^{\mathcal{M}} - 1}\right) \left(\frac{V_{REF}}{2^N - 1}\right) \\ &\approx \left(\frac{1}{2}\right) \left(\frac{V_{REF}}{2^N - 1}\right) \end{aligned} \quad (47)$$

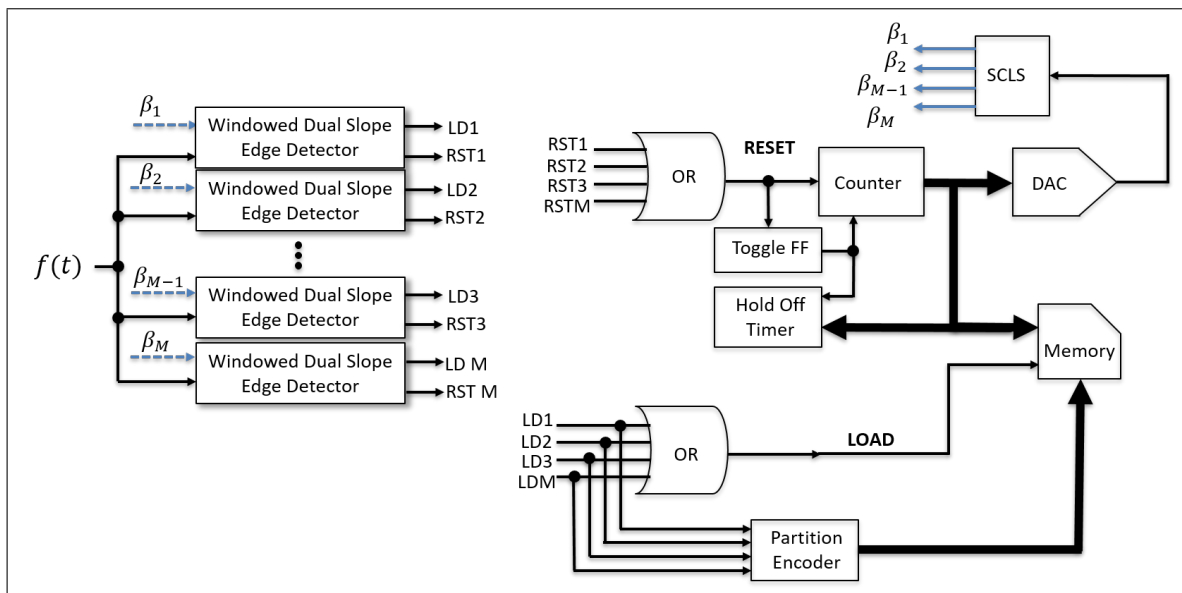


Figure 15: Parallel Digital Ramp ADC Overall Block Diagram

Using (46), we may write (47) as:

$$\Delta_M \approx \left(\frac{1}{4}\right) \left(\frac{2V_{REF}}{2^N - 1}\right) = \frac{2V_{REF}}{2^{(N+2)} - 4} \quad (48)$$

In a PDR data converter, with $L = 2^N$ levels per partition, (48) states, the PDR has gained approximately 2 bits of resolution compared to a uniform quantizer.

The results of (45) and (48) may be summarized as, given a PDR data converter with $L = 2^N$ levels, the SQNR of the system will be approximately equivalent to a uniform system with $L = 2^{(N+2)}$ levels and the PDR will attempt to maintain this performance for all input signal levels.

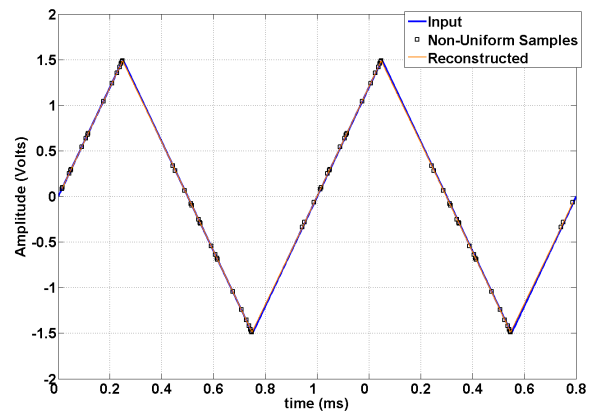


Figure 16: Reconstructed Linear

9 Simulation Results

The proposed parallel digital ramp ADC was modeled and simulated in Simulink® and the reconstruction performed in Matlab®. The overall block diagram of the new ADC is shown Figure 15.

9.1 Linearity

The linearity of the PDR was analyzed empirically using a triangle wave, as shown in Figure 16, since, by design, the sampled data is not equally spaced. This approach was taken because the data converter figure of merit, Differential Non-Linearity (DNL), is a function of the difference of consecutive samples with a constant step size, as given by (49), where for an ideal ADC, the $DNL = 0$ [18]. In the PDR, since large gaps appear in the data and the step size is not constant, linearity is more easily performed graphically.

$$DNL[k] = \frac{ADC_{out}[k] - ADC_{out}[k-1]}{\Delta_{LSB}} - 1 \quad (49)$$

9.2 Electrocardiogram (ECG) Reconstruction

To test the new PDR data converter to acquire and reconstruct an analog signal from its nonuniform samples, a simulated electro-cardiogram (ECG) signal [25] was used. These signals have a wide dynamic range and “contain the QRS complex, which ensures oscillations near the Nyquist rate” [26] and are thus useful in exercising the nonuniform sampling architecture of the PDR.

The simulated ECG signal was modeled with a 1mV peak amplitude with 0.3mV DC offset and zero mean Gaussian noise with noise variance $\sigma_{No} \approx 0.058nW$ was added to the signal.

A wide view of the simulated run time of 10 seconds is shown in Figure 17, which demonstrates the PDR’s ability to maintain the count. A zoom view of the reconstructed signal is shown in Figure 18, highlighting the features of the QRS pulse. When the signal is “idle” (DC like), and loiters near 0 Volts, the system continues to generate samples and is not adversely affected by the lack of signal dynamics.

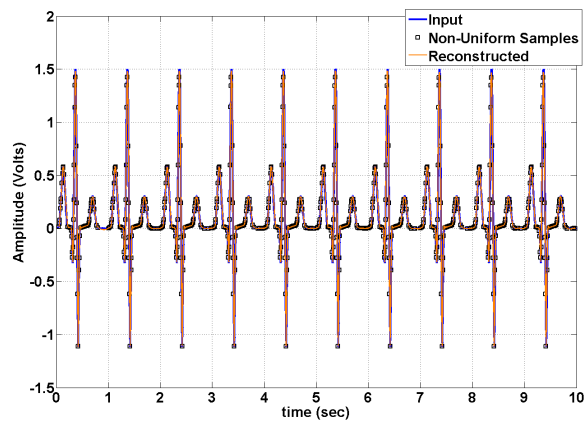


Figure 17: Reconstructed (simulated) ECG: Wide View

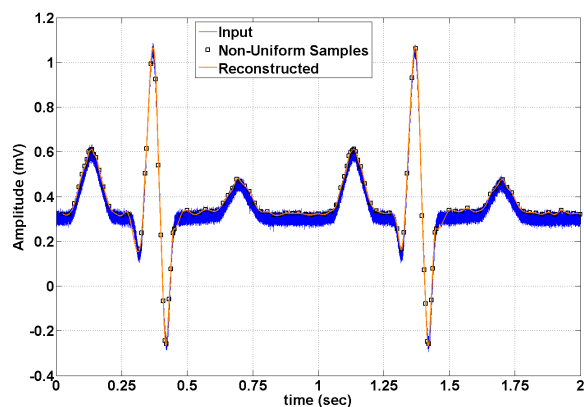


Figure 18: Reconstructed (simulated) Noisy ECG: Zoom View

10 Conclusion

A novel Analog-to-Digital Converter architecture based on partitioning the signal amplitude axis as a geometric series has been described. A detailed analysis of the design requirements to achieve the geometric partitioning has been provided and the essential circuits to realize the design presented. To extract the information content in each nonuniform digital sample, a proposed format of the nonuniform data was established, where it was shown that the partition number must be included in the digital word. Using reset Δ -PPM was shown to cause the system to satisfy the Nyquist requirement on average, and the geometric partitioning was shown to cause the SQNR to attempt to remain approximately constant. Lastly, the linearity of the PDR and the reconstruction of a simulated ECG signal were illustrated through simulation.

Conflict of Interest The authors declare no conflict of interest.

References

- [1] C. Pappas, "A New Non-Uniform ADC: Parallel Digital Ramp Pulse Position Modulation" in 31st IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Quebec Canada, May 2018. <https://doi.org/10.1109/CCECE.2018.8447844>
- [2] S. Maymon, A. Oppenheim, "Sinc Interpolation of Nonuniform Samples" *IEEE T SIGNAL PROCES*, 59(10), 4745–4758, 2011. <https://doi.org/10.1109/TSP.2011.2160054>
- [3] D. Donoho, P. Stark, "Uncertainty Principles and Signal Recovery" *SIAM J. Appl. Math.*, 49(3), 906–931, 1989. <https://doi.org/10.1137/0149053>
- [4] E. Candes, T. Tao, "Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies?" *IEEE T INFORM THEORY*, 52(12), 5406–5425, 2006. <https://doi.org/10.1109/TIT.2006.885507>
- [5] Y. Eldar, G. Kutyniok, *Compressed Sensing Theory and Applications*, Cambridge University Press, 2012.
- [6] F. Marvasti, *Nonuniform Sampling Theory and Practice*, Kluwer Academic/Plenum Publishers, 2001.
- [7] J. Mark, T. Todd, "A Nonuniform Sampling Approach to Data Compression" *IEEE T COMMUN*, 29(1), 24–32, 1981. <https://doi.org/10.1109/TCOM.1981.1094872>
- [8] N. Sayiner, H. Sorensen, T. Viswanathan, "A Level-Crossing Sampling Scheme for A/D Conversion" *IEEE T CIRCUITS SYST*, 43(4), 335–339, 1996. <https://doi.org/10.1109/82.488288>
- [9] E. Allier, J. Goulier, G. Sicard, A. Dezzani, E. Andre, M. Renaudin, "A 120nm Low Power Asynchronous ADC" in ISLPED '05. Proceedings of the 2005 International Symposium on Low Power Electronics and Design, San Diego USA, 2005. <https://doi.org/10.1145/1077603.1077619>
- [10] T. Wang, D. Wang, P. Hurst, B. Levey, S. Lewis, "A Level-Crossing Analog-to-Digital Converter with Triangular Dither" *IEEE T CIRCUITS SYST*, 56(9), 2089–2099, 2009. <https://doi.org/10.1109/TCSI.2008.2011586>
- [11] M. Malmirchegini, M. Kafashan, M. Ghassemian, F. Marvasti, "Non-uniform sampling based on an adaptive level-crossing scheme" *IET SIGNAL PROCESS*, 9(6), 484–490, 2015. <https://doi.org/10.1049/iet-spr.2014.0170>
- [12] S. Qaisar, M. Ben-Romdhane, O. Anwar, M. Tlili, A. Maalej, F. Rivet, C. Rebai, D. Dallet, "Time-domain characterization of a wireless ECG system event driven A/D converter" in 2017 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Turin Italy, 2017. <https://doi.org/10.1109/I2MTC.2017.7969682>
- [13] S. Naraghi, "Time-Based Analog to Digital Converters", Ph.D Thesis, University of Michigan, 2009.
- [14] P. Maechler, N. Felber, A. Burg, "Random Sampling ADC for Sparse Spectrum Sensing" in 2011 19th European Signal Processing Conference, Barcelona Spain, 2011.
- [15] S. Becker, "Practical Compressed Sensing: Modern Data Acquisition and Signal Processing", Ph.D Thesis, California Institute of Technology, 2011.
- [16] M. Wakin, S. Becker, E. Nakamura, M. Grant, E. Sovero, D. Ching, J. Yoo, J. Romberg, A. Emami-Neyestanak, E. Candes, "A Nonuniform Sampler for Wideband Spectrally-Sparse Environments" *IEEE J EM SEL TOP C*, 2(3), 516–529, 2012. <https://doi.org/10.1109/JETCAS.2012.2214635>
- [17] R. Gray, D. Neuhoff "Quantization" *IEEE T INFORM THEORY*, 44(6), 2325–2383, 1998. <https://doi.org/10.1109/18.720541>
- [18] W. Kestler (Editor), *The Data Conversion Handbook*, Analog Devices Inc., Elsevier-Newnes, 2005.
- [19] C. E. Shannon, "Communication in the Presence of Noise" *Proceedings of the IRE*, 37(1), 10–21, 1949. <https://doi.org/10.1109/JRPROC.1949.232969>
- [20] H. Shapiro and Richard A. Silverman, "Alias-Free Sampling of Random Noise", New York University - Institute of Mathematical Sciences, 1959.
- [21] F. Beutler, "Error-Free Recovery of Signals from Irregularly Spaced Samples" *SIAM J. Appl. Math.*, 8(3), 328–335, 1966. <https://doi.org/10.1137/1008065>
- [22] R. Baker, H. Li, and D. Boyce, *CMOS Circuit Design, Layout and Simulation*, IEEE Press Series on Microelectronic Systems, Wiley Interscience, 1998.

- [23] W.R. Bennett, "Spectra of Quantized Signals" *Bell Sys. Tech. Journal*, 27(3), 446–472, 1948. <https://doi.org/10.1002/j.1538-7305.1948.tb01340.x>
- [24] Subbarayan Pasupathy, *Minimum Shift Keying: A Spectrally Efficient Modulation*, *IEEE Communications Magazine*, Vol: 17, Issue: 4, July 1979
- [25] Sophocles J. Orfanidis, *Introduction to Signal Processing*, Pearson Education/Prentice Hall, 1996.
- [26] Marco A. Gurrola-Navarro, "Frequency-Domain Interpolation for Simultaneous Periodic Nonuniform Samples" in 2018 IEEE 9th Latin American Symposium on Circuits and Systems (LASCAS), Puerto Vallarta Mexico,, 2018. <https://doi.org/10.1109/LASCAS.2018.8399937>

Enhancing and Monitoring Patient Outcomes Through Customized Learning

Majed Almotairi, Mohammed Abdulkareem Alyami, Yeong-Tae Song*

Dept. of Computer & Information Sciences, Towson University, 21252, USA

ARTICLE INFO

Article history:

Received: 21 December, 2018

Accepted: 31 January, 2019

Online: 13 February, 2019

Keywords:

Patient Education

Customized learning

Dublin Core Metadata

Learning Object

Learning Object Repository

ABSTRACT

Chronic diseases such as heart disease, cancer, diabetes, and asthma continue to increase in the general public within the modern era. With careful observation of the symptoms potential diseases may be detected early and managed properly. For that to happen, the awareness of the symptoms and proper knowledge about the diseases may be needed for each patient. To acquire such knowledge, patients may need to gather essential health information from a variety of sources such as the Internet, articles, or some type of e-learning systems. However, the amount of available information, often too much, which discourages patients to continue. In response to such scenarios, we propose an approach that delivers only the relevant information that is specific to each patient's condition. In our approach, we utilized observed symptoms and vital signs to identify potential diseases of a patient. As they use the system, their profile may be constructed to deliver patient-specific set of learning materials called a study plan. To monitor and promote their study, we developed a mobile application that allows patients to view their study plan(s) and to study the customized learning materials. Such customized learning allows patients to take control of their symptoms and potential diseases, which eventually helps them to improve their outcomes as a result.

1 Introduction

Chronic diseases such as heart diseases, cancer, diabetes, and asthma are continuously increasing [1]. As reported by the Centers for Disease Control and Prevention (CDC) [2], 60% of adults in the US have one chronic disease, 40% have two or more chronic diseases, and almost 40% of adults were considered obese in 2015-2016. Annually, over 1.7 million people are diagnosed with cancer, and more than a third die from it. The US health care budget on cancer is incessantly increasing and anticipated to reach \$174 billion by 2020. Chronic diseases require continuous treatment, which decreases the quality of life and increase medical expenses. According to the CDC [1], about 90% of The US health care expenditure is spent on individuals with chronic diseases and mental health conditions. Hence, chronic diseases are considered highly costly compared to other health problems. However, these diseases may be prevented or improved by early detection and proper management. According to many

of the randomized controlled trials, educational techniques may help decrease pain, increase coping skills, and decrease the primary care visits, which saves time and money [3,4]. Patients also need to learn about how and when to take their prescription drugs, and other information such as side effects. If patients do not follow the instruction on prescription, this may lead to undesirable results such as the status of disease getting worse, or even death, which eventually increases health care budget in the United States [5]. Therefore, it may help improve patients' outcome when care givers ensure that patients have read and understood the drugs' instructions.

When patients suffer from some symptoms such as chest pain, they would look for the information related to that. However, getting the proper knowledge about that can be quite difficult as patients may be overwhelmed by the vast amount of information that is available from various sources. They do not know where to start to educate themselves, as shown

*Yeong-Tae Song, Email:ysong@towson.edu

in Fig. 1. CDC and WebMD, for instance, provide a vast amount of Internet accessible information regarding prescription drugs, diseases and their conditions, healthy living, public health statistics, and more that can be utilized by general public [2].

However, the efficient consumption of the available information by patients can be difficult due to the amounts and types of information. To overcome such issues, we propose an approach that filters out irrelevant information and provide only the necessary information for the patient, i.e., customized patient education. In our study, we deliver patient education by providing customized learning materials based on their prescription drugs, clinical conditions, and potential diseases. Therefore, only the relevant information that is specific to each patient is delivered. The learning materials are queried semantically and delivered to the patients from the cloud based learning object repository. To ensure effective learning, patients' levels of understanding in health information may be gained to control the difficulty level of learning materials. In particular, semantic queries can be performed based on a patient's clinical condition and learning needs. Then the obtained results from the queries are filtered and organized based on the patient's knowledge level and relevancy. Based on the query results, a study plan can be created and organized from the learning object repository, which satisfies the learners' needs. In addition, we employed an assessment technique to monitor the patient's level of understanding on the learning materials and to ensure that he/she has read and understood the content of the learning materials. The remainder of this paper is arranged as follows: Section 2 discusses the literature review. Section 3 gives an overview of the proposed Patient E- Learning System (PELS), discusses the role of the system components, and presents the implementation of the suggested system. Finally, we conclude our study in section 4.

2 Literature Review

In the modern era, there are numerous contributions and researches about e-learning systems in various domains. For example, In [6], the authors offered a virtual medical school as an e-learning system that provides a problem-based e-learning environment that enables medical students and residents to access clinical cases by using the Hospital Information System (HIS). Ouf et al. [7] have used ontology and Semantic Web Rule Language (SWRL) to develop a smart e-learning ecosystem that delivers the associated research articles from different sources, such as IEEE library, Science Direct and Springer Link, to learners according to their needs (for instance preferred learning activities). In addition, other studies concentrated on ubiquitous learning [8-10], language learning, sharing of learning resources and mobile learning [11,12]. Mesquita and Peres [13] offered a Customized x-Learning Environment model that delivers customized learning materials by allowing learning to take place anytime

/ anywhere, depending on the student's needs and characteristics. However, the majority of the studies mentioned above rely on the learner profile generally in order to improve learner's achievements. Additionally, there is a lack of conducted studies on utilizing e-learning systems for patients and clinicians to use. Consequently, the nucleus of our research is to develop learning management system for clinicians and patients to use in order to improve patients health knowledge with an aim to improve patients outcomes. Despite the fact that there are many available resources (e.g. websites and mobile applications) that are reachable online for example PatientsLikeMe, MAYO Clinic, WebMD, iTriage, and others, patients are overwhelmed by the amount of learning materials. For example, PatientsLikeMe enables patients with the similar disease(s) to share information regarding their treatments, medical experiences, medication, outcomes, and more. However, room for medical errors exists while using this site since the patient is the one who generates the data. There are several reasons why these medical errors may take place, amongst them misconstruction due to limitation of using medical standards (e.g. patients may use of diverse terms to describe their problem), lack of using other factors that might be related to disease such as vital signs and demographic data, and concentration on diseases only [14]. In contrast, in our study, we utilized medical standards (e.g. RxNorm, SNOMED CT and ICD), vital signs and demographic data in order to provide learning materials to the patient according to his/her prescription drugs, diseases and symptoms and eventually avoid medical errors. In our study, an innovative learning approach is proposed by using customized learning. Based on this, a learning model is designed to achieve personalized learning experiences for patients. We can achieve this by retrieving the most proper learning objects and then providing them to proper patients. This, in turn, may encourage patients to educate themselves and reduce the time of seeking such learning materials from different resources.

2.1 Related Standards

This section describes the most common related standards that are available in the industry, and how they can be utilized in the PELS.

2.1.1 Patient Profile

In the modern era, there are various learner profile standards, including, but not limited to, Felder-Silverman Model, Kolb's Experiential Learning Model and Myers-Briggs Type Indicator, each individual standard is suitable for a specific learning domain [15]. However, Luciana et al. [16], and Ioannis et al. [17] reported that the two widely held standards that are utilized to define the learner profile are: Public and Private Information (PAPI), introduced by Institute of Electrical and Electronics Engineers (IEEE) [18], and Learner Information Package (LIP), introduced

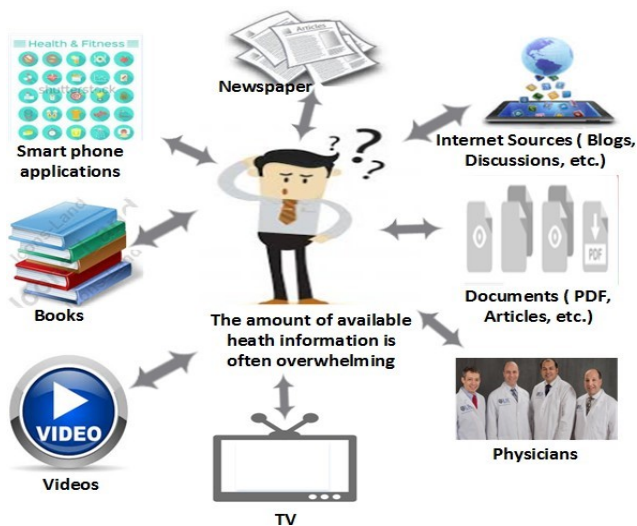


Figure 1: Patients are overwhelmed by the amount of available information

by Instructional Management System (IMS) [19]. The PAPI model contains six core classifications (personal information, security information, performance information, relation information, preferences information, and portfolio information) [18]. The IMS LIP model contains 11 core classifications (identification, affiliation, interest, accessibility, relationship, competency, activity, QCL [qualifications, certificates and license], goals, transcript, and security key) [19]. In addition, other extensions are available in each classification of these two standards that enable dealing with more possible characteristics [18,19]. Evangelou et al. [20] expresses that the PAPI model is generic and was developed to be utilized by variety types of systems or applications, but it does not afford information about dynamic characteristics of the learner profile. However, LIP is useful in the case that the system needs to be interoperable with other systems [20]. As mentioned above, the PAPI and LIP models comprise different categories to describe the learner profile, but none of these models entirely meet the PELS requirements, such as allowing clinicians to input their patients' symptoms and vital signs regularly. For that to happen, we developed a patient profile that involves specific characteristics to describe both dynamic and static information for each patient, which are utilized to gather patients' information in order to deliver customized learning materials to each individual.

2.1.2 Learning Object

A variety of learning object (LO) metadata standards are available in industry that are utilized to describe the LOs, which in turn, enables the ease of discovery and retrieval of LOs. An example of these standards, Learning Object Discovery and Exchange (LODE), was introduced by IMS Global (www.imsglobal.org). Another example of the standards, is called the Learning Object Metadata (LOM) developed by IEEE, which offers a technique to de-

scribe the profile of the LOs for use in e-learning systems [21]. Furthermore, one of the most common standards that is available is called the Dublin Core (DC), which also can be utilized to represent the LO profiles [22]. For generating and packaging the LOs, a well-known standard is called Common Cartridge (CC), which offers a standardized technique to create the LO content [23]. The CC also provides a standardized way to make interoperable e-learning systems and learning materials [23]. The CC utilizes the core components of the DC to describe the LO [22]. In our research, we utilized the CC standards that best fit our purpose. Most of the well-known e-learning systems use the CC to create the LOs as it offers standard methods to create LOs.

2.1.3 Medical Standards

Alyami et al. [24] summarizes the medical standards that are used at the present time. The ones that are related to our study are:

- Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT): a comprehensive standardized clinical terminology [24].
- International Classification of Diseases (ICD-9/ICD-10): used for epidemiology, health management and clinical purposes [24].
- RxNorm: used to provide normalized names for clinical drugs [24].

These three standards are used in the PELS for the clinicians use when they input data in the patients' profile (e.g. symptoms and prescription drugs).

2.1.4 Function Modeling Notations:

The US Air Force Program for Integrated Computer Aided Manufacturing (ICAM) presented various modeling notations that they have developed to provide better analyses and communication techniques in order to increase the manufacturing efficiency [25]. The main modeling notation that ICAM presented is called IDEF0 (ICAM Definition), which is based on Structured Analysis and Design Technique TM (SADTTM) [25]. IDEF0 is used to create a functional model that represents the functions, activities and processes that are involved in a system [25].

2.2 Limitations of Existing E-learning Systems

We examined the contemporary e-learning systems that are available for the public in industry, and the results show that there are more than 400 e-learning systems, either commercial, free or open source systems. In order to obtain functions that can be utilized for clinical education appropriately, we utilized the main functions from the free and open source well-known learning management systems, including Moodle, ATutor, Dokeos, ILIAS and Sakai. In addition, we took into

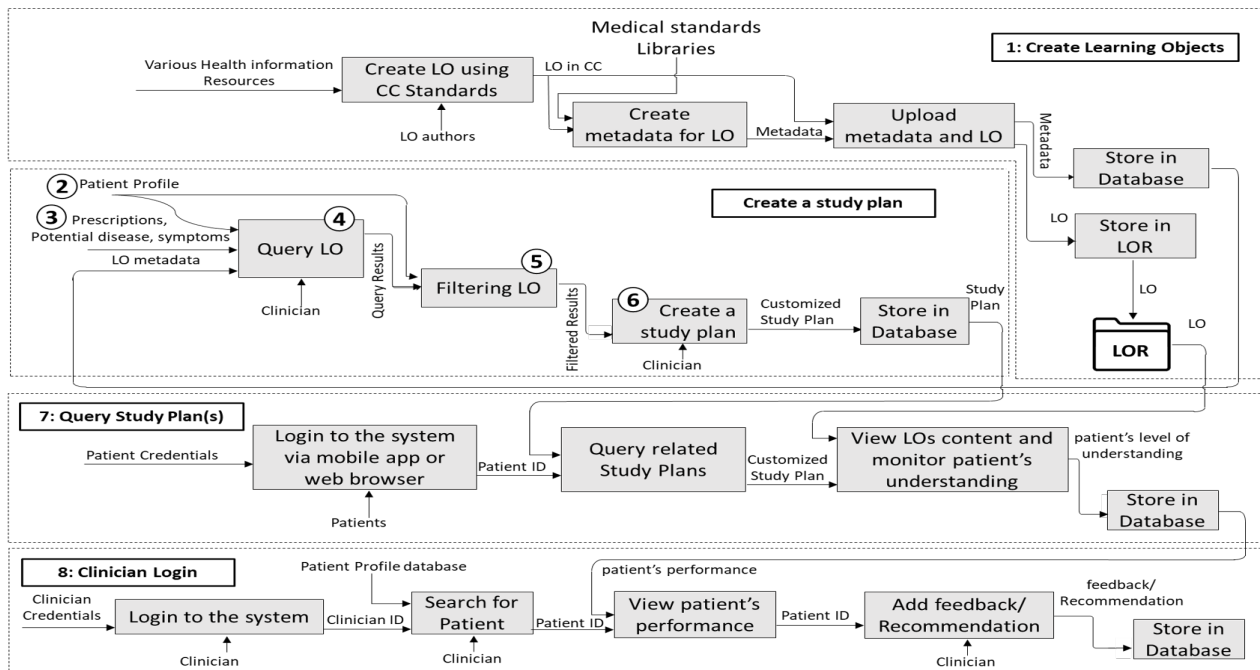


Figure 2: Main Functions of the Patient E-learning System (Numbers: 1, 2, etc. in this figure correspond to section 3)

account the other organizations that offer information for patients, such as CDC, WebMD and MAYO Clinic. The results indicate that there are many limitations of these systems, including:

- In the most of the e-learning systems, the instructor chooses particular content to be accessible by learners, which make the system a teacher-centered instruction.
 - Limited options are available for customized learning or filtering for learning materials that can be utilized to satisfy learners’ needs and goals. However, patients are required to use the search function within these e-learning systems to query the desired information.
 - Enrollment is required in most cases by each individual learner in a specific course to be able to contribute and view the course materials where the instructor is the person who evaluates the students’ performance.
- Limited development of learner profiles that may allow for saving and analyzing the current learner’s condition in order to offer learning materials based on their needs.
- MAYO and WebMD organizations have developed web applications that deliver clinical information to the users according to their inputs (e.g. prescription drugs and symptoms). However, they do not provide any assessment techniques to ensure that users have read and understood the content. In addition, they do not enable patients to save their information so they can monitor the changes of the patient outcomes.

To cope with these boundaries and limitations, we proposed new functions and improved some of the functions of existing applications to be used in the PELS as follows:

- To generate the learning object, we followed the CC standard for the creation of the learning objects, and we used the DC metadata to represent the learning object in order to facilitate the finding and retrieval of learning materials that are associated with the patients’ education.
- As none of the current learner profiles that exist in the current e-learning systems are appropriate for the patient education, we offered a patient profile with specified characteristics that can deal with static and dynamic information, such as patient’s vital signs and symptoms. We utilized patient profiles to enable the following:
 - To allow clinicians to save and update patients’ information (e.g. symptoms and potential disease[s]).
 - Dynamic retrieval of customized learning materials that meet patients’ needs according to the available information in patients’ profiles. Therefore, our proposed system is not similar to those that ask users to type keywords in the search textbox to retrieve their desired learning materials.
- Assessment techniques are conducted to ensure that patients have read and understand the learning materials.

Create a new exam

Select type of question

MultipleChoice

of Choices

4

Continue..

Figure 3: Exam creator tool for LO authors

- For the ease of use, we developed a mobile application that allows patients to view their learning materials straightforwardly.
- Although the PELS is an independent system, well-known standards (e.g. medical and meta-data standards) were utilized to record such data in our proposed system. Therefore, the data can be integrated with other systems if they utilize the same standards.

3 The Proposed Patient E-Learning System (PELS)

Patients and clinicians are the main users in our proposed PELS. For that reason, the limitations of the systems that provide information for patients and the current e-learning systems are taken into account. In addition to the functionality of our proposed system that was presented in [1], we also proposed more enhancements by adding more functions that are used to deliver learning materials to patients based on their need and capture their performance in order to insure that patients have learned the desired knowledge. To represent the functionality of the PELS, the IDEF0 notation [25] has been used to show the main components of the proposed system along with the main input(s) and output(s) of each component as shown in Figure 2. The following sections illustrate in detail how each component is used in the PELS.

3.1 Creating Learning Objects

We utilized the CC standard [23] in the process of creating the LOs which make the PELS organize and retrieve the LOs efficiently. In addition, the CC standard utilizes the DC standard for the metadata representation. The processes of creating LO are shown in the Figure 2 section 1. First, the LO authors create LO using

Confirmation!

The exam has been Created Successfully!
Please copy the link below and paste it at the end of the learning object content.

```
<iframe src="http://www.viasoftech.com/LMS/Exam.aspx?ExamCode=288070c3-7cf4-49c5-a94a-289dd7358dc2" width="100 % " style="height: 100vh; "></iframe>
```

Copy the link

Figure 4: Example of the generated embed code

any content authoring tool that create LO in CC standards. While creating the LO, most of the content authoring tools allow authors to create exams and quizzes for each LO to ensure that learners have read and understood the content of the LO.

However, by using these authoring tools, it is difficult for the PELS to capture and save patient's answers, as there is no option to integrate LOs with the PELS (or third party systems in general) that allow developers to capture learners' inputs and store them in a database. Therefore, we developed an exam creator tool that allows LO authors to create an exam that can be managed by each clinician to capture their patients' performance. However, patients may not be able to answer questions that require typing, therefore, we only focused on multiple-choice and True/False questions that are easy for patients to answer and for the clinician to receive the final score without reading through the responses.

To create an exam, the LO authors create a title for the exam as first step, and then they choose the type of question(s) which can be either multiple-choice or True/False. If the author chooses a multiple-choice question, another menu appears that allows them to specify the number of choices as shown in Figure 3. The LO author can repeat this step several times to add the desired number of questions they want. When they finish adding questions, the exam creator tool generates an embed code that allow LO authors to embed the exam with the LO as shown in Figure 4. Once LO authors complete the process of creation the LO, they need to add metadata based on the DC standard to represent the LO. DC consists of fifteen core elements that can be used to define the LO profile more precisely [22]. In this research, we used the DC metadata elements to describe clinical LOs efficiently, as shown in Table 1. The LO authors are required to fill most of these elements in order to gain metadata records that can be queried efficiently. An example of the learning objects' metadata that are stored in the database is shown in Figure 5. The World Health Organization (WHO), that provides International Classification of Diseases (ICD-9/ICD- 10), grouped the potential diseases based on the body systems and other factors. In our study, we

Table 1: Metadata schema for the LO

Element	Description
Title	Learning object's title
Creator	Learning object's author
Subject	Keywords that outlines the LO
Description	Detailed description about the LO
Relation	The related medical standard code(s) that is related to the topic of the LO, which also can be used to indicate whether the LO is specific or generic
Date	The creation date of the learning object
Type	The category of the LO (e.g. symptoms, causes, diagnosis and drugs)
Identifier	The ID of the learning material
Language	The spoken/written language used in the content of the LO (e.g. English and Spanish)
Rights	Access rights and copyright of the LO
Source	The source of the LO

utilized these classifications to determine whether the LO is generic (e.g. body system) or specific (e.g. specific disease).

The PELS utilizes the Learning Object Repository (LOR), which is a directory that is used to organize and store all learning materials. Accordingly, once the LO authors complete filling the metadata of a LO and click “upload metadata and LO”, the system creates a separate sub-directory inside the LOR for each new LO with a unique label and adds this unique label to the LO's metadata to be efficiently and effectively queried.

The content of the LOs can be collected from various recognized health information resources. However, the creation of the LO can be collected manually using any e-learning content authoring tool that generates LO(s) in CC standard. In addition, other organizational websites (e.g. CDC [2]) provide Syndication API, that collect an enormous number of learning materials from their repository. All learning materials are stored and organized in the LOR to allow the PELS to use them and retrieve appropriate ones for the patient efficiently. During the exploratory analysis, we collected content for variety types of diseases, symptoms and drugs from different well-known health information resources such as CDC [2] and created LOs based on the CC standard.

3.2 Patient Profile

The Patient profile is developed to store patients' information regularly. As patients' information may include clinical information such as prescriptions, symptoms and vital signs, clinicians are involved in the process of adding the clinical information to the patient profile. In this study, we categorized the patient profile into two groups: dynamic (clinical information) and static (demographic information) information for each patient. The PELS utilizes patient profiles to gather information about each patient in order to deliver customized learning materials to each individual patient. The components of the patient profile can be listed as

Id	Title	Description	Lang...
1	Genital Warts S...	Genital warts ...	EN
2	How to prevent ...	Pertussis (wh...	EN
3	Vitiligo: Get the F...	Vitiligo is a co...	EN
4	Strep Throat: Sy...	Strep throat is...	EN
5	Key Facts About ...	Influenza (als...	EN
6	Pink Eye Facts: I...	"Pink eye" is ...	EN
7	The Best and W...	Picking the ri...	EN
8	Acute Maxillary S...	Acute sinusiti...	EN
9	Scabies: Images,...	Scabies is a v...	EN
10	What Causes Kid...	Kidney stones...	EN

Figure 5: Portion of the learning objects metadata

follows (see Figure 6):

- Patient demographic information: consists of full name, date of birth, gender, address, username, password, email address and phone number.
- Patient level of knowledge and preference: Level of knowledge is used to determine the learning materials' difficulty level that patient can understand. Preference contains content types such as text, video, etc., preferred language and input / output device.
- Patient vital signs: updated frequently, which includes respiratory rate, pulse, temperature, blood oxygen saturation and blood pressure.
- Patient symptoms: includes symptoms that patients observe, which may be used to determine the potential disease(s). Potential diseases: includes disease(s) that patients may have based on their symptoms.
- Prescriptions: includes prescription drugs that patients obtained.
- Study Plan: a collection of customized LOs for each individual patient to study.

On the other hand, clinicians' profile consists of the following:

- Demographic information: consists of full name, date of birth, gender, address, username, password, email address and phone number.
- Patient list: a list of patients that the clinician deals with. In order for the clinician to add a patient in their list, the patient's approval is required.

By using PELS, a patient signs up as a new patient and starts adding his/her main information and preferences (static information). However, patients may not be able to deal with clinical information. Therefore,



Figure 6: Components of the Patient Profile

the PELS limits adding clinical information only to clinicians. Consequently, a clinician needs to search for a patient first using search criteria, and then sends a request to the patient to add him/her to their list. The patient then will have the option to approve/reject the clinician's request. Once the patient approves the clinician's request, the clinician then can add clinical information to the patient profile.

3.3 Potential Diseases and Prescriptions

With the aim of creating study plans, the PELS needs to determine the potential diseases that patients may have, observed symptoms that they are suffering from, or prescription drugs that clinician prescribed. To collect such data, potential diseases can be entered in two methods: directly and indirectly. From the clinician directly, if the clinician already know the (potential) disease that patient suffers from. Indirectly, which can be obtained by using the Health Decision Support System (HDSS) that we have developed [24] which relies on medical standards such as SNOMED CT and ICD-9/ICD-10. HDSS assists patients and clinicians in identifying such potential disease(s) based on symptom(s) they provided. For instance, if a patient utilizes HDSS to determine his/her potential disease and enters his/her symptoms through the system, PELS can store this information on the patient profile, as shown in Figure 7. On the other hand, the prescription drug(s) need to be determined in the patient profile in order to educate patients about how to use specific drug, side effects when using this drug, and more. The PELS utilizes the RxNorm standard that allows clinicians to utilize standard terms for clinical drugs when adding them to the patient profile.

3.4 Query Module

As patients may not know how to deal with clinical terms or what is the right learning materials they should start with, clinicians are involved in this process to make sure that patients receive learning materials based on their needs. To query the metadata, Microsoft offers the semantic search that built upon the full-text

Potential disease(s)			
ICD-10 CODE	Description	Date Created	Delete
B86	Scabies	2018-12-07	Delete
N20.0	Calculus of kidney	2018-03-17	Delete
H10.89	conjunctivitis	2018-03-14	Delete
A37.90	Whooping Cough	2018-01-28	Delete

symptom(s)			
SNOMED CT CODE	Description	Date Created	Delete
128870005	Crusted scabies (disorder)	2018-12-07	Delete
95570007	Kidney Stones	2018-03-17	Delete

Figure 7: List of potential diseases and symptoms in the patient profile

search feature in the SQL server [26]. It extracts and statistically indexes the relevant key phrases. These key phrases are used to find and index the similar or related results [26].

The SQL server provides mechanisms that are utilized to implement the full-text search, for instance [26]:

- Stop list: deals with list of words commonly taking place in strings that do not help the search which should to ignored when performing a query e.g. "in" or "an".
- Stemmer: deals with the inflectional forms of a word, for instance, in the word "play" the full-text searches for the stems of the word, such as played, playing and player.
- Thesaurus: deals with the synonyms of a word, for example bike and bicycle.
- Word breaker: deals with the word boundaries as in the words "multi-millions" and "multimillions".
- Replacements: deals with the most frequently misspelled words such as Calendar/ Calender.

The clinician manages this process in order to ensure that patient receives the desired learning materials. In this way, amongst variety kinds of LOs that are stored in the LOR, and based on the available information in the patient profile, the PELS analyzes this information through the query module in order to deliver the proper learning materials to the patient. To do so, this module systematically queries the most associated learning materials that the patient needs based on his/her conditions. The query results (retrieved LOs) go through to the filtering module.

3.5 Filtering Module

The filtering module gets called after the query module is executed in order to retrieve customized learning materials based on the patient's profile. It filters the learning materials according to the patient's

Learning Object Title	View Contents
How to prevent Pertussis (Whooping Cough) (Not Completed!) This Learning material is recommended to you because it is associated with the potential disease Whooping Cough that is listed in your profile	View Contents
Pink Eye Facts: Identify Symptoms and Treat Pink Eye (Conjunctivitis) (New!) This Learning material is recommended to you because it is associated with the potential disease Conjunctivitis that is listed in your profile	View Contents
Key Facts About Influenza (Flu) Seasonal Influenza (Flu) (Completed!) This Learning material is recommended to you because it is associated with the potential disease Influenza (disorder) that is listed in your profile	View Contents

Figure 8: Customized study plan based on patient profile

preferences and/or the level of knowledge. It also eliminates learning materials that the patient has already completed in the past. In this manner, the module retrieves only the related learning materials for an individual patient, i.e., customized learning contents, to the patient. Figure 5 and Figure 8 show the differences before and after filtering.

3.6 Creating Study Plans

The study plan is a collection of learning materials that patients can study to educate themselves about a certain topic [27]. In the PELS, the clinician is involved in this process to verify the process of selecting learning materials. The created study plan module automatically generates a collection of learning materials according to the results received from the query and filtering modules in order to provide it to the patient. Then clinicians can verify the choices of learning materials as they have the authorization of adding/deleting learning materials from the study plan. When the clinician confirms the creation of the study plan, the PELS generates a list of recommended learning materials as an XML that contains information about the recommended learning materials, and it points out the reason of recommending each learning material. Figure 8 shows how the patients view the content of the study plan in PELS. The PELS allows multiple creations of study plans. Therefore, each time that clinicians update patients' conditions (e.g. add new potential disease), this module allows clinicians to generate a new study plan that contains a new list of learning materials in order to provide efficient learning materials to each individual patient according to their up-to-date conditions.

3.7 Query Study Plan

This module is about how the PELS queries the study plan(s) and delivers them to the right patient. Patients may not like the way that they need to use the personal computer/laptop to be able to sign in to the PELS and study. Therefore, we make the use of the

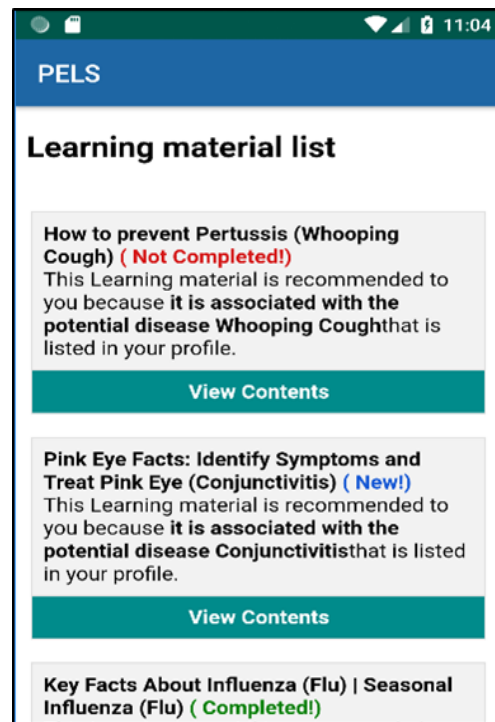


Figure 9: View study plan content through the mobile application

PELS even easier by developing a mobile application that allows them to view and navigate through their study plan(s) and display the content of each study plan. Figure 9 shows how patients view study plan(s) when they use the mobile application. In order for the PELS to query the patient's study plan(s), the patient first needs to log in using their credentials. Once the patient logs in successfully, the PELS queries the list of study plan(s) associated with this particular patient. Each study plan includes one or more learning material(s) that are approved by a clinician as shown in Figure 8. In order to monitor how patients consume the learning materials, the PELS categorizes three different statuses that appear next to the title of the learning object as shown in Figure 8. The time when each of these three statuses that appear can be illustrated as follows:

- **New:** appears when the patient still has not viewed the learning material yet.
- **Not Completed:** each learning material contains a quiz that needs to be completed by the patient to ensure that he/she has read and understand the content of the learning material. This status appears when the patient viewed the learning material but still has not attempt to complete the quiz.
- **Completed:** appears when the patient spent some time to reading the learning material and has completed answering the quiz.

If the status of the learning material remains new or not completed for a certain period of time (e.g. a week),

the system sends a reminder notification to the patient saying that the learning material needs to be completed. However, at the end of each learning material

Patient Name: *Amotairi, Maged*
 Study Plan Title: Flu, and other symptom
 Total Learning materials in the Study Plan: 3 Learning material(s)

Current Progress for each Learning material

LO Title:Key Facts About Influenza (Flu) | Seasonal Influenza (Flu)
 LO Status : Completed!
[View Exam Results](#)

LO Title:How to prevent Pertussis (Whooping Cough)
 LO Status : **The Patient still has not completed the quiz yet!**

LO Title:Strep Throat: Symptoms, Causes, Diagnosis, Treatment
 LO Status : **The patient has not read the Learning material yet!**

Figure 10: Patient’s performance from the clinician’s view

content, there is a quiz, which has been included to encourage patients to read the learning materials and make sure that they understand its content. Therefore, the PELS allows patients to take the quiz multiple times if needed. Accordingly, when the patient chooses the wrong answer, the PELS shows a hint to tell the patient that he/she needs to change their response and find the correct answer. On the other hand, when the patient views the content of the learning material, the PELS contains quantitative measures that capture the patient’s performance of metrics, for instance:

- Total time that the patient spent on reading the learning material.
- Total number of attempts that the patient made to complete the quiz.
- Total points scored on each attempt.

All the captured data are stored in the patient profile so the patient can see how he/she performed on each task. In addition, it can be reachable by the clinician so they can see how their patient performed in completing their tasks (see next section).

3.8 Clinician Login

Clinicians need to have their own access to the PELS so they can perform particular tasks, such as creating study plans and monitoring their patients’ performance. Therefore, the PELS provides certain views with their functionality that perform when the user signs in as a clinician. Accordingly, clinicians start signing in to the PELS using their credentials if they have already created their own account. If they want to deal with a new patient, they need to search for the patient using the search criteria and send a request to the patient to add him/her to their list. If the patient

approves the request, the patient will be added to the clinician list. Accordingly, the clinician can select any patient from their list in order to perform any task

Time spent on reading the Learning Material	00:05:06
Total attempts to complete the quiz	2 attempt(s)
Attempt # 1	
Submission Time	11/17/2018 9:19:27 PM
Total points scored	2 / out of 8
Attempt # 2	
Submission Time	11/17/2018 9:34:11 PM
Total points scored	8 / out of 8

Figure 11: Patient’s quiz result

such as adding a new potential disease(s) and recommending study plan(s). In addition, clinicians can view patients’ performance and see how they performed in a specific recommended study plan as shown in Figure 10. The PELS shows the status of each learning material whether or not the patient has read the learning material, has read the learning material but still has not completed the quiz or if the patient has completed reading the learning material and has answered the quiz. When the status is completed, the PELS shows an option to view the results, which includes the metrics that are mentioned in the previous section, shown in Figure 11. However, the PELS has an optional feature that allows clinicians to add feedback/recommendation for their patient when needed.

4 Conclusion

In this paper, we provide customized e-learning for patients with the aim of increasing their level of health knowledge and eventually improving their outcomes. For this to happen, we analyzed patient health data, such as observed symptoms and vital signs, to identify potential disease(s) for the given health data. Such factors are utilized to build patient profiles (such as patient preferences) that are used to provide only the relevant learning materials. Our proposed system, PELS, showed how to gather and organize heterogeneous learning materials using CC and DC metadata. To facilitate the retrieval of relevant LOs, the DC tags are utilized to retrieve such learning materials. In addition, we use an assessment technique to ensure that patients have read and understood the content of the learning materials. Moreover, we enhanced the usability of the PELS by developing a mobile application to facilitate the ease of use. Through this approach, we provide customized education that may help to improve patient outcomes. For future work, we plan to utilize an ontology-based patient profile analysis, which may increase the level of accuracy and relevancy of the search.

References

- [1] This paper is an extension of work originally presented in 2018 IEEE/ACIS 19th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel Distributed Computing (SNPD) Busan, 2018. <http://dx.doi.org/10.1109/SNPD.2018.8441054>.
- [2] CDC, "Centers for disease control and prevention," <https://www.cdc.gov/>, 2006, accessed: 2017-01-21
- [3] K. Jordan, N. Arden, M. Doherty, B. Bannwarth, J. Bijlsma, P. Dieppe, K. Gunther, H. Hauselmann, G. Herrero-Beaumont, P. Kaklamanis et al., "Eular recommendations 2003: an evidence based approach to the management of knee osteoarthritis: Report of a task force of the standing committee for international clinical studies including therapeutic trials (escisit)," *Annals of the rheumatic diseases*, vol. 62, no. 12, pp. 1145-1155, 2003. doi:10.1136/ard.2003.011742.
- [4] E. Superio-Cabuslay, M. M. Ward, K. R. Lorig et al., "Patient education interventions in osteoarthritis and rheumatoid arthritis: a meta-analytic comparison with nonsteroidal antiinflammatory drug treatment," *Arthritis Care and Research*, vol. 9, no. 4, pp. 292301, 1996. DOI: 10.1002/1529-0131(199608)9:43.0.CO;2-4.
- [5] M. L. McCarthy, R. Ding, N. K. Roderer, D. M. Steinwachs, M. J. Ortmann, J. C. Pham, E. S. Bessman, G. D. Kelen, W. Atha and R. Retezar, "Does providing prescription information or services improve medication adherence among patients discharged from the emergency department? A randomized controlled trial," *Annals of emergency medicine*, vol. 62, no. 3, pp. 212-223, 2013. DOI: 10.1016/j.annemergmed.2013.02.002.
- [6] F.-M. Shyu, Y.-F. Liang, W.-T. A. Hsu, J.-J. Luh, and H.-S. Chen, "A problem-based e-learning prototype system for clinical medical education," in *Medinfo*, 2004, pp. 983987. DOI: 10.3233/978-1-60750-949-3-983.
- [7] S. Ouf, M. A. Ellatif, S. E. Salama, and Y. Helmy, "A proposed paradigm for smart learning environment based on semantic web," *Computers in Human Behavior*, vol. 72, pp. 796818, 2017. <https://doi.org/10.1016/j.chb.2016.08.030>
- [8] S. Yu, X. Yang, G. Cheng, and M. Wang, "From learning object to learning cell: A resource organization model for ubiquitous learning," *Journal of Educational Technology & Society*, vol. 18, no. 2, p. 206, 2015.
- [9] G.-J. Hwang, C. Hui-Chun, S. Ju-Ling, S.-H. Huang, and T. Chin-Chung, "A decision-tree-oriented guidance mechanism for conducting nature science observation activities in a context-aware ubiquitous learning environment," *Journal of Educational Technology & Society*, vol. 13, no. 2, p. 53, 2010.
- [10] Y.-M. Huang and P.-S. Chiu, "The effectiveness of the meaningful learning-based evaluation for different achieving students in a ubiquitous learning context," *Computers & Education*, vol. 87, pp. 243253, 2015. <https://doi.org/10.1016/j.compedu.2015.06.009>.
- [11] Y.-J. Lan and Y.-T. Lin, "Mobile seamless technology enhanced oral communication." *Journal of Educational Technology & Society*, vol. 19, no. 3, 2016.
- [12] J. Abramson, M. Dawson, and J. Stevens, "An examination of the prior use of e-learning within an extended technology acceptance model and the factors that influence the behavioral intention of users to use m-learning," *SAGE Open*, vol. 5, no. 4, p. 2158244015621114, 2015. <https://doi.org/10.1177%2F2158244015621114>
- [13] F. Moreira, A. Mesquita, and P. Peres, "Customized x-learning environment: Social networks & knowledge-sharing tools," *Procedia Computer Science*, vol. 121, pp. 178185, 2017. <https://doi.org/10.1016/j.procs.2017.11.025>
- [14] J. H. Frost and M. P. Massagli, "Social uses of personal health information within patientslikeme, an online patient community: what can happen when patients have access to one another's data," *Journal of medical Internet research*, vol. 10, no. 3, 2008. <https://dx.doi.org/10.2196%2Fjmir.1053>
- [15] R. M. Felder and R. Brent, "Understanding student differences," *Journal of engineering education*, vol. 94, no. 1, pp. 5772, 2005. <https://doi.org/10.1002/j.2168-9830.2005.tb00829.x>
- [16] L. A. Zaina, J. F. Rodrigues Jr, and G. Bressan, "An approach to design the student interaction based on the recommendation of e-learning objects," in *Proceedings of the 28th ACM International Conference on Design of Communication*. ACM, 2010, pp. 223228. <https://doi.org/10.1145/1878450.1878488>
- [17] I. Panagiotopoulos, A. Kalou, C. Pierrakeas, and A. Kameas, "An ontology-based model for student representation in intelligent tutoring systems for distance learning," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2012, pp. 296305. https://doi.org/10.1007/978-3-642-33409-2_31
- [18] IEEE, "IEEE standard for learning technology - public and private information (papi) for learners," <http://metadata-standards.org/>, 2002, accessed: 2017-09-04.
- [19] IMS Global Learning Consortium Inc, "Ims learner information packaging information model specification," <http://www.imsglobal.org/profiles/lipinfo01.html>, 2001, accessed: 2017-09-07.
- [20] C. E. Evangelou, M. Tzagarakis, N. Karousos, G. Gkotsis, and D. Noutsia, "Augmenting collaboration with personalization services," *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, vol. 2, no. 3, pp. 7789, 2007. DOI: 10.4018/jwltt.2007070105
- [21] IEEE, "IEEE standard for learning object metadata," <https://standards.ieee.org/findstds/standard/1484.12.1-2002.html>, accessed: 2017-10-09.
- [22] DCMI, "innovation in metadata design, implementation & best practices," <http://www.dublincore.org/documents/dces>, accessed: 2017-10-02.
- [23] IMS Global, "Ims common cartridge specification," <http://www.imsglobal.org/activity/common-cartridge>, 2008, accessed: 2017-11-12.
- [24] M. A. Alyami, M. Almotairi, A. R. Yataco, and Y.-T. Song, "Health decision support system based on patient provided data for both patients and physicians use," Presented in *ACM IMCOM 2018, 12th International Conference on Ubiquitous Information Management Information Management and Communication*. <https://doi.org/10.1145/3164541.3164632>
- [25] F. PUBS, "Announcing the standard for integration definition for function modeling (idef0)," *Draft Federal Information Processing Standards Publication*, vol. 183, 1993.
- [26] Microsoft, "Semantic search (sql server)," <https://docs.microsoft.com/en-us/sql/relational-databases/search/semantic-search-sql-server>, accessed: 2017-06-12.
- [27] Y.-T. Song, K.-e. Park, and Y. Yoon, "Ontology based learner-centered smart e-learning system," in *Proceedings of VI International GUIDE Conference*, Athens, 2013.

Robot-Assisted Posture Emulation for Visually Impaired Children

Fang-Lin Chao^{*1}, Hung-Chi Chu², Liza Lee³

¹Department of Industrial Design, Chaoyang University of Technology, 436, Taiwan R.O.C.

²Department of Information and Communication Engineering, ³Department of Early Childhood Development & Education, Chaoyang University of Technology, 436, Taiwan R.O.C.

ARTICLE INFO

Article history:

Received: 20 October, 2018

Accepted: 08 February, 2019

Online : 14 February, 2019

Keywords:

Posture movement

Small robot

Touching

Visually impaired

ABSTRACT

This study proposes robot-assisted posture emulation for visually impaired children. The motor of a small robot (low torque) can be controlled using our palms. A user does not risk injury when the robotic hand is directly touched. The dimensions of the body of a commercially available small robot are different from those of a person. We adjusted the length of the upper arm to easily distinguish the movements of the upper and lower limbs. Adults and children were requested to perceive the robot's movements through touch and imitate its action. The study demonstrated that visually impaired subjects enjoyed playing with the robot and the frequency of body movements increased in robot-assisted guidance. The majority of the children could identify the main posture and imitate it. The scoring of continuous movement is medium. A stand-alone concept design was proposed. The design can present the main actions of the upper body and prevent body dumping when touched. Main torso and hand movements can reveal the body language of most users.

1. Introduction

The benefits of play facilitate integration, survival, and understanding. Play supports flexibility in thinking, adaptability, learning, and exploring the environment. These abilities are essential for developing social, emotional, and physical skills [1]. The ability to pretend play reveals a child's cognitive and social capacities [2].

1.1. Background

Effects exerted by visual impairment include psychological, social, mobility, and occupational effects. The psychological effect refers to negative consequences such as the loss of self-identity and confidence [3]. Play and exercise can create scenarios that promote social inclusion.

For parents, mobility is the main concern. Limited mobility affects a child's participation in daily physical activities. In the absence of guidance, children experience frustration [3]. An Indian dancer conducted a workshop for visually challenged and sighted underprivileged students at Acharya Sri Rakum School [4]. School children were observed to prefer such guidance and interactive teaching. For visually challenged children, the use of canes while moving is an additional psychological barrier [4].

*Fang Lin CHAO, Email: flin@cyut.edu.tw

Several observations were noted in the special school for visually impaired. As shown in Figure 1(a), many students preferred staying at their desk during the class break. They were concerned regarding potential risks that can arise in outdoor activities and therefore tended to stay at their desk.

Students did not exhibit considerable behavioral changes in a typical classroom. Figure 1(b) shows students with a downward looking posture while conversing and limited body expressions. The absence of body language results in a loss of a communication channel with others. Slow physical movements suggest low confidence. Figure 1(c) depicts children being group guided to reduce uncertainty. They supported each other and followed the footsteps of the previous student. In this case, students missed the opportunity to explore surroundings in a safe environment independently. Some students can be unfit, which causes a difficulty in physical balance and coordination. Individual guidance is required to encourage such students to exercise. Furthermore, it is necessary to provide accessories and an environment that can provide safe opportunities for physical exploration.

Gestures in this study included simple movements: moving the hand left and right indicated rejection and pressing down indicated affirmation. Silent gestures produced by English speakers enacted in motion elements [5]. Humans rely on motion events when they

convey those events without language. Research on gestures demonstrated a strong correlation between speech and body in language use [6].



(a)



(b)



(c)

Figure 1 Behavior observations of visually impaired children: (a) student in a regular classroom, (b) lack of posture expression during talk, (c) group guiding by teacher

Gestures play an essential role in language learning and development [7]. Gestures serve as a critical communication channel between people. Some dancers use hands to touch the body of an instructor to illustrate movements; this may cause uncomfot to the instructor. Blind children use gestures to convey thoughts and ideas [7]. Gesture is integral to the speaking process. Corresponding postures and other sensations can fulfill the learning gap.

1.2. Related existing studies

Physical exercise can reduce feelings of isolation in blind children and strengthen their friendship with each other. Figure 2 depicts a linear servo structure constructed using five motors [8] to emulate the main torso or hand posture. A solid bar was fixed with a servo motor by using a joint. Joints extend the flexibility of structures and can perform a curvature movement. A student requires two hands together to perceive a posture.

Robot dance has been developed since 2000. Many programmable humanoid robot dancers (Figure 3) are available. A dance-teaching robot was proposed as a platform (Figure 4) in

Japan. An estimation method for dance steps was developed in [10] by using time series data of force. A study [11] presented a physical human–robot interaction that combined cognitive and physical feedback of performance. Direct contact enables synchronized motion according to the partner’s movement. A role adaptation method was used for the human–robot shared control [12]. The robot can adjust its role according to the human’s intention by using the measured force.

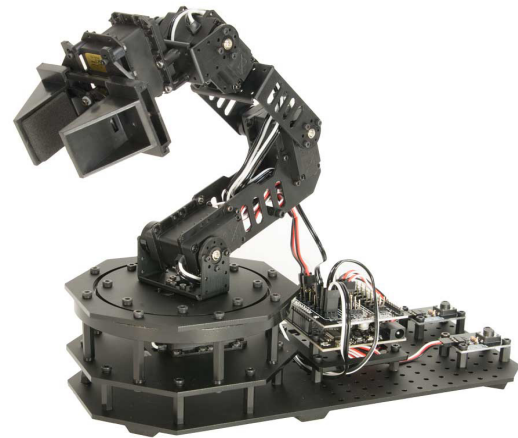


Figure 2. Linear servo structure emulates body posture with linkage bar [8].



Figure 3. Humanoid robot dancer [9]



Figure 4. “Robot Dance Teacher” of Tohoku University [10]

Teaching activities that require physical contact and complex motion were investigated in [13]. The methodology presents teaching through the following physical human–robot interaction: learning is enhanced by providing a constant feedback to users and by making a progressive change in the interaction of the controller. A robot was used by a special education teacher and physical therapist [14]. An external device encouraged a dancer to perform more energetically to the rhythm of upbeat songs. Children were not instructed to mimic the movements of a human or robot dancer, which is the common practice of a normal dancer.

2. Methodology

2.1. Motivated posture display

A small commercially available robot [15] (Figure 5) was used as motivation for posture expression. The trial was based on three considerations: (1) provide haptic attention, (2) encourage posture response, and (3) fit the available school budget. When children perceive the posture of a body, they imitate the movements of the body. The small size of robots was suitable for activities involving touching with the hands.



Figure 5. BeRobot [15]

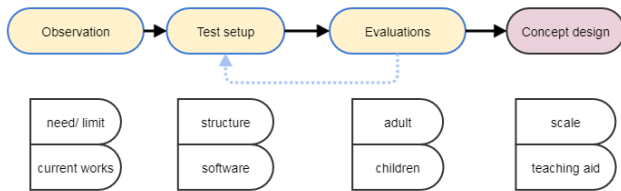


Figure 6. Study structure

The main controller was placed in the central square part. Two feet and hands with servomotors extended from the body; the lower legs were wide and thick. This bigfoot structure was intended to reduce the center of gravity. The length of the leg and arm was not consistent with the ratio of that of a human. Although the robot can emulate the body posture, adjustments are necessary to ensure proper interaction with children.

2.2. Structure of studies

Qualitative research involves interviews and observations. Students and a teacher joined the practice. Then, three observers filled observation forms accompanied by video recording. First, the test was conducted by a regular adult wearing an eye mask.

A supporting frame, which consisted of a steel frame attached with an adjustable clamp to hold a micro-robot's body, was provided to prevent the robot from falling (Figure 7). The height of the bracket was 30 cm, and hooks were fastened at the upper and lower ends of the body. Although the body was buckled, the limbs were free to move. Therefore, the user could safely touch the limbs without worrying about falling.

Figure 6 shows observation steps, evaluation procedures are:

- explain the procedure
- touching the upper part of the robot's body

- emulate the corresponding posture, evaluate the response
- correcting the response with vocal or physical guidance
- reaching the lower body, follow the posture
- perform whole-body posture

2.3. Motion edit

BeRobot is equipped with a robotic motion commander, which is a graphic interface that enables a user to control motions such as twisting, bending down, bending knees, walking, and standing. Figure 7 displays the motion commander interface.

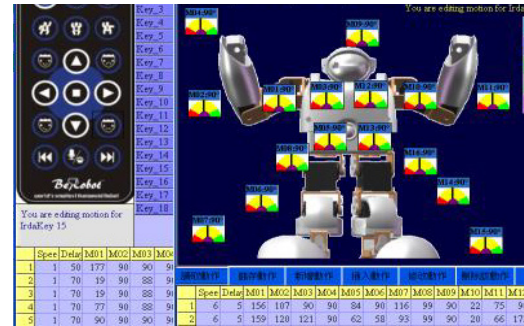


Figure 7. Motion commander of remote control mode and interactive mode

A program mode was included for the complex movement design. As depicted in Figure 8, we designed a simplified control interface with fewer parameters to enable the school teacher to overcome the barrier of motions. We used an USB interface to control the robot with a computer. As indicated in Figure 8, the section editor program was developed using Microsoft Visual Studio. By selecting the movement type, users produced desired simplified postures. The midpoint was the initial state of the pose. An ordinary action consists of a few simple hand and foot movements. When we adjust to the correct position, the angle of each servo motor can be transmitted through the USB port and stored in the memory. The user can manually control the timing by directly clicking to switch among different postures. Ordinary movements are achieved by combining simple postures.

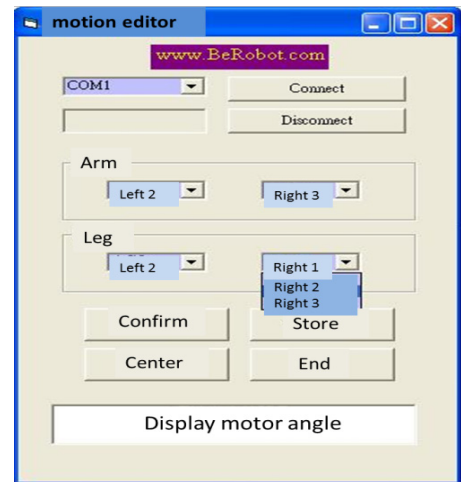


Figure 8 Motion editor: interface for interactive editing of posture

3. User test

3.1. Adults test: object preparation and task

The size of the robot (length × width × height) was 6.5 × 10.2 × 15.5 cm. To enhance movement detection, we increased the length of the upper arm. We used laser cutting to create a steel strip

extension. The dimensions of the upper and lower arms were close to those in the real situation.

We invited college students to detect a potential problem. The purpose of the test was to gauge the body sensitivity through touch. Observers were trained graduate students. All interview sessions were recorded on video. Observers viewed the videos and characterized them.

We first demonstrated a few postures (e.g., open hands and one hand straight down), and the subject attempted to replicate similar actions. We requested the students to touch robot in a mild manner, the achievement rate was high. During the test, the subject was allowed to verbally state their intentions, which allowed us to understand their feelings. Second, continuous motion was preferred. They expressed positive experiences and motivated physical movements.

- Adjusted the robot one movement at a time
- Extended the duration of touch, in which time to imitate each of the postures was long
- Changed the robot's orientation from back to front to ensure that a student's direction was the same as that of the guiding robot
- Regular pauses were added to ensure that subjects perceived the movement
- Vocal and physical guidance was provided during the tests.

3.2. Adult test results

College students wore eye masks to ensure that they perceived motion through touch. To avoid the scenario of testers influencing each other, we isolated subjects by testing one person at a time. The picture and video indicate that most of them were happy. Because their palms were large, they could perceive the overall posture. The majority of college students could repeat the presented posture.

During the continuous operation, interference between the palm and the robot affected the clarity of perception of a movement. When a switch occurs in the action, grip of the human with the robot may detach. Some subjects self-evolved according to the final posture. Therefore, direct touch was restricted during the practice of a static posture. Figure 9 depicts the interactions. Findings of experiments, in which a regular adult wore an eye mask, are as follows:

(a) A 22-year-old girl smiled. Although her palm was sufficiently large, she held the robot's hand. She seems to be looking forward to being led by the move.

(b) Girl-2 made a fist in one hand and moved smoothly backward in the opposite direction. Because the subject was face to face with the robot, she questioned whether she should mimic the robot in the same direction.

(c) Male subjects mimicked the movement of flying wings. Although the robot was small in size, it could perceive the basic posture of the flying movement.

A semi-structured acceptance interview was conducted, and its responses are presented as follows:

(1) Can you know the corresponding body parts of the robot?

Before beginning the activity, guiding tutors specified the robot's body parts by encouraging participants to touch the part of

the robot being introduced. Thus, users could identify body parts, such as the head and hands, of the robot. In the case of blind children, a tutor was attentive to rejection situations. Blind children are reluctant to touch unfamiliar objects. A guided story allows them to gradually relax during the test.



(a)



(b)



(c)

Figure 9 Supporting structure and its test situation where user interact with a programmed controlled robot: (a) touching, (b) the response, (c) response with both vocal and physical guidance

(2) Can you emulate the robot's movements?

Preliminary measurement results indicated that 80% of people could identify the movement. However, 20% of users were unable to emulate these actions in the allotted time. Although some participants could identify the parts, not all were able to identify the movement. Blind children require additional time to touch the robot.

(3) Would you recommend this activity for blind children? Any other proposals or ideas?

Approximately 80% of the measured responses indicated that the robot was small and the dimensions of the body did not match with those in a real situation. For example, the robot's head was relatively small and in contact with the main body. Participants indicated that the position of the robot's head should be clear to identify the corresponding body part.

3.3. Children's test: object preparation & task

Six visually impaired subjects aged between 7 and 12 years were selected. They are referred to as A, B, C, D, and E in this

discussion. All of them were students of a special education school for the visually impaired located in the mid-Taiwan region.

To avoid disrupting normal classes of students, we arranged to test six children during the classmate activity time. We conducted four teaching sessions for approximately 15 min for each student. A student's degree of acceptance and fluency of movement were recorded each week to measure the performance of the child after receiving physical aids. Researchers first narrated a story and invited children to experience the posture patterns of the robot (Figure 10). To avoid frustration, actions were arranged from simple to complex.

Compared with the test of an adult, visually impaired children were agitated when touching the stent. To reduce influence, we removed the stent and manually maintained the stability of the robot. The micro-robot can be fiddled in hand to increase the intimacy; the teacher familiarized the children with the robot.

Cautious touching extended the response time. An exercising behavior coding system was developed by modifying the system created by Moore, which consists of two behavioral categories: activity and emotion. Activity refers to a child's posture behavior and emotion refers to participation behavior. Specific items of observation were graded on a scale of 1–5, in which the following precautions were maintained:

- (1) Activity action: Touch the robot in a precise motion and appropriate manner without being reminded. You can touch the robot with fingers and palms rather than through arbitrary movements. Similar actions can be performed without assistance of robots.
- (2) Emotion: Emotion consists of psychological aspects such as excitement, feeling, cognitive processes, and behavioral responses to a situation. This refers to a child's happiness, anger, sadness, and joy.

Children can actively touch the robot and express positive emotions by using spoken language and body and facial expressions. From the beginning to the end of the activity, the children did not cry or use negative language. The children were fond of the event and appeared highly interested.

3.4. Children's Results

During the four weeks, both vocal and touch guidance were provided to enhance interactions. When removing the stent in the second week, a test required holding the robot to avoid the robot from tipping down. The records indicated the following aspects:

A-1 (student A, first week): When the student entered the classroom, he said he wanted to play with the robot. He touched the robot and smiled. When the music stopped, he stretched his hand to pull the robot's limbs, which caused changes in movement settings. After a reminder, he could imitate the action of "lifting both hands", "flying" and "swim."

D-1: The student did not imitate the robot's action and vigorously grabbed the robot's hand. After a reminder, he gently touched the robot with his hands and palms.

A-2: The facilitator held his hand, and the student touched the robot gently. However, the student gripped the robot once the facilitator left. The child could distinguish robot movements such as flat or low. However, he did not repeat every action and replied, "I won't."



(a)



(b)



(c)

Figure 10. Participant interacted with the robot through guidance.

D-2: The child gently touched the robot after being prompted. He followed the robot and moved around. When the music stopped, he emulated the robot's action. His facial expression was pleasant.

A-3: The student expressed that he wanted to touch the robot. After being prodded several times, he attempted to emulate the robot's movements. However, some of the actions were incorrect. He gently touched the robot and was able to identify the body part of the robot.

D-3: After verbal reminders, the student displayed initiative to imitate the correct robotic action. The subject had residual vision and could see some movements.

A-4: The student could grasp the arm of the robot with his hand and pulled the arm. He touched the robot lightly and attempted complex moves. He pulled the robot's arm, swayed around, smiled, and said, "Sister, look."

D-4: The student swung his limbs without assistance, and he could do the same action immediately after the robot stopped. He had many verbal interactions with the person recording the video and continued to smile.

The study indicated that visually impaired subjects enjoyed playing with the robot. Figure 11 presents the average results of independent observers. By manually analyzing recorded videos, we concluded that the frequency of bodily movements increased. The performance of emotional aspects was considerably higher than that of activity actions. Emotions were influenced by mentality. During the test, when children were familiar with teaching, the resistance considerably reduced. Activity action requires carefully distinguishing the details of the move which requires clearly tactile identification.

The demand for verbal guidance decreased week-by-week. The average value of the correct body movements gradually increased. However, the guide frequently reminded the subjects to properly touch the robot. Several students correctly imitated movements; the C-students could accurately and completely repeat the movement by the fourth week. In the emotional aspect, the student often displayed smiles, expressed their favorite robots, and wanted to touch the robots. The D-students displayed rich positive emotions in the fourth week.

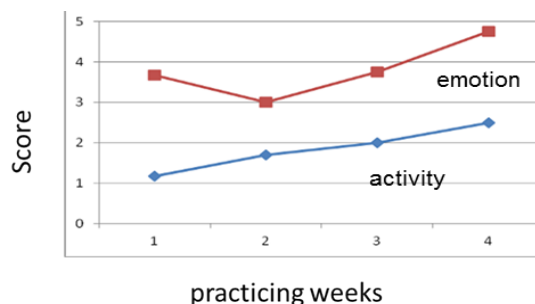


Figure 11. The evaluation results by four independent observers.

Two of the participants were absent for two weeks. The absence caused an analysis problem. Although analysis was performed in a small group, results revealed that the effectiveness of the robot depended on the teacher's guidance. The finding suggested that the interaction with a robot positively encouraged posture emulation behavior.

4. Discussion and Concept

In contrast with adults, visually impaired children exhibited emotional reactions or inadequate concentration because they were not able to view the setting. The mood affects the willingness to perform actions. Therefore, some children initially resisted and demonstrated a limited response in the first week. Therefore, teachers' encouragement was required to ensure children gradually participated. An auxiliary object of design plays an assisting role. Objects can assist teachers in initial activities, but they cannot replace the vital role of a teacher. "The guidance for properly touch the micro-robot" demonstrates the fragility of the micro-robot, and we propose the following design concepts for teaching.

4.1. Concept design

We selected a stationary base structure. Figure 12 illustrates changes in the body with the motor located on the torso and hands. The user attached the head to the upper cushion and held the robot's hands. Rotatable gears were present on the arm and elbow, and the simplest waist shaft requires five motors. The arm and the waist can be rotated sideways by adding more degrees of freedom, which reduces the cost of the robot. Because the magnitude of the movement was not large, the child will not hurt when he or she cautiously attached to the robot.

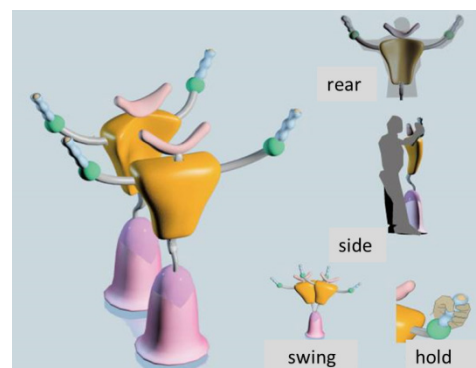


Figure 12. Perspective view of robot and user interaction, which emulates upper body posture, with a basement for stabilization

The user used this upper-body humanoid robot as a teacher's body. When a posture changes, the whole body can perceive the change. The motion control driver adjusts parameters to ease the learning of blind children and synchronize learning with music. This design incorporated motors with pressure sensors, which determined the speed of the response. Although the external force was large, it demonstrated that a user could keep up with the current action and slow down to avoid danger.

4.2. Comparison with existing work

Emulating the posture of the entity assists students in understanding the meaning of posture expression. The learning of correct postures aids communication with others. Most previous studies have focused on cutting-edge dance-teaching, which required complex control systems and mechanical constructions. For the vulnerable visually impaired group, it is difficult to obtain sufficient funds to procure such teaching aids. The visually impaired group requires a simple structure, which appropriately simplifies the design goal and transforms the complex dance into simple postures. Once familiar with gestures, movement learning can be expanded. This can be a useful tool for visually impaired people.

5. Conclusions

Dance is a continuous movement that involves a change in posture at different times. We evaluated posture displays through direct contact. The small robot motor has a low torque, which can be controlled by the strength of our palms. The user does not risk injury when the robotic hand is directly touched. The body's dimension of a small robot is different from those of human. We adjusted the length of the upper arm to easily distinguish the movements of the upper and lower limbs. This setup demonstrated positive results in visually impaired children. They attempted to emulate the posture of the robot. A standing concept of the torso and hand movements was presented as a teaching aid.

Acknowledgment

This research was supported by the National Science Council, Taiwan, R.O.C, under grant NSC 99-2221-E-324-026-MY2. We gratefully thank research students and the Special Education School in Taichung for arrangement of field test.

References

- [1] Case-Smith, Jane, Anne S. Allen, and Pat Nuse Pratt, eds. *Occupational Therapy for children*, St. Louis: Mosby, 2001.

- [2] Stagnitti, Karen, and Carolyn Unsworth. "The importance of pretend play in child development: An occupational therapy perspective" *British Journal of Occupational Therapy*, 63(3), 121-127, 2000.
- [3] Roe, Joao. "Social inclusion: meeting the socio-emotional needs of children with vision needs" *British Journal of Visual Impairment*, 26(2), 147-158, 2008.
- [4] Visually impaired children and dance, [Online]. 2011 Available: <https://youtu.be/hYalrsMfU90>
- [5] Özçalışkan, Şeyda, Ché Lucero, and Susan Goldin-Meadow. "Does language shape silent posture?" *Cognition* 148, 10-18, 2016.
- [6] Lowie, Wander, and Marjolijn Verspoor. "Variability and variation in second language acquisition orders: A dynamic reevaluation" *Language Learning*, 65(1), 63-88, 2015.
- [7] Iverson JM and Goldin-Meadow S. "Why people posture when they speak" *Nature*, 396(6708), 228, 1998.
- [8] [Online]. 2019 Available:<https://www.trossenrobotics.com/robotgeek-snapper-robotic-arm>
- [9] [Online]. 2019 Available:<https://www.videoblocks.com/video/humanoid-robot-dance>
- [10] Takeda, Takahiro, Yasuhisa Hirata, and Kazuhiro Kosuge. "Dance step estimation method based on HMM for dance partner robot." *IEEE Transactions on Industrial Electronics* 54.2 (2007): 699-706.
- [11] Granados, Diego Felipe Paez, et al. "Dance Teaching by a Robot: Combining Cognitive and Physical Human-Robot Interaction for Supporting the Skill Learning Process." *IEEE Robotics and Automation Letters* 2.3 (2017): 1452-1459.
- [12] Y. Li, K. P. Tee, W. L. Chan, R. Yan, Y. Chua, and D. K. Limbu, "Continuous Role Adaptation for Human-Robot Shared Control," *IEEE Transactions on Robotics*, vol. 31, no. 3, pp. 672–681, 2015.
- [13] Granados, Diego Felipe Paez, Jun Kinugawa, and Kazuhiro Kosuge. "A Robot Teacher: Progressive Learning Approach towards Teaching Physical Activities in Human-Robot Interaction." *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (ROMAN 2016)*. IEEE. 2016.
- [14] Barnes, Jaclyn A. *Musical Robot Dance Freeze for Children*. Diss. Michigan Technological University, 2017.
- [15] [Online]. 2019 Available: www.berobot.com

Development of Application Specific Electronic Nose for Monitoring the Atmospheric Hazards in Confined Space

Muhammad Aizat Bin Abu Bakar^{*1}, Abu Hassan Bin Abdullah², Fathinul Syahir Bin Ahmad Sa'ad²

¹Faculty of Engineering Technology, Universiti Malaysia Perlis, Kampus UniCITI Alam, 02100, Padang Besar, Perlis, Malaysia

²School of Mechatronic Engineering, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600, Arau, Perlis, Malaysia

ARTICLE INFO

Article history:

Received: 20 August, 2018

Accepted: 04 February, 2019

Online : 15 February, 2019

Keywords:

Confined Space

Atmospheric Hazards

Electronic Nose

Purging

Pre-processing

Baseline Manipulation

Normality Test

PCA

SVM

RBFNN

ABSTRACT

The presence of atmospheric hazards in confined space can contribute towards atmospheric hazards accidents that threaten the worker safety and industry progress. To avoid this, the environment needs to be observed. The air sample can be monitored using the integration of electronic nose (e-nose) and mobile robot. Current technology to monitor the atmospheric hazards is applied before entering confined spaces called pre-entry by using a gas detector. This work aims to develop an instrument to assist workers during pre-entry for atmosphere testing. The developed instrument using specific sensor arrays which were identified based on main hazardous gasses effective value. The instrument utilizes multivariate statistical analysis that is Principal Component Analysis (PCA) for discriminate the different concentrations of gases. The Support Vector Machine (SVM) and Artificial Neural Network (ANN) that is Radial Basis Function Neural Network (RBFNN) are used to classify the acquired data from the air sample. This will increase the instrument capability while the portability will minimize the size and operational complexity as well as increase user friendliness. The instrument was successfully developed, tested and calibrated using fixed concentrations of gases samples. The results proved that the developed instrument is able to discriminate an air sample using PCA with total variation for 99.42%, while the classifier success rate for SVM and RBFNN indicates at 99.28% for train performance and 98.33% for test performance. This will contribute significantly to acquiring a new and alternative method of using the instrument for monitoring the atmospheric hazards in confined space.

1. Introduction

This paper is an extension of work originally presented in IEEE 13th International Colloquium on Signal Processing & Its Applications (CSPA 2017) in title of Electronic Nose Purging Technique for Confined Space Application [1]. A confined space is large enough for workers to enter and perform work. It has a limited means of entry or exit and is not designed for continuous occupancy because it's could contribute towards atmospheric hazard accidents. Accidents do happen sometime in these areas and usually involve human fatalities or death. The Occupational Safety and Health Administration (OSHA) and National Institute of Occupational Safety and Health (NIOSH) state that the presence of atmospheric hazards in confined space are serious environmental problem that threatens the industry operation and safety of the workers [2].

The hazards in confined spaces can be classified into two categories which are physical hazards and atmospheric hazards [3]. The physical hazards can be visualized and avoided by taking initial safety precautions. The example of physical hazards includes unstable materials, moving parts of machinery, falling objects, a slippery surface and noise. Atmospheric hazards are more dangerous compared to physical hazards as they are unseen and come from oxygen deficiencies, hazardous gases, dust and welding fumes. The hazardous gases can interfere with the human body's ability to transport and utilize oxygen as well as cause negative toxicological effects. Usually the atmospheric testing in confined space is carried out during pre-entry by authorised person using gas detectors. The atmospheric hazards are rated into three stages which are High, Moderate and Low.

The atmospheric hazards workers are exposed to in confined spaces normally involve oxygen (O₂) too low or high or the

* Muhammad Aizat Bin Abu Bakar, Email: ijatbakar@gmail.com

presence of flammable and toxic gases [4]. The main flammable gas is methane (CH₄) while the toxic gases include hydrogen sulphide (H₂S) and carbon monoxide (CO). Before workers enter the confined space, a pre-entry for atmospheric testing is conducted by the authorised person for safety requirement. The atmospheric hazard conditions must be monitored before and while workers perform their activities inside the area. The hazards can cause serious health problems or death to the workers if not monitored properly [5].

At present, the pre-entry for atmospheric testing is done by using a direct-reading gas detector that need to carry by the tester towards specific location [3]. The worker (tester) is exposed with the atmospheric hazards directly during pre-entry testing activity is carried out. Even do, there are multi gas detector in the market, but the acquired measured data still not reliable due to the purging system that cause low repeatability [6]. The detector also shows the real time measurement at specific location only which not represent the whole confined space environment. Therefore, there is a need for a system like e-nose that is able to measure the hazardous gases and predicts the atmospheric hazards in confined space with high accuracy and repeatability [7].

An electronic nose (e-nose) is an instrument which comprises an array of electronic chemical sensors with partial sensitivity, an appropriate pattern recognition system and capable of recognising simple or complex odours [8]. Development of this instrument over the past decades is significant for its possible applications and achievements [9]. The application includes food quality assurance, work safety, medical diagnosis, plant disease detection and environmental monitoring [10]. An e-nose has shown a good potential for detecting and monitoring atmospheric hazards present in the confined space which could contribute to deadly accidents. A good e-nose must be able to produce the same pattern for a sample on the same array to maintain its repeatability [7].

The main aim of this work is to develop an instrument to assist workers during pre-entry for atmosphere testing in confined space (i.e. hospital mechanical room) and which address the following:

- i. To investigate and identify the atmospheric hazards main hazardous gases.
- ii. To design and fabricate e-nose system with multimodal sensor detection.
- iii. To integrate the system with optimum self-purging.
- iv. To test and validate the functionality of fabricate system in laboratory and field environment.

2. Atmospheric Hazards

The atmospheric hazards can only be detected by sense of smell. The main atmospheric hazards in confined space are oxygen (too much or too little), flammable and toxic atmosphere [4].

2.1. Oxygen Atmosphere

The normal air in the atmosphere is approximately composed of 21% oxygen and 79% nitrogen [4]. The oxygen deficiency in confined space may happen when metals rust, combustion engines run for a period time is replaced by other gases (i.e. welding gases)

or used by micro-organisms (i.e. fermentation vessels). The enriched or over limit of oxygen in confined space is also dangerous in that it increases the risk of fire or explosion. Materials would quickly and easily burn when there is a high level of oxygen. Table 1 shows the potential effects of oxygen deficient and enriched atmospheres [11].

Table 1: Potential effect of oxygen enriched and deficient in atmosphere

Oxygen content	Effects and symptoms
> 23.5%	Oxygen enriched, extreme fire hazard
20.9%	Oxygen concentration in normal air
19.5%	Minimum permissible oxygen level
15% to 19%	Decreased ability to work strenuously may impair coordination and may cause early symptoms for persons of coronary, pulmonary or circulatory problems
10% to 14%	Respiration further increases in rate and depth, poor judgment, blue lips
8% to 10%	Mental failure, fainting, unconsciousness, ashen face, nausea and vomiting.
6% to 8%	Recovery still possible after four to five minutes. 50% fatal after six minutes. Fatal after eight minutes
4% to 6%	Coma in 40 seconds, convulsions, respiration ceases, death

2.2. Flammable Atmosphere

Three elements that termed the fire triangle are necessary for a fire or explosion to occur in confined spaces which are oxygen, flammable material (gas or fuel) and source of ignition (spark or flame).

For flammable material, Figure 1 illustrates the relationship between Lower Explosive Limit (L.E.L) and Upper Explosive Limit (U.E.L). The ignition sources of fire may start from open flames, welding arcs, lightning, sparks from metal impact, arcing of electrical motor or a chemical reaction [4]. Several industrial processes that generate static charge such as steam cleaning, purging and ventilation also have potential of fire hazards.

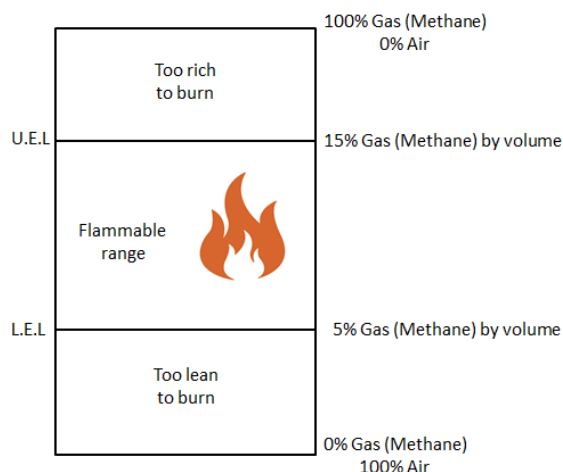


Figure 1: Lower Explosive Limit (L.E.L) and Upper Explosive Limit (U.E.L)

The L.E.L is the critical point at which ignition or explosion occurs [11]. Once the L.E.L is reached, the danger of a fire or

explosion is continuous all the way to U.E.L. For example the L.E.L of methane (CH₄) is 5% by volume and the U.E.L is 15% by volume. When a confined space reaches 2.5% of methane by volume this would be equal to 50% L.E.L which means that 5% methane by volume would be 100% L.E.L. Between 5 to 15% by volume, a spark could cause an explosion. Different gases have different percentage by volume concentrations to reach 100% L.E.L.

2.3. Toxic Atmosphere

The main toxic gases in confined space are carbon monoxide (CO) and hydrogen sulphide (H₂S) [4]. It may leak from products that are stored due to improper procedure or poor ventilation. The activities inside the area such as welding, painting, scraping and sanding may produce the toxic gases. The toxic fumes produced from nearby activities may flow and accumulate inside the area.

Workers expose to toxic gases in the atmosphere may be injured or killed by the gas. At certain concentrations, some toxic gases become too dangerous. At such levels, even a brief exposure can cause permanent health effects such as brain, heart and lung damage or the substance may render workers unconscious so that they cannot escape from the confined space. The Life-Threatening Effects of carbon monoxide and hydrogen sulphide are listed in Table 2 and Table 3 respectively [11].

Table 2: Effect of carbon monoxide exposed period

Exposure (ppm)	Time	Effects and symptoms
35	8 hour	Permissible Exposure Level
200	3 hour	Slight headache, discomfort
400	2 hour	Headache, discomfort
600	1 hour	Headache, discomfort
1000 to 2000	2 hour	Confusion, discomfort
2000 to 2500	30 minutes	Unconsciousness
4000	> 1 hour	Fatal

Table 3: Effect of hydrogen sulphide exposed period

Exposure (ppm)	Time	Effects and symptoms
10	8 hour	Permissible exposure level
50 to 100	1 hour	Mild eye and respiratory irritation
200 to 300	1 hour	Marked eye and respiratory irritation
500 to 700	½ to 1 hour	Unconsciousness, fatal
> 1000	1 minutes	Unconsciousness, fatal

3. Fundamental of Electronic Nose System

3.1. Electronic Nose

The e-nose as a device to mimic the discrimination of the mammalian olfactory system for smell is introduced in 1982 [12]. Initially the instrument used three different Metal Oxide Semiconductor (MOS) gas sensors to identify several chemical volatile compounds by using the response steady state signals. Then Gardner [13], described the e-nose development comprise of: (1) a matrix sensor to simulate the receptors of the human olfactory bulb, (2) a data processing unit that performs the same function as the olfactory bulb and (3) a pattern recognition system that would recognize the olfactory patterns of the substance being tested, which is a function performed by the brain in the human olfactory system. Figure 2 shows the similarity between human olfactory system and the e-nose system.

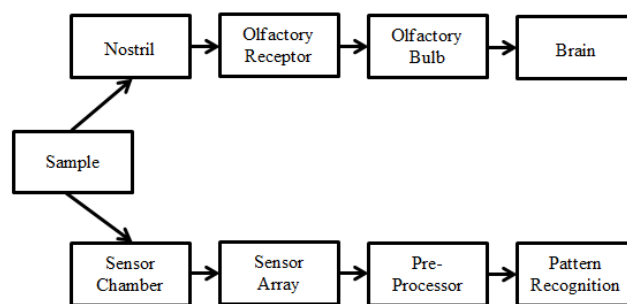


Figure 2: Similarity between human olfactory system and e-nose system

3.2. Sample Handling

The sample handling is a process to deliver the air sample to the sensor array. In the detection process, a sensor chamber is used to locate the sensor array where it interacts with the air sample. The data processing unit and pattern recognition algorithm will process the acquired data and classify them accordingly. The standard e-nose operating procedure includes determination of sample preparation and sampling method. The sample preparation procedure is determined by the type of sampling either static or dynamic method [14].

For static sampling, the sensor array and the sample is located at the same place in the chamber. The reading of sensor array is taken from the sample after the headspace in the chamber is homogenised. For dynamic sampling, the sensor array and the sample are at different or separate locations. The chamber is designed to locate a set of sensor arrays in one cavity and the sample is delivered into another cavity in order to be exposed to the sensors. This is also known as the Odour Capturing Module (OCM). The static sampling method is suitable to implement at a laboratory and not practical for outdoor use, while OCM is suitable to implement in both environments. However, the OCM must be purged to clean the chamber cavity after every sampling process to ensure sensor stability and repeatability for the next sampling process to produce consistent output readings [15].

Sample handling is a critical step affecting analysis by e-noses. The quality of the analysis can be greatly improved by adopting an appropriate sampling technique. To introduce the volatile compounds present in the headspace (HS) of the sample into the e-noses detection system, several sampling techniques have been used [16]. Various types of systems have been developed and are used to gather an air sample for analysis such an electric air pump which is used to suck the outside air sample into the sensor chamber [17]. A tube connected to a dedicated pump is generally used to direct the air sample into the e-nose. Other than that, the air sample delivery system was designed to provide seamless control over the operation by using a fan [18].

3.3. Sensor Chamber

A sensor chamber is used to accommodate the sensor array's interaction with the air sample. The chamber must be properly developed for optimum sensor response measurement. The design will emphasize optimum sensor response, stability, reproducibility and repeatability [19]. The air sample flow inside the chamber should be homogenous with low velocity to minimise the recirculating zones and stagnant volume [7]. This homogenous

flow is to ensure that the sensor arrays are exposed simultaneously to the air sample for optimum response measurement. The sensors positions in the chamber can be either in a series or parallel to the air sample. The different sensors position from the chamber inlet will cause a time shift in each sensor response to the air sample. Since the distance between the sensors is small, so the time shift effect can be ignored [20].

The symmetric structure will enable sensor arrays to respond simultaneously to the incoming air sample. However, this structure size is quite large to accommodate the sensor arrays in one area. The chamber development must focus on the structure geometry, material selection, type of flow and sensors position which will optimise its performance. The geometry of the chamber should be symmetrical to enable the air sample to be exposed to the sensor arrays efficiently. The flow of air sample is usually characterized using dimensionless ratio known as Reynolds number (Re). The Re is a measure of the kinetic forces to the viscous forces in a flowing fluid. The Navier-Stokes formula shown in Equation 1 is used to solve for air sample flow characteristics [21].

$$Re = \rho du / \mu \tag{1}$$

Where,

- ρ = density of the fluid (kg/m³)
- d = hydraulic diameter (m)
- u = fluid mean velocity (m/s)
- μ = dynamic viscosity of the fluid (kg/ms)

3.4. Gas Sensor

The gas sensor is used by the e-nose to interact with the air sample inside the sensor chamber. The interaction will cause a change in certain chemical and physical properties known as sensor responses that will convert into electrical signal. These sensor responses will be acquired by a laptop computer that links wirelessly with the e-nose to be used by an off-line signal processing method. The instrument sensors sensitivity should be suitable with the application's volatile compounds [12]. The sensors are also selected based on their characteristics which include fast response, stability, reproducibility and reversibility.

The principles of gas sensors are optical, thermal, electrochemical and gravimetric [22]. Most e-nose devices use electrochemical sensors to detect chemical substances to response. The operating principles are based on changes in the conductivity of sensing material by either adsorption or absorption of the gaseous molecules. The sensor conductivity depends on the material used and is relative to the sample concentration. The gas sensors performance is measured by several criteria and their behaviour include sensitivity, selectivity, stability, detection limit, responses time, recovery time, life cycle and operating temperature [23]. All these criteria are very important to take into consideration during the selection of the gas sensors. The MOS gas sensor was widely use selected based on its stable response in performance.

3.5. Microcontroller System

The e-nose normally uses an embedded controller which is a microcontroller with embedded software for operation, data acquisition and classification [24]. The control and system

software is embedded in the instrument memory that may perform concurrently during operation. A microcontroller is a component that is used to control the sampling process for acquiring the sensor response signals to be processed by a personal or laptop computer [25]. The microcontroller has good processing capabilities and a flexible interface to the instrument's components which make it the ideal choice as the controller. The microcontroller memories are RAM, Flash ROM and EEPROM. The Flash ROM memory is where the program is stored, also called program memory. RAM is used for the temporary data during run-time. The EEPROM is used for data memory that needs to be retained during power failure. This microcontroller memories ability makes it suitable for e-nose operation that consist multimodal sensor to send the signals concurrently.

Currently, the dsPIC33 microcontroller type from Microchip Technology Inc. as the embedded controller was used in e-nose system development [26]. The microcontroller was designed using Surface Mounted Technology (SMT) component to improve the performance by reducing the signal to noise ratio (SNR). The microcontroller also converts the acquired analogue signal to digital by using an on-board Analogue-to-Digital Converter (ADC).

3.6. Signal Conditioning

The e-nose signal conditioning frame work as shown in Figure 3 consists of an interface circuit, conditioning circuit and filter [27]. The interface circuit measures the sensor responses and converts it into output voltage that varies with conductive change in the sensing element [28]. A conditioning circuit is used to buffer and amplify the signal to suit the microcontroller requirement. A voltage follower is used as a buffer to isolate the signal and as impedance matching of the sensors output. The noise is removed from the signal by using a passive Low Pass Filter (LPF) because it is good for removing a small amount of high frequency noise (>2 kHz).

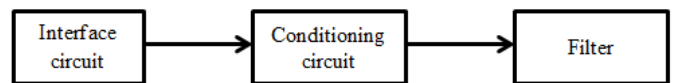


Figure 3: The signal conditioning frame work

3.7. Data Processing

The e-nose will produce a time-series sensor response corresponding to the air sample. The sensor response as data also depends on ambient air, temperature and humidity. The process consists of pre-processing, dimension reduction, pattern recognition and results as shown in Figure 4.



Figure 4: Electronic nose data processing block diagram

The chemo-metric processing technique that relates mathematical and statistical is used to extract the relevant information from the data [29]. This method uses multivariate analysis to discriminate the data simultaneously. The process uses either a statistical or biological classification method approach to classify or cluster the data qualitatively or quasi quantitatively. The

analysis methods are divided into parametric, non-parametric, supervised and unsupervised [30]. A parametric pattern is when the data are Gaussian and the distribution feature is normal. Supervised method develops a classification model using known data while unsupervised method uses unknown data for the development.

3.8. Pre-processing

The e-nose raw data is highly susceptible to noise, uncertain value, and inconsistency pattern. The quality of the data affects the classification model result. The data needs to be pre-processed to improve its quality which will in turn improve the classification process [31]. The process is critical for data processing that involves preparation and transformation of the initial raw data.

3.9. Feature Selection

The e-nose time series acquired data consists of dynamic or transient and steady state sensor response [32]. Figure 5 shows the transient and steady state sensor response. Feature selection is used to select a certain region of the sensor response that contains relevant sample information. Most of the e-noses extract the steady state region from the sensor response for the pattern recognition process.

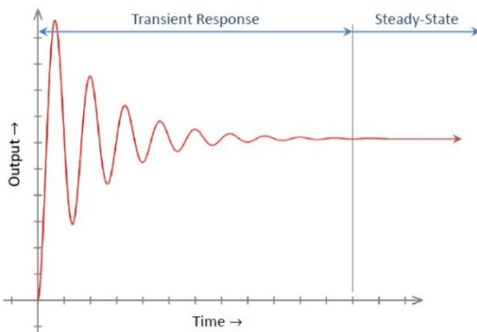


Figure 5: The transient and steady state sensor response

3.10. Baseline Manipulation

The baseline manipulation is a method that is based on the difference of the sensor response value between reference and sample [33]. The reference normally is the ambient air or nitrogen gas. Four baseline manipulation methods are commonly employed which are difference, relative, fractional and logarithm. The difference method is a widely used by directly subtracts the baseline and can be used to eliminate additive drift from the sensor response. Relative manipulation divides by the baseline, removing multiplicative drift and generating a dimensionless response. Fractional manipulation subtracts and divides by the baseline, generating dimensionless and normalized responses. The advantages of these processes will enhance the quality of data which contains the relevant sample information.

3.11. Normality Test

The procedure in assessing a sample of data has a normal distribution are graphical methods (histogram, boxplots and Q-plots), numerical methods (skewness and kurtosis indices) and normality test [34]. A normality test is a statistical process used to determine if a sample or any group of data fits a normal distribution or not to determine suitable types of data analysis.

A normality test can be performed mathematically for example the Shapiro-Wilk (SW), Kolmogorov- Smirnov (KS) or Anderson-Darling (AD) test [34]. The power of test method proportionally increased with sample size and level but was still low for a small sample. The Kolmogorov- Smirnov test is the best with high sensitivity in rejecting null hypothesis [35]. The Shapiro-Wilk is suggested being to be used for a normality test, but Kolmogorov-Smirnov is popular [36].

Basically the data skewness and outlier's features indicate non-normality distribution. The data non-normality distribution needs to be transforming in order to process using the parametric classification method [35]. But the alteration on the original non-normality data will create curvilinear relationships that will complicate the interpretation. The data should us a non-parametric classification method when the distribution is not normal.

3.12. Multivariate Statistical Analysis

The Multivariate Statistical Analysis is used to discriminate and classifier sample. It build discriminate and classifier models. Some of the multivariate analysis techniques are Principal Component Analysis (PCA) for discriminate and Support Vector Machine (SVM) for classifier. This technique is described in the following sub section.

3.13. Principal Component Analysis

The PCA is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. The PCA will transform the number of original data with correlated variables (number of sensors) into uncorrelated variables (Principal Components, PC) and loadings or weight of each original variable of variance percentage.

Traditionally, PCA is performed on a square symmetric matrix. It can be a pure sum of squares and cross product (SSCP) matrix, covariance matrix (scaled sums of squares and cross products) or correlation matrix (sums of squares and cross products from standardized data). The SSCP and covariance do not differ, since these objects only differ in a global scaling factor. A correlation matrix is used if the variances of individual variants differ much or if the units of measurement of the individual variants differ. Basically, PCA can be represented in five steps [37]:

- i. Normalize the data by subtracting the mean
- ii. Calculate the covariance matrix
- iii. Calculate the eigenvectors and eigenvalues of the covariance matrix
- iv. Choose components and form a feature vector
- v. Derive the new data set

The PCA result is shown by a two-dimensional (2D) or three-dimensional (3D) graphical plot that contains the original data's important information and shows the PC percentage and group cluster which is being analysed visually. The PCA is used as the dimension reduction technique as it is the most widely used in e-nose data analysis.

3.14. Support Vector Machine

The SVM is a supervised learning model that is based on statistical learning theory for data classification and regression

[38]. The SVM training used data with known class when designing a linear classifier hyper plane model for separate group. SVM is targeting clear maximum margins between the hyper plane and closest training. In addition, the SVM optimized classification decision function is based on structural risk minimization in order to avoid over fitting.

In more complicated for non-linear cases, SVMs employ the kernel trick where a positive definite kernel function is used to map the input data into a high dimensional transformed feature space and this method regularly used by the e-nose community that used multisensory system [39]. The SVM mathematical modelling and algorithm can be referring in [40]. Although SVMs were initially developed for solving two-class problems, but it is also can be easily adapted to tackle multiclass problems. The two common techniques consist of:

- i. One per class (OPC). This method is also known as “one against others”. OPC is simple and results in reasonable performance. K SVM classifiers are trained, each of which separates one class from the other (K-1) classes. For each measurement, x , to be classified, K SVM decision outputs $f_k(x)$, $1 \leq k \leq K$ are obtained. The class of measurement x , j , is determined as $f_j(x)/f_k(x) > 1$ for $(1 \leq k \leq K) / = j$.
- ii. Pairwise coupling (PWC). This method is also known as “one against one”. This method trains $(1/2)K(K-1)$ binary SVM classifiers. To classify a measurement, PWC combines the scores of these $(1/2)K(K-1)$ classifiers. Each of the $(1/2)K(K-1)$ binary SVM classifiers provides a partial decision for classifying a measurement. There are different methods of combining the obtained classifier and the most common is a simple voting scheme. When classifying a new instance each one of the base classifiers casts a vote for one of the two classes used in its training.

Since SVM does not require any estimation of statistical distributions of classes to accomplish the classification task, the application is widely used in many fields. For examples, the classifier performance is excellent in determination of green tea quality grades [41] and identification of selected features on Mexican coffee [42].

3.15. Artificial Neural Network

The non-linear e-nose sensor responses have to be converted to linear before they are classified using linear classifier methods. However, high volume of data from complex instrument’s sensor responses is difficult to produce an efficient classification model. The Artificial Neural Network (ANN) has good learning, generalization and noise tolerance that is suitable for the e-nose non-linear data [43]. It has the ability to learn the complex interaction between the array of sensors and the air sample [44]. The network’s ability to compensate the moderately sensitive sensors towards the air sample has improved the classification success rate.

The ANN is a mathematic model in which the nodes are interconnected with each other with weights and biases to form the network architecture. The architecture depends on the mathematical analysis of task criteria. The common network structure uses fully connected three-layer architecture models

which are made of input, hidden and output nodes. Figure 6 shows that an ANN model consists of the input layer, hidden layer and output layer.

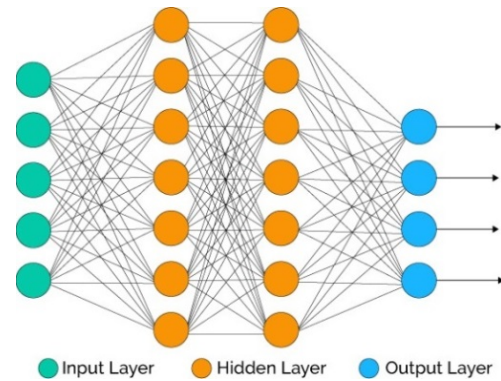


Figure 6: Basic neuron model showing the components and connections

Each neuron has a few weighted inputs and finally generates one output. The output is a function on these weighted inputs. Various types of functions can be used to calculate the output in a neuron. One type of neuron is called a perceptron which takes a vector of real-valued inputs and calculates a linear combination of these inputs. It outputs 1 if the result is greater than some threshold and -1 otherwise. Given inputs x_1, \dots, x_n , this perceptron function is shown in Equation 2 till Equation 5:

$$O(x_1, \dots, x_n) = \begin{cases} 1, & \text{if } w_1x_1 + w_2x_2 + \dots + w_nx_n \geq w_0 \\ -1, & \text{Otherwise} \end{cases} \quad (2)$$

Where,

- w_0 : a threshold
- w_i : real-valued constant weights
- x_n : inputs

An unthresholded perceptron is a linear unit, whose output is:

$$O(\bar{x}) = \bar{w} \cdot \bar{x} \quad (3)$$

Where,

- \bar{w} : weights vector
- \bar{x} : inputs vector.

Another commonly used neuron is sigmoid unit, whose output is:

$$O = g(\bar{w} \cdot \bar{x}) \quad (4)$$

Where,

- g : sigmoid function

The logistic form of sigmoid function is:

$$g(y) = \frac{1}{1 + e^{-y}} \quad (5)$$

A sigmoid unit is equal to a perceptron. Firstly, it calculates the linear combination of its inputs and then applies a threshold to the result. However, this threshold is continuous comparing with a

perceptron with a discontinuous threshold. The output of sigmoid function is between 0 and 1 (or between -1 and 1) and increases monotonically with its inputs. The sigmoid unit maps a very large input domain $(-\infty, \infty)$ onto a small range of outputs $(0, 1)$ or $(-1, 1)$.

A particular transfer function is chosen to satisfy some specification of the problem that the model is attempting to solve [45]. During training, the model learns from the input neuron and gradually adjusts its weight to reflect the desired outputs [46]. The network model is tested by using the randomised unused testing data. The process uses the unknown data to test the network's generalisation ability of the classification model. The performance of the network model is given by the Equation 6.

$$\text{Classification success rate} = \frac{\text{Correct classified data}}{\text{Total data}} \times 100 \quad (6)$$

Types of ANNs refer to the model learning methods: supervised, unsupervised and reinforcement. These include Multilayer Feed-Forward Perceptron (MLP), Fuzzy ARTmaps, Kohonen's Self-Organizing Maps (SOMs), Probabilistic Neural Network (PNN) and Radial Basis Function Neural Network (RBFNN). The MLP, PNN and RBFNN are supervised types of classifiers that are normally used for e-nose applications [47]. The classification results using RBFNN is the most stable with a high classification success rate [48].

3.16. Radial Basis Function Neural Network

The RBFNN consists of three layers includes input layer, hidden layer and output layer. The method approximation capabilities are based on the superposition of local models of the response system. The output layer only computes a linear combination of the activation of the neurons in the hidden layer. The input output multivariate relationship is given by using Equation 7 and also can be written in matrix form in [49]:

$$y_{ij}(x) = \sum_{f=1}^K \sum_{i=1}^N w_{mj} G(\|x_i - c_m\|) + b + e_{ij} \quad (7)$$

Where,

- y_{ij} is i, j -th element of the output matrix $Y_{N,K}$
- N is the number of observations (data)
- K is the number of outputs in the RBFNN (or responses)
- If denote M as the number of hidden neurons (or the RBF center's), w_{mj} is m, j -th element of the weight matrix $W_{M+1,K}$, $m = 1, 2, \dots, M$
- $M+1$ is the number of hidden neurons add bias (b)
- X_{NM} is the input patterns
- c_m is the square centroids matrix $M \times M$
- $E_{N,K}$ is the matrix of residuals of e_{ij}
- $Y_{N,K}(x) = d_{ij}$ is the process output

The activation of every neuron depends on the distance of the input vector to the prototype represented by the radial basis function and neuron parameter. The radial activation function network provides a nonlinear method of interpolating between numbers of different areas, time-series estimation and classification. The method always converges at the same point when trained with the orthogonal least squares algorithm. The

advantage of the classifier is that the network has no local minima problem [50].

4. Electronic Nose Development

4.1. System Structure

The e-nose system structure includes the sensing module, signal conditioning, microcontroller and embedded software. The instrument peripherals (i.e. an air pump, keypad, graphical Liquid Crystal Display (LCD), purging system, power supply and wireless Radio Frequency (RF) communication are controlled by the microcontroller) as shown by Figure 7.

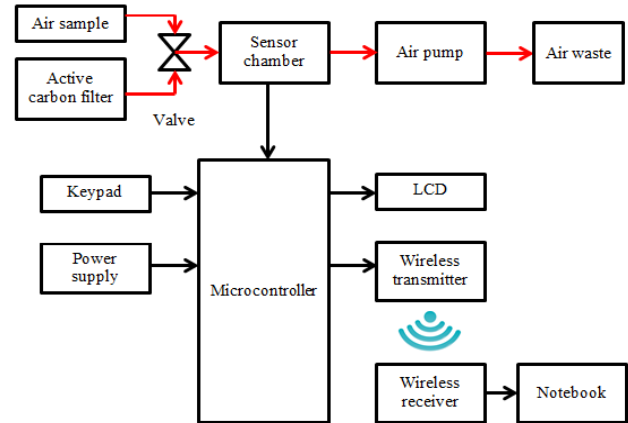


Figure 7: The e-nose block diagram

4.2. Sensing Module

The sensing module consists of sensor arrays for gas detection and a sensor chamber to locate the sensor arrays. Each gas sensor should have a different sensitivity profile over a range of hazardous gases for the confined space application. The process had selected suitable sensors with varying sensitivity to generate the responses known as smell prints. The process had reduced the instrument size, cost of the sensors, power consumption and will increase classification performance.

The gas sensors performance is measured by several criteria such as sensitivity, selectivity, stability, detection limit, response time, recovery time, life cycle and operating temperature [23]. All these criteria are very important to consider during selection of the gas sensors. The developed e-nose sensing module used four selected MOS gas sensors that are normally and widely used for environmental monitoring [51]. The sensors were selected based on each sensor varying sensitivity and selectivity that generated a smell print profile corresponding to the air sample. The selections were based on the confined space main hazardous gases effective value with different sensitivity manufactured by Figaro Inc. and Synkera Technologies Inc. as listed in Table 4. The sensors are able to detect main hazardous gases effective value to effect human for oxygen at between 0-30%, carbon monoxide at 35 ppm, hydrogen sulphide at 20 ppm and methane 5%. For monitoring the environmental conditions, the temperature and humidity sensor (SHT75) from Sensirion company was also located under the developed instrument due to its characteristics and ability (refer to appendix A). The sensor uses two wires through Serial Peripheral Interface (SPI) communication to communicate with the embedded controller.

Table 4: E-nose sensors selection

Category	Target Parameter	Sensor	Sensitivity (ppm)	Sensitivity (%)	Operating Temperature (°C)
Oxygen	Oxygen	Sk-25F	-	0 to 30	-10 to 50
Toxic	Carbon Monoxide	TGS 2442	30 to 1000	-	-10 to 50
	Hydrogen Sulphide	PN 714	1 to 100	-	-10 to 50
Flammable	Methane	TGS 2612	-	1 to 25	-10 to 40

The e-nose sensor chamber was designed and simulated using Computational Fluid Dynamic (CFD) SolidWorks version 2012 software. The SolidWorks software was used to create the Three Dimensional (3D) model of the chamber structure. Then the CFD was used to simulate and analyze for optimum air sample flow inside the chamber cavity to expose to the sensors. The analysis was based on the pressure, velocity and Reynolds (*Re*) number that were generated during the simulation.

The simulation finding was used to enhance the sample flow optimization, structure geometry, material and reduce the chamber dead zone [52]. The geometry should be symmetrical for efficient sample flow inside the chamber cavity. The simulation analyzed the best sensor position between series and parallel chamber design for optimum sample flow exposure to the sensor array and the series type was chosen shown in Figure 8.

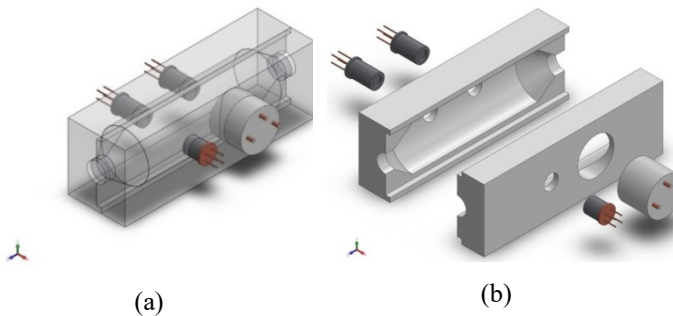


Figure 8: (a) Isometric and (b) explode view for chamber design

4.3. Signal Conditioning Board

The e-nose MOS gas sensors were attached to the signal conditioning board. The board is used to enhance the sensor responses signal suitable for the data acquisition by the microcontroller. The unit consists of the interface circuit, amplifier, filter and Analogue-to-Digital Converter (ADC) as shown in Figure 9.

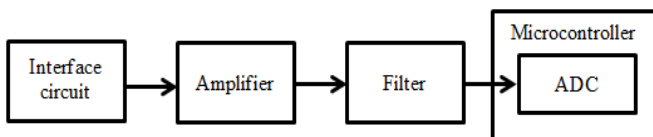


Figure 9: The signal conditioning block diagram

The interface circuit will measure the sensor responses and convert it into a constant current circuit. The sensor heater was used to increase the sensor temperature to around 300°C and the heater circuit is including in the sensor. The sensor response was calculated by using Equation 8.

$$R_s = \frac{V_C - V_{RL}}{V_{RL}} \times R_L \tag{8}$$

Where,

- R_s is the sensor resistance or responses
- V_c or V_{cc} is the circuit voltage
- V_{RL} is the voltage load resistor or voltage output (V_{out})
- R_L is the load resistor

The R_L value used to limit the sensor output voltage and V_{RL} up to five volts. The R_L were using 10 K Ω with 1% accuracy to enable the output voltage range suitable for the microcontroller. The signals from sensors are in analogue form and voltage divider rule are applied to the signal conditioning circuit. The sensor principle is impedance with low range (refer to appendix A) and the MCP602 Integrated Circuits (IC) unity gain voltage follower from Microchip Technology Inc. was used as the signal amplifier. The amplifiers high input and low output with good impedance matching will minimise the loading effects to the microcontroller ADC.

A passive Low Pass Filter (LPF) was used to filter the sensor responses signal from unwanted noise. The passive filter was used because of its fewer components and low noise as compared to active filter. The cut off frequency was calculated by Equation 9. The LPF was designed to allow frequencies below 2 KHz to pass, while blocking frequencies above it.

$$f_c = \frac{1}{2\pi RC} \tag{9}$$

Where,

- f_c is the cut off frequency
- R is the resistor value
- C is the capacitor value

Two signal conditioning circuit boards were designed and developed for sensors located at the left and right of the sensor chamber. The circuit boards were designed using Power Logic software. Then it was converted into a schematic diagram for the board etching process by using Power PCB software. Then both signal conditioning board is etching for components placement through soldering process. The developed sensing module includes sensor chamber and signal conditioning board. The left side board is for carbon monoxide and methane while oxygen and hydrogen sulphide sensors are on the right.

4.4. Controller Board

The controller Printed Circuit Board (PCB) was also designed using Power Logic and Power PCB software. The fabrication is as shown in Figure 10. The PCB main component uses Surface Mount Technology (SMT). The electronic components were laid out on a two-layer plated through hole PCB. All the ICs were decoupled to ground with capacitors to reduce noise from the circuits. The ground plane tracks at the bottom layer were kept to a minimum to reduce the signal to Noise Ratio (SNR). The acquired data is transmitting through Zigbee wireless Radio Frequency (RF) communication.

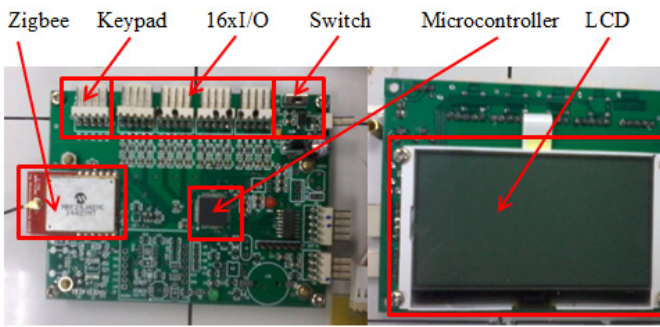


Figure 10: The controller board front and back view

4.5. Microcontroller

The e-nose system controls and data processing use the microcontroller with embedded software. The dsPIC33 microcontroller from Microchip Technology Inc. was selected because its 40 MIPS is good for high processing speed and its 16-bit data bus is suitable for the instrument operation in real time. The microprocessor flash memory is 256 KB and 30 KB for RAM are large enough and capable of handling the data acquisition operation and system control at the same time.

Furthermore, the dsPIC33's built-in 12-bit ADC conversion speed is capable of simultaneously sampling from all the analogue sensor responses signal. The digitized data is stored inside the microcontroller flash memory and sent to laptop computer which is linked through Zigbee wireless RF communication during the data acquisition process. The output digital signals are expressed in numbers ranging from 0 to 4095 and sensors voltage output is calculated by Equation 10.

$$V_{out} = \frac{ADC \times 5}{4096} \quad (10)$$

4.6. Wireless Communication

The e-nose wireless communication was developed using Zigbee (MRF24J40MC) module from Microchip Technology Inc. based on 2.4 Gigahertz (GHz) IEEE Std. 802.15.4™. This module has been chosen as a medium for communication because its communication range can reach up to 4000 feet and the temperature range is at -40 °C to + 85 °C which are suitable for confined space applications. It is also popular because its miniature size, simple circuit and low power. The Zigbee module has a 50Ω Ultra Miniature Coaxial (U.FL) connector to connect to an external 2.4 GHz antenna. The module interfaces to the microcontrollers through a four wire Serial Peripheral interface (SPI), interrupt, wake, reset, power and ground as shown in Figure 11.

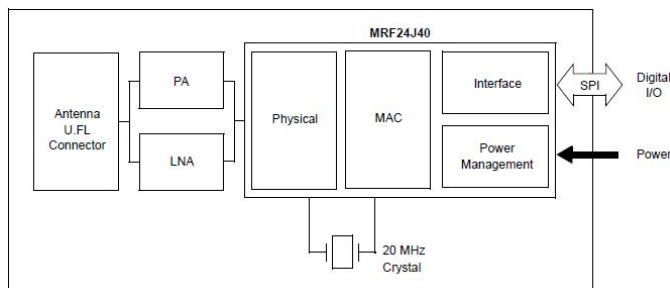


Figure 11: Wireless Zigbee (MRF24J40MC) module block diagram

4.7. Input and Output Devices

The e-nose main controller board was interfaced with a four-button keypad as the input device. The input is used as the user interface to select the instrument program menu. A 128x64 character graphic Liquid Crystal Display (LCD) from Topway Displays brand was selected as the instrument display unit (output). The LCD is used to display several instrument attributes such as operation, data acquisition and utilities.

4.8. Purging System

The sampling system will deliver the air sample to flow into the sensor chamber. The system is essential to ensure that all sensors are exposed to the air sample effectively during the sniff cycle. For this confined space application the e-nose has a self-purging function to ensure that the sensor chamber is purged accordingly [1].

Figure 12 shows a pneumatic system that has been designed as the e-nose self-purging system. The system has a 3-way solenoid electro-valve that was placed between instrument inlet and sensor chamber. During sniff cycle the electro-valve would allow the air sample from the confined space atmosphere to flow into the sensor chamber through channel one. While during purging cycle, the electro-valve would allow filtered air through the carbon filter to purge the sensor chamber cavity through channel two. The system is programmed and controlled by the instrument's microcontroller.

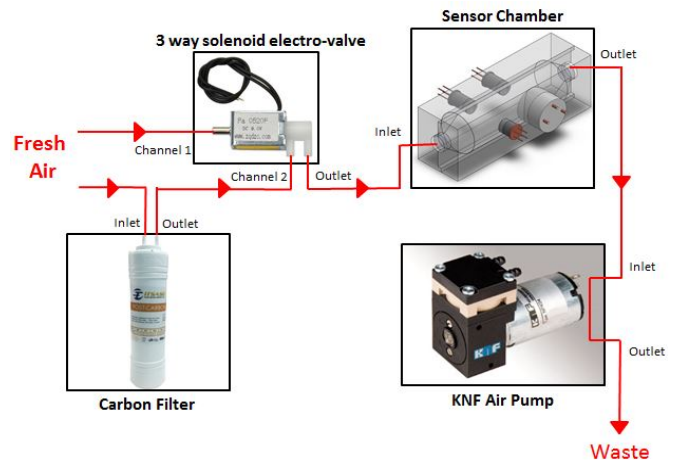


Figure 12: The e-nose purging system

4.9. Power Unit

The e-nose system uses a single 12 volt lead acid battery as the power supply. The power supply is used for the air pumps, electro-valve, signal conditioning boards and main controller board. The PCB consists of two unit voltage regulators (LM2575) used to provide 5 volts of power for the sensor measuring circuits, amplifiers, analogue multiplexer and relay.

4.10. Enclosure

The enclosure or casing for the e-nose was selected based on space for component placement and the material used. The plastic enclosure was used because of its non-flammable properties when exposed to high temperatures. It also did not interfere with data acquisition during the sniff operation. Figure 13 shows the enclosure with dimension of 22cm (L) x 30cm (W) x 14cm (H) and the instrument's complete hardware development.

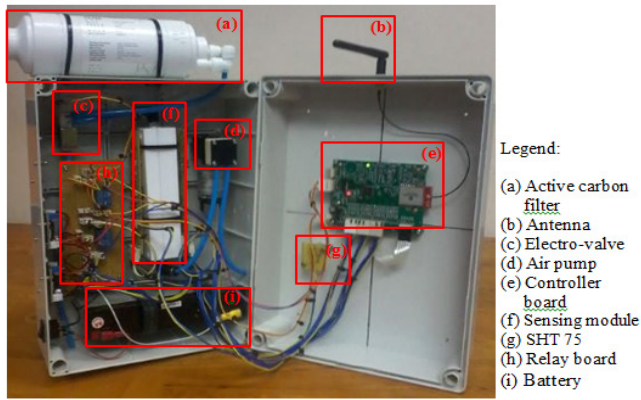


Figure 13: E-nose enclosure and hardware development

4.11. Software Development

The e-nose embedded software is used to integrate hardware to enable the instrument operate as a system. The acquired data transferred wirelessly from the instrument to a laptop computer that is received and displayed through Graphical User Interface (GUI). The data is processed and classified using an off-line method by utilizing MATLAB software.

4.12. Embedded Control Software

The e-nose embedded software has been developed on the laptop computer using dsPIC33 C compiler version 8.0 from Microchip Technology Inc. The compiler was bundled inside the MPLAB Integrated Development Environment (IDE). The menu driven program contains the instrument control, data acquisition, communication and utilities.

4.13. Graphical User Interface

A Graphic User Interface (GUI) was developed using Visual Basic 6.0 (VB6) software on a windows 7 platform. The GUI is used for receiving and recording the sensors output digital signal that is transmitted wirelessly by the e-nose to laptop computer. The GUI will display the sensor digital output signal that is plotted in a single bar graph according to their concentrations as shown in Figure 14. The instrument sensor digital output signals that are displayed by the GUI in real time are saved as a text file.

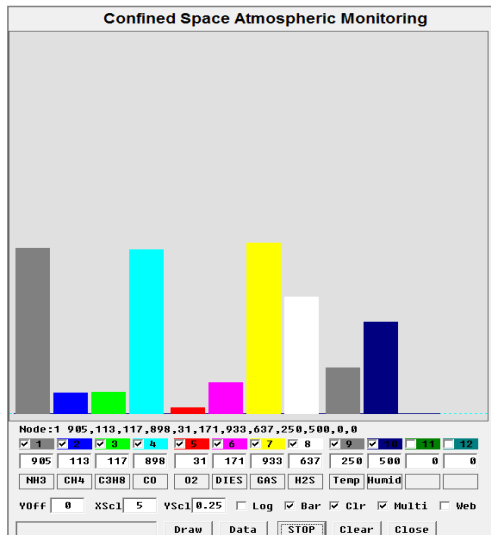


Figure 14: Graphic User Interface (GUI) for sensor responses

4.14. Converting ADC Data to parts per million

The developed GUI is also able to convert the e-nose sensor digital output signal as data in ADC value to parts per million (ppm) and percentage correspond to the air sample concentration. Based on the e-nose MOS gas sensors data sheet, the MOS gas sensors have a linear correlation between the sensor responses and air sample concentration. The sensor response gradient (m) and constant value (k) that intersect at $x = 1$ was calculated using Equation 11 and Equation 12.

$$m = \frac{\Delta(\log y)}{\Delta(\log x)} \quad (11)$$

$$k = \frac{y}{x^m} \quad (12)$$

Where,

- m is the gradient
- y is the sensors response (Ω) value
- x is the sample concentrations (ppm) value
- k is the intersection point at axis $x=1$

Based on the calculated value of the sensor response gradient and constant value, the sensor digital output signal in ppm value, x can be calculated by using Equation 13.

$$x = 10^{\left[\frac{\log(\frac{y}{k})}{m}\right]} \quad (13)$$

For the flammable gas (i.e. methane), the sensor digital output signal is measure in percentages by using Equation 14. The developed GUI for the instrument sensors digital output signal conversion is shown in Figure 15.

$$x_{\%} = \frac{x_{ppm}}{10000} \quad (14)$$

Where,

- $x_{\%}$ is the sample concentrations in percentage
- x_{ppm} is the sample concentrations in ppm

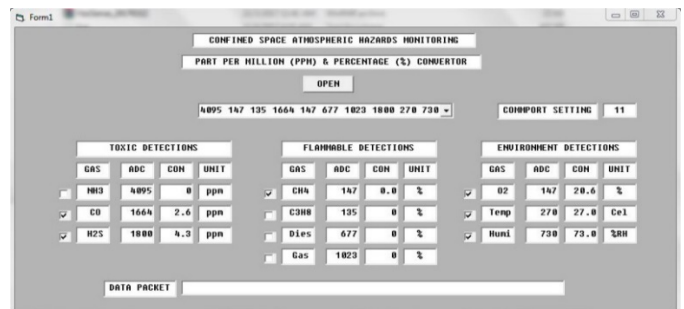


Figure 15: Graphic User Interface (GUI) for data conversion

5. Experimental Setup

The e-nose was tested in laboratory and field environments. This process will ensure the instrument functionality, performance and reliability were satisfied. The first experiment was conducted to test the performance of the purging system. After that, the instrument was calibrated to test its response to various air sample

concentrations. Lastly the instrument was tested in a hospital confined space (i.e. mechanical room) for field atmospheric hazards monitoring.

5.1. System Setup

Dynamic headspace technique was used for the e-nose sampling process. The instrument system setup was set according to the parameters listed in Table 5.

Table 5: The e-nose system setup parameter

Operation	Time (s)	Air pump	Electro-valve
Switch on e-nose and run GUI	30	OFF	OFF
Sensors pre-heat period	90	OFF	OFF
Purge cycle	60	ON	ON
Holding period after purge cycle	30	OFF	OFF
Sniff cycle	30	ON	OFF
Holding period after sniff cycle for data acquisition	200	OFF	OFF

Based on the system setup parameters, the e-nose uses the dynamic headspace sampling method includes “purge”, “sniff” and “hold”. The method will hold the air sample in sensor chamber to enhance the e-nose sensitivity. The setup parameter values were based on the selected sensors characteristics. The data acquisition experiment procedures are as follows:-

- i. Switch on the e-nose and run the GUI on a laptop computer to test the communication.
- ii. Pre-heat the instrument sensors to ensure the MOS gas sensors reach the optimum operation temperature.
- iii. Activate the self-purging procedure (purge cycle) by activating the 3-way solenoid electro-valve channel two. This would allow the ambient air to flow through the active carbon filter by using the air pump. The clear reference air will flow into the sensor chamber and flush out the previous air sample. Then the instrument will remain idle for about 30 seconds to allow sensors response to return to its basic value.
- iv. Sniff cycle, the 3-way solenoid electro-valve channel one is activated for the instrument expose to the air sample. The headspace air sample will flow into the sensor chamber by using the same air pump.
- v. The air sample will be held in chamber for 200 seconds that will expose the sensors for optimum interaction and sensor response reach steady state.
- vi. The steady state sensor response data is transfer wirelessly through the Zigbee module to a laptop computer that is displayed and saved by the GUI. The GUI will record 200 data at one data per second baud rate based on the sensor’s response.
- vii. Repeat steps (iii) to (vi) for another nine times that will generate 2000 data which is used for the data processing.

5.2. Purging Testing

The e-nose self-purging test experiment was conducted at the biomaterials laboratory, School of Mechatronics Engineering, Universiti Malaysia Perlis (UniMAP). The laboratory temperature was 25°C with 75% Relative Humidity. The room temperature and humidity was continually observed and recorded.

The experiment sampling process used the instrument system setup as explained in section 5.1. This experiment was conducted

to test its ability to flush out the air sample from the sensor chamber enabling gas sensor responses return to its basic value. The experiment had to follow the data acquisition procedure.

Initially the instrument purge cycle operation experiments were activated using laboratory ambient air as the reference. Then the instrument purge cycle was repeated by using the self-purging system which is activated by the three-way solenoid electro-valve channel two. This allows the ambient air to flow through the active carbon filter. The reference air will flow into the sensor chamber and enables sensors response to return to their basic value. Both purging method sensor responses are analyzed for their performance.

5.3. System Calibration and Validation

E-nose calibration and validation experiments were conducted also at the same laboratory. The laboratory temperature was 25°C with 75% Relative Humidity. The room temperature and humidity was continually observed and recorded.

The experiments were conducted in laboratory in a fume hood because it will absorb any leaking of the hazardous gases as shown in Figure 16. The e-nose sampling process involves two gas sampled cylinders with different concentration as listed in Table 6. The experiment sampling process used the e-nose system setup as explained in section 5.1. The experiment had to follow the data acquisition procedure.

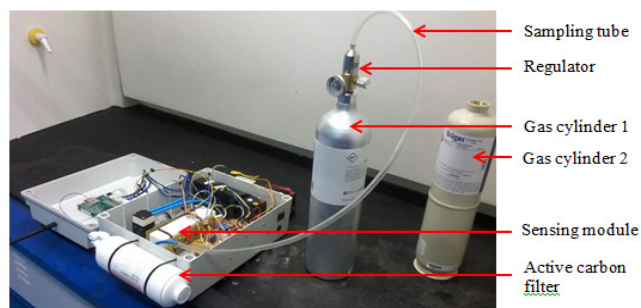


Figure 16: The e-nose calibration and validation testing

Table 6: The component and concentrations for gas cylinder one and gas cylinder two

Components	Concentrations	
	Cylinder 1	Cylinder 2
Hydrogen sulphide (H ₂ S)	10 ppm	< 1 ppm
Carbon monoxide (CO)	50 ppm	
Methane (CH ₄)	2.5 %	
Oxygen (O ₂)	18.0 %	20.8%

The gas in cylinder one which contains oxygen, hydrogen sulphide, carbon monoxide and methane used for the calibration while cylinder two which contains air with zero grades (<1 ppm) were used for the validation.

Initially, the experiments were conducted by using gas from cylinder one (a mixing of hydrogen sulphide, carbon monoxide, methane and oxygen) as the sample to identify and calculate the calibration value. Then the experiments were repeated by using gas cylinder two which contain air with zero grades. The experiments were used to verify the e-nose ability to responses with lowest concentration of air sample.

E-nose performance was validated with an Altair 5X Multi Gas Detector from MSA brand which is able to detect hydrogen sulphide, carbon monoxide, methane and oxygen [53]. The instrument temperature and humidity measurement value were also validated by using the Humidity Alert Detector from Extech as shown in Figure 17.

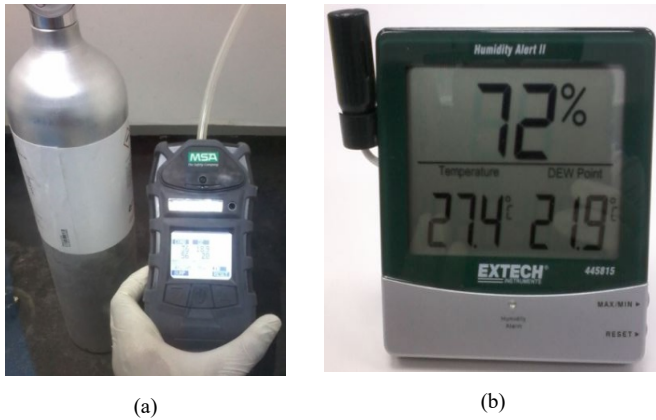


Figure 17: (a) Altair 5X Multi Gas Detector and (b) Humidity Alert detector

Initially both gas cylinders were used as air samples for the gas detector. The validate instrument measurement was used as the reference value. Then the experiments were conducted for the instrument using both gas cylinders. The different measurement values between the instrument and validate instrument which are known as percentage error (%_{error}) by using Equation 15 were adjusted to be less than 10% by modifying the embedded instrument data acquisition program [53]. The experiments were repeated to 10 times for the instrument and validate instrument using both gas cylinders for validation process.

$$\%_{Error} = \left| \frac{\text{Experimental value} - \text{Reference value}}{\text{Reference value}} \right| \times 100 \quad (15)$$

Where,

- %_{Error} is percentage error value
- Experimental value is e-nose sensor responses data
- Reference value is validate instruments data

E-nose temperature and humidity measurement values were also compared with the commercial temperature and humidity meter. The different measurement values were also adjusted to less than 10% by modifying the embedded instrument data acquisition program. The experiments were also repeated for the instrument to validate the temperature and humidity measurements values.

5.4. Field Environment Testing

E-nose field environment experiment was conducted in a confined space (i.e. mechanical room) at the Hospital Sultanah Bahiyah, Alor Setar, Kedah. There are two set generators placed in the mechanical room to generate electricity when electricity is suddenly cut off in the hospital building. This room is potentially exposed to the atmospheric hazards when there is a leak at the generator that uses diesel fuel to operate. During the operation, the carbon monoxide gas will release and oxygen level became limit because eliminate by this gas.

This mechanical room is also chosen because of its appropriate area and the distance that allows wireless communication devices to function fully without losing any data. The room size was 20m (L) x 10m (W) x 5m (H) and the room average temperature was 31°C with 63% Relative Humidity. This room temperature and humidity condition has no effect to the sensors ability in term of drift response and sensitivity changes after calibration experiment testing were done based on the sensors operating temperature. The room temperature and humidity were continually observed and recorded.

Four locations, ambient (outside room), location one, location two and location three were selected for data acquisition. The purpose of these locations was to prove the ability of the instrument to be able to classify the concentration of gas according to their location. The experiment started with data acquisition at outside room as reference air, followed by location one, two and lastly at location three. The generator and data acquisition location in the mechanical room are illustrated in Figure 18. The experiment sampling process used the e-nose system setup as explained in section 5.1.

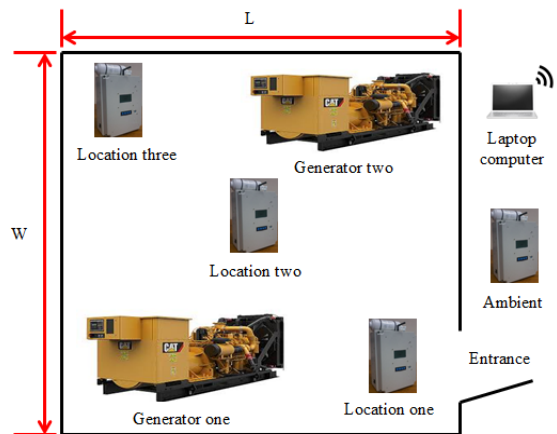


Figure 18: Data acquisition locations in mechanical room

6. Results and Discussion

6.1. Purging Testing

This section discusses the data analysis for the purging testing to compare reading between normal ambient air and the filtered ambient air. The analysis utilised the statistical multivariate technique to discriminate between two different samples.

E-nose sensor responses were a time series waveforms profile. The sensor responses purging cycle using laboratory ambient air and ambient air through an active carbon filter are shown in Figure 19 and Figure 20 respectively. The reference air flows into the sensor chamber and enables the sensor responses to return to their basic values.

The average for 10 times readings from sensor responses between the ambient air and filtered ambient air are listed in Table 7. The readings show that sensor responses for the ambient air through the active carbon filter is lower than ambient air without filter. This means that the active carbon filter is able to flush out the air sample from the sensor chamber better to enabling gas sensor responses return to its basic value and become stabilized during purge cycle.

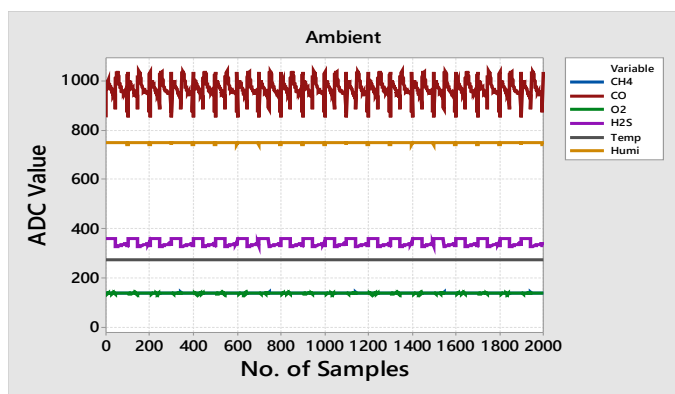


Figure 19: Sensor responses for ambient air

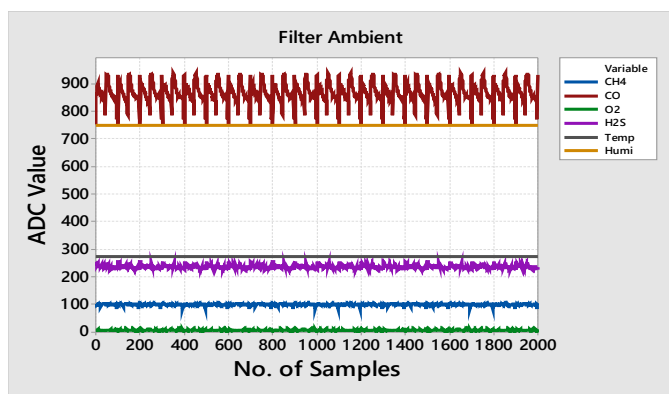


Figure 20: Sensor responses for ambient air through active carbon filter

Table 7: Average sensor responses for purging testing

Parameter	Methane (ADC value)	Carbon monoxide (ADC value)	Oxygen (ADC value)	Hydrogen sulphide (ADC value)
Ambient air	150	950	140	350
Active carbon filter	100	850	10	250

The PCA used the correlation matrix method because for all samples, the sensor responses data variances were completely different so the data variances were standardized. Figure 21 shows the PCA score plot for instrument purging testing.

Consists of six total principal component's (PC) that represent number of instrument sensors and a 2D-PCA score plot was used because the first two PC variation accounted for 99.43% (PC1 was 91.21% and PC2 was 8.22%) of the total data more than 90% which contain useful information [54]. The plot shows that data for ambient air and ambient air through an active carbon filter are successfully discriminate into two groups. This shows that the proposed self-purging system capability is successful in providing reference air for the instrument purge cycle operation.

6.2. Calibration and Validation

This section discusses the data analysis for the calibration process by calculating the calibration value. The validation between developed instrument and validated instrument is to verify its detection performance and reliability.

Sensor responses for the e-nose calibration which used gas cylinder one and gas cylinder two are shown in Figure 22 and

Figure 23. The plotted graph shows that the different detection reading between both gas cylinders. The feature selection process was applied to the raw data and 100 steady state responses were select for both gas cylinders detection.

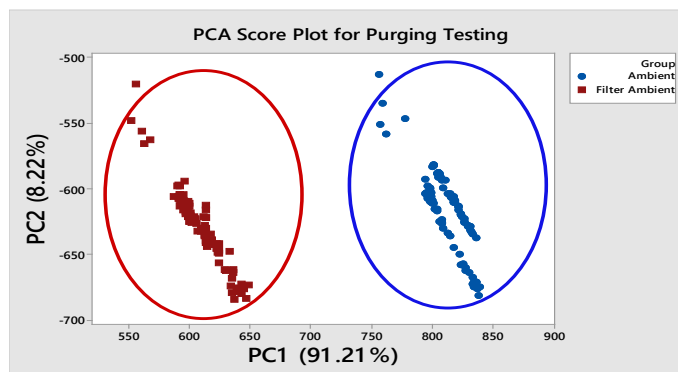


Figure 21: PCA score plot for purging testing

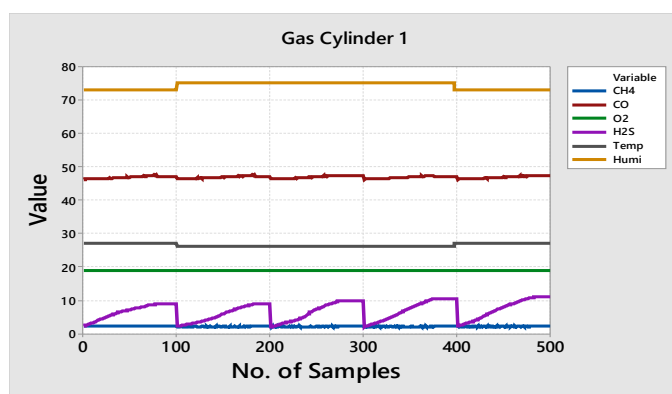


Figure 22: Sensor responses for gas exposure from cylinder one

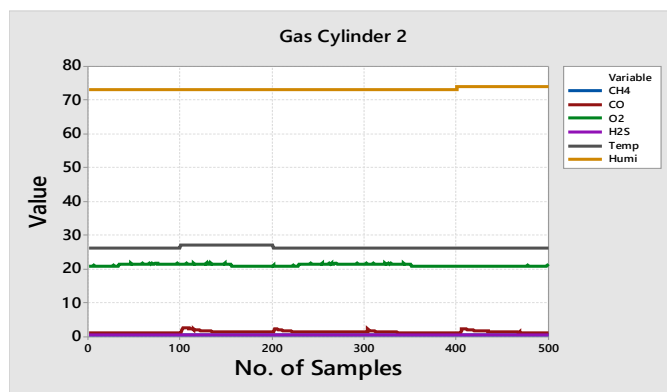


Figure 23: Sensor responses for gas exposure from cylinder two

The calibration value is the different reading between the e-nose average sniff and the commercial detector as a validated instrument. The percentage error ($\%Error$) calculation for instrument sensors when exposed to the concentrations of gas cylinder one is shown in Table 8. The instrument sniff reading was 2.03% for methane, 46.83 ppm for carbon monoxide, 18.68% for oxygen, 6.68 ppm for hydrogen sulphide, 26.40°C for temperature and 74.20% Relative Humidity for humidity sensors. The different values between instrument sniff and validate instrument shown by the methane is 0.47%, carbon monoxide is 3.17 ppm, oxygen is 0.12%, hydrogen sulphide is 3.32 ppm, temperature is 0.40°C and humidity is 0.20% Relative Humidity.

Results from these different values are used as the calibration values for each sensor when performing the next detections. The percentage error was then calculated and it shows 18.80% for methane, 6.34% for carbon monoxide, 0.63% for oxygen, 33.20% for hydrogen sulphide, 1.53% for temperature and 0.27% for humidity sensors.

Table 8: Calibration value and percentage error calculation before calibration

	Methane (%)	Carbon Monoxide (ppm)	Oxygen (%)	Hydrogen Sulphide (ppm)	Temperature (°C)	Humidity (%RH)
E-nose sniff	2.03 ±0.08	46.83 ±0.12	18.68 ±0.02	6.68 ±0.93	26.40 ±0.36	74.20 ±0.24
Altair 5x	2.50	50.00	18.80	10.00	-	-
Humidity Alert	-	-	-	-	26.00	74.00
Calibration value	0.47	3.17	0.12	3.32	0.40	0.20
Percentage error (%)	18.80	6.34	0.63	33.20	1.53	0.27

The same sampling method process is performed, but this time by adding the calibration value for the instrument sensors when exposing them to the concentrations of gas in cylinder one. The result is shown in Table 9. The different readings between instrument sniff and validate instrument shown by the methane is 0.04%, carbon monoxide is 1.18 ppm, oxygen is 0.24%, hydrogen sulphide is 0.50 ppm, temperature is 0.60°C and humidity is 0.40% Relative Humidity. The percentage error was then calculated and it shows 1.60% for methane, 0.36% for carbon monoxide, 0.21% for oxygen, 5.00% for hydrogen sulphide, 2.22% for temperature and 0.80% for humidity sensors. Since the percentage error is decreased to less than 5.00% which is within the standard total difference allowed (less than 10%), the results were considered as valid and acceptable [55]. This proved that the instrument is trusted and ready for real field environment testing.

Table 9: Calibration value and percentage error calculation after calibration

	Methane (%)	Carbon Monoxide (ppm)	Oxygen (%)	Hydrogen Sulphide (ppm)	Temperature (°C)	Humidity (%RH)
Average, sniff	2.46 ±0.09	49.82 ±0.11	18.76 ±0.03	9.50 ±0.86	26.40 ±0.30	75.60 ±0.21
Altair 5x	2.50	50.00	18.80	10.00	-	-
Humidity Alert	-	-	-	-	27.00	75.00
Calibration value	0.04	1.18	0.24	0.50	0.60	0.40
Percentage error (%)	1.60	0.36	0.21	5.00	2.22	0.80

Table 10: Results of e-nose sensors ability in lowest responses

	Methane (%)	Carbon Monoxide (ppm)	Oxygen (%)	Hydrogen Sulphide (ppm)	Temperature (°C)	Humidity (%RH)
Average, sniff	0.20 ±0.04	0.76 ±0.43	20.94 ±0.09	0.30 ±0.07	26.40 ±0.13	73.20 ±0.16
Altair 5x	0.00	0.00	20.80	0.00	-	-
Humidity Alert II	-	-	-	-	26.00	73.00

Table 10 shows the average lowest sniff readings for the instrument sensors during exposure to gas cylinder two. The lowest reading shown by the methane sensor is 0.20%, carbon

monoxide sensor is 0.76 ppm and hydrogen sulphide sensor is 0.30 ppm. The commercial gas detector readings are also recorded to validate and to prove that the concentrations from the gas sample that have been used are trusted and reliable. The results proved that the instrument sensors are able and functioning well to response less than one ppm when exposed to the air with zero grades.

6.3. Field Environment Testing

This section discusses the data analysis for the field environment testing to test the instrument performance in real confined space. The analysis utilised the statistical multivariate and the ANN techniques.

The sensor responses for e-nose field environment testing at location one, two, three and four. Figure 24 shows the instrument sensor responses for location one air sample which consists of a dynamic slope response (transient) and steady state response. The data is pre-processed to enhance its suitability for the analysis. Feature selection is used to extract 100 steady state sensor responses in a 100 second sampling rate for pattern recognition purpose. Figure 25 shows instrument sensor responses after the feature selection process. The same feature selection process applied to other samples which are location two, three and four.

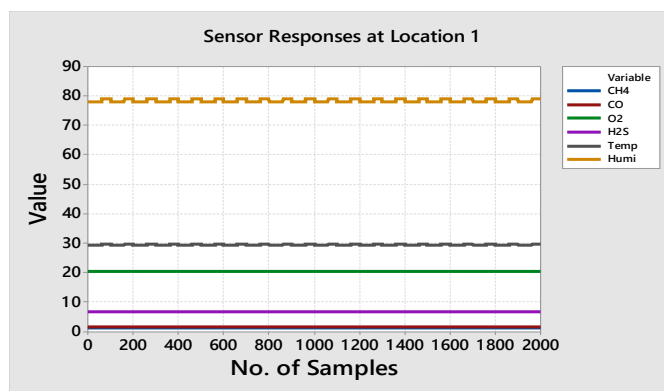


Figure 24: Location one sensor responses in field environment testing

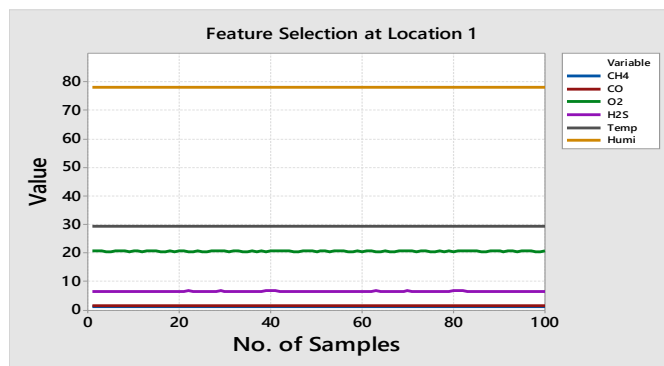


Figure 25: Location one feature selection in field environment testing

Differential baseline manipulation method was used for the air sample data. The baseline data is the different between the reference air and sample data. The process purpose is to remove the outliers and compensate the disturbance effect. This will enhance the data quality and improve sensitivity. Figure 26 shows the instrument baseline manipulation for the location one air sample.

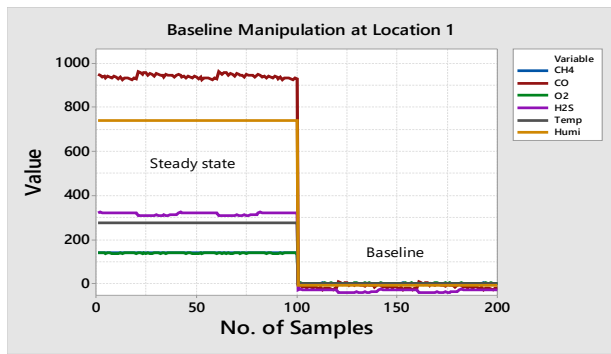


Figure 26: Location one baseline manipulation in field environment testing

A normality test was used to investigate the acquired data distribution pattern. The tests used Anderson-Darling (AD), Shapiro-Wilk (SW) and Kolmogorov-Smirnov (KS) methods. The hypothesis for the normality tests was set as:

Null hypothesis, H_0 : Data is normal

Alternative hypothesis, H_1 : Data is not normal

The significance level (alpha, α) for AD, SW and KS method was set at 0.05. The results shown in Table 11 indicate that the probability value (p -value) was less than α value for all three normality test methods. As the p -value is less than α , so it will reject the H_0 . Thus, the acquired data distribution is not normal and the appropriate method chosen for this data is non-parametric classification analysis.

Table 11: Normality test for AD, SW and KS in field environment testing

	AD	SW	KS
Ambient	<0.005	<0.010	<0.010
Location one	<0.005	<0.010	<0.010
Location two	<0.005	<0.010	<0.010
Location three	<0.005	<0.010	<0.010

The PCA score plot for field environment testing was shown in Figure 27. Consists of six total PC's that represent number of instrument sensors and a 2D PCA score plot was used because the first two PC's variables was 99.42% (PC1 was 97.01% and PC2 was 2.41%) which contain most of the useful information 54. The plot shows that the sample data were clustered into four groups: ambient, location one, location two and location three. The plot indicates the instrument field environment testing's ability to differentiate air samples at different locations.

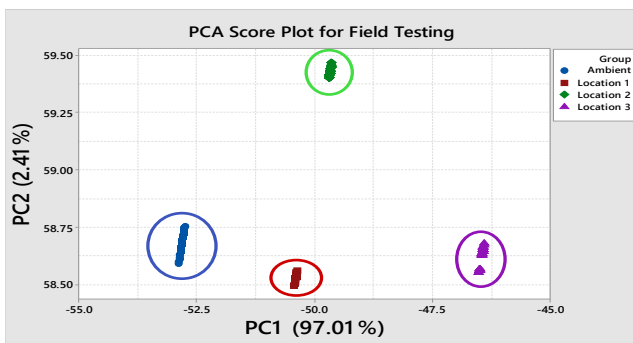


Figure 27: PCA score plot for field environment testing

The SVM is a statistical multivariate analysis method that used to test the developed instrument's capability. The methods robust features are suitable for analysing the sample data. The analysis

was conducted to differentiate and classify the samples between ambient (Amb), location one (L1), location two (L2) and location three (L3).

The SVM method used 400 sample data inputs, with 100 acquired from each of the four samples. The train class was set at 70% (280 data) while test class was set at 30% (120 data). The SVM utilised the linear kernel function with prediction speed at 9400 obs/sec, training time at 2.88 sec, automatic kernel scale, box constraint at level one and the multiclass method was one versus one.

The instrument in field environment test classification success rate is shown by the confusion matrix with 99.28% for train performance and 98.33% for test performance as shown in Table 12. The results indicate that the instrument is successfully being developed and has a high classification success rate in classifying the samples.

Table 12: SVM confusion matrix for e-nose field environment testing

Train Performance					Test Performance						
Class		Train Class				Class		Target Class			
		Amb	L1	L2	L3			Amb	L1	L2	L3
Amb		69	1	0	0	Amb	29	1	0	0	
L1		1	69	0	0	L1	1	29	0	0	
L2		0	0	70	0	L2	0	0	30	0	
L3		0	0	0	70	L3	0	0	0	30	
Success rate		99.28%				Success rate	98.33%				

Table 13: RBFNN confusion matrix for e-nose field environment testing

Train Performance					Test Performance						
Class		Train Class				Class		Target Class			
		L1	L2	L3	L3			Amb	L1	L2	L3
Amb		69	1	0	0	Amb	29	1	0	0	
L1		1	69	0	0	L2	1	29	0	0	
L2		0	0	70	0	L3	0	0	30	0	
L3		0	0	0	70	L4	0	0	0	30	
Success rate		99.28%				Success rate	98.33%				

The RBFNN analysis used 400 sample data inputs, with 100 acquired from each of the four samples. The network used 70% (280 data) of the data for training and the remaining 30% (120 data) for testing. The three-layer feed-forward classification model used six input nodes which corresponds to the instrument sensors, temperature and humidity sensor. The optimum hidden layer neuron also is set to four by using *trial and error* method. Four neurons were used as the output in sequences of 0001, 0010, 0100 and 1000 to help process the input which implemented the full classifications.

Table 13 shows the RBFNN classification confusion matrix for the instruments filed environment testing. The instrument classification success rate for classifying four different samples is 99.28% for train performance and 98.33% for test performance. The RBFNN classification result indicates that the instrument is successfully developed and functions accordingly.

7. Conclusion

This work has successfully developed portable e-nose for confined space application. The statistical multivariate analysis and ANN were used to model the air sample in providing potential solutions for the specific confined space application. The major achievement and contributions are summarised.

This work has successfully investigated and identified the atmospheric hazards main hazardous gases (i.e. oxygen, carbon monoxide, hydrogen sulphide and methane) that could contribute to confined space atmospheric accidents. The e-nose sensing module development was based on the main hazardous gases for confined space application.

The e-nose system was successfully fabricated with multimodal sensor detection. The instrument was able to function effectively in providing sample assessment. The instrument's key features are its portability, ease of operation and effectiveness for use in confined space application.

The e-nose system was successfully integrated with optimum self-purging. The instrument's self-purging system is able to supply the lowest sensors responses with filtered ambient air for the instrument purging cycle which is important for repeatability.

The e-nose was successfully tested its functionality in the laboratory environment through calibration and validation process using validated instruments includes Altair 5x Multi Gas Detector for concentrations of gases and Humidity Alert meter for temperature and humidity. The 5% percentage error readings between instrument and validated instruments proved its capabilities.

The e-nose was successfully tested in a real confined space environment (i.e. mechanical room) at Hospital Sultanah Bahiyah (HSB), Alor Setar, Kedah. The instrument is capable of recognising the concentrations of hazardous gasses at specific locations by discriminated using PCA with a total variation of 99.42%. While the classifiers success rate for SVM and RBFNN indicates of 99.28% for train performance and 98.33% for test performance. The instrument is able to display the atmospheric hazards concentrations in the room which is important during pre-entry testing.

Finally, based on the achievement of this research work, it is clear that the development of e-nose is able to compete existing devices to accurately and consistently measure atmospheric hazards in confine space applications.

8. Future Work

In order to enhance the instrument capabilities, a number of improvements need to be considered for future work. Firstly, improve the data conversion programming from GUI to embedded system. Secondly, improve the wireless communication distance suitable for a confined space that has a large area. It will ensure a smooth data acquisition process and prevent any data loss.

Thirdly, integrate the acquired data to send into an Internet of Things (IoT) system for real-time remote monitoring. Finally, integrate the instrument with a mobile robot for ease manoeuvring to various locations in the confined space [56]. This can eliminate an authorised person from entering into the confined space during pre-entry atmospheric testing. The mobile robot should have a complete system with a remote control and charge-coupled device (CCD) camera for any obstacle avoidance.

Acknowledgement

The author would like to acknowledge the support from the Sciencefund under grant number of 03-01-15-SF0257 from the Ministry of Science, Technology & Innovation Malaysia (MOSTI). The author also gratefully acknowledge to Universiti Malaysia Perlis (UniMAP) for the opportunities given to do the research.

References

- [1] M. A. A. Bakar, A. H. Abdullah, F. S. A. Saad, S. A. A. Shukor, M. S. Kamis, A. A. A. Razak, M. H. Mustafa, "Electronic Nose Purging Technique for Confined Space Application" in IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA), Penang, Malaysia, 2017. DOI: 10.1109/CSPA.2017.8064948.
- [2] A. Suruda, T. Pettit, G. Noonan, R. Ronk, "Deadly Rescue: The Confined Space Hazard" *Journal of Hazardous Materials*, vol. 36, no. 1, pp. 45–53, 1994. [https://doi.org/10.1016/0304-3894\(93\)E0051-3](https://doi.org/10.1016/0304-3894(93)E0051-3).
- [3] Hazards of Confined Spaces. Vancouver: WorkSafeBC, 2008.
- [4] H. Ye, "Atmosphere Identifying and Testing in Confined Space" in First International Conference on Instrumentation, Measurement, Computer, Communication and Control, Beijing, China, 2011. DOI: 10.1109.
- [5] N. B. A. Wahab, "Fatal Accident Case" Website Department of Occupational Safety and Health Malaysia, Available: <http://www.dosh.gov.my/index.php/en/component/content/article/352-osh-info/accident-case/955-accident-case>.
- [6] T. Suski, "Common Mistakes in Confined Space Monitoring" *EHS Today*, Available: https://www.ehstoday.com/safety/confinedspaces/ehs_imp_37605.
- [7] M. Mamat, S. A. Samad, "The Design and Testing of an Electronic Nose Prototype for Classification Problem" in International Conference on Computer Applications and Industrial Electronics, Kuala Lumpur, Malaysia, 2010. DOI: 10.1109/ICCAIE.2010.5735108.
- [8] W. Chansongkram, N. Nimsuk, "Development of a Wireless Electronic Nose Capable of Measuring Odors Both in Open and Closed Systems" *Procedia Computer Science*, vol. 86, pp. 192–195, 2016. DOI: 10.1016/j.procs.2016.05.060.
- [9] F. Röck, N. Barsan, U. Weimar, "Electronic Nose: Current Status and Future Trends" *Chemical Reviews*, vol. 108, no. 2, pp. 705–725, 2008. DOI: 10.1021/cr068121q.
- [10] A. Wilson, M. Baietto, "Applications and Advances in Electronic-Nose Technologies" *Sensors*, vol. 9, no. 7, pp. 5099–5148, 2009. DOI: 10.3390/s90705099.
- [11] RAE Systems by Honeywell, "Application Note 206 Guide to Atmospheric Testing in Confined Spaces" Available: https://www.raesystems.com/sites/default/files/content/resources/Application-Note-206_Guide-To-Atmospheric-Testing-In-Confined-Spaces_04-06.pdf.
- [12] L. Capelli, S. Sironi, R. D. Rosso, "Electronic Noses for Environmental Monitoring Applications" *Sensors*, vol. 14, no. 11, pp. 19979–20007, 2014. DOI: 10.3390/s141119979.
- [13] Y. Zou, H. Wan, X. Zhang, D. Ha, P. Wang, *Bioinspired Smell and Taste Sensors*, Dordrecht: Springer, 2015.
- [14] L. Capelli, S. Sironi, R. D. Rosso, "Odor Sampling: Techniques and Strategies for the Estimation of Odor Emission Rates from Different Source Types" *Sensors*, vol. 13, no. 1, pp. 938–955, 2013. DOI: 10.3390/s130100938.
- [15] N. E. Barbri, E. Llobet, N. E. Bari, X. Correig, B. Bouchikhi, "Application of a Portable Electronic Nose System to Assess the Freshness of Moroccan Sardines" *Materials Science and Engineering: C*, vol. 28, no. 5-6, pp. 666–670, 2008. DOI: 10.1016/j.msec.2007.10.056.

- [16] J. Gutiérrez, M. Horrillo, "Advances in Artificial Olfaction: Sensors and Applications," *Talanta*, vol. 124, pp. 95–105, 2014. <http://dx.doi.org/10.1016/j.talanta.2014.02.016>.
- [17] T. Pogfay, N. Watthanawisuth, W. Pimpao, A. Wisitsoraat, S. Mogpraneet, T. Lomas, M. Sangworasil, A. Tuantranont, "Development of Wireless Electronic Nose for Environment Quality Classification" in *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON2010)*, Chiang Mai, Thailand, 2010.
- [18] S. Panigrahi, S. Balasubramanian, H. Gu, C. Logue, M. Marchello, "Design and Development of a Metal Oxide Based Electronic Nose for Spoilage Classification of Beef" *Sensors and Actuators B: Chemical*, vol. 119, no. 1, pp. 2–14, 2006. DOI:10.1016/j.snb.2005.03.120.
- [19] F. Difrancesco, M. Falcitelli, L. Marano, G. Pioggia, "A Radially Symmetric Measurement Chamber for Electronic Noses" *Sensors and Actuators B: Chemical*, vol. 105, no. 2, pp. 295–303, 2005. DOI:10.1016/j.snb.2004.06.013.
- [20] A. H. Abdullah, A. H. Adom, A. Y. Shakaff, M. N. Ahmad, M. A. Saad, "Odour Sensor Chamber Development for Electronic Nose" in *4th International Workshop on Computer Science and Engineering-Winter (WCSE)*, Dubai, UAE, 2014.
- [21] S. M. Scott, D. James, Z. Ali, W. T. Ohare, "Optimising of the Sensing Chamber of an Array of a Volatile Detection System: Fluid Dynamic Simulation" *Journal of Thermal Analysis and Calorimetry*, vol. 76, no. 2, pp. 693–708, 2004. DOI: 10.1023/B:JTAN.0000034891.68585.4a.
- [22] D. James, S. M. Scoot, Z. Ali, W. T. O'Hare, "Chemical Sensors for Electronic Nose Systems" *Microchimica Acta*, vol. 149, pp. 1–17, 2005. DOI: 10.1007/s00604-004-0291-6.
- [23] G. H. Jain, "MOS Gas Sensors: What Determines Our Choice?" in *Fifth International Conference on Sensing Technology (ICST)*, Palmerston North, New Zealand, 2011. DOI: 10.1109/ICSensT.2011.6137067.
- [24] M. Nissfolk, *Development of an Electronic Nose-Tongue Data Acquisition System using a Microcontroller*, Uppsala University, 2009.
- [25] E. Noorsal, O. Sidek, J. M. Saleh, "Automated Odour Measurement in Electronic Nose System Using Microcontroller" in *International Conference on Man-Machine Systems 2006*, Langkawi, Malaysia, 2006.
- [26] A. H. Abdullah, A. H. Adom, A. Y. M. Shakaff, M. N. Ahmad, A. Zakaria, N. A. Fikri, O. Omar, "An Electronic Nose System for Aromatic Rice Classification" *Sensor Letters*, vol. 9, no. 2, pp. 850–855, 2011. DOI:10.1166/sl.2011.1629.
- [27] K.-T. Tang, S.-W. Chiu, M.-F. Chang, C.-C. Hsieh, J.-M. Shyu, "A Low-Power Electronic Nose Signal-Processing Chip for a Portable Artificial Olfaction System" *IEEE Transactions on Biomedical Circuits and Systems*, vol. 5, no. 4, pp. 380–390, 2011. DOI: 10.1109/TBCAS.2011.2116786.
- [28] S. Samadi, "Interface Design Techniques for Electronic Nose Sensors: A Survey" in *Sixth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services*, Venice, Italy, 2013.
- [29] K. M. Pierce, B. Kehimkar, L. C. Marney, J. C. Hoggard, R. E. Synovec, "Review of Chemometric Analysis Techniques for Comprehensive Two Dimensional Separations Data" *Journal of Chromatography A*, vol. 1255, pp. 3–11, 2012. <http://dx.doi.org/10.1016/j.chroma.2012.05.050>.
- [30] L. A. Berrueta, R. M. Alonso-Salces, K. Héberger, "Supervised Pattern Recognition in Food Analysis" *Journal of Chromatography A*, vol. 1158, no. 1-2, pp. 196–214, 2007. DOI:10.1016/j.chroma.2007.05.024.
- [31] J. Yan, X. Guo, S. Duan, P. Jia, L. Wang, C. Peng, S. Zhang, "Electronic Nose Feature Extraction Methods: A Review" *Sensors*, vol. 15, no. 11, pp. 27804–27831, 2015. DOI:10.3390/s151127804.
- [32] J. Lozano, J. P. Santos, M. C. Horrillo, "Enrichment Sampling Methods for Wine Discrimination with Gas Sensors" *Journal of Food Composition and Analysis*, vol. 21, no. 8, pp. 716–723, 2008. DOI:10.1016/j.jfca.2008.07.002.
- [33] R. Sheebha, M. Subadra, "A Survey on Various Signal Pre-Processing Methods for Sensor Response in E-Nose" *Researcher*, vol. 6, no. 11, pp. 76–80, 2014. <http://www.sciencepub.net/researcher>.
- [34] N. M. Razali, Y. B. Wah, "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests" *Journal of Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21–33, 2011.
- [35] N. A. Ahad, T. S. Yin, A. R. Othman, C. R. Yaacob, "Sensitivity of Normality Tests to Non-normal Data" *Sains Malaysiana*, vol. 40, no. 6, pp. 637–641, 2011.
- [36] A. Ghasemi, S. Zahediasl, "Normality Tests for Statistical Analysis: A Guide for Non Statisticians" *International Journal of Endocrinology and Metabolism*, vol. 10, no. 2, pp. 486–489, 2012. DOI: 10.5812/ijem.3505.
- [37] L. I. Smith, *A Tutorial on Principal Components Analysis Introduction*, 2002.
- [38] C. Cortes, V. Vapnik, "Support-Vector Networks" *Machine Learning*, vol. 20, no. 3, pp. 273–297, Kluwer Academic Publishers, 1995.
- [39] Z. Haddi, A. Amari, H. Alami, N. El Bari, E. Llobet, B. Bouchikhi, "A Portable Electronic Nose System for the Identification of Cannabis-based Drugs" *Sensors and Actuators B: Chemical*, vol. 155, no. 2, pp. 456–463, 2011. DOI: 10.1016/j.snb.2010.12.047.
- [40] L. Auria, R. A. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis" *SSRN Electronic Journal*, 2008. DOI: 10.2139/ssrn.1424949.
- [41] I.-C. Chen, J. K. Hill, R. Ohlemuller, D. B. Roy, C. D. Thomas, "Rapid Range Shifts of Species Associated with High Levels of Climate Warming" *Science*, vol. 333, no. 6045, pp. 1024–1026, 2011. DOI: <http://ssrn.com/abstract=1424949>.
- [42] R. Domínguez, L. Moreno-Barón, R. Muñoz, J. Gutiérrez, "Voltammetric Electronic Tongue and Support Vector Machines for Identification of Selected Features in Mexican Coffee" *Sensors*, vol. 14, no. 9, pp. 17770–17785, 2014. DOI: 10.3390/s140917770.
- [43] J. Brezmes, N. Canyellas, E. Llobet, X. Vilanova, X. Correig, "Application of Artificial Neural Networks to the Design and Implementation of Electronic Olfactory Systems" in *5th Seminar on Neural Network Applications in Electrical Engineering (NEUREL 2000)*, Belgrade, Yugoslavia, 2000. DOI: 10.1088/0957-0233/1/5/012.
- [44] T. Li, "The Application of Artificial Neural Networks in the Electronic Nose for Odour Measurement," M.Sc. Thesis, University of Manitoba, 2005.
- [45] K. Suzuki, *Artificial Neural Networks - Industrial and Control Engineering Applications*, Rijeka: InTech, 2011.
- [46] C. S. Kudarihal, M. Gupta, "Electronic Nose Based on Metal Oxide Semiconductor Sensors as an Alternative Technique for Perception of Odours" *International Journal of Advances in Engineering & Technology*, vol. 7, no. 1, pp. 206–216, 2014.
- [47] M. Ghasemi-Varnamkhasi, S. S. Mohtasebi, M. Siadat, S. Balasubramanian, "Meat Quality Assessment by Electronic Nose (Machine Olfaction Technology)" *Sensors*, vol. 9, no. 8, pp. 6058–6083, 2009. DOI: 10.3390/s90806058.
- [48] B. Tudu, A. Jana, A. Metla, D. Ghosh, N. Bhattacharyya, R. Bandyopadhyay, "Electronic Nose for Black Tea Quality Evaluation by an Incremental RBF Network" *Sensors and Actuators B: Chemical*, vol. 138, no. 1, pp. 90–95, 2009. DOI: 10.1016/j.snb.2009.02.025.
- [49] H. D. Leon-Delgado, R. J. Praga-Alejo, D. S. Gonzalez-Gonzalez, M. Cantú-Sifuentes, "Multivariate Statistical Inference in a Radial Basis Function Neural Network" *Expert Systems with Applications*, vol. 93, pp. 313–321, 2018. DOI: 10.1016/j.eswa.2017.10.024.
- [50] N. P. Thanh, Y.-S. Kung, S.-C. Chen, H.-H. Chou, "Digital Hardware Implementation of a Radial Basis Function Neural Network" *Computers & Electrical Engineering*, vol. 53, pp. 106–121, 2016. <http://dx.doi.org/10.1016/j.compeleceng.2015.11.017>.
- [51] K. Wetchakun, T. Samerjai, N. Tamaekong, C. Liewhiran, C. Siriwong, V. Kruefu, A. Wisitsoraat, A. Tuantranont, S. Phanichphant, "Semicustoming Metal Oxides as Sensors for Environmentally Hazardous Gases" *Sensors and Actuators B: Chemical*, vol. 160, no. 1, pp. 580–591, 2011. DOI: 10.1016/j.snb.2011.08.032.
- [52] M. A. A. Bakar, A. H. Abdullah, F. S. A. Sa'ad, S. A. A. Shukor, M. H. Mustafa, M. S. Kamis, A. A. A. Razak, S. A. Saad, "Electronic Nose Sensing Chamber Design for Confined Space Atmospheric Monitoring" in *International Conference on Mathematics, Engineering and Industrial Applications 2016 (ICoMEIA)*, Songkhla, Thailand, 2016. <http://dx.doi.org/10.1063/1.4965179>.
- [53] M.A.A. Bakar, A.H. Abdullah, F.S.A. Sa'ad, S.A.A. Shukor, A.A.A. Razak, M.H. Mustafa, "Electronic Nose Calibration Process for Monitoring Atmospheric Hazards in Confined Space Applications" *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 2016.
- [54] A. F. Najmuddin, I. M. Ibrahim, S. R. Ismail, "A Simulation Approach: Improving Patient Waiting Time for Multiphase Patient Flow of Obstetrics and Gynecology Department (O&G Department) in Local Specialist Centre" *WSEAS TRANSACTIONS ON MATHEMATICS*, vol. 9, no. 10, pp. 778–790, 2010.
- [55] K.-T. Tang, S.-W. Chiu, C.-H. Pan, H.-Y. Hsieh, Y.-S. Liang, S.-C. Liu, "Development of a Portable Electronic Nose System for the Detection and Classification of Fruity Odors" *Sensors*, vol. 10, no. 10, pp. 9179–9193, 2010. DOI: 10.3390/s101009179.
- [56] H. Ishida, Y. Wada, H. Matsukura, "Chemical Sensing in Robotic Applications: A Review" *IEEE Sensors Journal*, vol. 12, no. 11, pp. 3163–3173, 2012. DOI: 10.1109/JSEN.2012.2208740.

Trajectory Tracking Control Optimization with Neural Network for Autonomous Vehicles

Samuel Oludare Bamgbose*, Xiangfang Li, Lijun Qian

Department of Electrical and Computer Engineering and CREDIT Center, Prairie View A & M University, Texas A & M University System, Prairie View, Texas, 77446, USA

ARTICLE INFO

Article history:

Received: 05 January, 2019

Accepted: 03 February, 2019

Online : 15 February, 2019

Keywords:

Nonlinear control

Optimal control

Neural network

Adaptive control

Autonomous systems

Mobile robot

2-DOF helicopter

ABSTRACT

For mission-critical and time-sensitive navigation of autonomous vehicles, controller design must exhibit excellent tracking performance with respect to the speed of convergence to reference command and steady-state accuracy. In this article, a novel design integration of the neural network with the traditional control system is proposed to adaptively obtain optimized controller parameters resulting in improved transient and steady-state performance of motion and position control of autonomous vehicles. Application of the proposed intelligent control scheme to mobile robot navigation was presented for an eight-shaped trajectory by optimizing a Lyapunov-based nonlinear controller. Furthermore, a Linear Quadratic Regulator-based controller was optimized based on the proposed strategy to control the pitch and yaw angles of a 2-Degree-of-Freedom helicopter. The simulation results showed that the proposed scheme outperforms the traditional controllers in terms of the speed of convergence to the desired trajectory and overall error minimization.

1. Introduction

There have been growing demands for autonomous vehicles with excellent maneuvering capabilities both for commercial and military applications [1, 2]. Significant amount of work is ongoing to realize the commercial operation of autonomous cars [3, 4], and Wheeled Mobile Robots (WMRs) are finding increasing use in industrial and service applications [5]. Autonomous Surface Vehicles (ASV) have been utilized to improve port safety and for ecological as well as meteorological purposes [6], whereas Autonomous Underwater Vehicles (AUVs) are useful tools for underwater search and inspection [7].

Safe navigation will require both excellent path planning and path tracking strategies. Path planning involves the determination of an appropriate trajectory for the vehicle whereas path tracking, which is the focus of this study, is the following of a desired trajectory. Several schemes have been reported in the literature for path planning, including in [8], where a framework was proposed to synthesize sequences of maneuvers that are tracked using nonlinear controllers, and other strategies are presented in [9, 10]. To execute the maneuvers generated by path planners, control strategies have been proposed as well.

In [11], Fuzzy Logic Controller (FLC) was proposed for robot tracking and Proportional-Integral-Derivative (PID) controller was

utilized to control the speed of WMRs in [12]. The authors of [13] presented an observer-based approach, whereas backstepping control strategy was presented in [14, 15]. Sliding-mode control for WMR trajectory tracking with initial error was reported in [16], and modular-based method was proposed in [17]. The authors of [18] proposed a Model Predictive Control (MPC) approach and Kalman filter-based strategy was presented in [19]. Lyapunov-based scheme was reported in [20] and controller design by approximate linearization utilizing Taylor expansion was proposed in [21]. In order to enhance the performance of traditional control methods, Least Square Policy Iteration (LSPI) and Dynamic Heuristic Programming (DHP) algorithms were utilized in [22] for optimizing a Proportional-Derivative (PD) controller. Also, the authors of [23] utilized neural network (NN) to describe the inverse dynamics of a biped robot with respect to output errors for the control of level walking. However, application examples were limited to scenarios where the robot is initialized at the desired starting coordinate.

In order to execute maneuvers for unmanned Air Vehicles (UAVs), feedback linearization [24, 25] and sliding mode control [26] have been proposed, but they have failed for certain models due to the nonlinear dynamics of the vehicle [27]. For helicopter control, backstepping control strategy [28] and Linear Quadratic Regulator (LQR)-based control [29] have been proposed.

* Samuel Oludare Bamgbose Email: sobams77@yahoo.com

www.astesj.com

<https://dx.doi.org/10.25046/aj0401121>

In this study, a novel learning-based adaptive scheme utilizing the neural network is developed for autonomous vehicle trajectory tracking. Whereas, plant models predict the vehicular motion for a given control command, accuracy is limited by modelling errors and approximations. Also, for certain models with complex nonlinear dynamics, it is difficult to obtain a suitable controller. Since machine learning models provide a more powerful tool to describe nonlinear dynamics of a plant given example data of plant operation, this article presents a parameterized control law designed to achieve trajectory tracking of autonomous vehicles adaptively. Rather than using a single controller, a family of controllers is obtained utilizing the NN model to estimate the time varying controller parameters. The scheme was applied to optimize a Lyapunov-based nonlinear controller parameters used to execute an eight-shaped maneuver for a mobile robot. Furthermore, LQR-based controller parameters for 2-Degree-Of-Freedom (DOF) helicopter position control were optimized using the proposed scheme. Simulation results show that the scheme outperforms the traditional control strategies in terms of faster convergence to the desired trajectory and more accurate steady-state performance, regardless of the initial shift in starting coordinates. Also, because the scheme is sample-based, it can compensate for modeling errors.

The rest of the paper is organized as follows; Section 2 presents the problem formulation. Section 3 provides an illustrative application of the proposed scheme. Simulation results are discussed in Section 4, and Section 5 summarizes the study.

2. Problem Formulation

A dynamic system can generally be defined in state-space as,

$$\dot{x}(t) = f(x(t), u(t)) \quad (1)$$

The system state is represented by $x(t)$, the control input is denoted by $u(t)$, whereas f denotes a mathematical function and $\dot{x}(t)$ is the state derivative with respect to time. Given a target signal $r(t)$, a traditional feedback control law can be computed based on the difference between the target signal and the actual system output. Such difference can be denoted by $e(t)$, resulting in a control input defined as,

$$u(t) = K(t)e(t) \quad (2)$$

where $K(t)$ denote the control gain, which can be constant or time-varying depending on the dynamic description of the system. However, by utilizing certain control strategies, not only the error is fed back for control but the states also. For two-dimensional and higher order systems; the states, control input and error signals are vectors, whereas the control gain could be a vector or matrix depending on the control strategy used.

In order to optimize the controller performance for changing system dynamics or operation regions, a neural network-based method that adaptively determines the control gain is proposed as follows. For a traditional closed-loop control system, sample test run of the system is performed and for each time step k , the state variables and the control gain are measured and arranged into 3-tuple $\{x(t), x(t+k), K(t)\}$. The state measurement before the application of control input is denoted as $x(t)$ whereas the next state caused by the control action is $x(t+k)$. The control gain that caused the state transition is represented as $K(t)$.

Multiple samples of state transitions and control gains are measured for a sequence of operation, then the control gain or the control gain parameters for time-invariant or time-varying system respectively are varied for other sequences. Example of a sequence is the tracking of a particular course by a mobile robot or the change in angular orientation of a helicopter over a period of time. For training stability, the data set is normalized depending on the data range. And based on the accumulated data set as represented by the 3-tuple, a NN is trained using the state and next state as input and the control gain as the output. Hence,

$$K(t) = f(x(t), x(t+k)) \quad (3)$$

The NN model is made up of processing units called neurons with weighted interconnections among them [30], and the neurons with nonlinear activation functions are arranged into layers as shown in Figure 1.

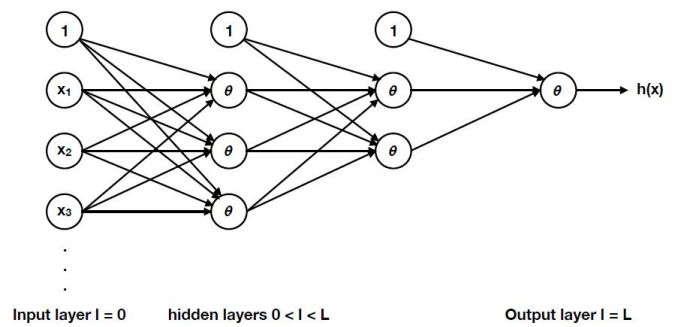


Figure 1: Feed forward neural network structure

The input layer is the vector of input variables, which are system states in the study. For a single hidden layer system, the hidden layer is a nonlinear combination of the input signals, and for multiple hidden layer system, subsequent layers are nonlinear combination of the previous layers. Knowledge is extracted from the output layer after model training. In supervised learning employed in this study, an error function based on the difference between the actual outputs called labels and the predicted outputs from the NN is minimized to adjust the interconnection weights among the neurons. After several iterations, the model is said to be trained. Whereas only a portion of the dataset is used for training, the remainder is used for testing the performance of the trained model. The mathematical description of the NN model as utilized in this study is,

$$K_n^{(l+1)} = \theta(s_n^{(l+1)}) \quad (4)$$

$$s_n^{(l+1)} = w_n^{(l+1)}K^l + b_n^{(l+1)} \quad (5)$$

where the output vector of a layer l is denoted by K^l . Hence, the input layer vector K^0 is equivalent to the state input $\{x(t), x(t+k)\}$. The activation function is denoted by θ and s^l represents the input vector to layer l . The interconnecting weights and biases from layer $l-1$ to unit n of layer l are represented by w_n^l and b_n^l respectively.

Following offline training as described above, the trained NN model is integrated as a control gain estimator in the traditional control system as shown in Figure 2.

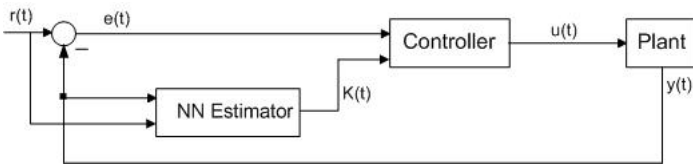


Figure 2: Neural network optimized adaptive control system

For online control gain estimation, the NN input due to the next state $x(t + k)$ is replaced with the reference signal $r(t)$. Hence,

$$K(t) = f(x(t), r(t)) \quad (6)$$

Then, (2) is transformed to,

$$u(t) = f(x(t), r(t)) * e(t) \quad (7)$$

3. Illustrative Examples

In order to prove the effectiveness of the proposed scheme, two application case studies are presented. One is the tracking of an eight-shaped trajectory by a mobile robot, and the other is the position tracking of a 2-DOF helicopter.

3.1. Mobile Robot Trajectory Tracking

The nonlinear kinetic model [31] of a non-holonomic wheeled mobile robot is described as,

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} \cos\theta & 0 \\ \sin\theta & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} \quad (8)$$

The robot's states x, y, θ are the cartesian x, y , and angular displacements respectively. Whereas \dot{x}, \dot{y} , and $\dot{\theta}$ are the state derivatives, v corresponds to the linear velocity and w represents the angular velocity around the vertical axis.

3.1.1. Traditional Control Design

Motion control of the mobile robot is achieved by computing appropriate linear and angular velocities to drive the robot in the desired trajectory. According to the procedure in [32], the state tracking error is generalized by defining it through a rotation matrix to obtain,

$$\begin{bmatrix} e_x \\ e_y \\ e_\theta \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_d - x \\ y_d - y \\ \theta_d - \theta \end{bmatrix} \quad (9)$$

where x_d, y_d , and θ_d are the target coordinates for the robot. Then the error dynamics in generalized coordinates is obtained by taking the derivative of (9) as,

$$\begin{bmatrix} \dot{e}_x \\ \dot{e}_y \\ \dot{e}_\theta \end{bmatrix} = \begin{bmatrix} 0 & wd & 0 \\ -wd & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} e_x \\ e_y \\ e_\theta \end{bmatrix} + \begin{bmatrix} vdcos\theta - v \\ vdsine\theta \\ wd - w \end{bmatrix} \quad (10)$$

Making the following substitutions,

$$u_1 = vdcos\theta - v \quad (11)$$

$$u_2 = wd - w \quad (12)$$

The error dynamics is transformed to the subsequent form,

$$\begin{bmatrix} \dot{e}_x \\ \dot{e}_y \\ \dot{e}_\theta \end{bmatrix} = \begin{bmatrix} 0 & w_d & 0 \\ -w_d & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} e_x \\ e_y \\ e_\theta \end{bmatrix} + \begin{bmatrix} 0 \\ sine\theta \\ 0 \end{bmatrix} v_d + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad (13)$$

To ensure global stability for the nonlinear dynamics, the following Lyapunov function is selected,

$$V = \frac{K_2}{2} (e_x^2 + e_y^2) + \frac{e_\theta^2}{2} \quad (14)$$

Taking the derivative of (14) gives

$$\dot{V} = K_2 (e_x \dot{e}_x + e_y \dot{e}_y) + e_\theta \dot{e}_\theta \quad (15)$$

By substituting for \dot{e}_x, \dot{e}_y and \dot{e}_θ from (13),

$$\dot{V} = K_2 (e_x u_1 + e_y v_d sine\theta) + e_\theta u_2 \quad (16)$$

Hence, the Lyapunov-based control law such that $\dot{V} < 0$ is guaranteed is obtained as,

$$u_1 = -K_1 e_x \quad (17)$$

$$u_2 = -K_2 v_d \frac{sine\theta}{e_\theta} e_y - K_3 e_\theta \quad (18)$$

The controller gains are defined as,

$$K_1(t) = K_3(t) = 2\zeta \sqrt{w_d^2(t) + cv_d^2(t)} \quad (19)$$

where $K_2 = c > 0$ is a constant, and ζ is the damping ratio.

The control law defined by (17) and (18) are composed of both the feedback and feedforward signals. The feedforward commands are determined from (8) as,

$$v_{d(t)} = \pm \sqrt{\dot{x}_d^2(t) + \dot{y}_d^2(t)} \quad (20)$$

$$w_d(t) = \frac{\dot{y}_d(t)\dot{x}_d(t) - \dot{x}_d(t)\dot{y}_d(t)}{x_d^2(t) + y_d^2(t)} \quad (21)$$

$$\theta_d(t) = atan2(\dot{y}_d(t), \dot{x}_d(t)) + \pi i \quad (22)$$

where i is 0 or 1 for forward or backward motion respectively. The feedback commands are adaptively computed based on NN estimation as described next.

3.1.2. Controller Parameter Estimation using the Neural Network

In order to optimize the controller for fast convergence to the desired trajectory and improved steady-state performance, the NN is utilized to adaptively estimate the feedback control parameters. Sample runs of the traditional closed-loop system are performed for 15 sequences of 1258 state transitions to make a total of 18870 data samples. The input features are the current state $[x(t) \ y(t) \ \theta(t)]$, the nominal velocities $[v_{d(t)} \ w_{d(t)}]$, and

the next state $[x(t+k) \ y(t+k) \ \theta(t+k)]$. The output labels are the control gains $[K_1(t) \ K_2(t) \ K_3(t)]$ responsible for the state transition. Data was normalized and the training was implemented using the MATLAB NN tool box using 70% of the dataset. The remaining dataset was divided into 15% validation and 15% testing. The Levenberg-Marquardt algorithm [33, 34, 35] was used for training using two hidden neurons. The performance of the NN estimator was evaluated using the Mean Square Error (MSE) plot shown in Figure 3. The figure shows good generalization performance as the test error is approximately equal to that of the training.

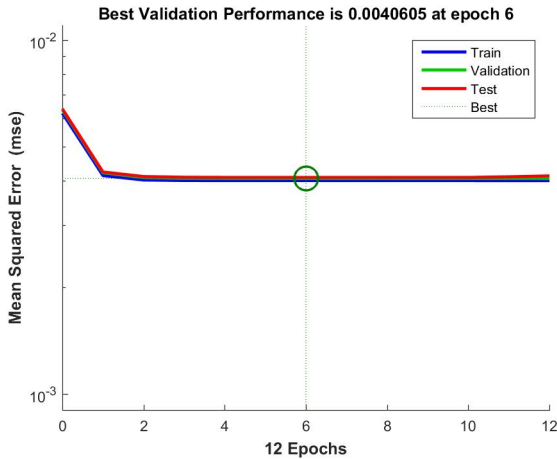


Figure 3: MSE of the NN estimator for robot control

The trained NN estimator is integrated in the second loop of Figure 2, where a normalizer and denormalizer are embedded for better control performance. Simulation results of the adaptive closed loop control system is presented in Section 4.

3.2. 2-DOF Helicopter Position Control

A Two-Input-Two-Output (TITO) Quanser 2D Helicopter model [35] is considered. The linearized model is obtained in [36] as follows.

$$\begin{bmatrix} \dot{\theta} \\ \dot{\phi} \\ \ddot{\theta} \\ \ddot{\phi} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1.0000 & 0 \\ 0 & 0 & 0 & 1.0000 \\ 0 & 0 & -0.0315 & 0 \\ 0 & 0 & 0 & -0.0673 \end{bmatrix} \begin{bmatrix} \theta \\ \phi \\ \dot{\theta} \\ \dot{\phi} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0.5293 & 0.0404 \\ 0.0372 & 0.1333 \end{bmatrix} \begin{bmatrix} V_{mp} \\ V_{my} \end{bmatrix} \quad (23)$$

where θ and ϕ are the pitch and yaw angles respectively. The first and second derivatives of the pitch and yaw angles are $\dot{\theta}$, $\ddot{\theta}$ and $\dot{\phi}$, $\ddot{\phi}$ respectively. The pitch and yaw propeller voltages V_{mp} and V_{my} are applied to the corresponding motors driving the propellers.

3.2.1. Traditional Control Design

The pitch and yaw angles of the 2-DOF helicopter were regulated using the optimal LQR-based control design presented in [29]. And the control law was obtained as,

$$u = K_x x(t) - K_e e(t) \quad (24)$$

The realized state feedback gains are,

$$K_x = \begin{bmatrix} 547.8 & -2.8 & 45.7 & -4.8 \\ 2.8 & 547.9 & -4.1 & 91 \end{bmatrix} \quad (25)$$

and the error feedback gains were obtained as,

$$K_e = \begin{bmatrix} -1.0000 & 0.0051 \\ -0.0051 & -1.0000 \end{bmatrix} \quad (26)$$

The system states are represented by $x(t)$ whereas $e(t)$ is the error feedback signal. The adaptive computation of the error feedback gains using the NN is subsequently described.

3.2.2. Controller Parameter Estimation using the Neural Network

In order to achieve better accuracy and faster tracking convergence, the NN is employed to optimize the error feedback control gains. In this case, six sequences of 700 state transitions totaling 4195 data samples were sufficient, since the plant has been linearized. The input features are the current angular orientation $[\theta(t) \ \phi(t)]$ along with the next angular orientation $[\theta(t+k) \ \phi(t+k)]$ of the 2-DOF helicopter. The output labels are the error feedback control gains $[K_{e11}(t) \ K_{e12}(t) \ K_{e21}(t) \ K_{e22}(t)]$ responsible for the angular orientation change. Similar procedures as proposed in Subsession 3.1.2 of this paper are followed, and the MSE plot is as shown in Figure 4. The figure shows good generalization of the estimator similar to observation in the previous section since the test error approximates the training error. The trained NN estimator is integrated in the second loop of Figure 2 to obtain optimized performance as presented next in Section 4.

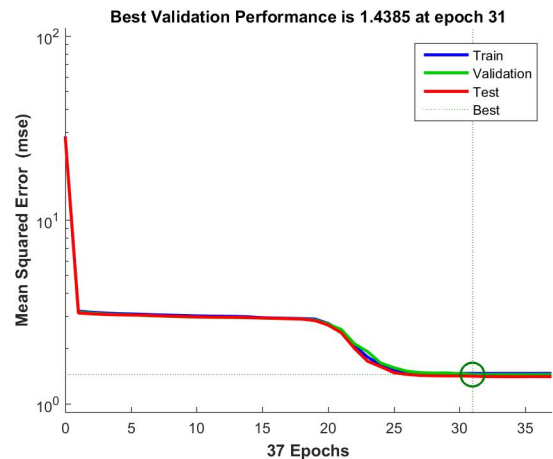


Figure 4: MSE of the NN estimator for helicopter control

4. Simulation Results

Closed-loop simulation was conducted in Matlab Simulink for both the traditional and the NN-optimized control system. The results for different initial state shifts were compared to prove the effectiveness of the proposed scheme. The desired eight-shaped trajectory for the mobile robot was defined as,

$$x_d(t) = \sin \frac{t}{10} \quad (25)$$

$$y_d(t) = \sin \frac{t}{20} \quad (26)$$

$$\theta d(t) = \text{atan2}(\dot{y}d(t), \dot{x}d(t)) \quad (27)$$

The feed-forward driving and steering velocities as described by (20) and (21) are shown in Figures 5 and 6 respectively.

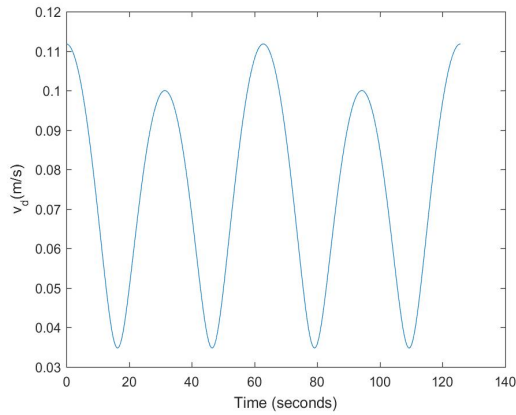


Figure 5: Robot feed-forward driving velocity input

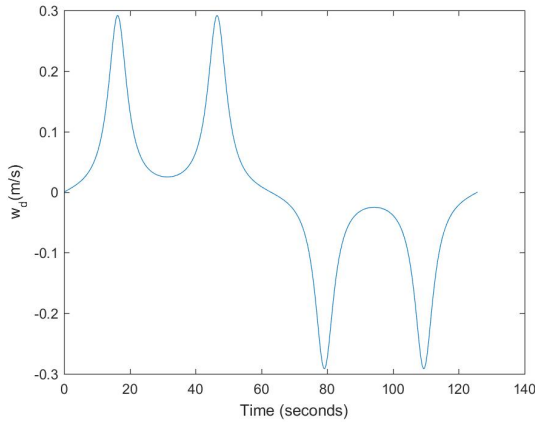


Figure 6: Robot feed-forward steering velocity input

Figures 7 – 9 is the trajectory tracking performance of the nonlinear Lyapunov-based controller for the mobile robot with initial state error $[0.1, 0.3, 0]$, where the linear displacement unit is meters and radians is the unit of the angular displacement. Figure 7 is the result of an underdamped control system due to low damping ratio, whereas Figure 8 is the result of an overdamped control system due to high damping ratio. The result of a critically damped control system with unity damping ratio is shown in Figure 9, whereas Figure 10 is the corresponding trajectory tracking output performance of the NN-optimized control system. It can be observed that the NN-optimized controller outperformed the different design modes of Lyapunov-based controller in terms of faster convergence to the desired trajectory and steady-state error minimization. The steady-state is the region starting at the point where the robot settles around the reference trajectory. Although, either fairly good convergence or steady-state performance can be obtained by varying the control parameters of the Lyapunov-based controller, but the improvement is mutually exclusive. However, the NN-optimized controller achieved both better transient and steady state performances concurrently by adaptively obtaining varying controller parameters.

Figures 11 - 14 are the control performance outputs for the mobile robot trajectory tracking with initial state error $[0.3, -0.5, \pi/6]$. Also in this case, linear and angular displacements are measured in meters and radians respectively. It can be observed, as the case is in the previous example, that the NN-optimized controller of Figure 14 showed better comparative control performance than the different variations of traditional Lyapunov-based controllers of Figures 11 – 13.

To further show the performance advantage of the NN-optimized controller, the average linear and angular tracking errors over the entire trajectory are presented in Table 1 as it applies to the different initial state errors. It can be seen that both the resultant linear and angular tracking errors are minimized using the NN-optimized controller.

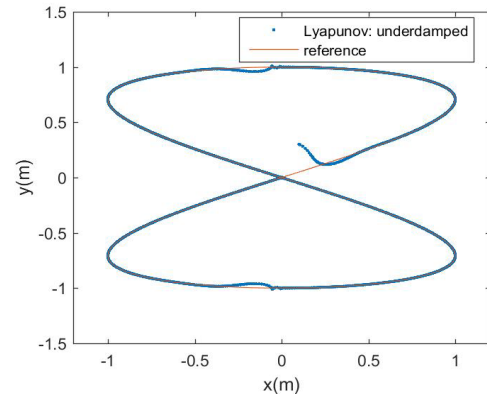


Figure 7: Underdamped robot Lyapunov-based control

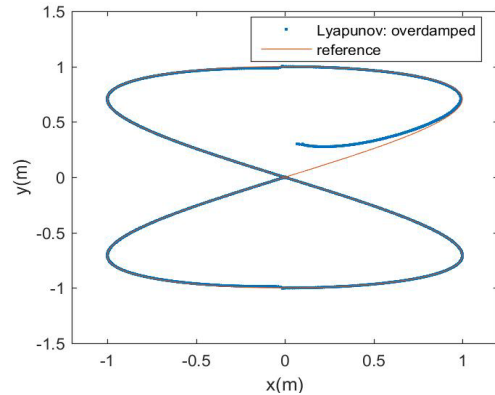


Figure 8: Overdamped robot Lyapunov-based control

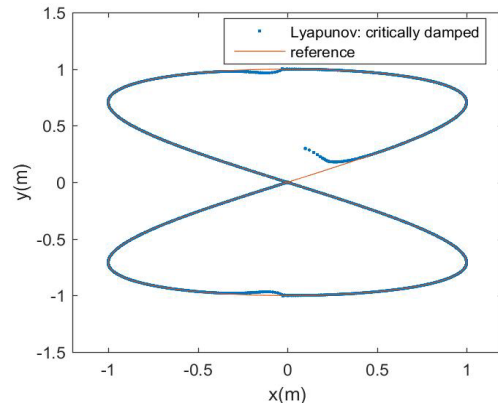


Figure 9: Critically damped robot Lyapunov-based control

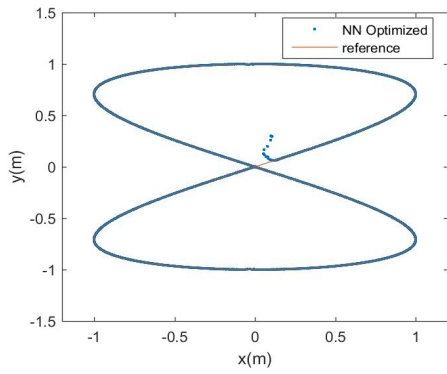


Figure 10: Robot NN-optimized control

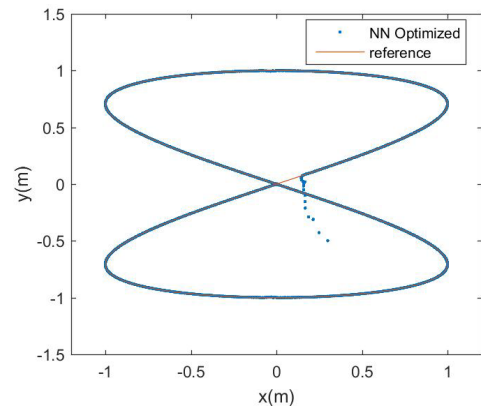


Figure 14: Robot NN-optimized control

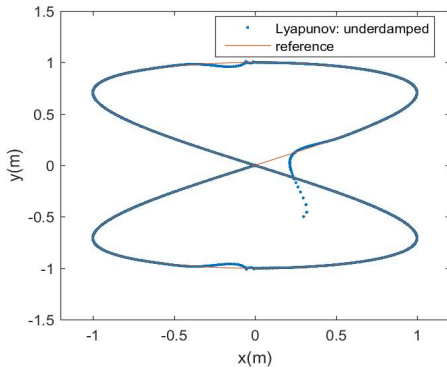


Figure 11: Underdamped robot Lyapunov-based control

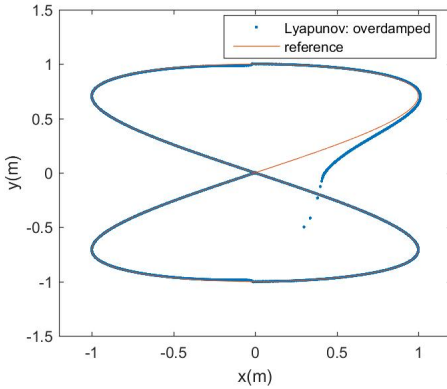


Figure 12: Overdamped robot Lyapunov-based control

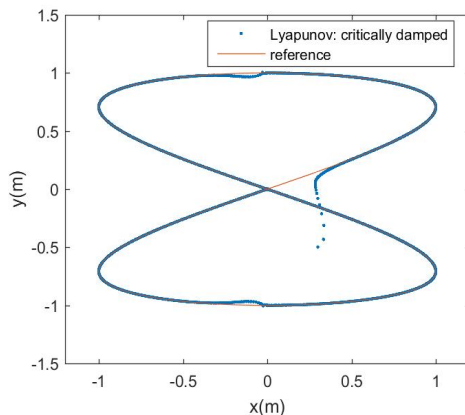


Figure 13: Critically damped robot Lyapunov-based control

Table 1: Robot mean tracking error based on control method

Control Method	Initial state error [m,m,rad]	Mean x error (m)	Mean y error (m)	Mean θ error (m)
Lyapunov_ underdamped	[0.1,0.3,0]	1.7×10^{-3}	-2.3×10^{-3}	2.4×10^{-2}
Lyapunov_ overdamped	[0.1,0.3,0]	-1.8×10^{-4}	-9.2×10^{-3}	2.3×10^{-2}
Lyapunov_ critically damped	[0.1,0.3,0]	2.8×10^{-4}	-3.8×10^{-3}	2.6×10^{-2}
Neural network optimized	[0.1,0.3,0]	3.5×10^{-4}	-2.5×10^{-4}	2.2×10^{-2}
Lyapunov_ underdamped	[0.3,-0.5, $\pi/6$]	3.5×10^{-3}	2.1×10^{-3}	-2.1×10^{-2}
Lyapunov_ overdamped	[0.3,-0.5, $\pi/6$]	4.4×10^{-4}	1.33×10^{-2}	-3.5×10^{-2}
Lyapunov_ critically damped	[0.3,-0.5, $\pi/6$]	1.4×10^{-3}	4.1×10^{-3}	-2.3×10^{-2}
Neural network optimized	[0.3,-0.5, $\pi/6$]	-5.1×10^{-4}	-1.3×10^{-4}	-3.3×10^{-2}

The desired positioning of the 2-DOF helicopter pitch and yaw angular orientations over time are defined as,

$$\theta_d(t) = \begin{cases} 0.8\tau^5 - 8\tau^3 + 10\tau + 6, & \tau = 0.05t, t \leq 47 \\ -16.9, & t > 47 \end{cases} \quad (28)$$

$$\varphi_d = 5 \quad (29)$$

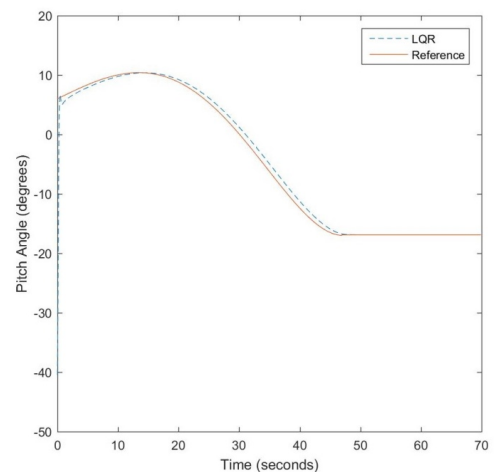


Figure 15: LQR-based pitch angle control

The LQR-based controller performance is shown in Figures 15 and 17 with initial pitch and yaw angular orientations of -40.5 and 0 degrees. The Q and R parameters were selected as $diag([3e5 \ 3e5 \ 1 \ 1 \ 1 \ 1])$ and $diag([1 \ 1])$ where Q is associated with both the state and error feedback signals. The performances of the NN-optimized controller are shown in Figures 16 and 18 for pitch and yaw angle control respectively. Since the system is linearized, modest performance improvement is observed with respect to response speed and steady-state error minimization. Table 2 shows the comparative tracking errors.

Table 2: Helicopter mean tracking error based on control method

Control Method	Initial state error (degrees)	Mean θ error (degrees)	Mean ϕ error (degrees)
LQR	[-40.5,0]	-0.18	0.08
Neural network optimized	[-40.5,0]	-0.06	0.04

5. Conclusion

A neural network optimized control system design for autonomous vehicle navigation has been proposed in this study. The design consists of an inner error loop integrated with an outer loop for estimating the controller parameters utilizing a neural network trained on samples from test navigation. Comparative studies of the proposed scheme with traditional methods were presented. In the first case, the neural network optimized control system was shown to outperform a Lyapunov-based controller in terms of faster convergence to the desired trajectory and better steady-state performance for an eight-shaped maneuver. A second illustrative simulation example was conducted for the control of pitch and yaw angles of a 2-Degree-Of-Freedom helicopter model, where improved transient and steady-state performances were observed for NN-optimized controller over a Linear Quadratic Regulator-based controller.

Because the neural network structure allows complex and nonlinear mapping of variables, the trained estimator is able to learn the behavior of both linear and nonlinear systems. Furthermore, since the proposed scheme is sample-based, it can compensate for plant modelling errors that degrades the performance of traditional controllers in real-life applications.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The research work presented in this article is supported in part by the US National Science Foundation award 1464387, 1736196 and by the U.S. Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under agreement number FA8750-15-2-0119. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the US National Science Foundation or the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) or the U.S. Government.

References

- [1] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffman et al, "Stanley: The robot that won the DARPA grand challenge". J. Field Robot, **23**(9), 661-692, 2006
- [2] D. Weatherington, Unmanned Aerial Vehicle Roadmap Report, The Information Warfare Site, 2003. <http://www.iwar.org.uk/news-archive/2003/03-18-5.htm>
- [3] T. Luettel, M. Himmelsbach, H. J. Wuensche, "Autonomous ground vehicles-concepts and a path to the future" in Proc. of the IEE, **100** (Special Centennial Issue), 1831-1839, 2012. <https://doi.org/10.1109/JPROC.2012.2189803>

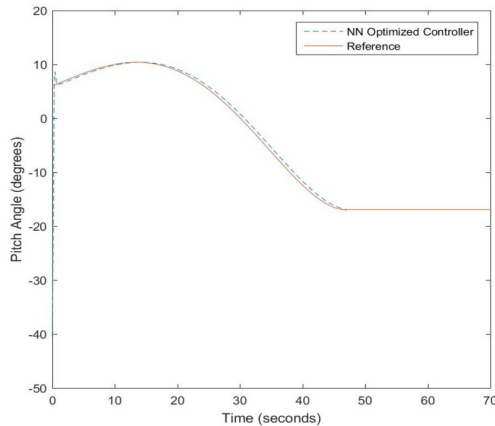


Figure 16: NN-optimized pitch angle control

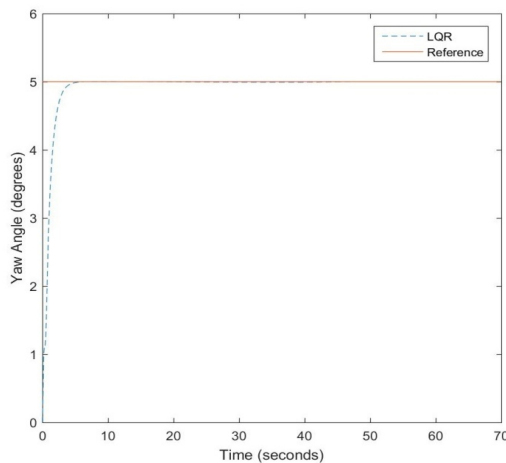


Figure 17: LQR-based yaw angle control

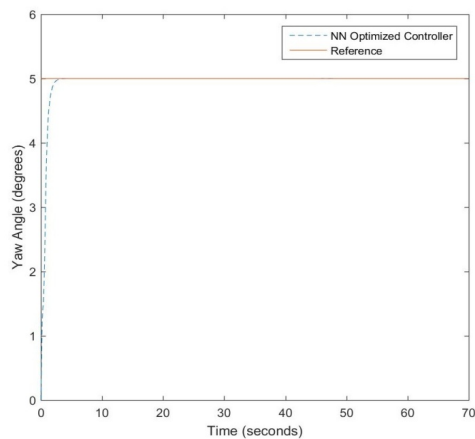


Figure 18: NN-optimized yaw angle control

- [4] T. Drage, J. Kalinowski, T. Braunl, "Integration of drive-by-wire with navigation control for a driverless electric race car" *IEEE Intelligent Transportation Systems Magazine*, **6**(4), 23-33, 2014. <https://doi.org/10.1109/MITS.2014.2327160>
- [5] R. D. Schratt, G. Schmierer, *Serviceboter*, Springer, Berlin, 1998
- [6] N. Miskovic, D. Nad, L. Rendulic, "Tracking divers: An autonomous marine surface vehicle to increase diverse safety" *IEEE Robotics Automation Magazine*, **22**(3), 72-84, 2015. <https://doi.org/10.1109/MRA.2015.2448851>
- [7] F. Plumet, C. Petres, M.A. Romero-Ramirez, B. Gas, S.H Long, "Toward and autonomous sailing boat" *IEEE Journal of Oceanic Engineering*, **40**(2), 397-407, 2015. <https://doi.org/10.1109/JOE.2014.2321714>
- [8] E. Frazzoli, M.A. Dahleh, E. Feron, "Maneuver based motion planning for nonlinear systems with symmetries" *IEEE Trans. Robot. Automat.*, **21**, 1077-1091, 2005. <https://doi.org/10.1109/TRO.2005.852260>
- [9] C. Dever, B. Mettler, E. Feron, J. Popovic, M. McConley, "Nonlinear trajectory generation for autonomous vehicles via parameterized maneuver classes" *J. Guid. Control Dyn.*, **29**(2), 289-302, 2006 <https://doi.org/10.2514.1.13400>
- [10] N. K. Ure, G. Inalhan, "Autonomous Control of unmanned combat air vehicles: design of a multimodel control and flight planning framework for agile maneuvering" *IEEE Control Systems Magazine*, **32**(5), 74-95, 2012. <https://doi.org/10.1109/MCS.2012.2205532>
- [11] S. Qiu, Z. Li, W. He, L. Zhang, C. Yang, C. Y. Su, "Brain-machine interface and visual compressive sensing-based teleoperation control of an exoskeleton robot" *IEEE Trans. On Fuzzy Systems*, **25**(1), 58-69, 2017. <https://doi.org/10.1109/TFUZZ.2016.2566676>
- [12] C. S. Shijin, K. Udayakumar, "Speed control of wheeled mobile robots using pid with dynamics and kinetic modelling" in 4th International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 1-7, 2017. <https://doi.org/10.1109/ICIIECS.2017.8275962>
- [13] M. Chen, "Disturbance attenuation tracking control for wheeled mobile robots with skidding and slipping" *IEEE Trans. On Industrial Electronics*, **64**(4), 3359-3368, 2017. <https://doi.org/10.1109/TLE.2016.2613839>
- [14] W. M. E. Mahgoub, I. M. H. Sanhoury, "Back stepping tracking controller for wheeled mobile robot" in Proc. of International Conference on Communications, Control, Computing and Electronics Engineering (ICCCCEE), 1-5, 2017. <https://doi.org/10.1109/ICCCCEE.2017.7867663>
- [15] O. Chwa, "Sliding-mode tracking control of nonholonomic wheeled mobile robots in planar coordinates" *IEEE Trans. Control Syst. Technol.*, **12**, 637-666, 2004. <https://doi.org/10.1109/TCST.2004.824953>
- [16] Z. Jiang, H. Nijmeijer, "Tracking control of mobile robots: a case study in backstepping" *Automatica*, **33**, 1393-1399, 1997 [https://doi.org/10.1016/S0005-1098\(97\)00055-1](https://doi.org/10.1016/S0005-1098(97)00055-1)
- [17] M. M. Michalek, "Cascade-like modular tracking controller for non-standard N-Trailers" *IEEE Trans. On Control Systems Technology* **25**(2), 619-627, 2017. <https://doi.org/10.1109/TCST.2016.2557232>
- [18] H. Yang, M. Guo, Y. Xia, L. Cheng, "Trajectory tracking for wheeled mobile robots via model predictive control with softening constraints". *IET Control Theory and Applications*, **12**(2), 206-214, 2018. <https://doi.org/10.1049/iet.cta.2017.0395>
- [19] M. Cui, W. Liu, H. Liu, X. Lu, "Unscented Kalman filter-based adaptive tracking control for wheeled mobile robots in the presence of wheel slipping" in Proc. 12th World Congress on Intelligent Control and Automation (WCICA), 3335-3340. <https://doi.org/10.1109/WCICA.2016.7578636>
- [20] Y. Wang, Z. Miao, H. Zhong, Q. Pan, "Simultaneous stabilization and tracking of nonholonomic mobile robots: a lyapunov-based approach" *IEEE Trans. On Control Systems Technology*, **23**(4), 1440-1450, 2015 <https://doi.org/10.1109/TCST.2014.2375812>
- [21] F. Kuhne, W. F. Lages, J. M. G. da Silva, L. Jono "Model predictive control of a mobile robot using linearization" in Proc. Mechatronics and Robotics, 2004.
- [22] H. Yang, Q. Guo, X. Xu, C. Lian, "Self-learning pd algorithms based on approximate dynamic programming for robot motion planning" in 2014 International Joint Conference on Neural Networks (IJCNN), 3663-3670, 2014. <https://doi.org/10.1109/IJCNN.2014.6889711>
- [23] J. K. Rai, V. P. Singh, R. P. Tawari, D. Chandra, "Artificial Neural Network controllers for biped robot" 2nd International Conference on Power, Control and Embedded Systems, 625-630, 2012. <https://doi.org/10.1109/ICPCES.2012.6508093>
- [24] A. Isidori, *Nonlinear Control Systems*, M. Thoma, Ed., Springer NewYork, 1995.
- [25] S. Sastry, *Nonlinear Systems*, J. E. Marsden, L. Sirovich, S. Wiggins, Eds., Springer New York, 1999.
- [26] V. I. Utkin, *Sliding Modes in Control and Optimization*, Springer Berlin, 1992
- [27] J. J. E Slobine, W. Li, *Applied Nonlinear Control*, Prentice Hall, Englewood Cliffs, NJ, 1991
- [28] E. Frazzoli, M. Dahleh, E. Feron, "Trajectory tracking control design for autonomous helicopters using backstepping algorithm" in Proc. American Control Cont., **6**, 4102-4107, 2000. <https://doi.org/10.1109/ACC.2000.876993>
- [29] X. Meng, Y. Zhang, J. Zhang, "Multiple input/output delays compensation for networked mimo systems" in 25th Chinese Control and Decision Conference (CCDC), 2545-2550, 2013. <https://doi.org/10.1109/CCDC.2013.6561369>
- [30] Y. S. Abu-Mostafa, M. Magdon-Ismael, H. Lin, *Learning from Data*, AMC book.com, USA, 2012.
- [31] S. Blazic, "Two approaches for nonlinear control of wheeled mobile robots" in 13th IEEE International Conference on Control Automation (ICCA), 946-951, 2017. <https://doi.org/10.1109/ICCA.8003188>
- [32] A. D. Luca, G. Oriolo, M. Venditteli, *Control of Wheeled Mobile Robots: An Experimental Overview* In: S. Nicosia, B. Siciliano, A. Bicchi, P. Valigi (eds) *Ramsete. Lecture Notes in Control and Information Sciences*, **270**, Springer, Berlin, Heidelberg, 2001.
- [33] K. Levenberg, "A method for the solution of certain problems in least square" *Quart. Appl. Math.*, **2**, 164-168, 1944. <https://doi.org/10.1109/qam/10666>
- [34] D. Marquardt, "An algorithm for least-square estimation of linear parameters" *SIAM J. Appl. Math.*, **11**(2), 431-441, 1963. <https://dx.doi.org/10.1137/0111030>
- [35] A. Ranganathan, "The Levenberg-Marquardt Algorithm". Honda Research Institute USA, 2004. <http://anath.in/Notes-files/lmtut.pdf>
- [36] Quanser 2 DOF Helicopter User and Control Manual. Quanser Inc., Canada, 2006.

PAPR and BER Performances of OFDM System with Novel Tone Reservation Technique Over Frequency Non-Selective Fading Channel

Moftah Ali*, Raveendra K. Rao, Vijay Parsa

Innovation Centre for Information Engineering (ICIE), Faculty of Engineering, Department of Electrical and Computer Engineering, University of Western Ontario, London, Ontario, N6A 5B9, Canada

ARTICLE INFO

Article history:

Received: 21 December, 2018

Accepted: 08 February, 2019

Online : 20 February, 2019

Keywords:

Orthogonal Frequency Division

Multiplexing

Tone Reservation

Pseudo-random Sequence

ABSTRACT

An Orthogonal Frequency Division Multiplexing (OFDM) system with Quadrature Phase Shift Keying (QPSK) mapper is considered. A novel low-complexity Tone Reservation (TR) technique is proposed for reduction of Peak-to-Average Power Ratio (PAPR) of the system. The technique is easy-to-implement and minimizes the search space of phases of reserved tones in the system. The ability of PAPR reduction of this proposed TR technique is assessed and compared with conventional TR technique that uses Pseudo Noise (PN) sequences to determine phases of reserved tones. The simulation results illustrate that the proposed TR technique is nearly the same as that of the conventional TR technique in terms of PAPR reduction capability but with reduced complexity. The Bit Error Rate (BER) performance of QPSK-OFDM system with the novel TR scheme over frequency non-selective Rayleigh fading channel is also determined and illustrated.

1. Introduction

Multicarrier modulation such as OFDM is extensively adopted in numerous communication systems and standards. For example, Digital Audio Broadcasting (DAB) [1], IEEE set of standards that includes wireless local area network 802.11a [2], wireless broadband standard 802.16 [3], and mobile broadband wireless access 802.20 [4]. In addition, OFDM is the core technique for 3GPP [5] and Large Term Evolution (LTE) [6] etc. OFDM systems are attractive due to: i) their immunity to multipath fading and impulse noise; ii) elimination of complex equalizer; and iii) efficient implementation using IFFT/FFT for modulation/demodulation. OFDM system has one significant drawback that is the transmitted signals exhibit high values of PAPR. As a result, a High Power Amplifier (HPA) with sufficient linear amplification range is required in the system; otherwise, the transmitted signal would be nonlinearly distorted resulting in loss of subcarrier orthogonality, and hence BER degradation of the system. Thus, it is important to consider an appropriate PAPR reduction technique in an OFDM for it to be cost effective as well as energy efficient. In the literature, several methods have been proposed and investigated, which can mitigate PAPR problem in an OFDM system, such as Clipping and Filtering (CF) [7], weighting [8], Selective Mapping (SLM) [9], [10], Active

Constellation Extension (ACE) [11], Partial Transmit Sequence [12], [13], Tone Reservation (TR) [14], Tone Injection [14], coding [15] etc. Each technique used in an OFDM system offers its own computational complexity, PAPR reduction capability, amount to side information required to be exchanged between transmitter and receiver, and BER performance of the system [16], [17], [18].

The focus in this paper is on TR technique for reduction of PAPR in an OFDM system that requires: i) fewer computations compared to TR techniques available in the literature and ii) no side information to be exchanged between transmitter and receiver. This method was briefly presented in [19]. In this paper, this work is extended and examined in more detail and is compared with the well-known SLM technique for PAPR reduction [18]. It is noted that SLM is a distortion-less technique and requires side information (SI) to be conveyed to receiver. In addition, the QPSK-OFDM system with TR method introduced in the paper is examined for BER performance over Rayleigh fading channel.

In the paper, a novel TR technique with low-complexity is presented for reduction of PAPR in an OFDM system; the proposed scheme is easy-to-use and minimizes the search space of phase sequences. The technique achieves approximately same performance as that of TR technique using conventional PN

*Corresponding Author: Moftah Ali, Email: mali254@uwo.ca

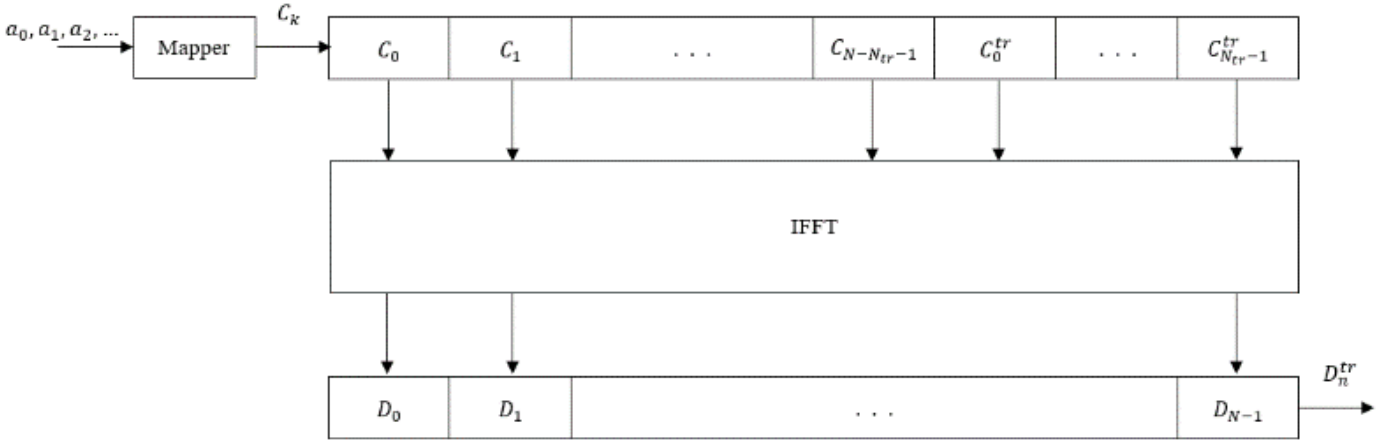


Figure 1: Block diagram for generation of OFDM signal with TR technique.

sequence. The simulation results are provided for a 128-subcarrier QPSK-OFDM system, as a function of numbers of tones reserved. The probability of bit error performance of the system is also presented.

In section 2, OFDM signal and conventional TR technique are explained. Section 3 describes the TR method introduced in the paper. In section 4, OFDM receiver is discussed and in section 5, BER of the system in frequency-non-selective Rayleigh fading channel is presented. In section 6, numerical results are presented and discussed. The paper concluded in section 7.

2. OFDM Signal Model and TR Technique

2.1. OFDM Signal and its PAPR

An OFDM signal is a summation of N (number of subcarriers or tones) independently modulated complex sinusoidal signals, and is given by:

$$D(t) = \frac{1}{N} \sum_{k=0}^{N-1} C_k e^{j2\pi f_k t}, \quad 0 \leq t \leq T_s \quad (1)$$

where $j = \sqrt{-1}$, T_s is the symbol duration and, $C_k, k = 0, 1, \dots, N - 1$, are the complex data. When QPSK mapper is considered, $C_k \in \{\pm 1 \pm j\}$ and the subcarrier frequencies are $f_k = k/T_s, k = 0, 1, \dots, N - 1$. When signal in (1) is sampled at Nyquist rate, discrete OFDM signal obtained and is given by:

$$D_n = D(n2T_b) = \frac{1}{N} \sum_{k=0}^{N-1} C_k e^{\frac{j\pi n k}{N}}, \quad n = 0, 1, \dots, N - 1 \quad (2)$$

The PAPR of signal in (1) is:

$$PAPR = \frac{\max_{0 \leq t \leq T_s} |D(t)|^2}{\frac{1}{T_s} \int_0^{T_s} |D(t)|^2 dt}, \quad (3)$$

Equivalently, PAPR in (3) is given by:

$$PAPR = \frac{\max\{|D_n|^2, n = 0, 1, \dots, LN - 1\}}{\frac{1}{NL} \sum_{n=0}^{NL-1} |D_n|^2}, \quad (4)$$

where $L (\geq 4)$ is the oversampling factor.

2.2. TR Technique

In TR method, N_{tr} of N tones, are reserved and used for PAPR reduction [20]. Figure 1 illustrates a block diagram for generation of OFDM signal with TR technique. The ratio $R = N_{tr}/N$ is typically kept small. Thus, the OFDM signal with tone reservation technique can be expressed as:

$$D_n^{TR} = D_n + D_n^{tr} \quad (5)$$

$$D_n^{TR} = \frac{1}{N} \sum_{k=0}^{N-1} (C_k + C_k^{tr}) e^{j2\pi k n / N}, \quad n = 0, 1, \dots, N - 1 \quad (6)$$

where $C_k = [C_0, C_1, \dots, C_{N-1}]$ and $C^{tr} = [C_0^{tr}, C_1^{tr}, \dots, C_{N-1}^{tr}]$ represent frequency domain vectors associated with data and reserved tones. The reserved tones' locations are denoted by the index $\mathcal{R}^{tr} = \{i_0, i_1, \dots, i_{N_{tr}-1}\}$ where $0 \leq i_0 \leq i_1 \leq \dots \leq i_{N_{tr}-1} \leq N - 1$. Let the complement of \mathcal{R}^{tr} be the index set \mathcal{R} in $\mathcal{N} = \{0, 1, \dots, N - 1\}$. Therefore, $C_k \equiv 0 \forall k \in \mathcal{R}^{tr}$ and $C_k^{tr} \equiv 0 \forall k \in \mathcal{R}$. This can be expressed as:

$$C_k + C_k^{tr} = \begin{cases} C_k, & k \in \mathcal{R} \\ C_k^{tr}, & k \in \mathcal{R}^{tr} \end{cases} \quad (7)$$

The PAPR is redefined as:

$$PAPR^{tr} = \frac{\max |D_n + D_n^{TR}|^2}{E\{|D_n|^2\}} \quad (8)$$

The objective in tone reservation technique is to obtain $\{C_k^{tr}, k = 0, 1, \dots, N - 1\}$ such that:

$$C^{tr(opt)} = \min_{(C^{tr} \in a_{mk})} \max\{|D_n^{TR}|^2, n = 0, 1, \dots, N - 1\} \quad (9)$$

3. Proposed Phase Sequence Scheme

The PN sequence is generated based on concatenating matrices. Denoting the seed matrix by \mathbf{A}_0 that can be expressed as:

$$\mathbf{A}_0 = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$$

where $a_i \in \{1, -1, +j, -j\}$, $i = 1, 2, 3, 4$ and \mathbf{A}_0 must include all the elements in the given set. Thus, one possible realization of \mathbf{A}_0 is:

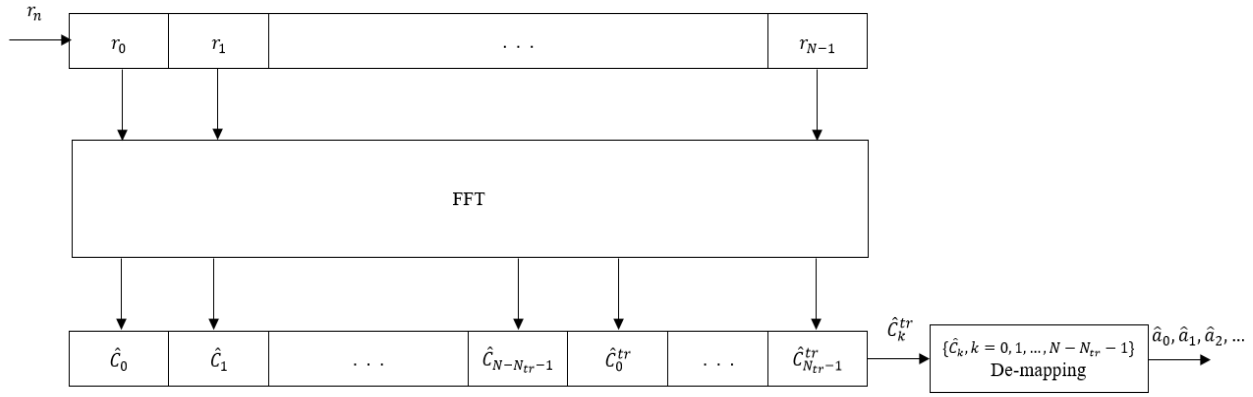


Figure 2: Block diagram for demodulation of OFDM signal with TR technique.

$$A_0 = \begin{bmatrix} 1 & j \\ -1 & -j \end{bmatrix}$$

It was reported in [19] that not all available patterns of A_0 will perform the same. Figure 3 demonstrates a sample of the possible patterns of the seed matrix A_0 .

a	b	c
$\begin{bmatrix} 1 & -j \\ -1 & j \end{bmatrix}$	$\begin{bmatrix} j & 1 \\ -1 & -j \end{bmatrix}$	$\begin{bmatrix} 1 & -j \\ j & -1 \end{bmatrix}$

Figure 3: Sample of the available patterns of A_0 .

In Figure 3, pattern (c) must be avoided for A_0 while Pattern (b) performs poorer than pattern (a).

It is noted that phase sequences in the proposed TR technique can be generated using seed matrix A_0 , which is controlled by:

$$A_m = \begin{bmatrix} A_{m-1} & A_{m-1} \\ A_{m-1} & A_{m-1}^T \end{bmatrix}, \quad m = \log_2 N - 1 \quad (10)$$

where, A_{m-1}^T is conjugate transpose of A_{m-1} and $N = 2^n$, $n = 2, 3, 4, \dots$

4. Receiver for OFDM Signal with the TR Technique

Figure 2 shows block diagram of receiver for detection of OFDM signal with TR method. The received signal, r_n , can be expressed as:

$$r_n = D_n^{TR} + w_n, \quad n = 0, 1, \dots, N - 1 \quad (11)$$

where w_n is AWGN with zero mean and a variance of $N_0/2$.

Thus, received complex symbols, \hat{C}_k^{tr} , are given by:

$$\hat{C}_k^{tr} = FFT\{r_n\} \quad n, k = 0, 1, \dots, N - 1 \quad (12)$$

If the AWGN noise term, w_n , in (11) is ignored, then (12) can be expressed as:

$$\hat{C}_k^{tr} = \sum_{n=0}^{N-1} D_n^{TR} e^{-j2\pi kn/N}, \quad k = 0, 1, \dots, N - 1 \quad (13)$$

The useful data symbols are denoted by \hat{C}_k , which are to be extracted from \hat{C}_k^{tr} and this can be expressed as:

$$\hat{C}_k = \hat{C}_k^{tr}, \quad k \in \mathcal{R} \quad (14)$$

5. BER in Frequency Non-Selective Rayleigh Fading Channel

Figure 4 shows OFDM signal received over frequency non-selective Rayleigh fading channel.

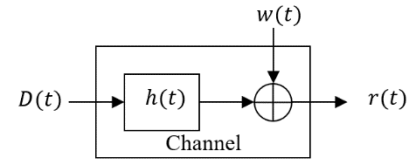


Figure 4: Model of frequency non-selective Rayleigh fading channel.

$D(t)$ in Figure 4 denotes the continuous OFDM signal of the OFDM discrete signal D_n^{TR} .

Hence, $D(t)$ is sent over the channel and therefore $r(t)$ can thus be represented as:

$$r(t) = h(t) * D(t) + w(t) \quad 0 \leq t \leq T_s \quad (15)$$

where $h(t) = \alpha(t)e^{j\varphi(t)}$, $0 \leq t \leq T_s$ is the channel impulse response, * denotes convolution, and $w(t)$ denotes AWGN.

For frequency non-selective Rayleigh fading channel, $h(t)$ can be modeled as a multiplicative distortion scaler to the transmitted signal $D(t)$. It is assumed that the channel is time-invariant and the impulse response of the channel is reasonably constant, at least during each symbol interval T_s of the transmitted signal $D(t)$. The condition of the frequency non-selective is expressed as:

$$B_s \ll B_c \text{ and } T_s \ll \sigma_\tau \quad (16)$$

where B_s denotes the bandwidth and T_s denotes the symbol duration of the transmitted signal, whereas, B_c denotes the coherence bandwidth and the RMS delay spread is denoted by σ_τ .

Thus, the channel can be expressed as:

$$h(t) = \alpha\delta(t), \quad 0 \leq t \leq T_s \quad (17)$$

where it is assumed that α is estimated perfectly. Using (17), in (15)

$$r(t) = \alpha D(t) + w(t) \tag{18}$$

where α is the fading parameter. If α is fixed then the probability of error rate P_b expression can thus be expressed as a function of the signal-to-noise ratio (γ_b). For OFDM-QPSK system, the P_b is given by [21]:

$$P_b(\gamma_b) = Q(\sqrt{2\gamma_b}) \tag{19}$$

where $\gamma_b = \alpha^2 \varepsilon_b / N_0$ and $Q(x)$ is the Q function [21].

The average error probability of OFDM-QPSK system can be shown to be given by:

$$P_{b_Rayleigh} = \frac{1}{2} \left(1 - \sqrt{\frac{\bar{\gamma}_b}{\bar{\gamma}_b + 1}} \right) \tag{20}$$

6. Numerical Results

QPSK-OFDM systems PAPR performance were observed when the proposed TR scheme is used. One hundred thousand blocks of OFDM symbols are considered for simulation when $N = 128$. Both the introduced PN sequence and the phase sequence presented in [10] are examined. The reserved tones' locations are chosen randomly.

Figures 5 (a-d) show the CCDFs plots of QPSK-OFDM systems PAPR performance. The PAPR performance provided are for number of subcarriers $N = 128$ and when reserved tones are $N_{tr} = 4, 8, 16,$ and 32 . Also, these results are tabulated in Table 1.

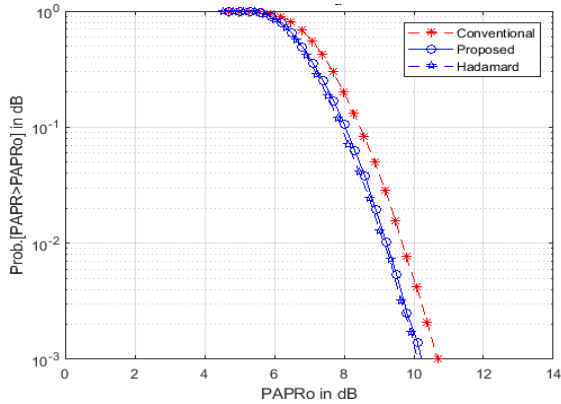


Figure 5 (a): PAPR performance of 128-subcarrier QPSK-OFDM system with 4 reserved tones

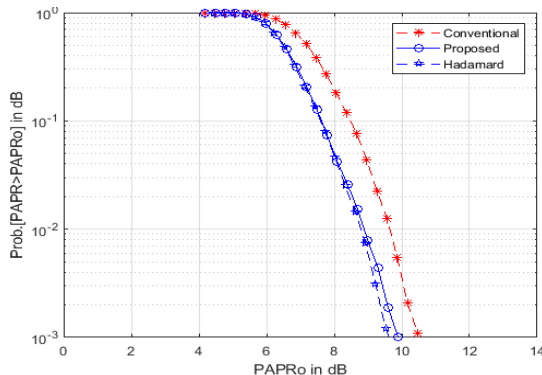


Figure 5 (b): PAPR performance of 128-subcarrier QPSK-OFDM system with 8 reserved tones

From Figures 5 (a-d) and Table 1, it is observed that the attainable PAPR performances for the two examined sequences are relatively equal. Also, it is noted that the reduction in PAPR is directly proportional to the number of reserved tones.

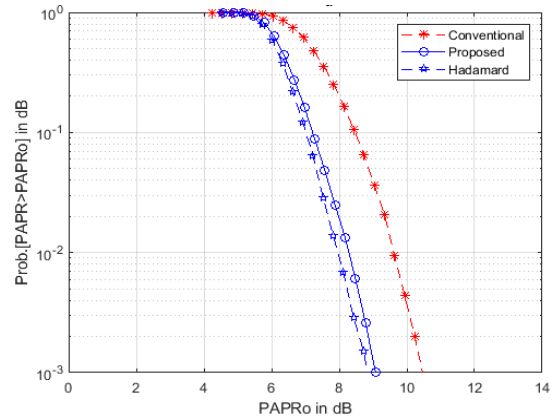


Figure 5 (c): PAPR performance of 128-subcarrier QPSK-OFDM system with 16 reserved tones

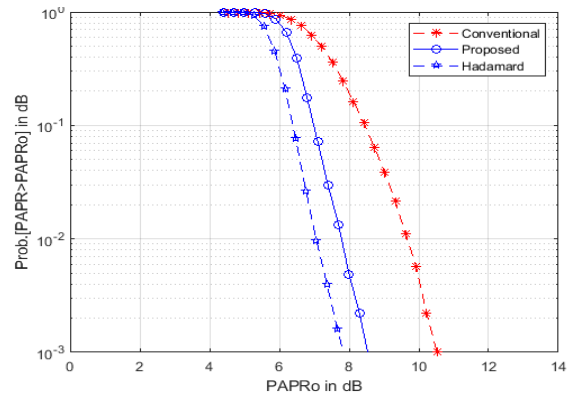


Figure 5 (d): PAPR performance of 128-subcarrier QPSK-OFDM system with 32 reserved tones

Table 1: Attainable PAPR for QPSK-OFDM system with the TR method using the proposed PN sequence and Hadamard phase sequence.

N_{tr}	Attainable PAPR Hadamard	Attainable PAPR Proposed	Attainable PAPR Conventional
4	10.286 dB	10.11 dB	10.659 dB
8	9.866 dB	10.082 dB	
16	8.992 dB	9.241 dB	
32	7.756 dB	8.457 dB	

Figures 6 and 7 illustrate the simulated BER versus SNR curves for 128-subcarrier QPSK-OFDM system with the TR technique when using the proposed PN sequence in AWGN channel and over a frequency non-selective Rayleigh fading channel respectively. Also, these results are tabulated in Table 2 and Table 3 accordingly.

Table 2: E_b/N_0 required to achieve BER= 10^{-3} for an OFDM system over the AWGN channel.

Sequence	Conventional	Hadamard	Proposed
BER	6.758 dB	6.758 dB	6.769 dB

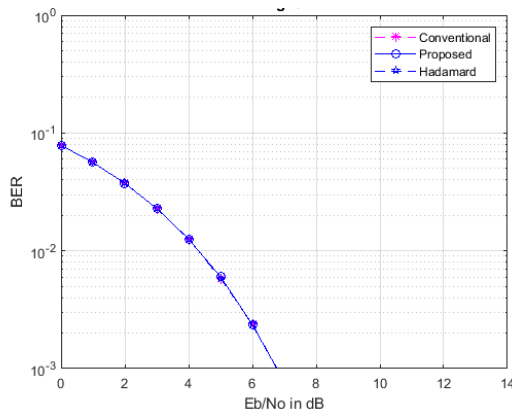


Figure 6: BER performance of 128-subcarrier QPSK-OFDM system with proposed TR technique in AWGN channel.

Figure 7 illustrates the simulated BER versus SNR curves for 128-subcarrier QPSK-OFDM system with the proposed TR technique when using the proposed PN sequence over the frequency non-selective Rayleigh channel. Ideal channel state information (CSI) is considered.

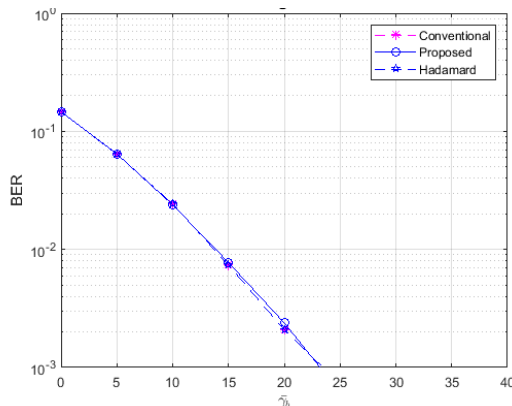


Figure 7: BER performance of 128-subcarrier QPSK-OFDM system with proposed TR technique over frequency non-selective Rayleigh fading channel.

Table 3: Average E_b/N_0 required to achieve BER= 10^{-3} OFDM system over the frequency non-selective channel.

Sequence	Conventional	Hadamard	Proposed
BER	22.936 dB	22.936 dB	23.504 dB

Figure 6 and Table 2 show that the BER performance over AWGN channel has not been degraded when the proposed PN sequence is used by the TR technique to reduce the PAPR in an OFDM system. In addition, an insignificant BER performance degradation is observed when the proposed PN sequence is used by the proposed TR method to reduce the PAPR in an OFDM system over frequency non-selective Rayleigh fading channel as shown in Figure 7 and tabulated in Table 3.

Figure 8 shows a PAPR performance comparison between QPSK-OFDM system with 16-reserved tones and QPSK-OFDM system with SLM method when the search for the optimum phase sequence is limited to 16.

Figure 8 and Table 4 illustrate that there is a 1.288 dB reduction in PAPR when TR technique is used compared to 2.497 dB when SLM technique is used. TR technique has only reserved

16-subcarriers for PAPR reduction however no SI is required at the receiver. In comparison to TR technique, SLM requires N multiplications in each iteration and it requires sending SI to the receiver.

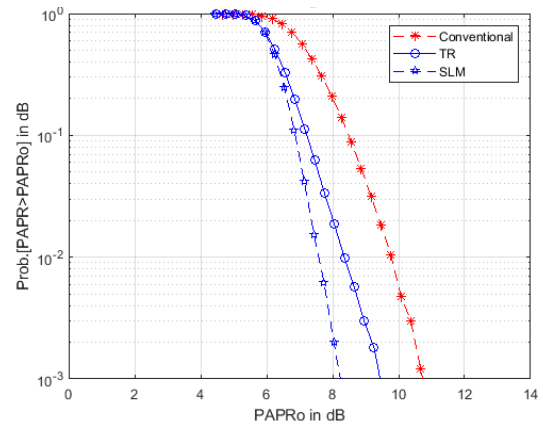


Figure 8: PAPR performance of 128-subcarrier QPSK-OFDM system with 16 reserved tones vs PAPR performance 128-subcarrier QPSK-OFDM system with SLM.

Table 4: A Comparison of the PAPR performance between 128-subcarriers QPSK-OFDM system with 16-reserved tones TR technique and 128-subcarriers QPSK-OFDM system with SLM technique

Sequence	Conventional	TR	SLM
PAPR	10.725 dB	9.437 dB	8.228 dB

7. Conclusion

In this paper, a novel low-complexity TR technique for reduction of PAPR in a QPSK-OFDM system is presented and investigated. The proposed PAPR reduction scheme is easy-to-implement and has low-complexity. Also, the proposed TR technique requires no SI to be conveyed to the receiver. The simulation results demonstrate that the PAPR performance of the proposed and conventional TR methods are the same. The BER performance of the proposed system over AWGN and frequency non-selective channel have not been compromised. When PAPR performance of the proposed TR method is compared with that of SLM technique, the SLM technique outperformed the proposed TR technique by approximately $\approx 1.2 \text{ dB}$. However, the complexity of SLM method is higher and requires sending SI to the receiver.

Conflict of Interest

The authors declare no conflict of interest.

References

- [1] U. Reimers, "Digital video broadcasting," *IEEE Communications Magazine*, 36 (6), 104-110, 1998. <https://doi.org/10.1109/35.685371>
- [2] R. van Nee, G. Awater, M. Morikura, H. Takanashi, M. Webster and K. W. Halford, "New high-rate wireless LAN standards," *IEEE Communications Magazine*, 37(12), 82-88, 1999. <https://doi.org/10.1109/35.809389>
- [3] I. Koffman and V. Roman, "Broadband wireless access solutions based on OFDM access in IEEE 802.16," *IEEE Communications Magazine*, 40(4), 96-103, 2002. <https://doi.org/10.1109/35.995857>
- [4] W. Bolton, Y. Xiao and M. Guizani, "IEEE 802.20: mobile broadband wireless access," *IEEE Wireless Communications*, 14(1), 84-95, 2007. <https://doi.org/10.1109/MWC.2007.314554>

- [5] Y. Tsai, G. Zhang, D. Grieco, F. Ozluturk and X. Wang, "Cell search in 3GPP long term evolution systems," *IEEE Vehi. Tech. Magazine*, 2(2), 23-29, 2007. <https://doi.org/10.1109/MVT.2007.912929>
- [6] A. Elnashar and M. A. El-Saidny, "Looking at LTE in Practice: A Performance Analysis of the LTE System Based on Field Test Results," *IEEE Vehicular Technology Magazine*, 8(3), 81-92, 2013. <https://doi.org/10.1109/MVT.2013.2268334>
- [7] X. Li and L. J. Cimini, "Effects of clipping and filtering on the performance of OFDM," *IEEE Communications Letters*, 2(5), 131- 133, 1998. <https://doi.org/10.1109/4234.673657>
- [8] S. Cha, M. Park, S. Lee, K. Bang and D. Hong, "A new PAPR reduction technique for OFDM systems using advanced peak windowing method," in *IEEE Transactions on Consumer Electronics*, 54(2), 405-410, 2008. <https://doi.org/10.1109/TCE.2008.4560106>
- [9] S. H. Müller and J. B. Huber, "A Comparison of Peak Power Reduction Schemes for OFDM," in IEEE GLOBECOM Conference, Phoenix, AZ, 1997. <https://doi.org/10.1109/GLOCOM.1997.632501>
- [10] R. W. Bauml, R. F. H. Fischer and J. B. Huber, "Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping," *Electronics Letters*, 32(22), 2056-2057, 1996. <https://doi.org/10.1049/el:19961384>
- [11] B. S. Krongold and D. L. Jones, "PAR reduction in OFDM via active constellation extension," *IEEE Transactions on Broadcasting*, vol. 49(3), 258-268, 2003. <https://doi.org/10.1109/TBC.2003.817088>
- [12] S. Muller and J. Huber, "OFDM with reduced peak-to-average power ratio by optimum combination of partial transmit sequences," *Electronics Letters*, 33(5), 368 - 369, 1997. <https://doi.org/10.1049/el:19970266>
- [13] L. Cimini and N. Sollenberger, "Peak-to-average power ratio reduction of an OFDM signal using partial transmit sequences," *IEEE Comm. Letters*, 4(3), 86 - 88, 2000. <https://doi.org/10.1109/4234.831033>
- [14] J. Tellado, "Peak to Average Power Reduction for Multicarrier Modulation," Ph.D Thesis, Stanford University, 2000.
- [15] A. Jones, T. Wilkinson and S. Barton, "Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes," *Electronics Letters*, 30(25), 2098 - 2099, 1994. <https://doi.org/10.1049/el:19941423>
- [16] Seung Hee Han and Jae Hong Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," in *IEEE Wireless Communications*, 12(2), 56-65, 2005. <https://doi.org/10.1109/MWC.2005.1421929>
- [17] T. Jiang and Y. Wu, "An Overview: Peak-to-Average Power Ratio Reduction Techniques for OFDM Signals," in *IEEE Transactions on Broadcasting*, 54(2), 257-268, 2008. <https://doi.org/10.1109/TBC.2008.915770>
- [18] Y. Rahmatallah and S. Mohan, "Peak-To-Average Power Ratio Reduction in OFDM Systems: A Survey And Taxonomy," in *IEEE Communications Surveys & Tutorials*, 15(4), 1567-1592, 2013. <https://doi.org/10.1109/SURV.2013.021313.00164>
- [19] M. Ali, R. K. Rao and V. Parsa, "PAPR Reduction in OFDM System Using New Method for Generating Pseudo-Random Sequence for SLM Technique" in 31st Annual IEEE Canadian Conference on Electrical & Computer Engineering (CCECE), Quebec City, Canada, 2018. <https://doi.org/10.1109/CCECE.2018.8447835>
- [20] L. Wang and C. Tellambura, "Analysis of Clipping Noise and Tone-Reservation Algorithms for Peak Reduction in OFDM Systems," in *IEEE Transactions on Vehicular Technology*, 57(3), 1675-1694, 2008. <https://doi.org/10.1109/TVT.2007.907282>
- [21] J. Proakis, *digital communications*, McGraw-Hill, 1995

Critical Embedded Systems Development Using Formal Methods and Statistical Reliability Metrics

Jonathan Lockhart*, Carla Purdy, Philip Wilsey

Department of Electrical Engineering and Computer Science, University of Cincinnati, 45221, USA

ARTICLE INFO

Article history:

Received: 21 December, 2018

Accepted: 09 February, 2019

Online: 23 February, 2019

Keywords:

Software Reliability

Trusted Systems

Hardware and Software
Co-Design

Safety Critical Embedded
Systems

Statistical Analysis

ABSTRACT

Trusted systems are becoming more integrated into everyday life. Security and reliability are at the forefront of trusted system design and are often directed at hardware-only solutions, especially for safety critical systems. This is because hardware has a well established process for achieving strong, precise, and reliable systems. These attributes have been achieved in the area of safety critical systems through the use of consistent and repeatable development processes, and a standardized metric for measuring reliability. However, due to the increase in complexity of systems and the looming end of Moore's Law, software is being incorporated more into the design of these trusted systems. Unfortunately, software typically uses agile development in modern design and uses unreliable metrics for illustrating reliability. This does not make it suitable for safety critical applications or for total system reliability in mixed hardware/software systems. Therefore, a comprehensive process of systems development needs to be utilized to allow for total system specification in the beginning and a comparable reliability metric in the end which covers software and hardware. Henceforth, we discuss an initial solution to these problems, leading to the establishment of a development process that allows for the proven correctness of a system specification via formal methods. This process also establishes a testing and error reporting process to allow software to be represented in a way that allows the application of reliability metrics similar to those used for hardware.

1 Notification of Intent

This paper is an extension of work originally presented at the 2017 National Aerospace & Electronics Conference (NAECON) [1]. In this work we illustrate the ability to utilize formal methods and automated theorem proving (ATP) to discover errors in specification development prior to implementation. A major concern addressed from feedback at NEACON 2017 is the prevention of errors in the proof environment. We expand on that work by showing that formal methods and ATPs can be used to find errors in the proof of function correctness. This allows for a system of checks and balances between specification and proof development for formal methods, verifying that both parts adhere to the original customer requirements.

This paper also extends work originally presented in 2016 at the Midwest Symposium on Circuits and Systems (MWSCAS) [2]. In that paper we addressed the use of statistical metrics to show the reliability of software over time. This process relies on the use of techniques we developed to define independent, random errors for injection into our elevator controller benchmark program. However, our presentation of those techniques was brief due to page limitations and the need to discuss the results of our testing procedure for software. To address the feedback on that work, we will be elaborating in depth on our technique for generating random, independent errors using real-world reported error results and random number generators (RNGs). This also includes a discussion of how we determine the best RNG for our needs.

*Jon Lockhart, EECS Department, 851 Woodside Drive, 812 Rhodes Hall, University of Cincinnati, Cincinnati, OH, 45221-0030, 513-556-2946 & lockhaja@mail.uc.edu

Finally, this paper extends work originally presented in 2018 at MWSCAS [3]. This was originally an expansion of our work in [2] using a different error rate and looking at possible better statistical models of reliability for software. Many of the commentators on our work suggested expanding the results to more programs and improving the results by looking into focusing the data, as some outliers had caused degradation in the results. We address both of these points by expanding the testing and analysis to five more programs, using characteristics and style similar to our original elevator controller benchmark, and then developing a process to remove outliers in the data through the use of statistical measures.

2 Introduction

Embedded systems in recent years have taken a leap in their integration with everyday society. They appear in various forms from wearables, to cellphones, to the vehicles we drive; it is important that these embedded systems function correctly. In many fields, such as medicine and aerospace, there is a strong focus on these systems being secure, reliable, and robust. These trusted and safety critical systems must work correctly as their failure could result in injury or, in the worst case, the loss of life of those that are dependent on them.

Many trusted systems encountered today utilize hardware specific embedded systems. Hardware has a tradition of reliability and safety, because of its use of industrial practices and standards. In VLSI design, developers can compare their working designs to standardized benchmarks (*e.g.*, MCNC [4], IWLS [5], and LEKO [6]) and this generates a baseline of confidence in the design. Beyond benchmarks, tool designers have incorporated Model Checking [7] into their development suites, allowing for reliability checking against industry determined “golden models” [8–11]. These techniques are invaluable in generating the statistical models of reliability for hardware systems.

Unfortunately, the constraints of Moore’s Law [12, 13] and the increased complexity of system function have required a change in the design approach of critical embedded systems. Some industry practice has attempted to maximize the resources available for a given space [14–17], but these techniques are fairly new, and VLSI tools and techniques are lagging behind in allowing for these techniques, slowing adoption. Other researchers are looking at maximizing the capabilities of current hardware by focusing on implementing machine learning in embedded systems [18]. However, this is difficult to achieve for safety critical systems with the large overhead that comes with high precision computing with machine learning [19].

Hence, software is an attractive alternative for meeting the demand for complex, real-time critical systems. Several recent disasters (*e.g.* [20–22]) have, however, indicated that the current process of testing to exhaustion and representing software reliability as defects

per thousand lines of code (ELOC) [23] is not sufficient. This is especially true when taking into account the fact that there is no industrial standard for counting “lines of code” and factors such as size play a role in misrepresenting results [24]. Software metrics similar to hardware would be ideal, and theoretical work in this area has shown software should be able to be modeled with statistics [25–30]. Government agencies have also released standards of development [31, 32] to provide a framework for illustrating reliability effectively. However, there is no specific development process outlined for achieving either of these improvements.

The work presented here is a complete development process that we propose for use in addressing the challenges described above. This development process allows for the establishment of a suite of benchmark programs, defining a baseline similar to hardware which other trusted systems utilizing software can be compared to. This suite is open ended and allows for the addition of more programs in the future. To accomplish this goal, we will address the area of specification design through the use of formal methods, and show how errors can be eliminated early in development with the use of ATP. This work is an expansion of work already presented in [1, 33]. We will also illustrate the development, use, and statistical results for reliable software in safety critical applications. This approach has been briefly presented previously in [2] and improved in [3]. We expand upon those findings to complete the suite of small embedded programs which establishes the initial set of well-defined benchmarks for software reliability and a baseline for software testing.

3 Related Work

There have been several groups researching the use of formal methods in the recent literature. This work has targeted two main areas. The first is the use of model checking with software, and the second is the improvement of ATPs. One area of use for model checking is for reverse engineering a system [34]. Since automated implementation is not available yet for specifications written in formal methods, the reverse engineering process uses the implementation to move back to a specification, verifying it is correct with model checking, and then verifying the original specification and reverse engineering specification are correct with respect to each other. Another area is related to the development of a tool called Evidential Tool Bus (ETB) that combines both model checking and ATPs into a single environment to allow better verification of specifications written in formal methods [35]. Finally, researchers are looking at the inclusion of model checking into development tools to use in verifying that object oriented programs are correctly implemented [36]. This process is similar to the current tools for model checking hardware designs provided by Synopsys [7].

The research in improving ATPs has involved making them more inclusive for developers. Currently

systems such as ProofPower [37], although powerful, require the user to know the formal method and ATP language for proof verification. Some work has been done to provide a more mathematical environment to work in, replacing the intermediate ATP language, *i.e.* SML, with mathematical notation. Some of this work includes the use of superposition calculus [38], modal logic [39], and standard calculus [40]. Other researchers have looked at changing the language of formal methods and ATPs together, working to use traditional programming languages from the beginning of specification design, and this work has focused primarily on the language LISP [41]. These changes would require a change in the current specification writing practices with formal methods [42, 43].

Current work in the field of software development has focused on improving the metrics of throughput and performance for large scale, big data [44] applications. This work includes generating heuristics for quality assurance (QA) [45] with probabilistic analysis, and attempting to predict where in the development life cycle a particular project will have the most errors [46, 47]. This allows developers to focus their efforts in those areas to mitigate the time and cost for a given project. Machine learning is assisting in this field, where it is used to predict the optimal release time of a given system based on the metrics of testing, cost, and errors produced [48]. Many of these techniques are being applied specifically to projects that rely on cloud computing [48] and open source software [49]. They are not being used to directly target the component level reliability which is required for safety critical systems and those relying on software in an embedded environment.

4 Background Knowledge

Formal methods are supported by many languages, such as Z [50] and VDM [51], that share in common a reliance on basic mathematical principles. These principles include axiomatic notation [52, 53], predicate calculus [54, 55], and set theory [56, 57]. When combined with the formal methods syntax, these principles allow us to develop specifications describing the functionality of a system [22, 58]. These principles are often taught in engineering course work, and formal methods often have manuals describing how the language operates [59], similar to a conventional programming language such as C or Java.

One important aspect, however, is the mechanisms through which formal methods have traditionally been used to show that a given method in a specification is correct. The traditional method is to prove a specification is correct through a process of refinement. This process requires “refining” a particular part of a specification written in formal methods to an implementation method, such as analog circuitry or Java, for a particular portion of the specification. This decision to refine to a particular implementation is required ahead of time so there is a goal to achieve for the re-

finement. Later changes to the implementation require that a new proof via refinement be performed. Understanding this process and why our work has moved in a more modern direction is important and we illustrate this process in the following example.

The example we are working with will utilize an elevator controller benchmark as originally described in [33] and referenced in Section 5. The elevator controller discussed here has requirements presented in Figure 8. For the refinement process, we will use Z for the formal specification language and C as the implementation target. Note that any formal method and implementation medium can be utilized for refinement as long as the characteristics of each are well understood and the proof can be completed. In contrast to the original requirements of [33], some additional constraints need to be in place before the refinement can proceed. In particular, the constraints are:

- The model will illustrate the action of users in the system; specifically we will model the entry of passengers onto the elevator
- People are unique individuals
- Entry into the elevator is sequential
- The elevator can move with or without people
- There is no max capacity
- The system contains one elevator

With these extra constraints on the system defined, the refinement proof can begin. The first step is to develop a definition of the possible group of users for the elevator. Since we used the wording in the constraints, we shall refer to these users as “People.” We use Z to define “People” as a “Set,” because sets are flexible and convenient for developer definitions in Z. Figure 1 illustrates the definition of “People” in Z.



Figure 1: Z definition of the set “People”

Now that the generic definition of “People” has been established, we need to define a subset of “People” who will be using the elevator. To ensure that the specification is as open as possible we will define the subset of “People” who are riding the elevator via a power set. This way any combination of members of “People” can become “riders.” Figure 2 captures the creation of this power set using the Z construct known as a state in ProofPower [59].

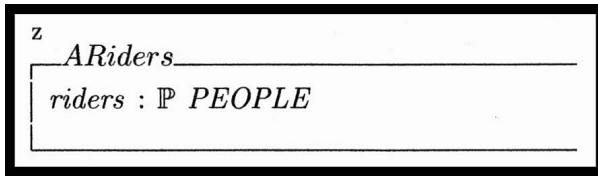


Figure 2: Generic state definition for the power set "Riders"

The final part of the initial definition is to describe the act of getting on the elevator. In Z this can be done using a function declaration, which will include a change in the state of the system and a member of the set "People" becoming a member of the power set "riders." Figure 3 shows the generic definition for this function.

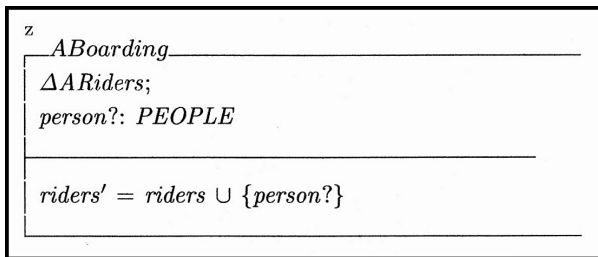


Figure 3: The generic function definition for boarding the elevator

Now we need to complete the same process but with the implementation medium we have decided on. As mentioned, we have chosen C for this task. The first step is to define the state of the system as it would be represented in C. We will reuse the same set "People" we defined previously for this task. In a programming language such as C, sets of items are usually stored in an array or a linked list. This is what we would use to store the members of the set "People" who board the elevator. However, Z does not have these constructs directly, though it does have a structure that an array or linked list can be modeled in, and that is a sequence. A sequence is a good comparison as a set is equal to the range of a sequence, as shown in (1).

$$s = \text{ran } ss \tag{1}$$

In this equation *s* represents the set, *ran* is the notation for range, and *ss* represents a sequence. Now we construct the state of the system with Z for C using the sequence term *seq*, and generating a new, more specific group called "riderss." Figure 4 shows the construction of this state.

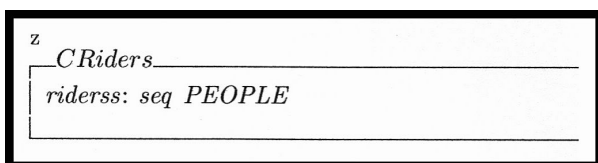


Figure 4: State for "Riderss" in C using Z

Finally, we complete this portion by describing the action of a new rider boarding the elevator in C. Since

we are dealing with a sequence, the mathematical notation we will use is a concatenation. Figure 5 shows the state for C as defined in Z.

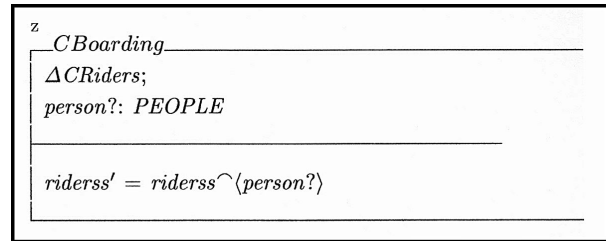


Figure 5: C function definition for boarding the elevator

With the set "People" we have now defined two versions of the system for boarding the elevator, a generic one which describes the basic functionality and a more specific one constrained by the properties of our implementation medium C. This gives us a starting and ending place. To bridge the gap between the two and what they describe, we need to develop a proof to show that the generic function can become the specific function. To start we build a proof implication, or goal, illustrating this transformation. This will be done using Equation 1 for some replacements for *riders* and *riders'* from the generic definition in Figure 3. Figure 6 gives the initial implication, showing the end requirement for the C definition from Figure 4 and using the previous set to sequence rule to rewrite *riders* and *riders'* in terms of the range for *riderss* and *riderss'*, respectively.

To complete the proof implication, we want to drive one side or the other to an absolute *true* which means everything on the other side must hold as defined. Since the outcome only has one statement, this is the easier of the sides to work on. Figure 7 shows the seven step proof indicating the implication holds and the refinement is *true*.

Using set theory and rewrites of variables taken from the specifications, we are able to show that the generic specification can be refined to the specification for C. This was a simple example and even then it took seven steps to complete, along with a strong knowledge of the constructs of Z, C, and set theory. To attempt to complete a definition for a more complex system and multiple functions would be very time consuming and may cost more than just using traditional testing methods due to the time involved. Furthermore, this process does not define the function in a useful manner that would assist with the conversion to an implementation, but rather shows the theoretical correctness of the system definitions for a particular implementation medium. Even when using an ATP to perform the refinement steps, the final implementation may not be correct in a functional sense, as the definitions are mathematical, not practical. Hence our work [1, 33] described in Section 5 and extended in Section 6 shows the advantages of utilizing a functional specification and proving that correct. It also provides the flexibility of being implementation neutral so a new proof does not need to be constructed if the implementation

$$riderss' = riderss \tilde{\langle person? \rangle} \wedge riders = ran\ riderss \wedge riders' = ran\ riderss' \Rightarrow riders' = riders \cup \{person?\}$$

Figure 6: Implication definition for showing the refinement steps

$$riderss' = riderss \tilde{\langle person? \rangle} \wedge riders = ran\ riderss \wedge riders' = ran\ riderss' \Rightarrow riders' = riders \cup \{person?\}$$

- 1 $\Leftrightarrow riders' = riders \cup \{person?\}$
- 2 $\Leftrightarrow ran\ riderss' = riders \cup \{person?\}$
- 3 $\Leftrightarrow ran\ (riderss \tilde{\langle person? \rangle}) = riders \cup \{person?\}$
- 4 $\Leftrightarrow ran\ riderss \cup ran\ \langle person? \rangle = riders \cup \{person?\}$ [Law: $ran\ (s \tilde{t}) = ran\ s \cup ran\ t$]
- 5 $\Leftrightarrow ran\ riderss \cup \{person?\} = riders \cup \{person?\}$ [Law: $ran\ \langle x? \rangle = \{x?\}$]
- 6 $\Leftrightarrow riders \cup \{person?\} = riders \cup \{person?\}$
- 7 $\Leftrightarrow true$

Figure 7: Proof showing the refinement of the specifications using Z

medium changes.

One final area of background work that we utilize for our error analysis is the theory that software errors can be modeled as a Non-Homogeneous Poisson Process (NHPP). A NHPP is defined well in the sources [60–63]. The important factor for our work is that software errors were shown to meet the requirements of a NHPP in [29], which originally had been theorized in [25, 26, 30], and shown to fit statistical models in [28]. The takeaway is that software errors have been shown to fit NHPP, implying that these errors occur at random intervals and that they are independent of one another. This allows us to define our error generation and testing procedure, discussed in Sections 7 and 8 and in the results presented in [2, 3].

5 Specification Development with Formal Methods

As shown in Section 4, an understanding of refinement is important for developing specifications and proofs for showing the system design is correct. However, this traditional methodology limits the scope of the system to a particular implementation method. The implementation chosen for the proof reduction may not be the best for a given system process later in development. If a new implementation is chosen, then a new proof will need to be created to drive the specification to that implementation, resulting in the development of new state diagrams and another refinement. This correction in the design would be costly in the time required for the refinement.

In previous work [33], we utilized a simple elevator controller, in the style of Moore and Mealy Machines, to show that an implementation independent specification could be developed and proven correct without the need for traditional refinement. This was done using “Z” [50] and the Automated Theorem Prover (ATP) “ProofPower” [37]. Figure 8 shows an excerpt from [33]

with the initial customer requirements for the simple elevator controller.

- Must know which direction it is going.
- Can only move up and down.
- Has only two inputs: the starting floor and ending floor.
- Must stop once it has reached the desired floor.
- Must have an emergency state to shut down the elevator.

Figure 8: Customer requirements for our elevator controller

From these requirements we are able to build a complete specification and proof illustrating our formal methods design process [33]. This added step in development does add time to the process, just like refinement, but there are strong benefits to its use over refinement. This includes focusing the functionality of the system, making it less likely errors will occur during the implementation step [33], since the specification is designed in a similar style to function definitions in software. This also eases the transition when building an implementation from customer requirements. Another benefit is finding development errors prior to implementation [1], which is usually a cost saving measure compared to finding errors later in development. There is also the added advantage of not having to refine the specification to a particular implementation, while maintaining the advantage of proving the system correct.

The error now is in the proof goal set by the independent tester, causing the goal statement to be $calling_floor? \geq starting_floor?$, which is shown in Figure 10.

6 Using Formal Methods to Eliminate Design Errors

An important advantage to using Formal Methods is to allow for design flaws to be determined prior to the

```

z
Moving_Up
Elevator_Operation

calling_floor? > 0; starting_floor? > 0; calling_floor? > starting_floor?;
counter' = calling_floor? - starting_floor?;
direction' = up
    
```

Figure 9: "Moving Up" function represented in Z

```

SML
set_goal ([], [zpre Elevator_1 ⇔ (calling_floor? ≥ starting_floor? ∧
calling_floor? > 0 ∧ starting_floor? > 0 ∧ counter ≥ 0)⊃]);
    
```

Figure 10: Proof goal for "Moving Up" with an incorrect math statement

system reaching implementation. Often a customer’s requirements may have conflicting or impractical goals which cannot be determined until they are built into the system. Formal Methods proves those requirements are sound prior to implementation, decreasing the time of development and the cost to remove errors later in development. However, human error can still plague the development life cycle.

Human error in development can come in many forms, but one such error in software design comes from the developer accidentally injecting an error into the system. For example, misspelling a variable name so it matches a similar variable could cause a drastic change in the functionality of the system. Another example is changing an addition to a subtraction, which could be accomplished with an errant key stroke. To combat these errors, independent verification is used in testing, helping to eliminate the bias a developer may put unknowingly into their own tests. This concept works equally well with Formal Methods.

In our work [1] we previously showed that independent verification during the proof of a specification enabled errors to be found and eliminated prior to implementation. Feedback on this process noted that the error could be propagated from the specification to the proof, causing it not to be caught. We emphasize that specification and proof development should be completed by separate individuals, allowing for independent verification. It is unlikely that two developers would make the same mistake by misinterpreting the customer requirements. This does however bring up a separate issue, *i.e.*, that the mistake could be with the proof and not with the specification. Would the proof environment be able to locate an error in the proof statement, given a correct specification? This is the question that was posed, and this is what we have set out to determine.

In [1] we utilized the "Moving Up" function as test and that is the function utilized here for consistency. Our work in [1] showed that a slight change in the specification over the customer requirements could cause a

dramatic change in how the system functioned. In this case the original requirements, as shown in Figure 8, require that, if the elevator moves up, the floor it is going to has to be a higher value than the floor the elevator starts on. This is represented in the specification as $calling_floor? \geq starting_floor?$, where \geq has replaced the correct term $>$.

Now, consider the reverse situation. The original requirements have remained the same from Figure 8, and now the specification for "Moving Up" is correct as it is in [33], shown in Figure 9.

With this mistake in the proof, the proof begins with a goal that is less strict than the specification, in that "greater than or equal to" is more inclusive than "greater than." Much like our previous use of ProofPower in [1, 33], we need to drive the proof to completion, attempting to resolve all goals defined in Figure 10 via a series of commands issued to the ATP. Figure 11 shows the current steps we will use to try to prove the specification is correct.

Step 10 in Figure 11 is where the proof fails. Figure 12 shows the output from ProofPower when the exception is thrown and the proof can not be completed. ProofPower is reporting that the original goal as stated can not be proven given the assumptions and constraints on the proof, which were set in the original specification.

Unlike the original error recognition test in [1], the semantics error takes longer to find when the error is in the proof goal; nevertheless the error was found, even when it was within the proof goal definition and was weaker than the original specification. Overall, the exception shows that there is an issue with the system, either with the specification or the proof. At this point it would be up to the proof writer and the developer to review their respective sections with regards to the customer requirements and to determine where the error occurred.

Finding errors early in development, as in this example or previously in [1], is invaluable to the development process. This allows you to find semantic errors

```

1 a (rewrite_tac [Elevator_1, Moving_Up, Elevator_Operation, Elevator_State]);
2 a (pure_rewrite_tac [z_get_spec [z(-≤-)"]);
3 a (rewrite_tac []);
4 a (REPEAT z_strip_tac);
5 a (asm_rewrite_tac []);
6 a (PC_T1 "z_lin_arith" asm_prove_tac []);
7 a (z_∃_tac [z(direction' ≐ up, counter' ≐ calling_floor? + ~ starting_floor?)]
   THEN rewrite_tac []);
8 a (asm_rewrite_tac [z_get_spec [DIRECTIONS"]]);
9 a (PC_T1 "z_library_ext" asm_rewrite_tac []);
10 a (PC_T1 "z_lin_arith" asm_prove_tac []);

```

Figure 11: The steps necessary to complete the proof to show the goal is correct

Table 1: PSP program line count and error results

Language	Program	Total Lines of Code	Total New Lines of Code	Number of Errors
C++	1	82	82	12
	2	233	233	31
	3	463	263	24
	4	236	129	15
	5	178	104	19
	6	568	299	40
	7	678	86	20
	8	458	298	47
	9	824	258	35
	10	1202	379	62
TOTAL		4922	2131	305
Eiffel	1	70	N/A	6
	2	82	N/A	17
	3	214	N/A	19
	4	220	N/A	8
	5	182	N/A	13
	6	396	N/A	28
	7	286	N/A	11
	8	444	N/A	20
	9	784	N/A	23
	10	766	N/A	36
TOTAL		3444	N/A	181

that are caused by developer mistakes prior to those errors being propagated into the implementation. If this were to occur, there is no guarantee that the error would be caught in the testing phase, and thus it could require a change to the system after release, costing extra time and money. Finding errors earlier in the development process is usually more cost effective, and utilizing a well developed ATP like ProofPower should help eliminate development errors prior to implementation.

7 Developing Repeatable Errors from Real World PSP Results

Before moving into implementation and testing, it is crucial to have repeatable results. In our case, this

means not repeatedly injecting the exact same system errors. This is important as the premise of some of our work goes back to the work done by Yamada [29, 64], as explained in Section 4. Our errors need to be random and independent of one another for them to fit into the model of an NHPP. This will not be the case if the same errors are used for every evaluation of our procedure. To that end we turned to the work of Victor Putz [65] who created examples using PSP [66] to illustrate its ability to assist a developer in becoming more proficient in software development. Through the posted results we are able to catalog the findings, developing metrics to allow us to generate error rates and determine errors for replication in our test system.

Putz performed a self-exploratory study through PSP, publishing the information for each step in the training process, along with the code he developed, his

```

Output Continued
...
:) a (PC_T1 "z_lin_arith" asm_prove_tac []);
Cannot derive a contradiction from the following assumptions using the linear
arithmetic proof procedure:
┌ z_Z $"calling_floor?" + ~ (z_Z $"starting_floor?") ≤ NZ 0 ┐
┌ ~ (z_Z counter) ≤ NZ 0 ┐
┌ ~ (z_Z $"starting_floor?") ≤ ~ (NZ 1) ┐
┌ ~ (z_Z $"calling_floor?") ≤ ~ (NZ 1) ┐
┌ ~ (z_Z $"calling_floor?") + z_Z $"starting_floor?" ≤ NZ 0 ┐
Assigning variables to terms as follows:
X1 = ┌ z_Z counter ┐
X2 = ┌ z_Z $"starting_floor?" ┐
X3 = ┌ z_Z $"calling_floor?" ┐
Gives the satisfiable system:
1: -X2 + X3 ≤ 0
2: -X1 ≤ 0
3: -X2 ≤ -1
4: -X3 ≤ -1
5: +X2 - X3 ≤ 0
Exception Fail * Could not prove theorem with conclusion  $\frac{z}{z}(0 \leq \text{calling\_floor?} \wedge$ 
 $0 \leq \text{starting\_floor?}) \wedge 0 \leq \text{calling\_floor?} + \sim (\text{starting\_floor?} + 1)$  (* System
is satisfiable [Z_lin_arith_tac.82110] *) [z_lin_arith_prove_tac.82200] *
raised
    
```

Figure 12: "ProofPower" output from failing step ten in the reduction process

thoughts, issues, experience, and results [67]. He used a chart system to categorize each error that he made for each program, and an example of this report can be found in [68].

There are altogether ten programs developed through the PSP training, and many of the previous programs build on one another. In Putz's analysis of the training, he went through the process using two different languages, C++ and Eiffel. Each language had its own unique set of code [67], and statistics for each based on the program developed. C++ did have better reporting as compared to Eiffel, but regardless, based on the information and programs reported for each chapter, we were able to collect data from the study. Table 1 shows the results we collected from what was presented by Putz.

Between the two programs, C++ had total lines developed and total new lines added (for programs that were expanded upon), whereas Eiffel just had the lines for each program developed. Despite this we are able to determine several error rates, two from C++ and one from Eiffel. Eiffel's error rate was calculated by taking the total number of errors committed and dividing it by the number of developed lines. This rate is 5.26%. C++ programs yielded two error rates, one from total developed lines and one from total new lines added, with both being divided into the total number of errors committed in the ten programs. This gives us two very different rates for C++, with one being below ten errors per lines of code (ELOC) and one over. The high error

rate for C++ is 14.31% and the low error rate is 6.2%.

Now that we have rates to use to inject errors into a system per one hundred lines of code, we next need to establish what error groups were reported. From the result tables, like the one in [68], we can see that each error is coded, and a text description of the error and/or its resolution is provided. We refer to the chart Putz posted on his site to get the categories and their meaning [69]. Using this information we are able to generate a list of the errors that appeared during his independent study of PSP. From the error categories and definitions, along with examining the code, we were able to break the errors into groups. We then determined the errors made for each type in each program for C and Eiffel. Table 2 lists the errors of each of the types for each program along with their totals.

Based on the categories of errors presented, we determined what errors could and could not be replicated using automation. The errors categorized as "Algorithm Alteration," "Missing Algorithm," "Specification Error," or "Unknown" are more overarching, sometimes requiring a redesign of a whole algorithm, and would be too difficult to reproduce automatically. Those errors should occur anyway in the implementation and testing process for a developer, so our programs should have representatives of these errors included in them naturally. The remaining errors could be replicated via selection, which is addressed in Section 8, because they affect a specific point in the code, requiring up to only a few changes to correct. Table 3

shows the final selection of errors chosen for replication along with the rate at which they occur during the PSP process. These rates were calculated by taking the total found for a specific error and dividing it by the total number of errors for error types we can replicate.

Table 2: Total errors made by type for C++ and Eiffel

Error Type	C++	Eiffel	Total
Spelling	29	26	55
Used Wrong Type	52	26	78
Missing Header Information	21	6	27
Incorrect Math Operation	28	33	61
Incorrect String Operation	11	10	21
Error Handling	13	2	15
Passing Parameters	12	3	15
Conditional Error	10	11	21
Incorrect Method	16	16	32
Missing Block/End Line Character	30	12	42
Algorithm Alteration	31	18	49
Missing Algorithm	37	15	52
Specification Error	11	2	13
Unknown	4	1	5

Table 3: Percentage of Errors Made by Type

Error Type	Percentage
Spelling	14.99
Used Wrong Type	21.25
Missing Header Information	7.36
Incorrect Math Operation	16.62
Incorrect String Operation	5.72
Error Handling	4.09
Passing Parameters	4.09
Conditional Error	5.72
Incorrect Method	8.72
Missing Block/End Line Character	11.44

From the diagnosis of Putz’s results we have been able to develop error rates to use for injecting errors into our system, and rates for error types that we can replicate. This will be valuable and completes the first step, namely the need for independent errors. The next step will be formulating a way to determine which errors will occur, how many will occur, and where each will be injected into the system.

8 Generating Random Errors for Insertion Into a System

Previously in [2] we briefly discussed an analysis to determine a good RNG for use in replicating random errors. A critique from the results presented was the need to test the RNGs at all. The most critical response to this is that the errors being generated need to be random and independent of one another such that the errors will fit the aforementioned NHPP. Unfortunately, RNGs for general computing, such as `rand()` that comes with C/C++, are only pseudo-random. This means that there is a pattern to the numbers being generated and it could be predicted. If the values are predictable then the numbers generated are not random and therefore will not fit NHPP.

Tests have been generated over the years to determine the randomness of an RNG, e.g., by Knuth [70,71] and Marsaglia [72]. However, these tests are singular and require the tester to input parameters to constrain the test. This allows an RNG developer to design their generator to defeat one or two of the tests chosen to show randomness. Fortunately L’Ecuyer has provided a solution to this issue with the `TestU01` package [73]. This C package combines multiple tests from multiple authors, and has been calibrated in such a way that each test is given the correct parameters for the level of testing being performed [74]. For this work we will be utilizing the test suite L’Ecuyer calls `SmallCrush` as it is described as sufficient for determining if an RNG is suitable for general computation [73]. All inputs and test generations for `TestU01` have been determined by L’Ecuyer using statistical analysis [74], and in the case of `SmallCrush` the total number of values generated for each test is approximately 5 million values, which we validated through testing.

Using `SmallCrush` we can determine an RNG that can be considered sufficiently random and allow for meeting the requirements for NHPP. `Rand()` was decided upon as a baseline for utilizing `SmallCrush` and generating results to see what tests it would pass. `Rand()` was also chosen as the baseline since it is the default RNG included with the C/C++ library. Testing with `SmallCrush` was completed three times for consistency, on integer and decimal values, and was initially seeded at the start of each test with the current time to prevent starting at the same value for each test. Figure 13 shows the an example of the output from `SmallCrush` using `rand()` to generate integer values, and Figure 14 shows the final results for `rand()` generating integer values.

The results show that `rand()` for C/C++ fails twelve of the total fifteen tests (passing only three). This is consistent for the next two runs, failing the same twelve tests each time. Now with this baseline we can move on to see if we can find a better RNG, needing one that can pass all fifteen tests to be considered “random.”

We want to minimize the development of tools and inclusion of extra libraries, so finding a generator that

can be random for both integer and decimal values is optimal. Researching the literature, one in particular that appears to meet these requirements is PCG32 by O’Neal [75]. This RNG has a small extended footprint with its included libraries, and the results show that PCG32 has already passed SmallCrush. We will independently re-evaluate these claims. Other RNG designs were considered, such as SBoNG [76], but they have already been shown not to be “random” [77] via SmallCrush.

```

*****
Test sknuth_collision calling smultin_Multinomial
*****
HOST =
cRandInt

smultin_Multinomial test:
-----
N = 1, n = 5000000, r = 0, d = 65536, t = 2,
sparse = TRUE

GenerCell = smultin_GenerCellSerial
Number of cells = d/t = 4294967296
Expected number per cell = 1 / 858.99346
EColl = n^2 / (2k) = 2910.383046
Hashing = TRUE

Collision test, Mu = 2909.2534, Sigma = 53.8954
-----
Test Results for collisions
Expected number of collisions = Mu : 2909.25
Observed number of collisions : 4999999
p-value of test : eps *****

Total number of cells containing j balls
-----
j = 0 : 4294967295
j = 1 : 0
j = 2 : 0
j = 3 : 0
j = 4 : 0
j = 5 : 0
-----
CPU time used : 00:00:00.84
    
```

Figure 13: Output from SmallCrush testing rand() RNG

```

===== Summary results of SmallCrush =====
Version: TestU01 1.2.3
Generator: cRandInt
Number of statistics: 15
Total CPU time: 00:00:13.22
The following tests gave p-values outside [0.001, 0.9990]:
(eps means a value < 1.0e-300):
(eps1 means a value < 1.0e-015):
-----
Test p-value
-----
1 Birthdays spacings eps
2 Collision eps
4 SimpPoker 6.8e-13
5 CouponCollector eps
6 Maxoft eps
6 Maxoft AD 1 - eps1
7 WeightDistrib eps
10 Randomwalk1 H eps
10 Randomwalk1 M eps
10 Randomwalk1 J eps
10 Randomwalk1 R eps
10 Randomwalk1 C eps
-----
All other tests were passed
    
```

Figure 14: Final results for rand() when tested with SmallCrush

```

===== Summary results of SmallCrush =====
Version: TestU01 1.2.3
Generator: pcg32Int
Number of statistics: 15
Total CPU time: 00:00:13.93
-----
All tests were passed
    
```

Figure 15: Final results for PCG32 when tested with SmallCrush

PCG32 was tested three times each for consistency, just like previously with rand(). Figure 15 shows the

results for one run of PCG32 generating integer values. All three runs through SmallCrush resulted in the same outcome.

The results from this testing are impressive. The results show that PCG32 performs as claimed [75], passing each of the tests for all three runs. It has proven to be sufficiently random for general computation and allows us to generate the random values necessary for compliance with NHPP.

9 Random Error Generation

Before we can evaluate our testing procedure, we need to generate random errors for the system using the error rates we determined in Section 7. This is subdivided into three parts:

1. Count the lines of code for each benchmark program in a consistent manner
2. Determine the possible injection points for each error we will replicate; this will need to be done for each benchmark program
3. Randomly select the number of errors to occur, which errors will occur, and where in the program using the derived data from the PSP analysis

Item 1 is the first to be addressed. An automated method for determining the number of lines for each program is critical for the count to be consistent and accurate for each benchmark program. Unfortunately there is no industry standard for counting lines. Many reports have proprietary methods of counting lines, such as compiled compared to written, or counting spaces versus not counting spaces. This is typically used to help improve the error rate for ELOC representation, as mentioned earlier. In our case, in reviewing the lines reported by Putz, and reviewing our own benchmark programs, it was determined that the fairest and most consistent measure would be to count the lines of written code that would be compiled. This means only complete lines would be looked at, and unnecessary spaces and trialing lines with braces would need to be eliminated. This would compress the written code down into lines to be compiled while allowing each required statement to be on its own line. Figure 16 shows an excerpt of code from our elevator controller benchmark after it has been compressed by the line counter.

The results show the extent of the function of our counter, removing unnecessary space for the functions of the code (usually they are for clarification), and comments (which are not necessary for function and skipped by the compiler). Also, braces have been moved to the last complete line that requires them (meaning that some lines have multiple braces from nesting). Overall, this makes for a compact and fair result. Each of our six benchmark programs was processed through the counter. Table 4 shows the final lines counted for each program.

```
#include <stdio.h>
#include <stdbool.h>
enum DIRECTIONS = {up, down, stop, off};
int main() {
    int floors = [1, 2, 3, 4, 5];
    int startingFloor = 1;
    int callingFloor = 1;
    DIRECTIONS direction = stop;
    bool failure = false;
    char text(20);
    print("Starting up elevator...\n");
    sleep(1.5);
    print("opening Doors...\n");
    sleep(2);
    print("Initial Passengers Boarding...\n");
    sleep(5);
    do {
        fputs("Enter Destination: ", stdout);
        fflush(stdout);
        if(fgets(text, sizeof text, stdin)) {
            if (sscanf(text, "%d", &callingFloor) == 1) {
                printf("Number Entered: %d\n", callingFloor);
                if ((callingFloor <= 0 || callingFloor > 5) && callingFloor != -1) {
                    failure = true;
                    emergencyShutdown();
                }
                else if (callingFloor == -1) {}
                else {
                    direction = determineDirection(startingFloor, callingFloor, direction);
                    if (direction == up) {
                        movingUp(startingFloor, callingFloor);
                    }
                    else {
                        movingDown(startingFloor, callingFloor);
                    }
                }
            }
        }
        while (callingFloor != -1 || !failure);
        if (failure) {
            print("Maintenance Password Entered Properly!\n");
            print("Elevator can now be Shutdown for Service!\n");
            getchar();
            print("Shutting Down!\n");
            sleep(5);
        }
        else {
            print("Program is over...\n");
            print("Returning Elevator to Bottom Floor...\n");
            movingDown(startingFloor, 1);
            print("Please Press Enter to End Program!\n");
            getchar();
            print("Program will End in 3 Seconds...\n");
            sleep(3);
        }
    }
}
```

Figure 16: Elevator controller code after being parsed by our line counter

Table 4: The line counter results

Program	Total Lines Counter
Elevator Controller	96
Automatic Door Controller	62
Security System	89
Streetlight	83
Tollbooth	54
Vending Machine	61

The next step is Item 2. This proved difficult to automate, as the C library is vast, and developing a parser to incorporate all the possible injection sites for each of the error types could not be completed. In the future we plan to automate this process, but for now this was completed by hand through scanning the code for each error type and marking an injection site with a value, and then counting the number of places an error could be injected. Some errors had overlap, so care will need to be taken to incorporate both errors if there is overlap when injection sites are randomly selected.

Finally, Item 3 is completed. Our program allows us to input the number of lines determined from the line counter program, and the number of places each error can be injected. The error selector is hard coded with the C++ high and low error rates determined in Section 7, as those will not change. The algorithm uses the rate, which is a percentage, the average errors seen per one hundred lines of code, and then determines the number of errors that should be injected into the current program. For example, if the program had two hundred lines of code, then, based on the high error rate, 28.62 errors should be injected into this program. Since we can not generate fractional errors, the system rounds up to the nearest integer for rates $x.y$ where

$.y \geq .5$ and rounds down to the nearest integer for rates $x.y$ where $.y < .5$. In a similar fashion, using the total number of injected errors just calculated, the program calculates the number of errors of each type to generate. This is done by having PCG32 generate a decimal value z within the range $0 \leq z \leq 100$. The previous error percentages are now on a scale from 0 to 100, with subdivisions ending with their rate plus the previous rate. Table 5 shows the range for each error.

Table 5: Repeatable error ranges for random generation

Error Type	Decimal Range
Spelling	$0 \leq z \leq 0.1499$
Used Wrong Type	$0.1499 < z \leq 0.3624$
Missing Header Info	$0.3624 < z \leq 0.436$
Incorrect Math Operation	$0.436 < z \leq 0.6022$
Incorrect String Operation	$0.6022 < z \leq 0.6594$
Error Handling	$0.6594 < z \leq 0.7003$
Passing Parameters	$0.7003 < z \leq 0.7412$
Conditional Error	$0.7412 < z \leq 0.7984$
Incorrect Method	$0.7984 < z \leq 0.8856$
Missing Block/End of Line Characters	$0.8856 < z \leq 1$

The system repeats the process of using PCG32 to generate a value and comparing it to the scale in Table 5 until all errors for the system have been generated. Table 6 shows an example of the number of errors of each type selected randomly for a single run of the program for the Elevator Controller for the high error rate.

Finally, the last step is to determine the locations for the errors to be placed for the number of each error selected. These locations are also randomly selected using PCG32. For example, Table 6 shows that, for the Elevator Controller, 3 spelling errors are to be made. Using the selected number of errors by type, and the total number of places in the code that error can occur,

the system randomly selects the locations for injection into the elevator code. The system is configured not to select the same injection site more than once.

Table 6: Error selection for the first elevator test run

Error Type	Total Generated
Spelling	3
Used Wrong Type	2
Missing Header Info	1
Incorrect Math Operation	1
Incorrect String Operation	1
Error Handling	1
Passing Parameters	0
Conditional Error	2
Incorrect Method	0
Missing Block/End of Line Characters	3

For multiple testing runs, the program is run multiple times to generate a unique set of error injections for each run of a program, with no two runs being alike. For our testing, each program is going to be put through the testing procedure ten times, which requires ten selections of error types and locations to be processed. This process is then repeated for the low error rate as well, giving us a combined twenty unique error profiles for our program. Therefore, with this process, no program will have the same errors from testing iteration to testing iteration, allowing for a random error removal process as errors are discovered.

10 Data Collection Using Elevator Controller

Previously in [2] we illustrated a four state testing procedure. To re-cap, this process occurs as follows: (i) Compilation, (ii) Static Analysis, (iii) Compilation After Static Analysis (CASA), and (iv) Testing. From each of these stages, errors are determined and removed, and the time taken to locate all errors is recorded for each stage. Each stage’s time and error count is combined, respectively, giving us total errors found during a testing iteration. Testing iterations are repeated until no more errors are located in all four stages.

A database is utilized to store the errors found. We built our database in Microsoft Access as it comes configured to allow the use of style sheets for inputting data entries, and it was easy to build a single repository to include all runs for each error rate for each program. By incorporating slots for time, location in the code, and the categorization method from Putz [69], we have created a comprehensive form that shows the discovery and removal of each error in the system being tested. Figure 17 shows the entry form for the database, and Figure 18 shows an excerpt from the table of entries after errors have been entered.

This digital recording method is faster than trying to record errors by hand and allows the developer to focus on finding and removing errors in a timely manner. This helps to reduce bias in the data and exclude

as much as possible extra time that might be taken up by writing the findings down or using other recording methods. This also allows the developer to have a digital repository of their error findings, allowing them to see possible improvement from program to program as they utilize our testing procedure.

11 Results and Data Analysis

Using the testing procedure described in Section 10 and explained in [2], we completed ten runs for each error rate on our starting example the elevator controller. The results [2, 3] showed that using our complete development process, from formal specification through testing, demonstrates that software systems can use error metrics that fit statistical models, as originally proposed by Musa [25, 26] and Drake [30]. Further refined models [29] were explored in [3] but were inconclusive as to whether they were better than the base exponential model. This required further testing.

Now we have completed the same testing procedure on our remaining benchmark programs, using feedback garnered from our work [3]. Critique of the work suggested improving the fit of the models. As the models are non-linear, traditional R-squared metrics do not apply, as these are just for linear regression models [78]. We did use the pseudo adjusted R-squared value along with the average error of regression, S , to determine the best fit for our models in [3]. To enhance these values while not degrading the accuracy of the results, we looked into mathematical methods for improvement. The best method seems to be the removal of outliers from the data set. Previously we had attempted the removal of outliers with the elevator controller; however this relied on what is considered the “eyeball test,” looking for data that did not fit with the rest, and that is not robust. For our current approach at eliminating outliers, we focused on the recorded time to find all errors in the first testing iteration. To do this we take the total time for iteration one of all tests completed, ten values in total, and calculate the mean and standard deviation of these values. From these values we generate a range of one standard deviation centered around the mean. Any test run whose starting time falls plus or minus one standard deviation outside of the mean has the entire run purged from the results. In most of the benchmarks this results in one to two strings of data being removed, with no more than three being removed in the most extreme case. The test runs removed also vary in number of iterations completed to reach “zero” errors. For the Automated Door results, we can see from Table 7 that two data sets are removed from the results.

This process modifies the results without compromising the randomness of the data or the model. With this new subset of results, we calculate the exponential, modified exponential, and multinomial exponential models as done in [3]. Table 8 shows the S and pseudo adjusted R-squared for the base results and then the three model types after the data improvement.

Time Defect Found:	<input type="text"/>
Defect Type:	<input type="text"/>
Defect Reason:	<input type="text"/>
Module:	<input type="text"/>
Line Number:	<input type="text"/>
Analysis Tools:	<input type="text"/>
Time Defect Removed:	<input type="text"/>
Comments:	<input type="text"/>

Figure 17: Form used to input errors into the database

Time Defect Found	Defect Type	Defect Reason	Module	Line Number	Analysis Tools	Time Defect Removed
12/3/2015 - 14:58:00.00						
12/3/2015 - 15:04	IU	KN	elevatorV1.	23	Compiler	12/3/2015 - 15:30
12/3/2015 - 15:05	WN	TY	elevatorV1.	25	Compiler	12/3/2015 - 15:38
12/3/2015 - 15:06	IU	OM	elevatorV1.	28	Compiler	12/3/2015 - 15:42
12/3/2015 - 15:07	IU	KN	elevatorV1.	31	Compiler	12/3/2015 - 16:02
12/3/2015 - 15:07	MI	OM	elevatorV1.	32	Compiler	12/3/2015 - 16:04
12/3/2015 - 15:08	WE	TY	elevatorV1.	33	Compiler	12/3/2015 - 16:09
12/3/2015 - 15:09	WE	TY	elevatorV1.	45	Compiler	12/3/2015 - 16:09
12/3/2015 - 15:09	WN	TY	elevatorV1.	59	Compiler	12/3/2015 - 16:13
12/3/2015 - 15:10	WN	TY	elevatorV1.	70	Compiler	12/3/2015 - 16:14
12/3/2015 - 15:11	MI	TY	elevatorV1.	84	Compiler	12/3/2015 - 16:18
12/3/2015 - 15:12	WT	KN	elevatorV1.	90	Compiler	12/3/2015 - 17:08
12/3/2015 - 15:12	WE	OM	elevatorV1.	91	Compiler	12/3/2015 - 17:11
12/3/2015 - 15:13	WN	TY	elevatorV1.	111	Compiler	12/3/2015 - 17:13
12/3/2015 - 15:14	WE	OM	elevatorV1.	135	Compiler	12/3/2015 - 17:16
12/3/2015 - 15:14	WT	IG	elevatorV1.	145	Compiler	12/3/2015 - 17:19
12/3/2015 - 15:15	WN	TY	elevatorV1.	148	Compiler	12/3/2015 - 17:20
12/3/2015 - 15:15	WA	TY	elevatorV1.	149	Compiler	12/3/2015 - 17:22
12/3/2015 - 15:15	MI	OM	elevatorV1.	159	Compiler	12/3/2015 - 17:24
12/3/2015 - 15:16	WN	TY	elevatorV1.	160	Compiler	12/3/2015 - 17:28
12/3/2015 - 17:30	MI	OM	elevatorV1.	161	Compiler	12/3/2015 - 17:31
12/3/2015 - 17:33	WN	TY	elevatorV1.	35	Compiler	12/3/2015 - 17:34

Figure 18: Excerpt from the table holding the errors discovered in run one of testing the elevator controller

Table 7: The automated door outlier removal process

First Iteration Starting Times	Mean of Starting Times	Standard Deviation of Starting Times	One Standard Deviation	Starting Times Without Outliers
47.005	27.525	8.35235	35.877	22.292
22.292			19.1727	24.296
24.296				28.060
28.606				28.223
28.223				28.846
28.846				34.313
34.313				19.502
19.502				24.636
24.636				
18.077				

Table 8: The final results for our benchmark programs

Error Rate	Program Tested	Data Model	S	Pseudo Adjusted R-Squared
High	Automated Door	Base Exponential	5.6586	0.4267
		Exponential	4.8513	0.5957
		Modified Exponential	4.7571	0.5789
		Multinomial Exponential	3.8516	0.7452
	Security System	Base Exponential	5.6457	0.6351
		Exponential	4.7965	0.7297
		Modified Exponential	4.7162	0.7201
		Multinomial Exponential	N/A	N/A
	Streetlight	Base Exponential	4.4259	0.5411
		Exponential	3.7957	0.6425
		Modified Exponential	3.9823	0.6356
		Multinomial Exponential	3.8650	0.6293
	Tollbooth	Base Exponential	5.2394	0.4518
		Exponential	4.9228	0.5655
		Modified Exponential	4.8191	0.5316
		Multinomial Exponential	4.5043	0.6149
Vending Machine	Base Exponential	2.4275	0.8491	
	Exponential	1.7077	0.9306	
	Modified Exponential	1.8919	0.9248	
	Multinomial Exponential	1.7077	0.9306	
Low	Automated Door	Base Exponential	4.5366	0.4411
		Exponential	3.8664	0.6091
		Modified Exponential	4.1704	0.5866
		Multinomial Exponential	3.9120	0.6091
	Security System	Base Exponential	2.6590	0.8513
		Exponential	1.5579	0.9488
		Modified Exponential	1.5175	0.9457
		Multinomial Exponential	N/A	N/A
	Streetlight	Base Exponential	2.6988	0.7028
		Exponential	2.4116	0.7594
		Modified Exponential	2.3674	0.7488
		Multinomial Exponential	2.4524	0.7512
	Tollbooth	Base Exponential	3.0550	0.6349
		Exponential	1.9577	0.8601
		Modified Exponential	2.1923	0.8496
		Multinomial Exponential	1.1234	0.9539
Vending Machine	Base Exponential	2.4878	0.6034	
	Exponential	1.2017	0.8980	
	Modified Exponential	1.3557	0.8902	
	Multinomial Exponential	0.8980	1.2017	

As we can see from the results, there is a definitive increase in the model fitment compared to the base, unmodified exponential model when the outliers are removed. The results show that in some cases the standard exponential model fits the data the best, in others the modified exponential, and in others the multinomial exponential. For the Vending Machine results, in both the high and low error rate results, the multinomial is the base exponential, as the best exponent has only the “x” term in it. From the results we can see that there is not any way to predict at this time which model works best for which rate, as each group has examples where each is the best for that program. In the case of the Security System, the multinomial could not be processed for both error rates, as the multinomial equation that fit best diverged to infinity when used as the exponent to an exponential.

12 Conclusion

The work presented here shows our complete design and testing process used in creating an initial benchmark suite for software reliability. Error analysis discovery with formal methods was shown to be more widely applicable than originally reported in [1], as we showed that errors could also be found by ATPs in the proof, not just the specification. We illustrated our RNG testing and random error generation that previously was mentioned in [2] but was not demonstrated in that work. This was utilized to help enhance our results in [3] and expand our tested programs. This expanded program suite illustrates that software errors, using a proper development process, can be shown to be modeled statistically similar to hardware. We have shown through the removal of outliers that the results can be improved further still. This develop-

ment process and the full results set a benchmark, so other development processes can be researched and even better software analysis can be achieved.

13 Future Work

Based on the work presented here, the first major area of expansion is the benchmarks themselves. We have shown the development and testing process allows small scale software programs to meet the statistical models that were originally theorized. This is great for embedded systems, but the process may be applicable to large system on chip solutions, or even large scale, big data applications. To begin this expansion, we will develop larger, more complex versions of the six benchmark programs already developed to see if the results hold when scaled up.

Similar to this area is the need to improve the development process with good, well developed software modules. In hardware, most systems are incorporated with ICs that come from manufacturers with known reliability. The same should be true of software. If modules have already been vetted through the development process, then they should be correct and accurate, allowing their reuse as a black box, and improving the error results of a new system prior to testing. This would be similar to pulling an IC of AND or NOT gates from the parts bin, knowing what it does and expecting it to work, without seeing the intricacies inside the black box.

The other side of safety critical systems is determining trust. Work in this space has looked into ways of showing trust for both hardware and software. Our development process could aid in this area, being two-fold with reliability, if measures of security and trust can be incorporated into the formal method specification. In the future we will use the original elevator specification and redesign it, incorporating new security measures, while maintaining our current reliability measures. We will attempt to show the proofs can be developed to show security and reliability in one step, allowing for their incorporation from the beginning of design.

One final area of expansion is the integration of the software designed with this process into a targeted application. This requires further improvement of the software reliability in the system. Unlike hardware, which can use redundancy of circuitry to improve reliability, copying software with an undiscovered bug just propagates that bug to all versions. One solution may be to develop a hardware based monitor, possibly with machine learning, to determine when a software anomaly has been seen, and correct it. AI could be trained on the state and inputs of the system to know what outputs should be provided, and if the wrong output is generated from a known state, then a safety protocol could be enacted for a trusted system, or just replaced with the correct value for a non-critical system.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgments The authors would like to thank L-3 Fuzing and Ordnance Systems. A portion of this work was funded in part by their generous grant.

References

- [1] J. Lockhart, C. Purdy, and P. A. Wilsey, "The use of automated theorem proving for error analysis and removal in safety critical embedded system specifications," in *2017 IEEE National Aerospace and Electronics Conference (NAECON)*, June 2017, pp. 358–361.
- [2] J. Lockhart, C. Purdy, and P. A. Wilsey, "Error analysis and reliability metrics for software in safety critical systems," in *2016 IEEE 59th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Oct 2016, pp. 1–4.
- [3] J. Lockhart, C. Purdy, and P. A. Wilsey, "Error analysis and reliability metrics for software in safety critical systems," in *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2018.
- [4] A. Mishchenko, S. Chatterjee, and R. Brayton, "Dag-aware aig rewriting: a fresh look at combinational logic synthesis," in *2006 43rd ACM/IEEE Design Automation Conference*, July 2006, pp. 532–535.
- [5] C. Albrecht, "Iwls 2005 benchmarks," Tech. Rep., Jun 2005.
- [6] J. Cong and K. Minkovich, "Optimality study of logic synthesis for lut-based fpgas," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 26, no. 2, pp. 230–239, Feb 2007.
- [7] Static & Formal Verification. [Online]. Available: <https://www.synopsys.com/verification/static-and-formal-verification.html> (Accessed 2017-3-17).
- [8] A. J. Hu. Formal hardware verification with bdds: An introduction. [Online]. Available: <https://www.cs.ox.ac.uk/files/4309/97H1.pdf> (Accessed 2017-4-22).
- [9] A. J. Hu, "Formal hardware verification with bdds: an introduction," in *1997 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, PACRIM. 10 Years Networking the Pacific Rim, 1987-1997*, vol. 2, Aug 1997, pp. 677–682 vol.2.
- [10] Arvind, N. Dave, and M. Katelman, "Getting formal verification into design flow," in *Proceedings of the 15th International Symposium on Formal Methods*, ser. FM '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 12–32. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-68237-0_2
- [11] Arvind, N. Dave, and M. Katelman. Getting Formal Verification into Design Flow. [Online]. Available: <http://people.csail.mit.edu/ndave/Research/fm.2008.pdf> (Accessed 2017-4-22).
- [12] J. Markoff. (2016) Moore's law running out of room, tech looks for a successor. [Online]. Available: <http://tinyurl.com/MooresLawOutOfRoom> (Accessed 2016-5-6).
- [13] J. Hruska. (2016, July) Moore's law scaling dead by 2021, to be replaced by 3d integration. [Online]. Available: <https://tinyurl.com/EXT-MooresLaw> (Accessed 2018-3-17).
- [14] R. Smith. (2015) Amd dives deep on high bandwidth memory - what will hmb bring amd. [Online]. Available: <http://www.anandtech.com/show/9266/amd-hbm-deep-dive> (Accessed 2016-4-29).
- [15] AMD. (2015) High bandwidth memory — reinventing memory technology. [Online]. Available: <http://www.amd.com/en-us/innovations/software-technologies/hbm> (Accessed 2016-4-29).
- [16] J. Hruska. (2017, July) Mit develops 3d chip that integrates cpu, memory. [Online]. Available: <https://tinyurl.com/EXT-MIT-3D-Chip> (Accessed 2018-3-17).

- [17] H. Knight. (2017, July) New 3-d chip combines computing and data storage. [Online]. Available: <http://news.mit.edu/2017/new-3-d-chip-combines-computing-and-data-storage-0705> (Accessed 2018-3-17).
- [18] M. Suri, D. Querlioz, O. Bichler, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Bio-inspired stochastic computing using binary cbram synapses," *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2402-2409, July 2013.
- [19] K. Alrawashdeh and C. Purdy, "Fast hardware assisted on-line learning using unsupervised deep learning structure for anomaly detection," in *International Conference on Information and Computer Technologies (ICICT 2018)*. IEEE, 2018, p. In Press.
- [20] T. Griggs and D. Wakabayashi. (2018, March) How a self-driving uber killed a pedestrian in arizona. [Online]. Available: <https://tinyurl.com/ybvcgfu4> (Accessed 2018-3-28).
- [21] R. Marosi. (2010) Runaway prius driver: 'i was laying on the brakes but it wasn't slowing down'. [Online]. Available: <http://articles.latimes.com/2010/mar/10/business/la-fi-toyota-prius10-2010mar10> (Accessed 2014-1-26).
- [22] A. Haxthausen, "An introduction to formal methods for the development of safety-critical applications," Technical University of Denmark, Lyngby, Denmark, Tech. Rep., August 2010.
- [23] G. Finzer. (2014) How many defects are too many? [Online]. Available: <http://labs.sogeti.com/how-many-defects-are-too-many/> (Accessed 2016-4-29).
- [24] V. R. Basili and B. T. Perricone, "Software errors and complexity: An empirical investigation," *Commun. ACM*, vol. 27, no. 1, pp. 42-52, Jan. 1984. [Online]. Available: <http://doi.acm.org/10.1145/69605.2085>
- [25] J. D. Musa, "A theory of software reliability and its application," *IEEE Transactions on Software Engineering*, vol. SE-1, no. 3, pp. 312-327, Sept 1975.
- [26] J. D. Musa and A. F. Ackerman, "Quantifying software validation: when to stop testing?" *IEEE Software*, vol. 6, no. 3, pp. 19-27, May 1989.
- [27] S. Yamada, *Software reliability modeling: fundamentals and applications*. Springer, 2014. [Online]. Available: <https://link.springer.com/content/pdf/10.1007/978-4-431-54565-1.pdf>
- [28] S. Yamada and S. Osaki, "Reliability growth models for hardware and software systems based on nonhomogeneous poisson processes: A survey," *Microelectronics Reliability*, vol. 23, no. 1, pp. 91 - 112, 1983. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/0026271483913720>
- [29] S. Yamada, *Stochastic Models in Reliability and Maintenance*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, ch. Software Reliability Models, pp. 253-280. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-24808-8_10
- [30] H. D. Drake and D. E. Wolting, "Reliability theory applied to software testing," *Hewlett-Packard Journal*, vol. 38, pp. 35-39, 1987.
- [31] RTCA and EUROCAE, "Software considerations in airborne systems and equipment certification," RTCA, Inc., Washington, D.C., USA, Tech. Rep., 1992.
- [32] Federal Aviation Administration, "Advisory circular 20-174," U.S. Department of Transportation, Tech. Rep., 2011. [Online]. Available: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_20-174.pdf
- [33] J. Lockhart, C. Purdy, and P. Wilsey, "Formal methods for safety critical system specification," in *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug 2014, pp. 201-204.
- [34] M. Popovic, V. Kovacevic, and I. Velikic, "A formal software verification concept based on automated theorem proving and reverse engineering," in *Engineering of Computer-Based Systems, 2002. Proceedings. Ninth Annual IEEE International Conference and Workshop on the*, 2002, pp. 59-66.
- [35] J. Rushby, "Harnessing disruptive innovation in formal verification," in *Software Engineering and Formal Methods, 2006. SEFM 2006. Fourth IEEE International Conference on*, Sept 2006, pp. 21-30.
- [36] W. Visser, K. Havelund, G. Brat, and S. Park, "Model checking programs," in *Automated Software Engineering, 2000. Proceedings ASE 2000. The Fifteenth IEEE International Conference on*, 2000, pp. 3-11.
- [37] Lemma 1, "Getting started," in *ProofPower: Tutorial*. Berkshire, United Kingdom: Lemma 1 Ltd, 2006, ch. 1, pp. 5-10.
- [38] H. Ganzinger and V. Sofronie-Stokkermans, "Chaining techniques for automated theorem proving in many-valued logics," in *Multiple-Valued Logic, 2000. (ISMVL 2000) Proceedings. 30th IEEE International Symposium on*, 2000, pp. 337-344.
- [39] C. Morgan, "Methods for automated theorem proving in non-classical logics," *Computers, IEEE Transactions on*, vol. C-25, no. 8, pp. 852-862, Aug 1976.
- [40] A. A. Larionov, E. A. Cherkashin, and A. V. Davydov, "Theorem proving software, based on method of positively-constructed formulae," in *MIPRO, 2011 Proceedings of the 34th International Convention*, May 2011, pp. 965-968.
- [41] M. Kaufmann and J. S. Moore, "Some key research problems in automated theorem proving for hardware and software verification," in *RACSAM. Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, vol. 98, no. 1, Oct 2004, pp. 181-195.
- [42] J. Bowen and V. Stavridou, "Safety-critical systems, formal methods and standards," *Software Engineering Journal*, vol. 8, no. 4, pp. 189-209, Jul 1993.
- [43] E. M. Clarke and J. M. Wing, "Formal methods: State of the art and future directions," *ACM Comput. Surv.*, vol. 28, no. 4, pp. 626-643, Dec 1996.
- [44] Z. Li, S. Lu, S. Myagmar, and Y. Zhou, "Cp-miner: finding copy-paste and related bugs in large-scale software code," *IEEE Transactions on Software Engineering*, vol. 32, no. 3, pp. 176-192, March 2006.
- [45] S. Morasca, "A probability-based approach for measuring external attributes of software artifacts," in *2009 3rd International Symposium on Empirical Software Engineering and Measurement*, Oct 2009, pp. 44-55.
- [46] A. Varshney, R. Majumdar, C. Choudhary, and A. Srivastava, "Role of parameter estimation & prediction during development of software using srgm," in *Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015 4th International Conference on*, Sept 2015, pp. 1-6.
- [47] D. Tang and M. Hecht, "Evaluation of software dependability based on stability test data," in *Fault-Tolerant Computing, 1995. FTCS-25. Digest of Papers., Twenty-Fifth International Symposium on*, June 1995, pp. 434-443.
- [48] Y. Tamura and S. Yamada, "Reliability analysis based on a jump diffusion model with two wiener processes for cloud computing with big data," *Entropy*, vol. 17, no. 7, pp. 4533-4546, 2015. [Online]. Available: <http://www.mdpi.com/1099-4300/17/7/4533>
- [49] S. Yamada and M. Yamaguchi, "A method of statistical process control for successful open source software projects and its application to determining the development period," *International Journal of Reliability, Quality and Safety Engineering*, vol. 23, no. 05, p. 1650018, 2016. [Online]. Available: <https://www.worldscientific.com/doi/abs/10.1142/S0218539316500182>
- [50] ISO/IEC, "Information technology: Z formal specification notation : syntax, type system, and semantics," International Organization for Standardization and International Electrotechnical Commission, Geneva, Switzerland, Standard 13568, 2002.
- [51] ISO/IEC, "Information technology, programming languages, their environments and system software interfaces, vienna development method, specification language," International Organization for Standardization and International Electrotechnical Commission, Geneva, Switzerland, Standard 13817-1, 1996.
- [52] M. Huth and M. Ryan, "Propositional logic," in *Logic in Computer Science: Modelling and Reasoning about Systems*, 1st ed. Cambridge, United Kingdom: Cambridge University Press, 2000, ch. 1, pp. 1-88.

- [53] J. Davies and J. Woodcock, "Propositional logic," in *Using Z: Specification, Refinement, and Proof*, 1st ed. Hertfordshire, United Kingdom: Prentice Hall Europe, 1996, ch. 2, pp. 9–26.
- [54] M. Huth and M. Ryan, "Predicate logic," in *Logic in Computer Science: Modelling and Reasoning about Systems*, 1st ed. Cambridge, United Kingdom: Cambridge University Press, 2000, ch. 2, pp. 90–146.
- [55] J. Davies and J. Woodcock, "Predicate logic," in *Using Z: Specification, Refinement, and Proof*, 1st ed. Hertfordshire, United Kingdom: Prentice Hall Europe, 1996, ch. 3, pp. 37–2012 44.
- [56] J. M. Spivey, "Basic concepts," in *Understanding Z: A Specification Language and its Formal Semantics*, 1st ed. New York, New York: Cambridge University Press, 1988, ch. 2, pp. 18–24.
- [57] J. Davies and J. Woodcock, "Sets," in *Using Z: Specification, Refinement, and Proof*, 1st ed. Hertfordshire, United Kingdom: Prentice Hall Europe, 1996, ch. 5, pp. 57–72.
- [58] J. M. Spivey, "Studies in z style," in *Understanding Z: A Specification Language and its Formal Semantics*, 1st ed. New York, New York: Cambridge University Press, 1988, ch. 5, pp. 98–113.
- [59] Lemma 1, "Solutions to exercises," in *ProofPower: Z Tutorial*. Berkshire, United Kingdom: Lemma 1 Ltd, 2006, ch. 8, pp. 143–149.
- [60] H. Pishro-Nik. ("", "") 11.1.4 nonhomogeneous poisson processes. [Online]. Available: <https://tinyurl.com/yb5kbspn> (Accessed 2018-12-7).
- [61] K. Siegrist. ("", "") 6. non-homogeneous poisson processes. [Online]. Available: <https://tinyurl.com/yc9y9vch> (Accessed 2018-12-7).
- [62] MIT OpenCourseWare. (2011, Spring) Chapter 2 poisson processes. [Online]. Available: <https://tinyurl.com/ycg6r3n4> (Accessed 2018-12-7).
- [63] NIST. ("", "") 8.1.7.2. non-homogeneous poisson process (nhpp) - power law. [Online]. Available: <https://tinyurl.com/yclorwqd> (Accessed 2018-12-7).
- [64] S. Yamada, "Software reliability models," in *Stochastic Models in Reliability and Maintenance*, S. Osaki, Ed. Springer Berlin Heidelberg, 2002, pp. 253–280.
- [65] V. Putz. (2000) The personal software process: an independent study. [Online]. Available: http://www.nyx.net/~vputz/psp_index/book1.html (Accessed 2013-10-30).
- [66] H. Watts, "The personal software process (psp)," Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU/SEI-2000-TR-022, 2000. [Online]. Available: <http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=5283>
- [67] V. Putz. (2000) Program 1a. [Online]. Available: http://www.nyx.net/~vputz/psp_index/x341.html (Accessed 2017-02-16).
- [68] V. Putz. (2000) Program 2a: Simple loc counter. [Online]. Available: http://www.nyx.net/~vputz/psp_index/x1547.html (Accessed 2017-01-26).
- [69] V. Putz. (2000) Chapter 2. lesson 2: Planning and measurement. [Online]. Available: http://www.nyx.net/~vputz/psp_index/c1028.html (Accessed 2013-10-30).
- [70] D. E. Knuth, *The Art of Computer Programming, Vol. 1*, 3rd ed. Addison-Wesley, 1997.
- [71] D. E. Knuth, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd ed. Addison-Wesley, 1998.
- [72] Marsaglia, "Note on a proposed test for random number generators," *IEEE Transactions on Computers*, vol. C-34, no. 8, pp. 756–758, Aug 1985.
- [73] P. L'Ecuyer and R. Simard, "Testu01 a software library in ansi c for emperical testing of random number generators," Département d'Informatique et de Recherche Opérationnelle Université de Montréal, Tech. Rep., 2013. [Online]. Available: <http://simul.iro.umontreal.ca/testu01/guideshorttestu01.pdf>
- [74] P. L'Ecuyer and R. Simard, "Testu01: A c library for empirical testing of random number generators," *ACM Trans. Math. Softw.*, vol. 33, no. 4, Aug. 2007. [Online]. Available: <http://doi.acm.org/10.1145/1268776.1268777>
- [75] M. E. O'Neill, "Pcg: A family of simple fast space-efficient statistically good algorithms for random number generation," Harvey Mudd College, Claremont, CA, Tech. Rep. HMC-CS-2014-0905, Sep. 2014.
- [76] F. Neugebauer, I. Polian, and J. P. Hayes, "Building a better random number generator for stochastic computing," in *2017 Euromicro Conference on Digital System Design (DSD)*, Aug 2017, pp. 1–8.
- [77] J. Lockhart, K. A. Rawashdeh, and C. Purdy, "Verification of random number generators for embedded machine learning," in *2018 IEEE National Aerospace and Electronics Conference (NAECON)*, In Press 2018.
- [78] UCLA: Institute for Digital Research and Education. (2011) Faq: What are pseudo r-squareds? [Online]. Available: <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/> (Accessed 2018-6-9).

Observing and Forecasting the Trajectory of the Thrown Body with use of Genetic Programming

Konstantin Mironov^{*1,2}, Ruslan Gayanov², Dmiriy Kurennov¹

¹Institute of New Materials and Technologies, Ural Federal University, 620002, Russia

²Faculty of Computer Science and Robotics, Ufa State Aviation Technical University, 450008, Russia

ARTICLE INFO

Article history:

Received: 06 January, 2019

Accepted: 28 January, 2019

Online : 20 February, 2019

Keywords:

Robotic Catching

Genetic Programming

Forecasting

Machine Vision

Machine Learning

ABSTRACT

Robotic catching of thrown objects is one of the common robotic tasks, which is explored in a number of papers. This task includes subtask of tracking and forecasting the trajectory of the thrown object. Here we propose an algorithm for estimating future trajectory based on video signal from two cameras. Most of existing implementations use deterministic trajectory prediction and several are based on machine learning. We propose a combined forecasting algorithm where the deterministic motion model for each trajectory is generated via the genetic programming algorithm. Object trajectory is extracted from video sequence by the image processing algorithm, which include Canny edge detection, Random Sample Consensus circle recognition and stereo triangulation. After that trajectory is forecasted using proposed method. Numerical experiments with real trajectories of the thrown tennis ball show that the algorithm is able to forecast the trajectory accurately.

1. Introduction

This paper is an extension of the article presented at the IEEE International Symposium on Signal Processing and Information Technology [1]. Here and there we address the task of robotic catching of thrown objects or, more precisely the subtask of observing and predicting the trajectory of the thrown object.

With the development of robotics, mechanical systems acquire more and more features that were previously only available to humans. One of these possibilities is the ability to catch objects thrown in the air. Initially, a robotic capture of objects thrown in the air was described in 1991 in [2]. Later, this task was considered several times in a number of articles [1,3,11].

In addition to theoretical value, such a skill can have a practical application. For example, in light industry, the task of transporting workpieces between machine tools processing them often arises. The traditional solution to this problem is the use of various conveyor systems. Robotic transfer as a method of such transportation was proposed in 2006 by Frank [12]. This application was developed in [13-21]. Transportation of an object from some point of departure A to some destination B is as follows: the robot thrower located in A throws the object in direction B and notifies about it via the communication line, and the robot catcher located in B, having received the notification, carries out object capture on the fly.

The authors of [12,15] specify the following potential advantages of robotic throw compared to traditional conveyor-based systems:

- Greater flexibility. Flexibility is understood as the ability to quickly deploy, collapse and redevelop a transport network with an arbitrary topology, or to use it in production facilities with an arbitrary layout.
- Higher speed of object transportation.
- Reduced energy consumption.

The share of successful captures in most existing systems does not exceed 80% (two exceptions are described in [11] and [16]; in the first article, a high proportion of successful captures is provided by large linear dimensions of the gripping device; in the second, by throwing cylindrical objects of high aerodynamic stability), which is not sufficient for use in a real industrial environment. Thus, the practical implementation of transportation by robotic throw is a complex and relevant scientific task.

For a successful capture, it is necessary to know in which point of space the object will be at the moment of capture, and at what speed it will move at the same time [15]. The point in space and time where the capture is carried out is selected among the set of points that the object passes when it flies through the working space of the capture device. Their combination forms the trajectory of the object in the working space of the capture device. This

*Corresponding Author: Konstantin Mironov, mironovconst@gmail.com

trajectory must be predicted in advance so that the robot catcher has time to complete the capture [15]. Prediction is based on measuring the object trajectory immediately after the throw. In general, in [15], the following four subtasks are distinguished when ensuring the transportation of objects by robotic transfer:

- throw;
- capture;
- forecasting;
- tracking.

Here we consider last two subtasks. Trajectory forecasting is needed in order to provide the catcher with the information about object trajectory within the workspace of the gripper. Most of the trajectory forecasting algorithms are based on ballistic modeling of the flight. These models include the influence of gravitation only (e. g. [2,3,7]; in this case forecasting may be implemented by fitting a parabola to the reference of measured positions) or gravitation and air drag (e. g. [4,8,9]). This modeling requires preliminary knowledge about ballistic properties of thrown objects. However, human children do not need such a knowledge to catch the ball successfully. They do it only based on the previous experience. This circumstance motivated the development of learning-based forecasting algorithms, such as neural network trajectory predictor [18] and k nearest neighbor's trajectory predictor [19-21]. Learning-based techniques require collecting the sampling of trajectories in order to train the predictor. Here we propose the method, which lies between model-based and learning-based model. The predictor is using equations to define future positions of the object, but the parameters of these equations are obtained by the learning procedure of genetic programming. Learning does not require a sampling of past trajectories: the parameters of the model are learned from the initial part of the current trajectory.

We do not consider the task of providing correct throwing and catching movement in this article. This is a complex control task, solved by various works in the field of robotics and mechatronics, e.g. Implementation of robotic control within our project is discussed in [17].

A tennis ball is considered as the object to be thrown. On one hand, this object is quite complex and unstable aerodynamically [22] so that its trajectory cannot be accurately predicted using simple models; on the other hand, its aerodynamic characteristics are investigated in sufficient detail ([22] provides a detailed overview of its characteristics completed in 50 years) so that the aerodynamic model can be used to verify the accuracy of the algorithm functioning.

2. Extraction of the spatial coordinates from video signal

Tracking the trajectory of a moving object is a task that often arises in machine vision applications. Following examples could be mentioned: In our case, it is considered for the following conditions: the object is a sphere thrown at a speed of several meters per second at certain angle to the horizon. Such conditions are determined by the task of robotic capture of a thrown object in the system of transportation of objects by transfer.

Since monitoring is performed through a camera, tracking an object becomes the task of processing images and video. Positioning the flying ball in space is performed using stereo

vision. The spatial position of a certain point is determined on the basis of its pixel coordinates on images from two cameras and on the basis of the system parameters: the relative location of the cameras, their focal lengths, etc. [23]. Camera parameters are configured using Zhang's calibration procedure [24-27].

The study of the stereo positioning accuracy is poorly described in the literature. Most of the articles describe positioning of static objects, for example, [28]. When positioning an object, errors inevitably occur. According to the classification proposed by Lee [28], they are divided into three types:

- Calibration errors. They are related to errors by calibration, i.e., in determining the parameters of the camera system. These errors are systematic and amount to no more than one millimeter per meter of distance.
- Quantization errors. They are associated with the transition from pixel to metric coordinates. The set of pixel coordinates does not correspond in space to a point, but to a certain area, the size of which increases with increasing distance from cameras. At a distance of up to two meters, the magnitude of quantization errors is small; at a greater distance, it becomes significant.
- Image processing errors. These errors are related to incorrect operation of image processing algorithms which are used to determine the position of a pixel point.

The influence of these errors on the positioning of a static spherical object was investigated in [26]. The object is tracked using two IDS uEye UI-3370CP [29] video cameras combined into a stereo pair. The resolution of each camera is 2048 by 2048 pixels. They are installed at a distance of several tens of centimeters. Studies of calibration and quantization errors showed that standard deviations due to calibration and quantization errors are less than 1.5 millimeters (ranges from 1 to 1.4 mm, with an increase in the range from 0.5 to 2.5 mm). Errors of image processing were more significant: the total standard deviation of the positioning of the sphere is up to 2.2 mm. Errors in the positioning of a thrown object in flight were analyzed in [27]. The algorithm described below was implemented in C++ using the CUDA library, which allows deparallelizing of the calculations for their execution on the graphics processor [30]. Some extensions of the algorithm help improve positioning accuracy.

The procedure for determining the spatial coordinates of an object is illustrated in Figure 1 (the original images are shown in the first row). It includes the following steps:

- Background subtraction. The results are shown in the second row of Figure 1.
- Selecting boundaries using the Canny algorithm [31]. Border images are shown in the third row on Figure 1.
- Circle detection on border images. The result of this stage are pixel coordinates of the circle center on each image. In [30], two methods of such prediction are compared: the Hough transformation [32,33] and the RANSAC [34] method. As a result, an algorithm based on the RANSAC method was chosen. It has the same accuracy as the Hough transformation, but requires fewer resources [30]. The algorithm selects three random points in the image, builds a circle on their basis and checks whether other points of the boundary image fit into this circle. This action is repeated sequentially until a circle is found that fits well

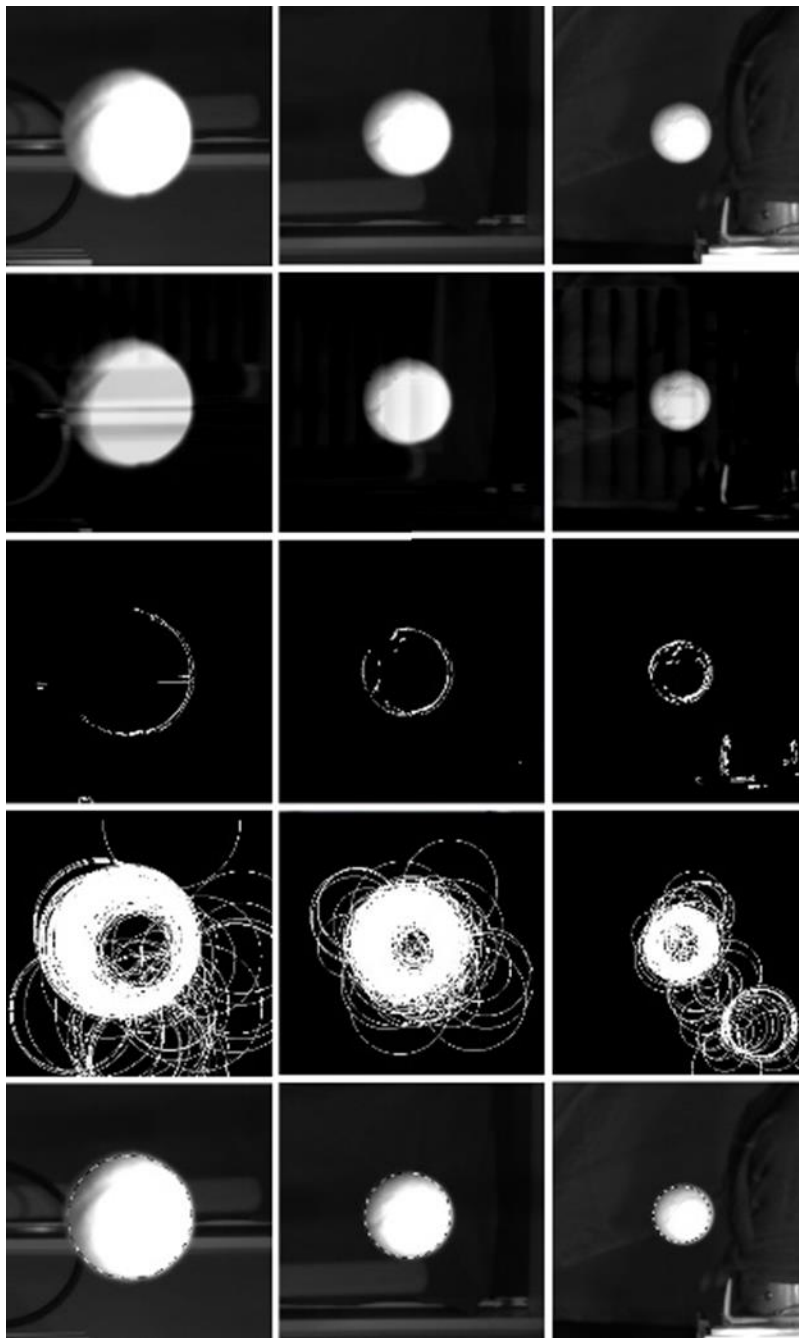


Figure 1. Circle recognition with RANSAC for three images.

with the points of the boundary image. The fourth row in Figure 1 shows the hypothetical circles generated by the algorithm, and the fifth row shows the selected circle projected onto the original image.

- Stereo triangulation. This is the operation of determining spatial coordinates for the center of an object based on its pixel coordinates in two images and using camera calibration parameters.

Coordinates obtained as a result of stereo-triangulation are then transferred to the system defined as follows:

- The center of coordinates coincides with the position of the object at the time of the throw.

- One of the axes is directed vertically upwards.
- The second axis is aligned with the horizontal projection of the direction of the throw.
- The transfer of coordinates into such a system provides a two main advantages. First, three-dimensional coordinates can be replaced by two-dimensional ones. Second, approximation of the trajectory by the plane allows you to identify outliers, i.e., filter out frames on which the position of the object is measured incorrectly. Image processing errors are associated with incorrect results in the first two steps, while calibration and quantization errors affect the result of stereo triangulation. Coordinate transform is described more precisely in the end of this section.

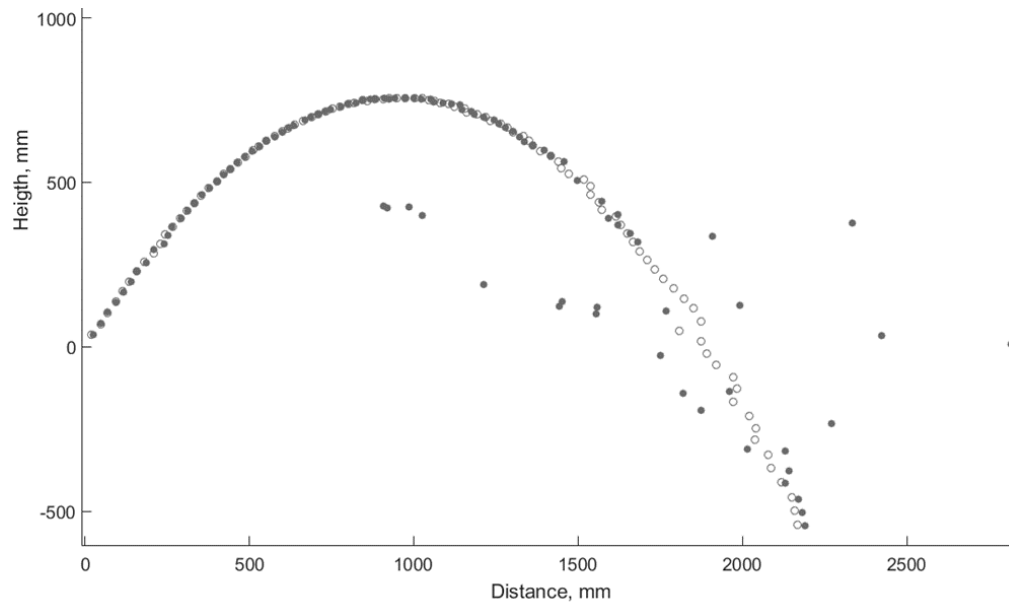


Figure 2. Graph of the trajectory measured with subtraction of the background (circles) and without it (ring)

The experiments described in [27] mainly concern the theoretical assessment of positioning accuracy. In [28], the author analyzed the errors found in the real situation. This analysis is complex, since there is no true data on the position of a real object at any given time. Many errors can be detected because they distort the smoothness of the path. This curve cannot be accurately determined analytically, but it is smooth [28]. Another way to assess errors is to approximate the measured values by a simplified model of the object movement. These models are not accurate, however, if they are more accurate than the vision system, the quality of the approximation can provide information about the accuracy of the observer.

Errors of individual processing steps can be detected by visual analysis of intermediate images. The quality of border selection can be assessed by comparing the found boundaries with the boundaries of the ball on the original image. The quality of circle recognition can be estimated by projecting the circles found on the original images. For example, a visual analysis of the images in Figure 1 shows that the border detection algorithm introduces some noise, but the RANSAC assessment gives plausible results. The disadvantage of this visual analysis is that it is performed by humans and cannot transmit objective information. However, it does detect some obvious tracking errors.

Subtracting the background before running the Kenny algorithm is an optional step, but in practice, this step is necessary for correct positioning of the object at a great distance. If background subtraction is not applied, the deviations of the measured values increase significantly when the distance from the camera to the object exceeds 1.5 meters. The effect is illustrated in Figure 2. Charts are shown for the same trajectory extracted by the RANSAC algorithm with and without background subtraction. You can see that at a distance of about 1.5 meters, the measurements almost coincide, and the trajectory looks like a second-order curve. Measurements with background subtraction (rings) retain this view afterwards, but measurements without subtracting the background (circles) become chaotic. This behavior is typical of most trajectories in a dataset.

Another way to obtain more accurate data for comparison and verification is typical for RANSAC. Since RANSAC does not provide the same results for different starts, several starts give several hypotheses about the position of the ball center. A correct statistical estimate based on these hypotheses is more accurate than the result of a single run of RANSAC. The results of multiple measurements are noisy and are not supported by the model of true motion and previous statistical knowledge, for example, the probability density function. According to [35], the least squares estimate shall be used under those conditions. Such an estimate for a static parameter with unknown random noise is equal to the average measurement result. In this paper, the mean value is replaced by the median. The median and average scores give similar results, but the median score is more resistant to emissions. The median of 1000 RANSAC launches was used in compiling the training base of trajectories; a further increase in the number of launches does not change the results of the median estimate. The use of such an estimate in real time is impossible due to the large amount of computation. An existing graphics processor can perform one run of RANSAC in real time (i.e., less than 9 ms for two images and less than 1 second for the entire trajectory). It takes about 10 minutes to run the RANSAC algorithm 1000 times.

The numerical evaluation of errors is given in Table 1. Here, the coordinates extracted by a single RANSAC run are compared with the results of the median estimate for 1000 runs. Differences are considered "errors." These numbers are not equal to real positioning errors, but they can be used to perceive the dispersion of measurements. Based on these differences, the standard deviation is calculated for each frame when the ball was thrown. In the table, each frames are combined into block to save space. Standard deviations are summarized based on 111 trajectories.

It can be seen that after the 65th frame, the parameter begins to increase strongly, and this growth is more impressive for the variant of the algorithm without subtracting the background. The reason for this increased stability at the beginning is that for the initial frames the size of the ball is larger and almost completely covers the image (compare the first and third columns in Figure 1). Therefore, the background borders make smaller distortions by the results of the border detection.

Table 1. Comparing the difference in millimeters between the measured 3D positions based on one RANSAC run and the median of 1000 RANSAC runs, for variants of the algorithm with and without background subtraction.

Frame Number	Standard Deviation		Median Error	
	Without Background Subtraction	With Background Subtraction	Without Background Subtraction	With Background Subtraction
1..5	7.9	6.4	1.6	0.8
6..10	4.0	1.9	1.8	0.9
11..15	3.7	2.1	2.0	1.1
16..20	2.8	1.9	1.6	0.5
21..25	2.1	1.8	1.4	0.3
26..30	4.0	2.2	2.2	0.5
31..35	22.1	17.2	2.9	2.2
36..40	10.9	4.0	3.4	0.4
41..45	24.8	3.2	3.7	0.4
46..50	29.8	3.9	4.5	0.7
61..65	63.9	14.8	5.5	1.0
66..70	187.4	41.3	10.5	4.2
71..75	305.6	138.5	20.4	5.4
76..80	520.4	242.2	208.2	7.3
81..85	897.0	229.3	171.9	8.0
86..90	1361.6	197.5	163.9	8.4
91..95	1450.0	212.1	176.1	9.3

It can be seen that even for the option with background subtraction, the standard deviation after the 70th frame reaches very high values. Standard deviation may not be the best option, as it has low emission resistance. Therefore, columns in the right-hand part of the table show median differences for the same blocks. The median results look the same as for standard deviations, but they are more detailed. For the algorithm without background subtraction, the average error lies at 3σ interval for static spheres, estimated as 6.75 mm [26], up to the 60th frame. For the algorithm with background subtraction, this property is preserved up to the 80th frame. In the version without background subtraction, an average value of more than 20 cm is reached after the 75th frame. This means that most frames are outliers in this area. Thus, measurements without background subtraction are practically useless.

Position measurements can be divided into inliers and outliers. Outliers are defined as measurements that are completely useless, even harmful for trajectory restoring. Inliers may be wrong, but they help to improve the score. Obviously, it is not possible to determine with 100% certainty whether a measurement is an inlier or outlier. The huge difference between the standard deviation and the median error at the end of the trajectory shows that outliers make up a large proportion of the measurements.

Trajectory construction demonstrates specific properties of these errors. Figure 3 shows three graphs describing the trajectory: relationship between the height of the object and the distance from the camera (upper graph), dependence of height from the frame

number (bottom left) and the dependence of distance from the frame number (bottom right). It is easy to see that the first and third graphs appear to be noisy on the right side, and the height--time dependence retains the appearance of a smooth curve of second order. In other words, errors are mainly related to distance measurement.

The reason for error localization in one dimension is that when the distance to the object exceeds the distance between the cameras of the stereo system, one pixel has a greater influence on the measurement of distance from the camera than on the measurement of other coordinates. An illustration of this property is shown in Figure 4. Experiments show that the error value at large distances is significant for tracking in terms of measuring distance. This problem can be overcome by increasing the number of cameras used for tracking, but this can be costly.

The cameras must be located so that the effect of large distance errors on the quality of the system function is minimized. The following question should be answered. In which part of the path is accurate positioning the most important? During the first experiments, the cameras were located opposite the throwing device. In this situation, positioning in the first frames is the least accurate. It was possible to accurately position the ball from the 10th or 12th frame. In further experiments, the cameras were moved to the side of the throwing device. In this case, positioning in the initial part of the trajectory is quite accurate, but the final part of the trajectory is measured with higher errors. High accuracy in the initial part of the trajectory and lower accuracy in the final

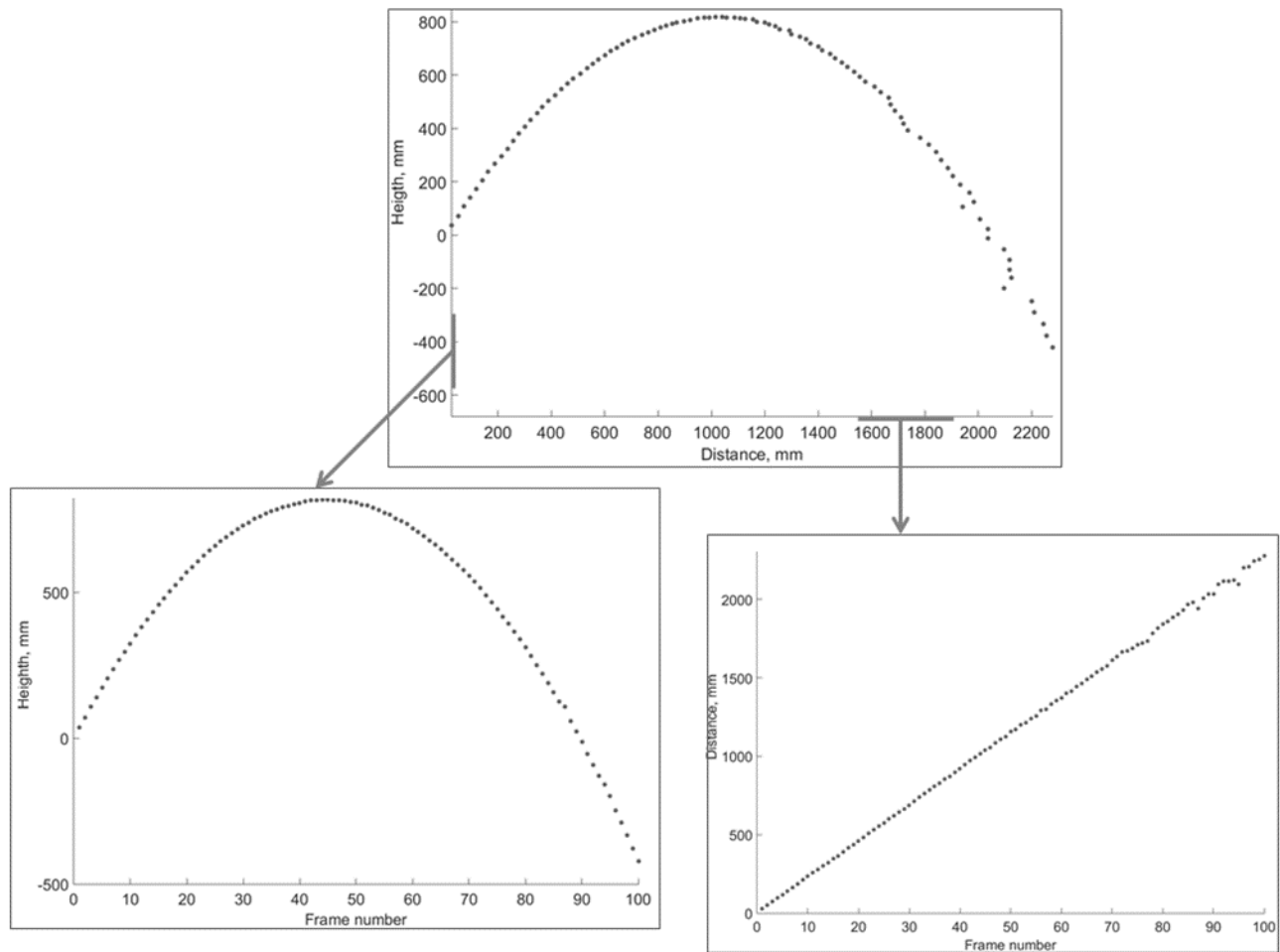


Figure 3. Dependence between height of the object and distance from the camera (upper graph), dependence of height on the frame number (bottom left) and dependence of the distance on the frame number (bottom right).

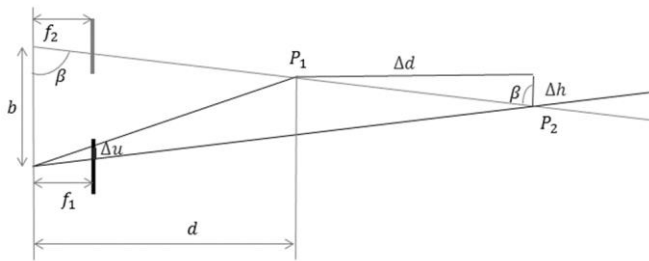


Figure 4. The effect of pixel error Δu on 3D positioning errors when measuring the height Δh and the distance Δd of an object.

stage is preferable than vice versa, due to the following factor. The final part of the trajectory is not processed in real time. Under actual transportation conditions, the ball will already be in the gripping workspace. Tracking the final part of the trajectory is used only for development of a trajectory training base; therefore, its accuracy can be improved by applying 1000 runs of RANSAC to the data. Accurate positioning is necessary to measure launch parameters: speed, throw angle, position in the first frame, etc. Another factor is ball capture; the measurement of the ball position in the final region will not be accurate in any case. The robot moving in the field of view generates excessive distortions in the functioning of the algorithm. Because of these factors, the location of the camera on the side of the thrower is more likely than vice

versa. It would also be possible to arrange the cameras in a different way: at a greater distance from each other or not parallel to the direction of the trajectory. However, as a result, the measurement error will not be localized in dimension coinciding with the direction of the object motion, as shown in Figure 3. In fact, this localization is very useful for correcting errors. In this measurement, the object moves at an almost constant speed, and the movement can be approximated by a second-order polynomial.

Since it is undesirable to use analytical models of object movement, approximation is applied only for the distance from the camera to the object and only at the final stage of the trajectory (starting from the 60th frame). The graph of the measured and approximated values of the distance to the object is shown in Figure 5. From a visual point of view, the results of the approximation look believable.

The stereo system used to track the trajectory of a thrown object measures its position in the coordinate system associated with the optical center of the left camera. In principle, trajectory prediction can be made in this form as well, but then the degree of trajectory proximities will be determined not only by the similarity of their shape, but also by the direction of the throw and the position of the point from which the throw is made. We have proposed to transform the coordinates of the thrown object into such a system where the trajectories can be compared and predicted based solely on the shape of the trajectory.

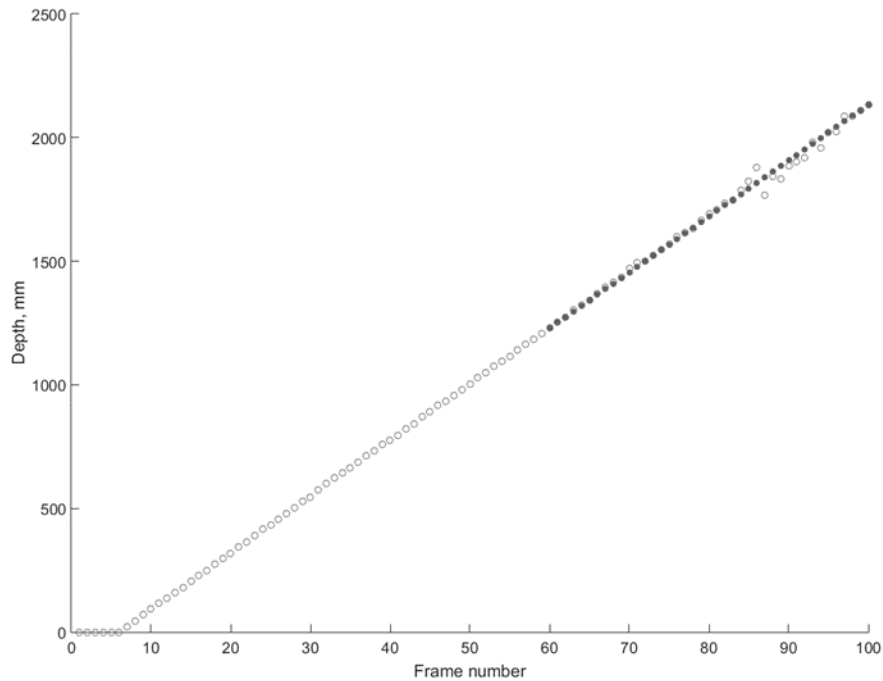


Figure 5. The difference between the measured (red circles) and approximated (blue points) values of the distance to the object

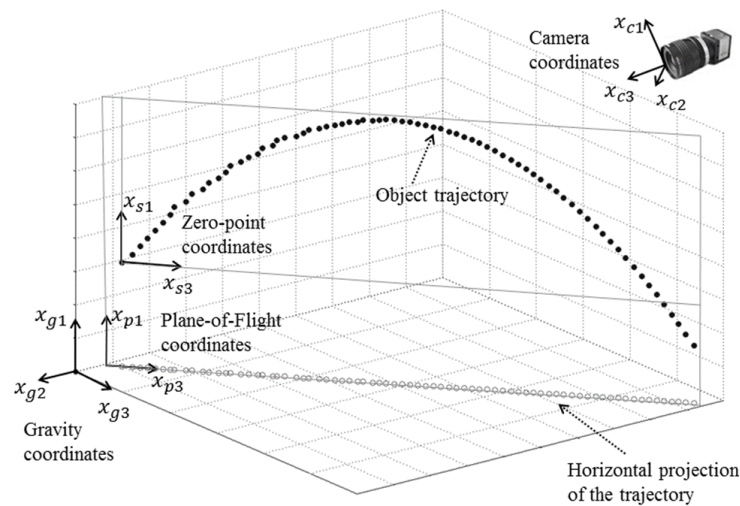


Figure 6. Mutual arrangement of coordinate systems

The purpose of coordinate transformations is to present the trajectories of thrown objects in a form in which it will be convenient to compare them. For this, the following sequence of coordinate transformations is proposed:- The three-dimensional system $x_{c1} O x_{c2} x_{c3}$, in which the point O coincides with the optical center of the left stereo pair camera, and the axis x_{c3} is aligned with the optical axis of the camera. In this coordinate system, the position of an object is measured by the stereo pair.

– The three-dimensional system $x_{g1} O x_{g2} x_{g3}$, in which the axis x_{g1} is aligned with the gravity vector, and the plane formed by the other two axes is, respectively, horizontal. This coordinate system allows you to localize the effect of gravity in one spatial dimension. As will be shown below, the transfer from such a system to the $x_{p1} O x_{p3}$ flight plane is simpler than from the $x_{c1} O x_{c2} x_{c3}$ system. The transition matrix from x_{c1}

$O x_{c2} x_{c3}$ to $x_{g1} O x_{g2} x_{g3}$ is determined during stereo system calibration. The gravity vector in the $x_{c1} O x_{c2} x_{c3}$ system can be determined by hanging the load on the thread: in equilibrium, the thread is parallel to the desired vector.

– The two-dimensional $x_{p1} O x_{p3}$ system (flight plane), in which the x_{p1} axis is aligned with the x_{g1} axis, and the horizontal projection of the object velocity lies on the x_{p3} axis. In the event that lateral forces do not act in flight on the body (they can be associated, for example, with the action of the wind or with the Magnus effect), the flight path lies in such a plane. Experiments conducted in [20] showed that the influence of lateral forces on the flight of an object can be neglected. Since the real direction of the throw in each case will be different from the others, then the transition matrix between the systems will be different for each case. The two-dimensional system $x_{s1} O x_{s3}$ in which the

directions of the axes coincide with those in x_{p1} Ox_{p3} , and the center is located on one of the first points of the trajectory. Transferring the trajectory to such a coordinate system ensures result independence from the spatial location of the point from which the throw was made. An analysis of the measurement accuracy carried out in [28] showed that the accuracy in the first few frames is slightly worse than in the subsequent frames. Therefore, the sixth point on the trajectory was chosen as the common center of coordinates.

The mutual arrangement of the coordinate systems is shown in Figure 6. At the stage of predictor learning, all trajectories in the database are converted into the x_{s1} Ox_{s3} system and saved in this form. In the process of predictor's work, the current XC trajectory is converted to the x_{s1} Ox_{s3} system; prediction is performed in this system, and then the result YC is converted back to the original coordinate system.

This was the procedure of extracting information about tracking the thrown tennis ball. Considering more complex-shaped objects will require use of more specific image processing algorithm. The procedure of stereo triangulation will be the same, while the question of how to define pixel coordinates of the object's center must be answered by other means instead of circle recognition. Various methods for object positioning task were developed such as rule-based algorithms, pixel-based classification, analysis of brightness distribution, convolutional neural networks, and other techniques. Development of image processing approach for complex-shaped objects is a subject of future work.

3. Predicting the trajectory of a thrown object

From the point of view of the subtask of trajectory prediction, the existing systems of robotic object capture on the fly can be divided into three groups:

- Accurate throw systems. The high accuracy of the throw (that is, the small deviation of the initial velocity and direction of flight from the given value) makes it possible to ensure that the trajectories of the thrown objects turn out to be almost identical. In this case, there is no need to predict the trajectory anew after each throw. It is enough to make a throw once, to track the trajectory of the object and to develop the trajectory of the capture device based on the results. This approach is applicable to objects with high aerodynamic stability (for example, cylindrical objects in [16]). If dropped objects do not have the required aerodynamic stability (studies in [22] and [16] show that even objects that are as simple in shape as a tennis ball and a hollow metal cylinder, respectively, do not possess it), this approach ceases to be useful.
- Interactive capture systems. In such systems, prediction is not used: movement of the capture device is determined by the current position of the object. For example, in [6], the movement of the working body is set in such a way as to maintain a constant value of the angle of view for an object in an image from a camera attached to a capture device. In [5], at each moment in time, the movement of the robot is set in the direction of the current position of the object. The implementation of such systems requires a high response speed of a robotic capture device and a high efficiency of obtaining information about object movement (for example, in [5] a vision system was used, in which the

frame rate was reduced to 1 kHz by directly connecting video matrix elements to the processor). The approach is not applicable if the throw is made from a long distance and you need to choose in which area of the working space to place the gripping device (just as the football player-goalskeeper first chooses which corner to defend and then catches the ball). Since transportation of objects in an industrial environment involves throwing over a distance of several meters, this method is not suitable for such systems.

- Systems with long-term forecasting. These include the ones described in [1-4,7-11] as well as the system discussed in this section. A more detailed overview of such systems is given below.

The majority of the ballistic trajectory predictors is based on the modeling of forces acting on the body. In the simplest case, it is assumed that the only such force is gravity (as if the body was moving in a vacuum). Such a model was considered in [2,3,7]. Prediction of the trajectory is carried out by approximation of the measured values in a parabola. In [10], the model was extended to predict the trajectory of asymmetric objects (in the experiments, empty and half-filled plastic bottles, hammers, tennis rackets and boxes were used). Prediction was made on the basis of the assumption that the acceleration vector of the object was constant over all six degrees of freedom. Strictly speaking, this assumption is wrong. The body movement under the action of gravity and air resistance is given by a differential equation which has no analytical solution and is solved in practice by numerical methods. This approach was used to predict the trajectory in a number of works [4,8,9]. Further complication of analytical models leads to a significant increase in the volume of computation [15].

On the other hand, people acquire the ability to catch thrown objects at an early age without any knowledge of aerodynamics. We catch a thrown ball based on our previous experience. Because of this, it was suggested [18,19] to use a trajectory predictor based on previous experience. In [18], a neural network predictor was proposed as a means of prediction, but it did not provide adequate forecast accuracy. Moreover, the results of prediction by neural networks are difficult to interpret; therefore, it was later proposed to apply a more transparent method of k nearest neighbors [19]. The development of individual details of this method is described in the articles of [20,21].

Here we propose the method, which lies between model-based and learning-based model. The predictor is using equations to define future positions of the object, but the parameters of these equations are obtained by the learning procedure of genetic programming. Learning does not require a sampling of past trajectories: the parameters of the model are learned from the initial part of the current trajectory. Genetic programming (proposed by Cramer [36] and developed by Koza [37]) is not a synonym of genetic algorithm. Genetic programming is an application of the principles of genetic algorithms to automatic generation of a program code. In many applications including this research genetic programming is used for generating equations, which represent the process with unknown parameters. Target program (equation) is defined as a tree consisting of nodes and arcs. The nodes are operations and the arcs are operands. Initial versions of the tree are modified via the genetic operations (mutation, crossover,

selection), which are similar to the respective processes in genetics [37]. For the forecasting of trajectory, the task is to define the function for calculating future values of the coordinates based on its previous known coordinates. Genetic operations aim to define recurrent equation for trajectory forecasting. Genetic Programming OLS MATLAB toolbox was used to execute the genetic operations.

Table 2. Results of numerical experiments for 20 trajectories

#	Equation	MSE, mm
1	$y(k-1) + (-0.038706) * (x(k-1)) + (0.033084)$	6
2	$y(k-1) + (-0.039329) * (x(k-1)) + (0.033553)$	6
3	$y(k-1) + (-0.038554) * (x(k-1)) + (0.031962)$	5
4	$y(k-1) + (-0.038689) * (x(k-1)) + (0.035645)$	2
5	$y(k-1) + (-0.038774) * (x(k-1)) + (0.033805)$	7
6	$y(k-1) + (-0.038546) * (x(k-1)) + (0.030245)$	7
7	$y(k-1) + (-0.038526) * (x(k-1)) + (0.032994)$	4
8	$y(k-1) + (-0.038740) * (x(k-1)) + (0.035160)$	6
9	$y(k-1) + (-0.038601) * (x(k-1)) + (0.033345)$	3
10	$y(k-1) + (-0.038454) * (x(k-1)) + (0.032473)$	4
11	$y(k-1) + (-0.038388) * (x(k-1)) + (0.031648)$	6
12	$y(k-1) + (-0.038501) * (x(k-1)) + (0.034544)$	6
13	$y(k-1) + (-0.038134) * (x(k-1)) + (0.033116)$	9
14	$y(k-1) + (-0.037313) * (x(k-1)) + (0.036494)$	8
15	$y(k-1) + (-0.038036) * (x(k-1)) + (0.032546)$	4
16	$y(k-1) + (-0.038347) * (x(k-1)) + (0.034728)$	4
17	$y(k-1) + (-0.038149) * (x(k-1)) + (0.034948)$	3
18	$y(k-1) + (-0.037972) * (x(k-1)) + (0.033884)$	4
19	$y(k-1) + (-0.037583) * (x(k-1)) + (0.032284)$	7
20	$y(k-1) + (-0.037841) * (x(k-1)) + (0.036221)$	7

Numerical experiment on trajectory prediction was conducted using the tool [38] in two stages. On the first stage, we tried to define the common trajectory equations by learning from various trajectories. This try failed: the trajectories are different from each other and equation may be good for one trajectory and useless for another. Therefore we changed our strategy. On the second stage of experiment the initial part of each trajectory (first 50 frames) was used for learning the recurrent formula of this trajectory. Then the accuracy of this formula was checked on frames from 60 to 80. Learning procedure was applied to 20 trajectories acquired during the throwing experiments. The results are presented in table 2.

Each row shows the results for one trajectory. First column is trajectory number. The second one show equations, defining the height of the object y on frame number k as a function from its

height y and distance x on previous frames. The right column show the standard deviation of the predicted ball position from the real one. It may be seen that the standard deviations of predicted height do not exceed 10 mm. According to the three-sigma rule the errors of prediction do not exceed 30 mm with high probability. All equations generated by the algorithm are relatively simple and have the same type: linear dependence from both coordinates of the previous frame. At the beginning of learning genetic operations often generate more complicated equations. These equations may include coordinate values from several previous frames.

4. Conclusion

We have proposed an algorithm for extraction and prediction of the ballistic trajectory based on video signal. Object trajectory is extracted from video sequence by the image processing algorithm, which include Canny edge detection, RANSAC circle recognition and stereo triangulation. The model of object motion is defined by the genetic programming. The result of the exploration is a recurrent formula for calculation of the object's position. Numerical experiments with real trajectories of the thrown tennis ball showed that the algorithm is able to forecast the trajectory accurately.

References

- [1] Gayanov, R., Mironov, K., Kurenov D.: Estimating the trajectory of a thrown object from video signal with use of genetic programming, 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Bilbao, Spain, pp. 134 to 138, December 2017.
- [2] Hove, B., Slotine, J.-J.: Experiments in Robotic Catching, American Control Conference, Boston, USA, pp. 381 to 386, 1991.
- [3] Nishiwaki, K., Konno, A., Nagashima, K., Inaba, M., Inoue, H.: The Humanoid Saika that Catches a Thrown Ball, IEEE International Workshop on Robot and Human Communication, Sendai, Japan, pp. 94 to 99, October 1997.
- [4] Frese, U., Baeuml, B., Haidacher, S., Schreiber, G., Schaefer, I., Haehnle, M., Hirzinger, G.: On-the-Shelf Vision for a Robotic Ball Catcher, IEEE/RSJ International Conference on Intelligent Robots and Systems, Maui, Hawaii, USA, pp. 591 to 596, November 2001.
- [5] Namiki, A., Ishikawa, M.: Robotic Catching Using a Direct Mapping from Visual Information to Motor Command, IEEE International Conference on Robotics & Automation, Taipei, Taiwan, pp. 2400 to 2405, September 2003.
- [6] Mori, R., Hashimoto, K., Miyazaki, F.: Tracking and Catching of 3D Flying Target based on GAG Strategy, IEEE International Conference on Robotics & Automation, New Orleans, USA, pp. 5189 to 5194, April 2004.
- [7] Herrejon, R., Kagami, S., Hashimoto, K.: Position Based Visual Servoing for Catching a 3-D Flying Object Using RLS Trajectory Estimation from a Monocular Image Sequence, IEEE International Conference on Robotics and Biomimetics, Guilin, China, pp. 665 to 670, December 2009.
- [8] Baeuml, B., Wimboeck, T., Hirzinger, G.: Kinematically Optimal Catching a Flying Ball with a Hand-Arm-System, IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, pp. 2592 to 2599, October 2010.
- [9] Baeuml, B., Birbach, O., Wimboeck, T., Frese, U., Dietrich, A., Hirzinger, G.: Catching Flying Balls with a Mobile Humanoid: System Overview and Design Considerations, IEEE-RAS International Conference on Humanoid Robots, Bled, Slovenia, pp. 513 to 520, October 2011.
- [10] Kim, S., Shukla, A., Billard, A.: Catching Objects in Flight, IEEE Transactions on Robotics, Vol. 30, No. 5, pp. 1049 to 1065, May 2014.
- [11] Cigliano P., Lippiello V., Ruggiero F., Siciliano B.: Robotic Ball Catching with an Eye-in-Hand Single-Camera System // IEEE Transactions on Control System Technology, 2015. Vol. 23. No. 5. P. 1657-1671.
- [12] Frank, H., Wellerdick-Wojtasik, N., Hagebeuker, B., Novak, G., Mahlknecht, S.: Throwing Objects: a bio-inspired Approach for the Transportation of Parts, IEEE International Conference on Robotics and Biomimetics, Kunming, China, pp. 91 to 96, December 2006.
- [13] Pongratz, M., Kupzog, F., Frank, H., Barteit, D.: Transport by Throwing - a bio-inspired Approach, IEEE International Conference on Industrial Informatics, Osaka, Japan, pp. 685 to 689, July 2010.
- [14] Barteit, D.: Tracking of Thrown Objects, Dissertation, Faculty of Electrical Engineering, Vienna University of Technology, December 2011.

- [15] Pongratz, M., Pollhammer, K., Szep, A.: KOROS Initiative: Automatized Throwing and Catching for Material Transportation, ISO/FA 2011 Workshops, pp. 136 to 143, 2012.
- [16] Frank, T., Janoske, U., Mittnacht, A., Schroedter, C.: Automated Throwing and Capturing of Cylinder-Shaped Objects, IEEE International Conference on Robotic and Automation, Saint Paul, Minnesota, USA, pp. 5264-5270, May 2012.
- [17] Pongratz, M., Mironov, K. V., Bauer F.: A soft-catching strategy for transport by throwing and catching, Vestnik UGATU, Vol. 17, No. 6(59), pp. 28 to 32, December 2013.
- [18] Mironov, K. V., Pongratz, M.: Applying neural networks for prediction of flying objects trajectory Vestnik UGATU, Vol. 17, No. 6(59) pp. 33 to 37, December 2013.
- [19] Mironov, K., Pongratz, M., Dietrich, D.: Predicting the Trajectory of a Flying Body Based on Weighted Nearest Neighbors, International Work-Conference on Time Series, Granada, Spain, pp. 699 to 710, June 2014.
- [20] Mironov, K., Vladimirova, I., Pongratz, M.: Processing and Forecasting the Trajectory of a Thrown Object Measured by the Stereo Vision System, IFAC International Conference on Modelling, Identification and Control in Non-Linear Systems, St.-Petersburg, Russian Federation, June 2015.
- [21] Mironov K. V., Pongratz M.: Fast kNN-based Prediction for the Trajectory of a Thrown Body, 24th Mediterranean Conference on Control and Automation MED 2016. Athens, Greece. pp. 512 to 517. July 2016.
- [22] Mehta, R., Alam, F., Subic, A.: Review of tennis ball aerodynamics, Sports technology review, John Wiley and Sons Asia Pte Ltd, 2008, No. 1, pp. 7 to 16, January 2008.
- [23] Szeliski, R.: Computer Vision: Algorithms and Applications, Springer, September 2010
- [24] Zhang, Z.: A flexible new technique for camera calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 11, pp. 1330 to 1334, November 2000
- [25] Camera Calibration Toolbox for Matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/, visited on September 18, 2018.
- [26] Pongratz, M., Mironov, K. V.: Accuracy of Positioning Spherical Objects with Stereo Camera System, IEEE International Conference on Industrial Technology, Seville, Spain, pp. 1608 to 1612, March 2015.
- [27] Mironov K. V.: Transport by robotic throwing and catching: Accurate stereo tracking of the spherical object, 2017 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM).
- [28] Lee, J.H., Akiyama, T., Hashimoto, H.: Study on Optimal Camera Arrangement for Positioning People in Intelligent Space, IEEE/RSJ international conference on intelligent robots and systems, Lausanne, Switzerland, pp. 220 to 225, October 2002.
- [29] UI-3370CP – USB 3 Cameras – CAMERAFINDER – Products, <http://en.ids-imaging.com/store/ui-3370cp.html>, visited on September 25, 2014.
- [30] Goetzinger, M.: Object Detection and Flightpath Prediction, Diploma Thesis, Faculty of Electrical Engineering, Vienna University of Technology, June 2015.
- [31] Canny, J.: A Computational Approach to Edge Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 6, November 1986.
- [32] Hough, P.: A method and means for recognizing complex patterns, U.S. Patent No. 3,069,654, December 1962.
- [33] Scaramuzza, D., Pagnotelli, S., Valligi, P.: Ball Detection and Predictive Ball Following Based on a Stereo-scopic Vision System, IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp. 1573-1578, April 2005.
- [34] Fischler, M. A., Bolles, R. C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, Communications of the Association for Computing Machinery, Vol. 24, No. 6, pp. 381 to 395, June 1981.
- [35] Hlawatsch, F.: Parameter Estimation Methods: Lecture Notes, Grafisches Zentrum HTU GmbH, Vienna, Austria, March 2012.
- [36] Cramer, N. A.: A Representation for the Adaptive Generation of Simple Sequential Programs, International Conference on Genetic Algorithms and their Applications, Pittsburgh, USA, pp. 183 to 187, July 1985.
- [37] Koza, J.R.: Genetic Programming: A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems, Stanford University Computer Science Department technical report STAN-CS-90-1314, June 1990.
- [38] GP-OLS MATLAB Toolbox [WebSite]. – http://www.abonyilab.com/software-and-data/gp_index/gpols, visited on 04.09.2017.

Performance Investigation of Semiconductor Devices using Commutation-speed based methodology for the application of Boost Power Factor Correction

Barkha Parkash^{*1,2}, Ajay Poonjal Pai¹, Wei Tian², Ralph Kennel²

¹Infineon Technologies AG, Automotive High Power, 85579, Germany,

²Technical University of Munich, Chair of Electrical Drive Systems and Power Electronics, 80333, Germany

ARTICLE INFO

Article history:

Received: 06 December, 2018

Accepted: 03 February, 2019

Online : 20 February, 2019

Keywords:

Commutation-speed

Power losses

Switching losses

Analytical model

ABSTRACT

In this paper, behavioral approach has been adopted for the calculation of total power losses that has been further used to derive an analytical model for the conduction and switching losses in a boost Power Factor Correction (PFC) stage of an On-board Charger (OBC). Detailed investigation of power losses can help in finding out ways to improve efficiency and for this purpose, commutation-speed based methodology has been used to split total power losses into their root causes. This gives opportunity to find the impact that an individual part creates on total losses which can serve as a starting point for efficiency improvement. For the analysis, two devices (IGBT with Si diode and IGBT with SiC diode) are used in the considered topology of PFC and a reduction of 40% was calculated when SiC diode was used instead of Si with the same IGBT. Hence it was found that the implemented method proves to be significantly useful in the optimization of efficiency.

1. Introduction

In electric vehicles (xEVs), it is desirable for the on-board charger to operate under maximum efficiency. Therefore, it is of prime importance to accurately calculate the power losses at the system level. A significant part of the power losses arise from the semiconductor devices, being used. A number of different methods are available for the calculation of power losses in semiconductor devices. One of the methods available is physics-based, which requires precise models of the power semiconductor devices and circuits under consideration are simulated numerically with the help of specially designed programs [1]. However in this method, the simulation time tend to be very long which doesn't seem to be a practical option for system simulations.. Another widely used approach which has been adopted in this paper, is the behavioral modelling of the power losses in which behaviour of the semiconductor device is captured under different operating conditions which is further used to develop simple equations to model the losses. This approach has been proven to give accurate results as discussed in [2,3]. It not only supports simplified calculation but also allows the application of commutation-speed based method in which all devices are switching at the same speed. This methodology makes it possible to split total power losses into their root causes, e.g. capacitive effect, tail current, reverse

recovery, forward conduction, reverse conduction [4]. The accurate split-up of power losses give the opportunity to assess and optimize the switching behaviour of power semiconductors at the converter level.

2. Loss Modelling

Power loss modelling involves creation of a generic model that helps in the accurate calculation of total power losses. The main losses associated with a semiconductor device are conduction, switching and gate driving losses but gate driving losses are not discussed in this paper. In general, the loss modelling is done using behavioural model of the device. For the behavioural part, several static and dynamic characterization measurements are conducted that gives insight into the behaviour of the device under different operating conditions. Characterization measurements are performed on the double pulse setup as discussed in [5]. The used setup is shown in Figure 1. For the characterization, one parameter is varied at a time while all other parameters are kept constant at their nominal values. For a higher degree of accuracy, the power loss equations are derived using a model that calculates losses as a function of device current I_{cf} , DC link voltage V_{dc} , junction temperature T_j , gate resistance R_g , area of device A_{di} and gate voltage V_{ge} as elaborated in [2]. The derived equations are used for the calculation of power losses in a boost Power Factor Correction (PFC) stage of an On-board Charger (OBC).

*Corresponding Author: Barkha Parkash, barkhaparkash2@gmail.com

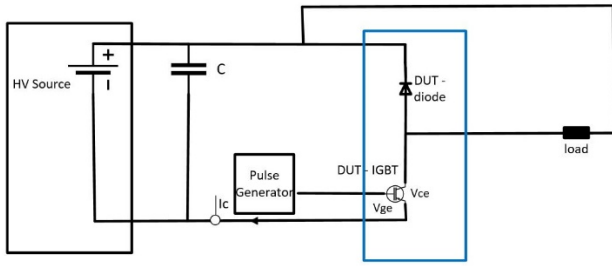


Figure 1 Double pulse setup for device characterization

2.1. Conduction Losses

Conduction losses are determined using instantaneous current and corresponding forward voltage. The simplest and most widely used way to model forward voltage is first order linear approximation consisting of threshold voltage and the drop across series resistance [6]. However, this model is accurate at the nominal operating point whereas the accuracy reduces at all other non-nominal points. A better approach to model forward voltage, which is also used in this work, is to consider the dependency between the considered parameters and voltage as quadratic functions instead of linear. The main advantage of using this approach is its accuracy at non-nominal points as well.

Mathematically, it can be written as [2];

$$V_{ce} = f(J_c) \cdot f(T_j) \cdot f(V_{ge}) \quad (1)$$

Where,

V_{ce} = Collector to emitter voltage in the case of an IGBT

J_c = Current density in the chip

$$f(J_c) = A_{11} \cdot J_c^2 + A_{12} \cdot J_c + A_{13} \quad (2)$$

$$f(T_j) = A_{21} \cdot T_j^2 + A_{22} \cdot T_j + A_{23} \quad (3)$$

$$f(V_{ge}) = A_{31} \cdot V_{ge}^2 + A_{32} \cdot V_{ge} + A_{33} \quad (4)$$

Similarly, the forward voltage of diode V_f is modelled as;

$$V_f = f(J_d) \cdot f(T_j) \quad (5)$$

Where,

J_d = Current density of the diode

2.2. Switching Losses

The most accurate way to calculate switching losses is to integrate the area where both voltage and current are overlapping in transition region, but this requires detailed knowledge of transient curves, which isn't readily available in most cases. A work around of this approach is to use the commutation time in which device is turned on or off and to take the corresponding voltage and currents. This approach is comparatively easier because this information is given in the datasheet [2]. However, these models are fitted at the nominal point and there is increasing error at all other points. That's why, this paper proposes to use the model presented in [2] where the switching energies as a product of quadratic functions of the considered parameters.

$$E_{on} = f(J_c) \cdot f(T_j) \cdot f(V_{dc}) \cdot f(R_g) \cdot f(A_1) \cdot f(A_D) \quad (6)$$

Where,

$$f(J_c) = C_{11} \cdot J_c^2 + C_{12} \cdot J_c + C_{13} \quad (7)$$

$$f(T_j) = C_{21} \cdot T_j^2 + C_{22} \cdot T_j + C_{23} \quad (8)$$

$$f(V_{dc}) = C_{31} \cdot V_{dc}^2 + C_{32} \cdot V_{dc} + C_{33} \quad (9)$$

$$f(R_g) = C_{41} \cdot R_g^2 + C_{42} \cdot R_g + C_{43} \quad (10)$$

$$f(A_1) = C_{51} \cdot A_1^2 + C_{52} \cdot A_1 + C_{53} \quad (11)$$

$$f(A_D) = C_{61} \cdot A_D^2 + C_{62} \cdot A_D + C_{63} \quad (12)$$

Similarly, turn off and reverse recovery energy are given as following;

$$E_{off} = f(J_c) \cdot f(T_j) \cdot f(V_{dc}) \cdot f(R_g) \cdot f(A_1) \quad (13)$$

$$E_{rec} = f(J_r) \cdot f(T_j) \cdot f(V_{dc}) \cdot f(A_D) \quad (14)$$

2.3. Extraction of polynomials for the behavioral model of device

These coefficients ($A_{11}, A_{12}, A_{13} \dots C_{63}$) are determined by fitting a quadratic curve to the characterization measurements. They represent the behaviour of the device and are used to determine the switching and conduction losses at any operating point.

3. Analytical Model for the losses in Boost PFC

Figure 2 shows boost PFC, which is a commonly used PFC topology in an on-board charger. In a boost PFC, there are two sub-sections; rectifier followed by the boost stage that is controlled to ensure unity power factor. Based on the discussion in section 2, an analytical model for the switching and conduction losses of the switching devices in a boost PFC are derived.

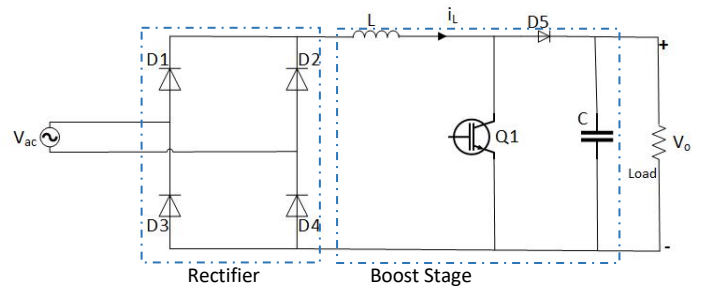


Figure 2 Boost PFC stage of an On-board Charger

The inductor current is sinusoidal; therefore, it can be written as a function of time.

$$i_L = I_{pk} \cdot \sin \alpha \quad (15)$$

Where I_{pk} is the peak value of current and α is given as;

$$\alpha = \omega \cdot t \quad (16)$$

The duty cycle for the switch $\delta(t)$ in boost PFC is given by;

$$\delta(t) = 1 - \frac{V_{pk} \sin \alpha}{V_o} \quad (17)$$

V_{pk} is the peak value of input voltage and V_o is the output voltage which is equal to the V_{dc} . Whereas the duty cycle for the boost diode $\delta_{diode}(t)$ is given by;

$$\delta_{\text{diode}}(t) = \frac{V_{\text{pk}} \sin \alpha}{V_o} \quad (18)$$

So, the instantaneous switch i_Q and diode current i_{diode} are given as;

$$i_Q = I_{\text{pk}} \cdot \sin \alpha \cdot \left\{ 1 - \left(\frac{V_{\text{pk}} \sin \alpha}{V_o} \right) \right\} \quad (19)$$

$$i_{\text{diode}} = I_{\text{pk}} \cdot \sin \alpha \cdot \left(\frac{V_{\text{pk}} \sin \alpha}{V_o} \right) \quad (20)$$

The instantaneous conduction losses for the switch and diode in boost PFC are as follows;

$$P_{\text{scond}} = V_{\text{ce}} \cdot i_Q \quad (21)$$

$$P_{\text{sdiode}} = V_f \cdot i_{\text{diode}} \quad (22)$$

where,

$$V_{\text{ce}} = (A_{11} \cdot J_c^2 + A_{12} \cdot J_c + A_{13}) \cdot (A_{21} \cdot T_j^2 + A_{22} \cdot T_j + A_{23}) \cdot (A_{31} \cdot V_{\text{ge}}^2 + A_{32} \cdot V_{\text{ge}} + A_{33}) \quad (23)$$

$$V_f = (B_{11} \cdot J_f^2 + B_{12} \cdot J_f + B_{13}) \cdot (B_{21} \cdot T_d^2 + B_{22} \cdot T_d + B_{23}) \quad (24)$$

The average conduction loss in the switch P_{scond} is calculated as given in paper [7]. Integration is performed over half the fundamental cycle and due to symmetrical nature, the integration result is directly multiplied with 2 and then averaged over the fundamental cycle. Mathematically, it is given as;

$$P_{\text{scond}} = \frac{2}{T} \int_0^{\frac{T}{2}} P_{\text{scond}} dt \quad (25)$$

Where, T is the fundamental period. On substitution, following expression is obtained.

$$P_{\text{scond}} = \frac{2}{2\pi} \left\{ \int_0^{\pi} ((A_{11} \cdot J_c^2 + A_{12} \cdot J_c + A_{13}) \cdot (A_{21} \cdot T_j^2 + A_{22} \cdot T_j + A_{23}) \cdot (A_{31} \cdot V_{\text{ge}}^2 + A_{32} \cdot V_{\text{ge}} + A_{33})) \cdot I_{\text{pk}} \cdot \sin \alpha \cdot \left(1 - \frac{V_{\text{pk}} \sin \alpha}{V_o} \right) d\alpha \right\} \quad (26)$$

Following result is obtained on performing integration.

$$P_{\text{scond}} = \frac{K_c \cdot J_{\text{cpk}} \cdot A_i}{\pi} \left[\frac{4 \cdot A_{11} \cdot J_{\text{cpk}}^2}{3} + \frac{\pi \cdot A_{12} \cdot J_{\text{cpk}}}{2} + 2 \cdot A_{13} - \frac{3 \cdot \pi \cdot A_{11} \cdot J_{\text{cpk}}^2 \cdot V_{\text{pk}}}{8 \cdot V_o} - \frac{4 \cdot A_{12} \cdot J_{\text{cpk}} \cdot V_{\text{pk}}}{3 \cdot V_o} - \frac{A_{13} \cdot V_{\text{pk}} \cdot \pi}{2 \cdot V_o} \right] \quad (27)$$

Where, K_c is a constant which is given by;

$$K_c = (A_{21} \cdot T_j^2 + A_{22} \cdot T_j + A_{23}) \cdot (A_{31} \cdot V_{\text{ge}}^2 + A_{32} \cdot V_{\text{ge}} + A_{33}) \quad (28)$$

Similarly, the diode conduction losses are derived as;

$$P_{\text{Dcond}} = \frac{K_d \cdot J_{\text{dpk}} \cdot A_d}{\pi} \left[\frac{3 \cdot \pi \cdot B_{11} \cdot J_{\text{dpk}}^2 \cdot V_{\text{pk}}}{8 \cdot V_o} + \frac{4 \cdot B_{12} \cdot J_{\text{dpk}} \cdot V_{\text{pk}}}{3 \cdot V_o} + \frac{B_{13} \cdot V_{\text{pk}} \cdot \pi}{2 \cdot V_o} \right] \quad (29)$$

Where, constant K_d is given as;

$$K_d = (B_{21} \cdot T_d^2 + B_{22} \cdot T_d + B_{23}) \quad (30)$$

The average switching losses is the sum of turn on, turn off and recovery loss. Integrating switching energies as given by (5), (12) and (13); following results are obtained.

$$P_{\text{on}} = f_{\text{sw}} \left[\left(\frac{C_{11} \cdot K_{\text{on}}}{2} \right) \cdot J_{\text{cpk}}^2 + (C_{12} \cdot K_{\text{on}}) \cdot \frac{2 \cdot J_{\text{cpk}}}{\pi} + (C_{13} \cdot K_{\text{on}}) \right] \quad (31)$$

$$P_{\text{off}} = f_{\text{sw}} \left[\left(\frac{D_{11} \cdot K_{\text{off}}}{2} \right) \cdot J_{\text{cpk}}^2 + (D_{12} \cdot K_{\text{off}}) \cdot \frac{2 \cdot J_{\text{cpk}}}{\pi} + (D_{13} \cdot K_{\text{off}}) \right] \quad (32)$$

$$P_{\text{rec}} = f_{\text{sw}} \cdot K_{\text{rec}} \left[\frac{E_{11} \cdot J_{\text{dpk}}^2}{2} + \frac{2 \cdot E_{12} \cdot J_{\text{dpk}}}{\pi} + E_{13} \right] \quad (33)$$

Where, f_{sw} is the switching frequency and constants K_{on} , K_{off} and K_{rec} are given as;

$$K_{\text{on}} = (C_{21} \cdot T_j^2 + C_{22} \cdot T_j + C_{23}) \cdot (C_{31} \cdot V_{\text{dc}}^2 + C_{32} \cdot V_{\text{dc}} + C_{33}) \cdot (C_{41} R_g^2 + C_{42} R_g + C_{43}) \cdot (C_{51} \cdot A_1^2 + C_{52} \cdot A_1 + C_{53}) \cdot (C_{61} \cdot A_D^2 + C_{62} \cdot A_D + C_{63}) \quad (34)$$

$$K_{\text{off}} = (D_{21} \cdot T_j^2 + D_{22} \cdot T_j + D_{23}) \cdot (D_{31} \cdot V_{\text{dc}}^2 + D_{32} \cdot V_{\text{dc}} + D_{33}) \cdot (D_{41} R_g^2 + D_{42} R_g + D_{43}) \cdot (D_{51} \cdot A_1^2 + D_{52} \cdot A_1 + D_{53}) \quad (35)$$

$$K_{\text{rec}} = (E_{21} \cdot T_j^2 + E_{22} \cdot T_j + E_{23}) \cdot (E_{31} \cdot V_{\text{dc}}^2 + E_{32} \cdot V_{\text{dc}} + E_{33}) \cdot (E_{41} \cdot A_D^2 + E_{42} \cdot A_D + E_{43}) \quad (36)$$

4. Splitting of Switching Losses

Experimental measurements give a complete picture of the total losses in the device however it doesn't explain the contribution of various factors that make up the total losses. To get a deeper insight into the switching losses, the switching events (turn on and turn off) are divided into various regions as explained in [4].

4.1. Turn On

The turning on of the device is divided into two regions as shown in Figure 3; commutation region where the actual current commutation occurs from the diode to the active switch and the reverse recovery region which is influenced by the type of freewheeling diode (FWD), being used.

4.1.1 Region A

The split-up can be achieved by utilizing the device current and voltage waveforms obtained by measurements. Considering device current and voltage shown in Figure 3, the commutation region starts when the gate voltage goes higher than its threshold and the collector current just starts increasing which is marked as t_o . This region continues till the collector current has reached its steady state value, which is equal to the load current and this time instant is indicated in the Figure as t_A . The area from time t_o to t_A is taken as Region A. During the current rise, the collector to emitter voltage drops and this voltage drop is due to the stray inductance of the commutation path, L_σ and the rate of change of current $\frac{di_{\text{ce}}}{dt}$ as described in [8];

$$V_{\text{ce}} = V_{\text{dc}} - L_\sigma \cdot \frac{di_{\text{ce}}}{dt} \quad (36)$$

4.1.2 Region B

The second region, reverse recovery part starts when the current through free-wheeling diode (FWD) reaches zero but the diode cannot take the entire voltage yet because of the presence of plasma [9]. Therefore, the current through the active switch continues to increase. The region B constitutes the time when collector current continues to rise due to plasma of FWD till the complete removal of excess carriers and falls subsequently to the steady state value i.e. at time instant t_B . The maximum value of collector current depends on the design of diode, gate driver that influences the rate of change of current, the junction temperature and the dc bus voltage [8]. The area from time t_A to t_B is the Region B.

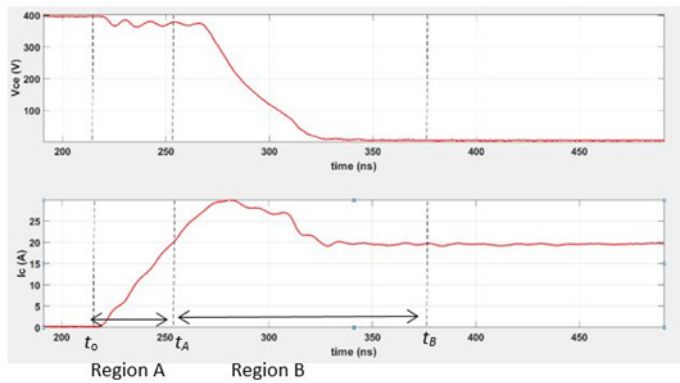


Figure 3 Split-up of the turn on event

4.2. Turn Off

The turn off event shown in Figure 4 is divided into three regions for an IGBT; capacitive region, commutation region and tail current and the speed at which turn off occurs depends on the gate resistance.

4.2.1 Region A

For the IGBT to turn off, it is important that diode takes up the load current which is not possible till voltage across diode has reduced to zero i.e. its junction capacitance needs to be discharged. The current that is used for the discharging of junction capacitance of the diode causes a gradual reduction in active switch's current i.e. collector current in the case of IGBT. This forms the capacitive region which can be seen in the Figure where current starts falling at time t_0 and the active switch takes up the full voltage at time t_A .

4.2.2 Region B

After discharging of junction capacitance, the collector current starts falling rapidly at time t_A and that's when the actual commutation takes place from the active switch to the diode.

4.2.3 Region C

The bipolar devices like IGBT do not turn off till the minority charge carriers have been removed completely. The sweeping of minority carriers from the device takes some time resulting in tail current which starts at t_B and continues till current reaches zero, marked in the Figure as t_C . The tail current doesn't exist in the case of MOSFET because it is a unipolar device.

The different regions of turn on, turn off and diode switching losses add up to form total switching losses of the switch. If the

contribution of each part on an individual level is known then it can help in the assessment of the impact that they create. This information can be very beneficial in optimizing the efficiency.

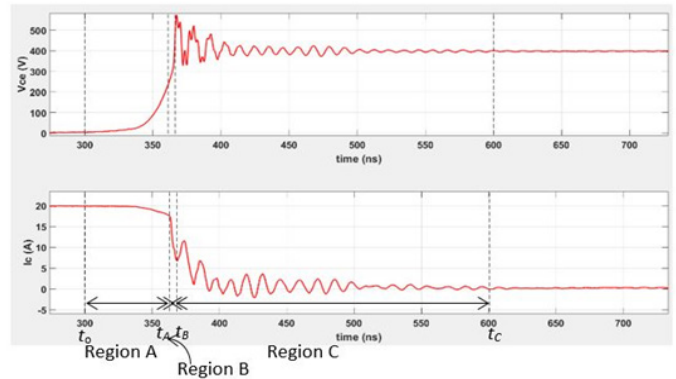


Figure 4 Split-up of the turn off event

5. Compared Devices

To understand the essence of this methodology, it has been implemented on two different discrete devices. One of the devices is Trenchtop-5 IGBT with Rapid Si diode from Infineon [10] and the second device is produced by replacing the Si diode of first device with SiC Schottky diode from Infineon whereas same IGBT is used for both the devices. In this paper, the first device will be referred to as "Full-Si IGBT" and the second device pair as "Hybrid-Si/SiC IGBT". The nominal values for both the devices are given in Table 1.

Table 1 Nominal values for both the considered devices

Nominal Values	I_c	T_j	V_{dc}	R_g
Full-Si IGBT	20 A	25°C	400 V	2.3 Ω
Hybrid-Si/SiC IGBT	20 A	25°C	400 V	14.6 Ω

6. Effect of Parameters on Split-wise Losses

All the parameters considered here; device current, junction temperature, gate resistance, DC link voltage, area of the device and gate voltage affect the switching losses. To get a better understanding as how each parameter effects the power losses in the devices, these devices are characterized using double pulse setup in which each of these parameter is varied at a time and corresponding total turn on and turn off energies are calculated. The data from characterization measurement is simulated on a specially written program that splits the total turn on and turn off energies into their respective parts as discussed in section 4. The split-up is done for both the considered devices. The results of the simulation are discussed as below where Full-Si IGBT is taken as the reference and the trends where Hybrid-Si/SiC IGBT differs are also explained. The comparison to study the effect of different parameters on switching energies isn't one to one because the nominal gate resistance is different for the devices.

6.1. Turn On

6.1.1 Effect of Device Current

It is observed that with increasing device currents, the total power losses due to turn on event increase. The trend of turn on energy as a function of device current for Full-Si IGBT and

Hybrid-Si/SiC IGBT are shown in Figure 5 and 6 respectively. It can be seen in these figures that the losses due to region A and region B are increasing with increasing current levels. As the value of device current is increased then the rate of change of current $\frac{di_{ce}}{dt}$ decreases i.e. current rises slowly as shown in Figure 7. For a current increase from 5A to 20A, the corresponding rate of change of current is from 0.65 kA/ μ s to 0.582 kA/ μ s. This indicates that the losses due to commutation (region A) increase with increasing current. The reason for the increase in losses is due to the direct relation between the value of current and the time taken by the device to reach the nominal value. For higher values of current, device takes more time to rise to that value.

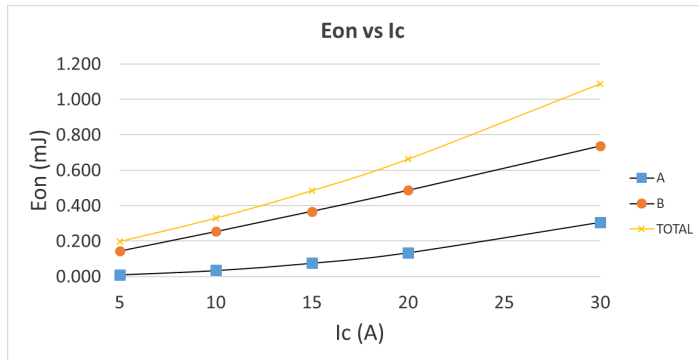


Figure 5 Turn on energy arising from its different root causes as a function of device current (Full-Si IGBT) at $V_{dc} = 400$ V, $T_j = 25^\circ\text{C}$ and $R_g = 14.6 \Omega$ A = commutation, B = reverse recovery

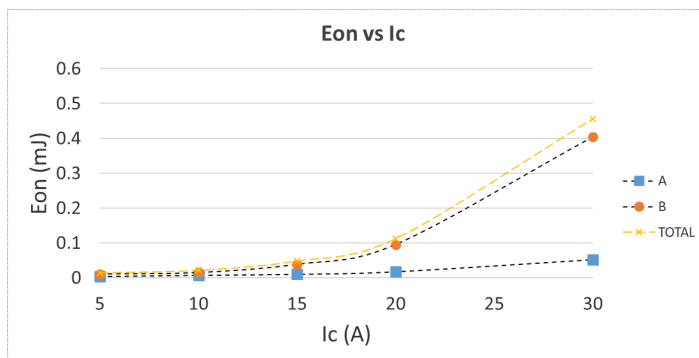


Figure 6 Turn on energy arising from its different root causes as a function of device current (Hybrid-Si/SiC IGBT) at $V_{dc} = 400$ V, $T_j = 25^\circ\text{C}$ and $R_g = 2.3 \Omega$ A = commutation, B = reverse recovery

Losses due to reverse recovery depends on the type of freewheeling diode. In the case of Si diode, which is a bipolar device, the losses due to this region are predominant because Si diode cannot support voltage till the removal of plasma from its PN junction. Therefore, the voltage at IGBT stays higher and allows collector current to rise beyond its nominal point which initiates the sweeping out of plasma from the junction [8]. Therefore, the losses due to reverse recovery (region B) increases with increase in current. Higher current means more number of charge carriers so the time taken by the device to sweep out the excess carriers from PN junction of freewheeling diode is greater as well which leads to increase in losses. On the other hand in the case of unipolar diode like Schottky SiC diode, the losses due to this region are very low because plasma doesn't exist here and the losses are due to the junction capacitance only.

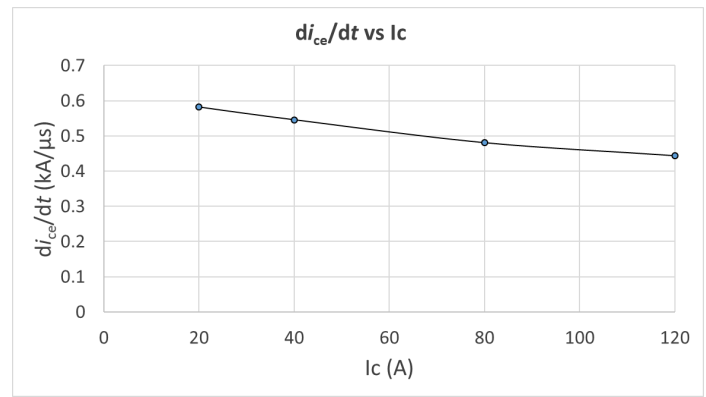


Figure 7 Rate of change of current as function of current (Full-Si IGBT) at $V_{dc} = 400$ V, $T_j = 25^\circ\text{C}$ and $R_g = 14.6 \Omega$

It is to note here that in turn on losses, major contribution of losses is due to region B which corresponds to reverse recovery for Full-Si IGBT and charging of junction capacitance in the case of Hybrid-Si/SiC IGBT.

6.1.2 Effect of Junction Temperature

With increase in temperature, the threshold voltage at which the IGBT turns on decreases, which leads to faster commutation compared to lower temperatures. However, this has a small effect on the power losses. Increment in temperature leads to a slight decrement in power losses due to region A.

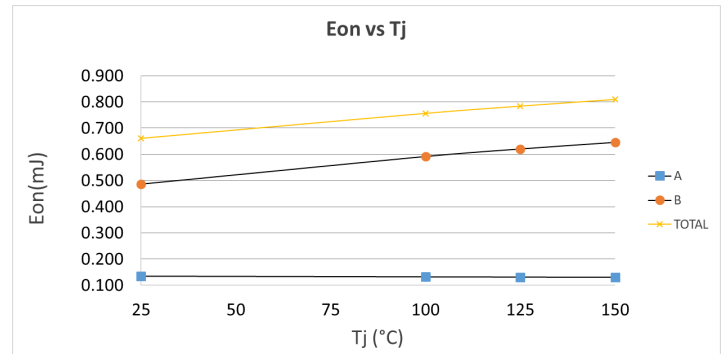


Figure 8 Turn on energy arising from its different root causes as a function of junction temperature (Full-Si IGBT) at $V_{dc} = 400$ V, $I_c = 20$ A and $R_g = 14.6 \Omega$ A = commutation, B = reverse recovery

As it can be seen in Figure 8, the losses due to region B increase with increase in temperature. This is because of the fact that the mobility of charge carriers decreases at higher temperature, which slows down the sweeping process of excess carriers from PN junction of diode, leading to increment in losses due to reverse recovery. So, it can be said that at higher temperature, major portion of turn on losses is due to the region B.

Figure 9 shows the dependency of temperature on turn on energy for Hybrid-Si/SiC IGBT. It is seen that the losses due to region A follow the same trend like Full-Si IGBT. This is because the region A is related with the rising current in the IGBT and same IGBT is being used for both the devices. However, the losses due to region B don't change with temperature. This is because in a SiC diode, the losses in this region do not come from reverse recovery but from the capacitive charge which is independent of the temperature.

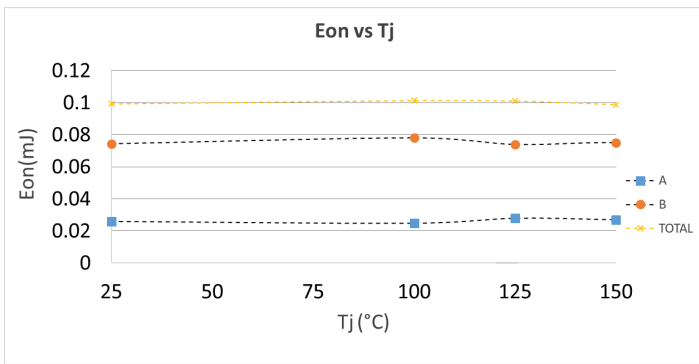


Figure 9 Turn on energy arising from its different root causes as a function of device current (Hybrid-Si/SiC IGBT) at $V_{dc} = 400$ V, $I_c = 20$ A and $R_g = 14.6$ Ω A = commutation, B = reverse recovery

6.1.3 Effect of DC link Voltage

During region A of turn on, the fall of collector to emitter voltage, V_{ce} is given by (36). The stray inductance L_σ is constant and $\frac{di_{ce}}{dt}$ doesn't change significantly at the nominal current and gate resistance. The only variable is V_{dc} which means that amplitude of V_{ce} increases with the increase in DC link voltage.

In general, the power losses are calculated as integration of the product of V_{ce} and i_{ce} . As V_{ce} is higher for higher values of V_{dc} so the power losses due to region A and B increase. However, the power losses due to region B increase by bigger margin compared to region A as shown in Figure 10. This happens because of the fact that the region B continues till the voltage has reduced to zero, so the time taken by the active switch for its voltage to drop to zero will be longer for higher values of V_{dc} .

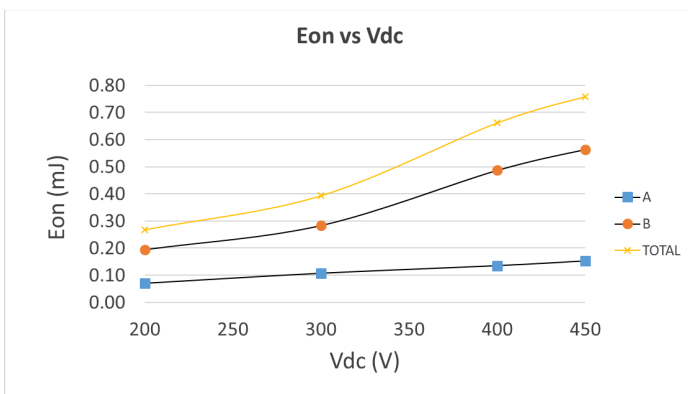


Figure 10 Turn on energy arising from its different root causes as a function of DC link voltage (Full-Si IGBT) at $T_j = 25^\circ\text{C}$, $I_c = 20$ A and $R_g = 14.6$ Ω A = commutation, B = reverse recovery

6.1.4 Effect of Gate Resistance

The charging of the gate capacitance determines the switching behavior of the semiconductor device and this is controlled by an external gate resistance. The gate resistance has influence on switching losses; therefore, optimal value of gate resistance should be selected according to the application [11].

A smaller gate resistance leads to shorter rise time which means device switches faster. So, a big gate resistance leads to increment in current rise time i.e. rate of change of current reduces

as indicated in Figure 11. This leads to increment in losses due to region A as shown in Figure 12.

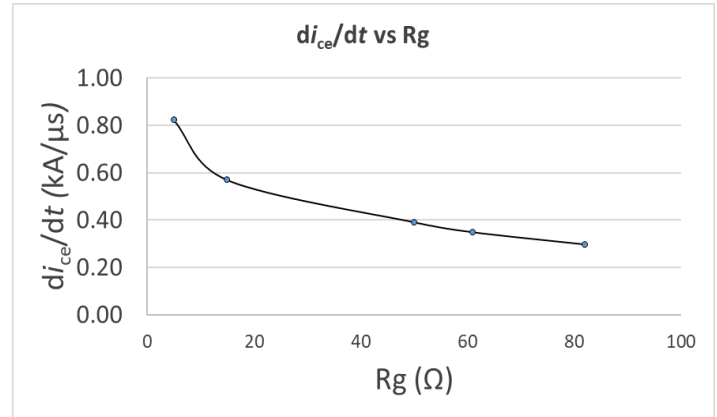


Figure 11 Variation in rate of change of current for different gate resistances at $T_j = 25^\circ\text{C}$, $I_c = 20$ A and $R_g = 14.6$ Ω

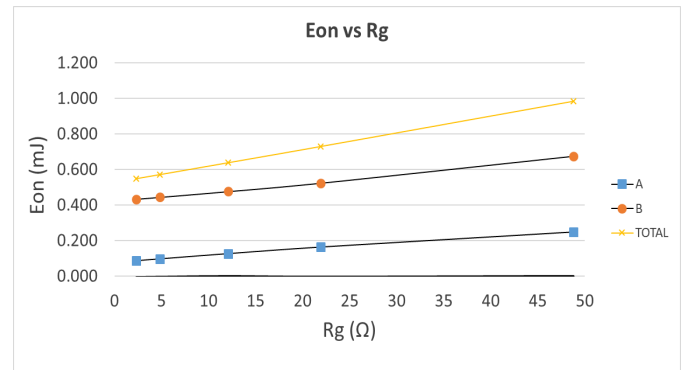


Figure 12 Turn on energy arising from its different root causes as a function of gate resistance (Full-Si IGBT) at $V_{dc} = 400$ V, $I_c = 20$ A and $T_j = 25^\circ\text{C}$ A = commutation, B = reverse recovery

With increasing gate resistance, there are two effects as shown in Figure 13; one is that the reverse recovery peak decreases and secondly $\frac{dV_{ce}}{dt}$ decreases which indicates that the time taken by the active switch for its voltage to drop to zero is longer compared to a smaller gate resistance. Out of these two effects, the predominant effect is the reduction in $\frac{dV_{ce}}{dt}$ which causes increment in the losses due to region B.

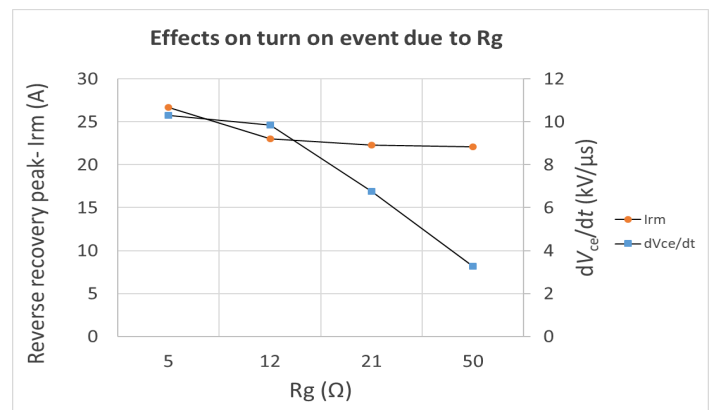


Figure 13 Effects on turn on event due to gate resistance (Full-Si IGBT) at $V_{dc} = 400$ V, $I_c = 20$ A and $T_j = 25^\circ\text{C}$

6.2. Turn Off

6.2.1 Effect of Device Current

With increasing current, the losses due to all regions increase shown in Figure 14. The increment in losses is simply because of the higher value of current. The first region of turn off event, capacitive region, is determined by the rate of change of voltage $\frac{dV_{ce}}{dt}$. If $\frac{dV_{ce}}{dt}$ is higher, it means that the junction capacitance of the freewheeling diode discharges at a faster rate and that the diode is able to take the load current faster.

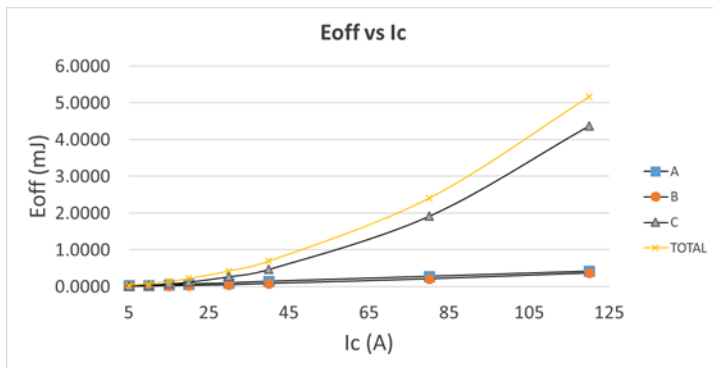


Figure 14 Turn off energy arising from its different root causes as a function of device current (Full-Si IGBT) at $V_{dc} = 400$ V, $R_g = 14.6 \Omega$ and $T_j = 25^\circ\text{C}$ A = capacitive, B = commutation, C = tail current

Here, with increasing current, there is no significant change in the rate of change of voltage as shown in Figure 15; therefore, the losses due to region A increases by small margin. This small increment is due to the beginning of integration from higher value of current.

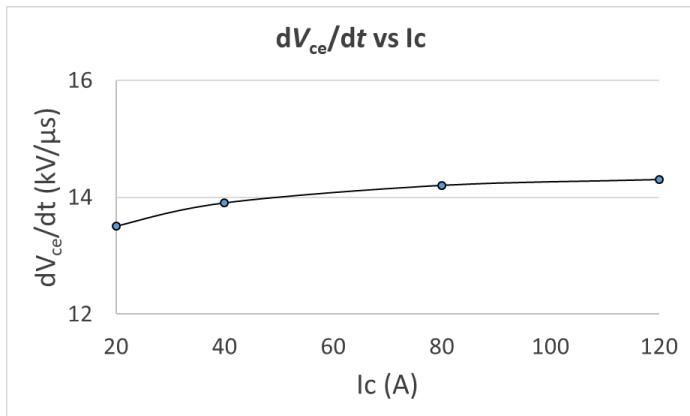


Figure 15 Variation in rate of change of voltage for different current levels (Full-Si IGBT)

It is clear from the Figure 14 that the turn off losses in an IGBT has a big part of losses coming from the tail current.

6.2.2 Effect of Junction Temperature

As temperature increases, the total turn off losses increases as shown in Figure 16. This happens because of the decrement in threshold voltage with increase in temperature so the time taken for the gate voltage to go lower than threshold voltage increases. This leads to slower switching speed; therefore, the power losses due to region B increases with increasing temperature.

The losses due to part C increase with increase in temperature as minority charge carriers take longer to recombine at higher temperatures.

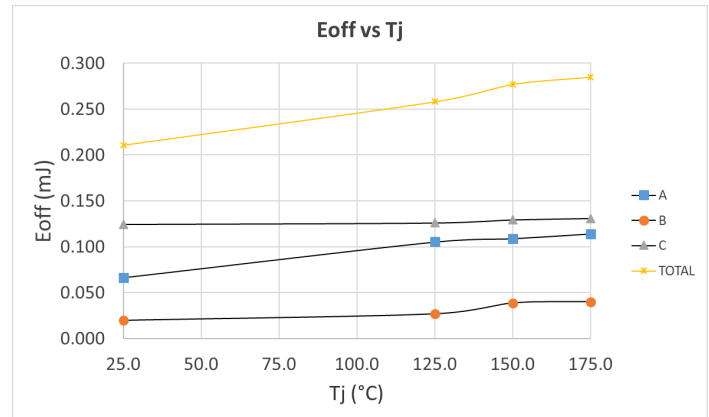


Figure 16 Turn off energy arising from its different root causes as a function junction temperature (Full-Si IGBT) at $V_{dc} = 400$ V, $R_g = 14.6 \Omega$ and $I_c = 20$ A A = capacitive, B = commutation, C = tail current

6.2.3 Effect of DC link Voltage

For higher DC link voltage, the turn off losses due to region A, B and C increase. This is because for higher V_{dc} , integration starts from higher value leading to increment in losses.

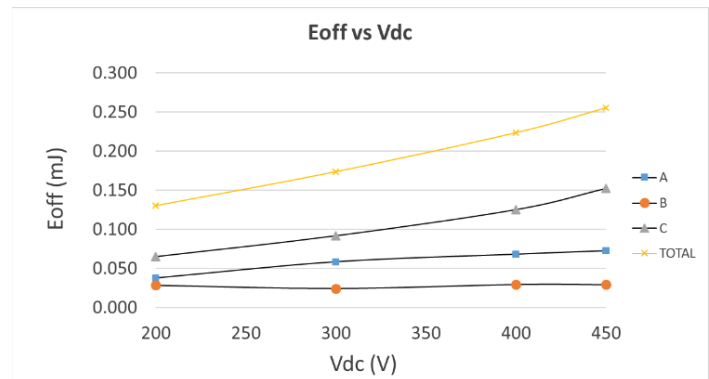


Figure 17 Turn off energy arising from its different root causes as a function DC link voltage (Full-Si IGBT) at $V_{dc} = 400$ V, $R_g = 14.6 \Omega$ and $I_c = 20$ A A = capacitive, B = commutation, C = tail current

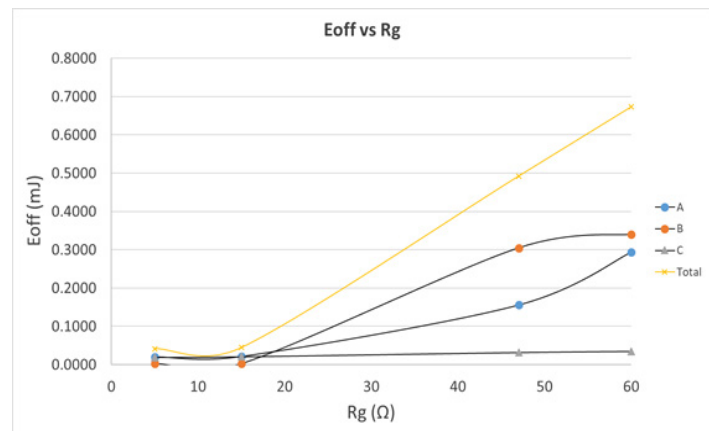


Figure 18 Turn off energy arising from its different root causes as a function gate resistance (Full-Si IGBT) at $V_{dc} = 400$ V, $T_j = 25^\circ\text{C}$ and $I_c = 20$ A A = capacitive, B = commutation, C = tail current

6.2.4 Effect of Gate Resistance

If higher value of gate resistance is chosen, then the rate at which device switches reduces. For higher gate resistance, $\frac{dV_{ce}}{dt}$ reduces and charging of junction capacitance is slower. This leads to increase in losses due to region A. Because of slower switching speed, the rate at which current falls down decreases leading to increment in losses due to region B. Since the tail current is specific to the IGBT; therefore, the losses due to part C do not change with the change in gate resistance [12].

7. Comparison of Switching Energies for the Considered Devices

Following the discussion of section 6, it is understood that all parameters affect both the devices in certain way. Figure 19 shows the turn on energies as a function of current at the same commutation speed with gate resistance of 5 Ω. At 30A, turn on energy reduces by 38% for Hybrid-Si/SiC IGBT. This is mainly due to the usage of unipolar Schottky SiC diode which doesn't have reverse recovery, leading to reduction in E_{on} by such a big margin.

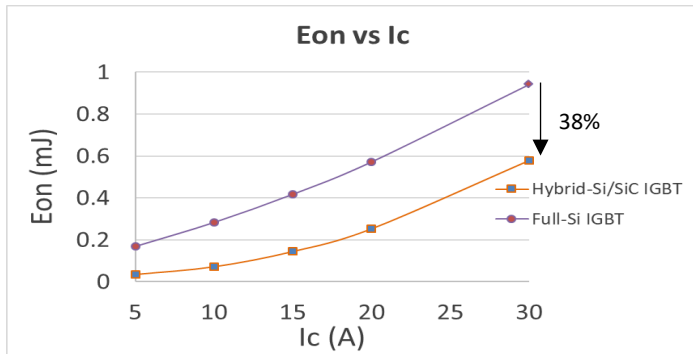


Figure 19 Turn on energy as a function of device current for Full-Si IGBT and Hybrid-Si/SiC IGBT at R_g = 5 Ω, V_{dc} = 400 V, T_j = 25°C

Figure 20 shows reverse recovery of the diode E_{rec} for different current levels. In the case of Full-Si IGBT, E_{rec} increases with increase in current. However, for Hybrid-Si/SiC IGBT, there is no significant change in E_{rec} , owing to the fact the losses for SiC diode do not originate from reverse recovery but due to the removal of charge carriers from junction capacitance.

8. Application Example

The commutation-speed based method can be applied to any application for the determination of the split-up of power losses at the system level. If the source contributing the most to the total losses is known, then steps can be taken in that direction to improve the efficiency. This method can help in the selection of most suitable device for the specific application. The essence of this method is demonstrated by applying it to previously discussed Boost PFC topology. The specifications of the implemented boost PFC are given in Table 2.

The power losses in this application are due to the rectifier stage, power switches used in boost stage and the losses due to passives. The losses in the power switches are modelled using the behavioral approach as discussed in section 2. Firstly, boost PFC with Full-Si IGBT is simulated in MATLAB where switching based simulation is performed i.e. losses at every switching instant

is calculated, summed up and then averaged. This indicates that the simulation follows the same approach for loss calculation as discussed in section 3.

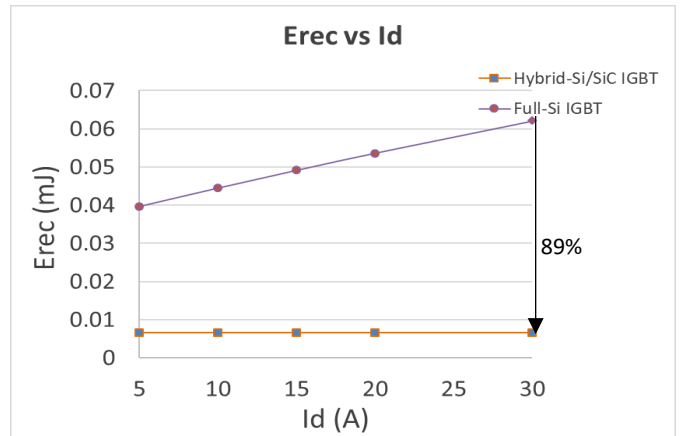


Figure 20 Reverse recovery energy as a function of device current for Full-Si IGBT and Hybrid-Si/SiC IGBT at R_g = 5 Ω, V_{dc} = 400 V, T_j = 25°C

Table 2 Parameters of the simulated boost PFC topology

Parameters	Values
V _{ac}	230V
P _o	3.3kW
V _{dc}	400V
L	263.8uH
C _o	1300uF
f _{sw}	80kHz
R _g	12Ω

Figure 21 shows the contribution of different sources of power losses in the boost PFC which are obtained from the simulation. It is seen in the Figure that there are significant losses due to the rectifier and passives in the boost PFC. However, to reduce losses of rectifier, the appropriate option is to use a different topology having less number of diodes in the rectifier stage, which will reduce the conduction losses in the rectifier.

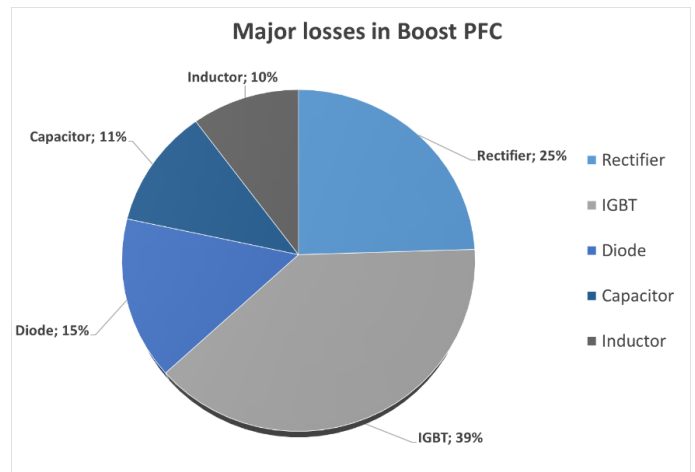


Figure 21 Average power losses in a boost PFC with Full-Si IGBT obtained from simulation at R_g = 12 Ω, V_{dc} = 400 V, P_o = 3.3kW and f_{sw} = 80kHz

The major contribution of losses is coming from the active switch i.e. IGBT. With just this knowledge, it is not easy to

determine the efforts that should be taken for the reduction of the power losses. At this point, the commutation-speed based method can be helpful as by splitting the switching losses, one can get a deeper insight into the losses that can help in the identification of the major cause behind the losses. This information can be used to figure out the steps that may be taken to reduce the losses.

This methodology is implemented on the considered two devices, the losses due to rectifier and passives are not considered further and the entire focus is on static and dynamic losses of the semiconductor switches.

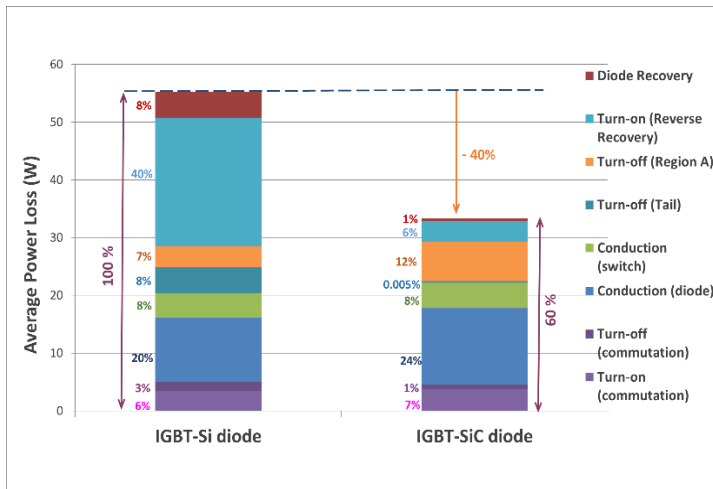


Figure 22 Split-up of the average power losses of the boost PFC for Full-Si IGBT and Hybrid-Si/SiC IGBT at $R_g = 12 \Omega$, $V_{dc} = 400 V$

Figure 22 shows the split-up of average power losses for the considered devices according to the regions described in section 4. The contribution of each of these regions help in the assessment of the impact that each region has. Full-Si IGBT is taken as the reference and is marked as 100%. The contribution of all the regions are shown in the Figure as percentages of the total loss of Full-Si IGBT. In Full-Si IGBT, about 40% of power losses are coming from the reverse recovery part of the freewheeling diode (FWD) during turn on event. As discussed earlier, the reverse recovery losses depend on the type of FWD and replacing Full-Si IGBT with Hybrid-Si/SiC IGBT, the losses due to turn on region B decreases by 84% as there is no reverse recovery in a SiC diode and only capacitive losses exist. Besides, a reduction of 89% is observed in the losses due to the diode recovery because there are no excess carriers in a SiC diode. The other major part of losses is due to the tail current, but tail current cannot be eliminated in an IGBT. The only way to eliminate the losses due to tail current is to replace the IGBT with a MOSFET. Another important conclusion that can be drawn from Figure 22 is that conduction losses are slightly greater for SiC diode compared to Si diode. This is because of the forward characteristics of SiC diode.

On the other hand, it is observed that the turn off power losses due to region A are higher for Hybrid-Si/SiC IGBT as compared to Full-Si IGBT. This is mainly because of the junction capacitance of the diode which is bigger for SiC diode so it's discharging takes longer, leading to higher power losses. This is shown in Figure 23 where under same conditions, Hybrid-Si/SiC IGBT takes 43.2 ns for the discharge of junction capacitance of diode where as Full-Si IGBT takes 36.6 ns.

Figure 24 shows the performance of both the devices for boost PFC at different power levels and it is seen that the efficiency of Hybrid-Si/SiC IGBT is greater than the Full-Si IGBT which is mainly due to the reduction in the losses in diode recovery and reverse recovery part during turn on event. For output power of 3.3kW, efficiency increases by 0.7% for Hybrid-Si/SiC IGBT.

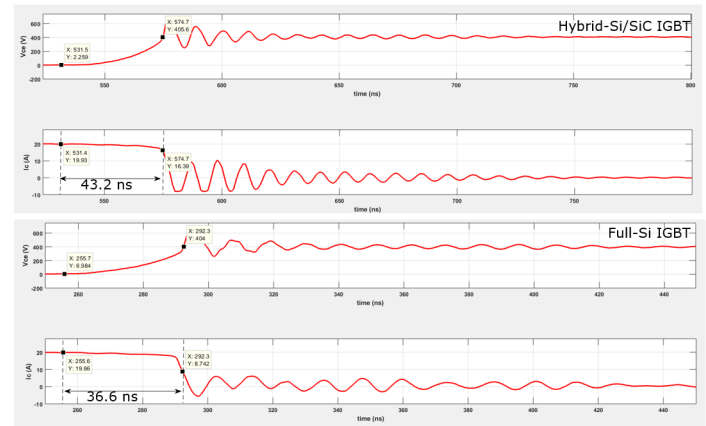


Figure 23 Turn off current for Full-Si IGBT and Hybrid-Si/SiC IGBT

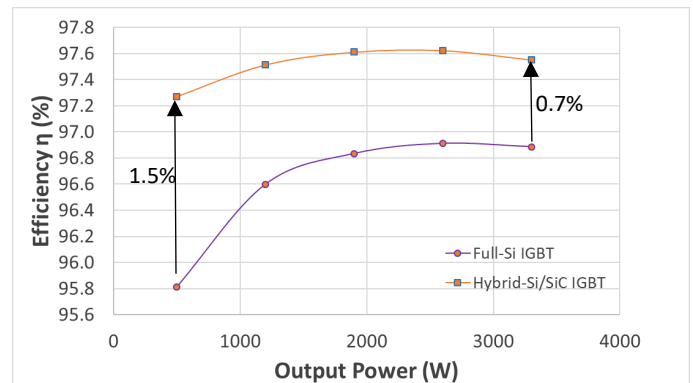


Figure 24 Efficiency comparison of considered devices for different power levels at $V_{dc} = 400 V$, $R_g = 12 \Omega$, $f_{sw} = 80kHz$

9. Conclusion

Traditional power loss calculation gives limited insight into the root cause of the power losses. The commutation-speed based method was applied for the calculation of power losses in the boost PFC of an on-board charger. In this paper, the total switching losses for two discrete devices were split up into their root causes e.g. tail current, reverse recovery. The two combination of devices were IGBT with Si diode and IGBT with SiC diode. The split-up of the considered devices was studied in detail as how the variation in device current $I_{c/f}$, DC link voltage V_{dc} , junction temperature T_j and gate resistance R_g affect the losses. Static and dynamic characterization measurements were performed to model the two devices behaviorally which were then used in boost PFC to calculate the power losses at the system level. It was found that the losses in the switch reduce by 40% just by replacing Si diode with SiC diode for the same IGBT. This reduction in the losses is due to the decrement in the losses in diode recovery and reverse recovery during turn on event indicating that the replacement of just the diode from Si to SiC reduces the power losses by considerable margin. Hence, it can be concluded that this

methodology can be applied to any application for the identification of ways to reduce power losses and eventually leading to improvement in efficiency.

Conflict of Interest

The authors declare that there is no conflict of interest

Acknowledgement

We would like to express our gratitude to Rafael Garcia for sharing his knowledge during the course of this research work

References

- [1] Gebhardt, Fabian, Hauke Vach, and Friedrich W.Fuchs, "Analytical Derivation of Power Semiconductor Losses in MOSFET Multilevel Inverters", Power Electronics and Motion Control Conference (EPE/PEMC), 2012 15th International. IEEE, 2012.
- [2] Ajay Poonjal Pai Poonjal, Tomas Reiter and Martin Maerz, "A New Behavioral Model for Accurate Loss Calculations in Power Semiconductors", PCIM Europe 2016; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management; Proceedings of. VDE, 2016.
- [3] Ajay Poonjal Pai, "Impact of Silicon Carbide based Power Modules on Mission Profile Efficiency of Automotive Traction Inverters", PhD Thesis to be published at Friedrich Alexander Universität Erlangen, (Internally available at Infineon).
- [4] Ajay Poonjal Pai, Tomas Reiter and Martin Maerz, "Efficiency Investigation of Full-SiC versus Si-based Automotive Inverter Power Modules at Equal Commutation Speed", PCIM Europe 2018; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management. VDE, 2018.
- [5] Infineon, "HybridPACK Drive, F2660R08A6P2 Power Module with 750V 660A EDT2 Chipset, Hints for Efficient IGBT Switching", Application Note, 2015, (Internally available at Infineon).
- [6] Kihyun Lee, Kyungsub Jung, Seunghoo Song, Yongsug Suh, Changwoo Kim, Hyoyol Yoo and Sunsoon Park, "Analysis and comparison of high power semiconductor device losses in 5MW PMSG MV wind turbines", Power Electronics Conference (IPEC-Hiroshima 2014-ECCE-ASIA), 2014 International. IEEE, 2014.
- [7] Ajay Poonjal Pai, Tomas Reiter and Martin Maerz, "An Improved Behavioral Model for Loss Calculations in Automotive Inverter", EEHE 2016 Wiesloch; Proceedings of 2016.
- [8] Dennis Ward, Iqbal Hussain, Carlos Castro, Andreas Volke, Michael Hornkamp, "Fundamentals of Semiconductors for Hybrid-Electric Powertrain", Published by Infineon Technologies AG.
- [9] Stefan Linder, "Power Semiconductors", EPFL Press, 2006.
- [10] Infineon, "TRENCHSTOP™ 5 H5 (High Speed 5) IGBTs", Datasheet
- [11] Markus Hermwille, "Gate Resistors-Principles and Application", Semikron International Application Notes.
- [12] B. Maurice, L. Wuidart, "Drive Circuits for Power MOSFETs and IGBTs", ST Microelectronics Application Notes.

Strategies of the Level-By-Level Approach to the Minimal Route

Nikolay Starostin^{*,1}, Konstantin Mironov^{1,2}

¹Institute of New Materials and Technologies, Ural Federal University, 620002, Russia

²Faculty of Computer Science and Robotics, Ufa State Aviation Technical University, 450008, Russia

ARTICLE INFO

Article history:

Received: 20 December, 2018

Accepted: 07 February, 2019

Online : 20 February, 2019

Keywords:

Optimization

Minimum route

Traveling salesman problem

TSP

CNC

ABSTRACT

The task of optimal path planning for drilling tool with numerical control is considered. Such tools are used in the production of printed circuit boards. The algorithm modification of level-by-level route construction for the approximate solution of the traveling salesman problem is discussed. Bypass objects can be specified either as a table of distances or values represented by a symmetric matrix, or as Cartesian coordinates (in applied cases of using numerically controlled equipment). The algorithm was tested on many different examples. As a result of the calculations performed for examples from the TSPLIB library (defined through a full distance matrix or Cartesian coordinates) with a dimension of up to 100 cities, the ability of the algorithm to construct the optimal route was confirmed. For examples from other sources or for artificially constructed ones (up to 130 objects), in the testing of the algorithm, the declared record values of the route minimum length were also achieved or improved.

1. Introduction

This paper is an extension of the article presented at the IEEE International Symposium on Signal Processing and Information Technology [1]. Initially, the problem of the rapid acquisition of an optimal route for moving the tool between processing zones on computer numerical control (CNC) equipment was considered, which is close in form to the symmetric traveling salesman problem with the Euclidean metric. For a limited number of processing zones, a solution [2, 3] belonging to the category of composite methods according to classification [4] was proposed. Construction of a convex hull, use of branch and boundary method elements, level-by-level analysis and route options design. This made it relatively easy to synthesize the optimal result for a small number of processing objects. Subsequently, the need to increase the effectiveness of this approach was revealed, where an important role, just like in the branch and boundary method, is played by a successful choice of upper bounds for the solution for which heuristic algorithms are used [5].

In the process of developing the required approximate algorithm, it was natural to use state of the art techniques from the exact method [3], while somewhat broadening the general formulation of the problem. Namely (in terms of the traveling salesman problem), to find the shortest closed route passing

through n cities, visiting each city exactly once. The distance between pairs of cities is determined by a symmetric, not metric in the general case, matrix of real numbers $C=|c_{i,j}|$. In [6] a natural version of the transition to the approximate method of searching for the minimum route in this formulation is presented, while the basic procedure, the basic calculation schemes and possible directions of development are described. In [7 – 9], the calculation procedures are detailed and studied, the results of testing are given, along with options for constructing the algorithm based on different schemes from the combined levels. For all benchmark tests (examples from [10] up to 100 cities inclusive, given by a complete table of distances or by Cartesian coordinates), optimal routes are obtained. Also, for all of several dozen random tests, with reported results (dimension in the range of 30-60 cities), for example, as in [11], record values of the minimum length of the route were reached, or these figures improved.

2. Basic procedure and trivial calculation scheme

If we rely on the algorithms' classification from [12], then the level-by-level approximation (considered and tested in [7,9], respectively, for the matrix and coordinate representation) can be attributed to the symbiosis of "tour construction algorithms" and "monotonic algorithms for improving the tour." Specifically, such one when the procedure for identifying improvements to the tour is involved at different levels of the multi-stage building process. According construction details, the proposed algorithm is closest

*Corresponding Author: Konstantin Mironov, mironovconst@gmail.com

to a truncated combination of three algorithms from the classification [12]. This is **C (arbitrary connection)**. «Let (i_1, i_2, \dots, i_n) be an order of vertices. Form a two-member tour (i_1, i_2) . The vertex i_l ($l = 3, 4, \dots, n$) is to be included in the $(l-1)$ -segment (pod) in such a place that the increase in the tour ... is minimal».

CR (most remote connection). «Act like in C. At step l ($l = 3, 4, \dots, n$) in the $(l-1)$ -st tour include the vertex i_l , for which the minimum distance to already included vertices is the largest among all vertices not yet included».

CE (most economic connection). «Act like in C. At step l ($l=3, 4, \dots, n$) in the $(l-1)$ -st tour include the vertex i_l , for which the increase in the tour is minimal among all the vertices not yet included. However, this combination emphasizes somewhat different points. It should also be noted that the proposed algorithm is multi-pass. Conscious departure from the seemingly natural use of coordinates for CNC equipment should also be underscored. This is done because different metrics can be used for different types of equipment and, accordingly, of technological processes (calculations of distances between treatment zones), therefore for the sake of greater generality we will not rely on the advantages of using metric space.

The essence of the basic algorithm is as follows. Let i_1, i_2, \dots, i_n be a certain order of processing zones, vertices or cities in terms of the traveling salesman problem, where for definiteness i_1 and i_2 are the most distant cities from each other. Build a tour from i_1 to i_2 and back, i.e., a two-member closed tour (i_1, i_2, i_1) . Next, include the city i_k ($k = 3$) in the existing two-member tour. This can be done in 2 ways: either (i_1, i_k, i_2, i_1) or (i_1, i_2, i_k, i_1) . Since they are equivalent, we will explain the process using the first one as an example. So, for each of the two three-member tours, it is necessary to perform consecutive actions starting from the step $l = 4$. In step l ($l = 4, 5, \dots, n$), in each of the available $(l - 1)$ -member tours, include one of the remaining cities i_l ($l = 4, 5, 6, \dots, n$) in that place of the tour, where it will provide a minimum tour increase based on the inclusion of this city. At the same time, the city of inclusion must be such that the value of the minimum increment of the tour from its inclusion is the maximum among the similar values of the minimum increments from other cities of the potential inclusion. Thus, at the last step ($l = n$) we obtain two complete routes. To fix one of them with the smallest length (in general, their length is the same, and the routes themselves coincide, the difference is only in the direction of the detour in the forward or reverse direction). Thus, the first of the approximate routes (Hamiltonian cycle) is obtained, which corresponds to its progenitor - a three-member tour i_1, i_k, i_2, i_1 . ($k = 3$). Then go back to the original two-member tour (i_1, i_2, i_1) and perform similar calculations for the remaining cities i_k ($k = 4, 5, \dots, n$). As the result, we will obtain $n-2$ approximate route options. Each route is determined by uniform constructions from its initial three-member tour i_1, i_k, i_2, i_1 , ($k = 3, 4, \dots, n$). We call the routes formed in this way from the original three-member tours, routes of level 0 and, for convenience, assign them indices $(00, 01, 02, \dots)$ in accordance with the order of their length increase.

It can be noted that the level 00 route (s), with any significant number of cities, rarely reach optimum or are close enough to it. Therefore, it is required to continue similar constructions in order to obtain level 1 routes. They are similarly constructed for each of the $n-2$ initial three-member tours i_1, i_k, i_2, i_1 , ($k = 3, 4, \dots, n$). At the same time, at least one of the new routes (for each of $k = 3, 4, \dots, n$) will be no longer by construction than its related route of level 0.

Let us explain this calculation by the example of constructing the first of the routes for level 1. In the initial tour (i_1, i_k, i_2, i_1) , ($k = 3$), include the city i_m (for definiteness $m = 4$) alternately between pairs of adjacent cities. Accordingly, obtain three 4-membered closed tours $(i_1, i_m, i_k, i_2, i_1)$, $(i_1, i_k, i_m, i_2, i_1)$, $(i_1, i_k, i_2, i_m, i_1)$. In step l ($l = 5, 6, \dots, n$), in each of the available $(l - 1)$ -member tours, include one of the remaining cities i_l ($l = 5, 6, \dots, n$) in the place of the tour where a minimum tour increase will be provided from the inclusion of this city. At the same time, the city of inclusion must be such that the value of the minimum increment of the tour from its inclusion is the maximum among the similar values of the minimum increments from other cities of the potential inclusion. Thus, at the last step ($l = n$) we get three complete routes. Fix one of them with the shortest length. This is the first of the approximate level 1 routes. Together with it, the initial 4-member tour is fixed as well (the value of m , k and the location of the corresponding cities in the tour).

To form the remaining routes of this level (with $k = 3$), we need to return to the original three-member tour (i_1, i_k, i_2, i_1) and perform similar actions with other valid values i_m ($m = 5, 6, \dots, n$), as in the construction of the first route of level 1. As a result, for the value of k under consideration, we obtain $n-3$ Hamiltonian cycles, i.e., approximate routes of level 1.

To form the remaining routes of this level ($n-3$ for each new k , where $k = 4, 5, \dots, n$), we need to return to the original three-member tours (i_1, i_k, i_2, i_1) . Perform similar actions for each of k with admissible values of i_m , as in the construction of the first group of routes of level 1. As a result, for each of k ($k = 4, 5, \dots, n$) we get $n-3$ full routes of level 1. All received routes in each group (fixed value k) are assigned indices $(10, 11, 12, \dots)$ in accordance with the order of increasing their lengths (several routes can have the same index if their lengths are equal to each other).

Similarly, routes of levels 2, 3 and subsequent are built, respectively, from 4, 5 etc. member source tours identified at the previous level.

But if nothing restricts this process, then we will have an avalanche-like growth of calculations. At the same time, when moving inward through the levels, a situation similar to the "thermal modeling" algorithms [12, 13], may be observed, where the process of improving the complete route is attenuated. For greater certainty, in [9], based on the accumulated statistical data, examples of different dimensionality of the level number, beyond which constructions are hardly advisable, are roughly indicated. When the level-by-level constructions are terminated according to the conditioned criterion, then in accordance with the multipass algorithm, we must again go to the two-member closed tour (i_1, i_2) for successive inclusion of the vertex i_k ($k = 4, 5, \dots, n$), as in the sample presented above (for the case $k = 3$). As a result, the route, the smallest of all constructed, is determined. It is also a potential contender for correspondence to the optimal route. This correspondence is consistently confirmed when testing the examples of a relatively small (up to 60 elements) dimension. Additionally, in practice, for achieving records from the original tripartite tours (i_1, i_k, i_2) , in tests from [10, 11, 14] with a dimension of 25-60 vertices, it was sufficient to perform calculations not for all but only for several vertices i_k where the values of index k correspond to those inclusion vertices that form zero-level routes with lower indices [7, 8]. In addition, for an overall reduction in the calculation of tests with a dimension of 60-100 vertices, extended schemes can be used, so that to reach a

record we would also be limited to calculations with only lower indices from the starting extended level [9].

In general, starting from this basic procedure with a trivial calculation scheme, it is necessary to form strategies rational for practical use to create complete routes and check the criteria for stopping calculations.

3. Possible directions in organizing calculations.

First, we return to the problems of CNC machine tools. Drilling holes in printed circuit boards is one of two sources of TSP instances that are associated with practical application [16]. From an applied point of view, local search algorithms are best for solving the traveling salesman problem [17]. Nevertheless, practical problems of this kind can bring certain difficulties for local search, even with a small number of vertices. So, for example, one of such effective algorithms with the Lin-Kernigan neighborhood implemented in the Concord program [17], does an excellent job with large-dimension tests from [10], but gives, for example, a fragment of a printed circuit board with a dimension of 15 vertices, which is not an important result (approximately 103% from optimum [1]). To obtain the optimal route (Figure 1 on the right), the program needs to change the default settings, which may not help at all in more complex cases. In addition, metaheuristic methods are very popular for practical problems. [18]. They do not guarantee that the best solution to the problem will be achieved, but with a successful setup they can find a “good enough” result. Their disadvantage is the presence of a different number of control parameters [20], which makes it difficult to set up algorithms for solving specific problems.

Now let us consider possible approaches for the organization of calculations in the framework of our basic algorithm for the level-based approximation.

3.1. Calculation on the basis of not deteriorating routes

Theoretically, each of the $n-2$ initial 3-member tours i_1, i_k, i_2 , ($k = 3, 4, \dots, n$), from which level 1 routes are formed, can be the progenitor of the optimal route. That, however, can no longer be said about each of the 4-member (and subsequent) tours, from which routes of level 2 and further are formed. Therefore, after constructing routes of level 1 (for each of the $n-2$ three-member tours), the route (s) with the index zero (10) is (are) allocated to each of the received $n-2$ groups and the corresponding four-member tour is fixed for them, from which the subsequent construction will go (routes with a different level index and the corresponding components are not considered further). Level 2 routes are similarly constructed on this base (level 2), after which only route (s) with index zero (20) and so on is (are) also considered (that is, at subsequent levels everything related to non-zero indices routes is not participating). Thus, routes of non-zero level J ($J = 2, 3, 4, \dots$) are constructed similarly to routes of level 1, from their initial $(J + 2)$ -fold tours, which correspond to full routes of levels $(J-1) 0$. The whole process of level building can be conditionally represented by the following scheme:

$$(1) \quad \text{level } 0_i - \text{level } 1_0 - \text{level } 2_0 \dots (i=0,1,2,\dots)$$

The shortest length among all level 1 routes will be the route with a zero index (level 10), which is the result of the initial embedding of a certain city i_z ($z \neq 1,2, k$) into a certain place of the tour (i_1, i_k, i_2, i_1) . By construction, its length cannot be greater

than the length of the “parent” route of level 0, resulting from embedding in the tour (i_1, i_2, i_1) of city i_k . The same applies to all derived routes with a zero index of subsequent levels, i.e., the best routes of the level cannot be longer than their ancestors. Moreover, when building at the initial levels (when moving from a lower level number to a larger number), the length of the descendant route decreases; as practice shows, this is a frequent case. However, as the level numbers increase, this happens less frequently and then fades out.

Therefore, the process of constructing routes in ascending level numbers continues, approximately as in “thermal modeling” algorithms, until a certain equilibrium is reached, understood as “no significant deviations from the found tour length are observed for a long time” [12, 13], i.e., until the termination of reducing the current length of the route. After that, the level constructions in this direction are interrupted. In the end (when the calculation is stopped in all available directions) the route is determined, the smallest of all constructed. It is also a potential contender for compliance with the optimal route. In [6, 7], the main results of testing in the framework of the considered scheme (1) for several dozens of tasks with cities given by a symmetric table of distances up to 60 objects in dimension are presented. At the same time, in all calculations, the reported record results were obtained or improved.

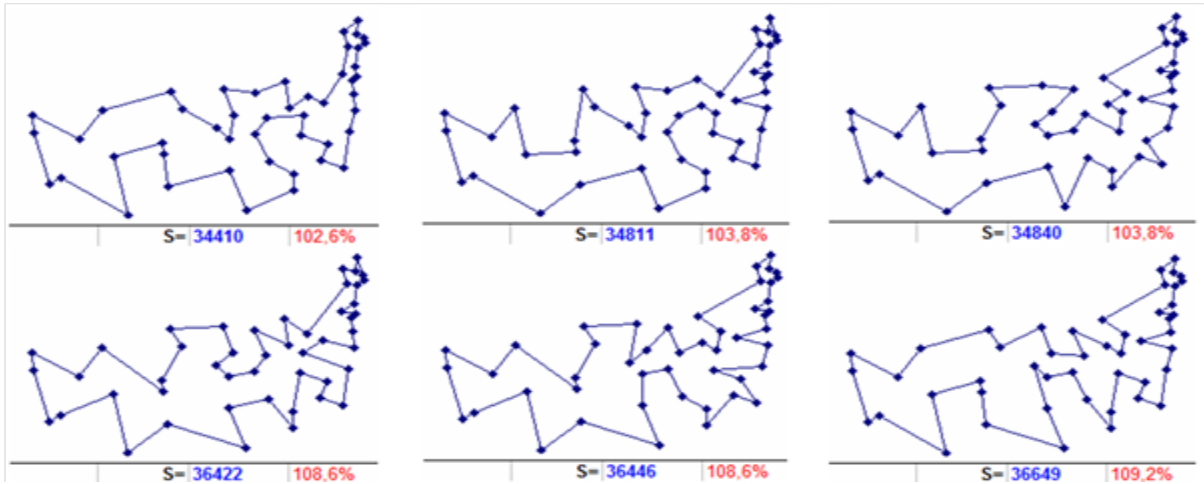
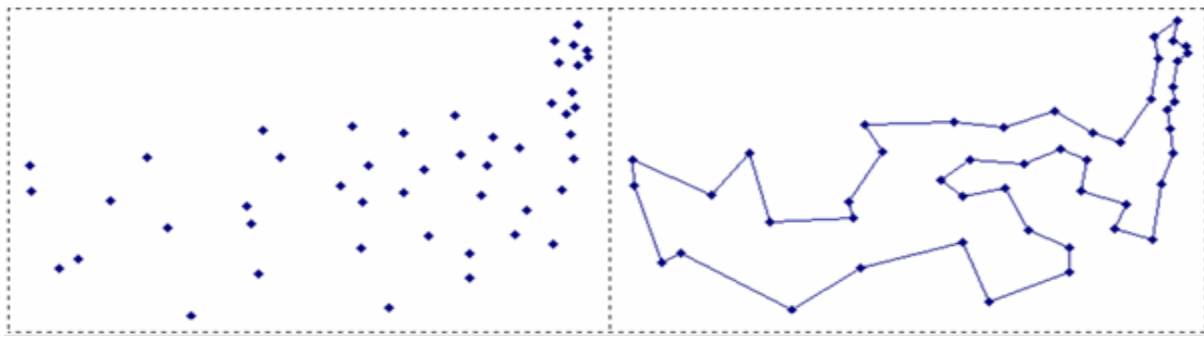
To explain the essence of the material presented in [1], a simple example was selected with 29 bypass objects from the TSPLIB library [10], and to demonstrate the results of innovations, a test (file *xqf131.tsp*) measuring 131 elements from the VLSI TSPs collection [20] created in based on SBIS data sets investigated at the University of Bonn.

Here, as a simple illustrative material, we take a classic sample containing 48 objects (*att48_d.txt* file) from a library with TSP data [21]. This is one of the key examples put on the cover of a famous book [22], which was used in it to illustrate existing approaches to solving the problem in question. To demonstrate the results of innovations, we will again use the same test of 131 elements.

So, let us explain the organization of the process for finding the minimum route on the selected example with 48 US cities (*att48_d.txt* file), given by the complete distance table. For graphical representations of the constructed routes, the Cartesian coordinates of cities from the *att48.tsp* [10] file will be used (in Fig.1, coordinates are on the left, and the known optimal route with a length of 33551 is on the right).

If we number the rows of the original distance table starting from 0, then the most distant cities will have numbers 3 and 16. Route level 00 (Fig. 2 upper left) built from the initial two-member tour (by cities 3, 16, 3) has a length of 34410 and a percentage ratio of 102.6% with the optimal route (hereinafter, percentages are indicated to the nearest tenth). The same figure shows the routes of level 01, level 02 (continuation of the first row of Fig. 2) and in the second row - the three longest routes of level 0 (which have the largest indices).

Sample construction 1. Route level 00 (Fig. 2 left upper) in one copy. It assumes the inclusion of the city with the number 30 in the two-member tour (3, 16, 3) (other cities with similar embedding give a greater length of the resulting route of the zero level, and, accordingly, such routes will have a larger index.



The route of level 10 (00–10), built from the initial three-member tour (3, 30, 16, 3) is also one, its length is 33961 (Fig. 3 on the left). It implies embedding city 46 in the tour (3, 30, 16, 3) between cities 16 and 3.

Level 20 routes (00–10–20) are already 15, their length has not improved, and has remained the same (33961) as at the previous (parent) level 10. Each of them involves embedding one of the 15 cities: 5, 6, 7, 9, 14, 17, 18, 26, 27, 29, 34, 36, 37, 43, 44 into the initial tour in a certain place (3, 30, 16, 46, 3). Next, for each newly formed tour routes of level 30 (00 - 10 - 20 - 30) are built, of which there are already about 152, which also do not show the change in the length of the best tour. And only at level 40 there is a decrease in length in certain directions (Fig. 3 in the center and to the right).

First, we shall consider the direction corresponding to the shorter route length (33614 Fig. 3 on the right). Let us call it “branching 1”. When building routes of level 50 from it, there will be several of them again, i.e., 22, from each of which at level 60 will be somewhere along 23-27 similar new ones, and their length does not change (it remains equal to 33614). The situation is similar at the level of 70 and 80. That is, there is a circumstance that “for a long time no changes are observed” of the lengths of routes with zero indices and therefore it is necessary to stop further constructions in this direction.

For the future, we should agree on the criteria for stopping calculations in the current direction. Let it be, for example, the following: “threefold repetition of the route length value with a zero index during the transition from level to level in the process of forming hereditary directions».

That is, in our case, calculations for this direction (branching 1) should be stopped after the formation of level 70 routes, despite the fact that we would reach the optimum (length of the route is 33551 cm. Fig. 1) already while calculating from level 90, continuing calculations for this direction. For example, when calculating from the initial closed tour 3, 1, 33, 7, 30, 16, 32, 19, 46, 31, 4, 3 (when you include city 40 between cities 33 and 7). Or from another source tour 3, 1, 33, 40, 7, 30, 16, 32, 46, 31, 4, 3 (with the inclusion of the city 19 between cities 32 and 46).

Let us now consider the second direction (“branching 2”), corresponding to the greater length of the route (33741 Fig. 3 in the center). When building from it, 4 routes of level 50 are obtained with a length that has changed to a value of 33614. From each of the corresponding four routes, the eight-member initial tours are obtained at level of 60, with about 21-24 new routes in them, while their length does not change (it remains equal to 33614). At level 70, another average of 23 routes is added. This is about 4 * 232 new calculation directions in grand total. And if during the formation of the level 80 routes there is no change in the length of at least one of them, then, according to the entered criterion, the calculations in this direction (“branching 2”) should be stopped, and therefore, in general, the entire “parent” direction of level 00 should cease. However, in reality this does not happen. When forming routes of level 80 from the initial closed tour 3, 40, 2, 8, 7, 30, 6, 16, 46, 3 (when you include city 32 between cities 16 and 46 - constructing at level 70), a route of length 33551 (Fig 1) results, which is an optimal one.

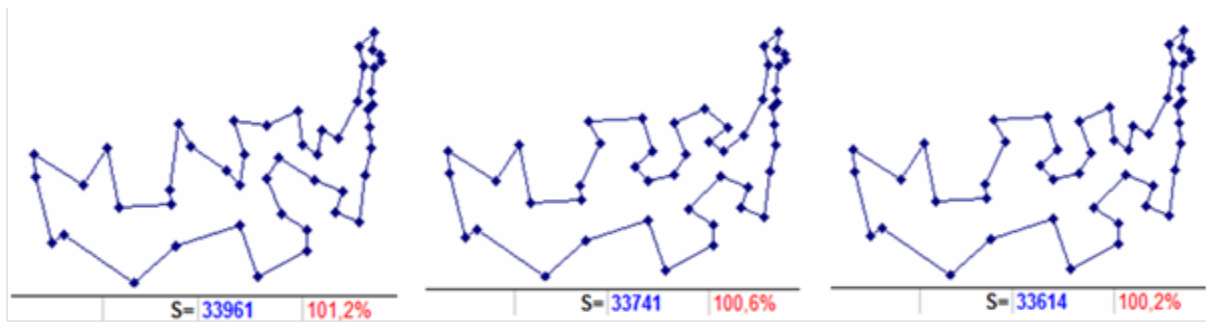
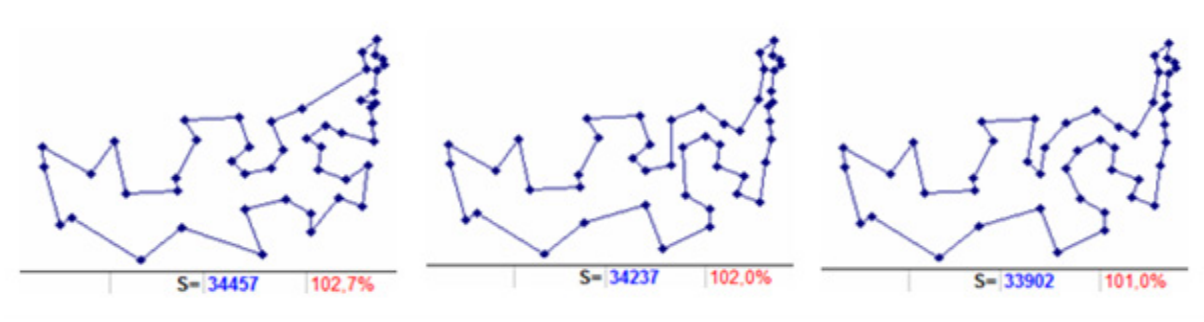


Figure 3. Route of level 10 (left), and of level 40 (center and right). Additionally shown are the length and deviation relative to the optimum.



Despite the fact that the result within the declared scheme (1) has been achieved, we will continue our constructions from the 0i ($i \neq 0$) levels in order to identify less labor-intensive directions. And such options are, for example, the following ones.

. This route is one copy. It assumes the inclusion in the two-member tour (3, 16, 3) of the city with the number 32. There are three routes of level $l_0(0_{24} - l_0)$, constructed from the original three-member tour (3, 32, 16, 3), their length is 34457 (Fig. 4 left). They suggest embedding in the tour (3, 32, 16, 3) cities 10, 19 or 34 between cities 16 and 3.

We will begin further constructions from city 10 (having a smaller sequence number). When calculating from the tour (3, 32, 16, 10, 3), the only route of level 20 is 34237 long (Fig. 4 in the center). It assumes embedding in the tour (3, 32, 16, 10, 3) city 43 between cities 16 and 10. By performing this operation, we also get a single route of level 30 with a length of 33902 (Fig. 4 on the right). This, in turn, is assuming that the only embedding in the tour (3, 32, 16, 43, 10, 3) shall be city 22 between cities 3 and 10. This embedding, in turn, results in obtaining an optimal route of length 33551 (Fig. 1). As a result, the entire calculation is significantly less time consuming than the calculations from level 00 in sample construction 1. The remaining two branches (those from cities 19 and 34) are more computationally expensive.

Here we shall note that it is possible to dramatically increase the number of ways to build optimal routes through the use of extended schemes for non-deteriorating routes.

In [8] it was shown that within the framework of the scheme (1), as a rule, there are quite a few options for constructing optimal routes; however, in some cases it is not possible to quickly reach

the optimal result, therefore in [7], alternatively, extended schemes are proposed. They allow, for example, in problems with a dimension of up to 50 elements, to obtain an optimum often even at the initial state. Either they open new directions in which the optimal result is achieved in the absence of significant branching of the process, or at earlier than in scheme levels (1), which may already be more relevant for tasks in the range of dimensions, from 50 to 100 elements [9].

The first version of the extended calculation scheme:

$$(1) \quad \text{level } (0 - 1)_i - \text{level } 2_0 - \text{level } 3_0 \dots \\ (i=0,1,2,\dots)$$

Here, level (0 - 1) represents some combination of level 0 and level 1 from the previous scheme. Full routes (Hamiltonian cycles) from level 0 are not built, and each of the $n-2$ source 3-member tours $i_1, i_k, i_2, i_1, (k = 3, 4, \dots, n)$ is converted into 3 four-member one, from which the full routes are already formed from the combined level (0-1). Here, they are presented in the number of $(n-2) * (n-3) * 3$, which is a broader basis for the formation of full routes of levels 20 and subsequent levels. At the same time, within the framework of the scheme (1) for the formation of level 20 routes, there is, as practice shows, in all directions $(n-2)$ or several more closed source routes, that is, a small subset of the analog from the scheme (2).

Let us explain the scheme (2) in our current example. Among the many variants of the initial routes corresponding to the level (0 - 1) i , we take two (with some average level indicators of the full route length) and at the same time non-existent (unrealizable) within the framework of the scheme (1).

Sample of construction 3. Let the first of them be the original route that implies embedding into a two-member tour 3, 16, 3 pairs of cities 15 and 30 between cities 3 and 16. As a result, we have tour 3, 15, 30, 16, 3, and the only route from it is on level 20, involving the incorporation of city 37 between cities 30 and 15, with a total length 34410, corresponding to the upper left in Fig. 2

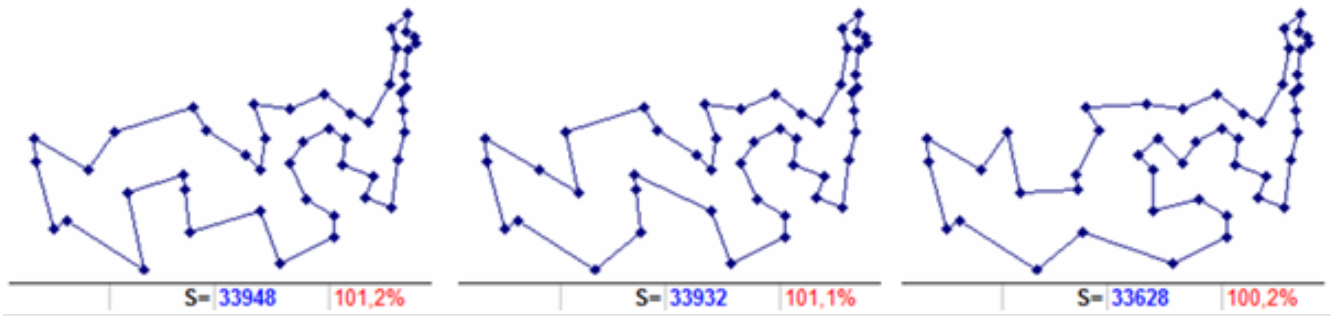


Figure 5. Intermediate routes of level 30 (left), level 40 (center), level 50 (right) as part of the implementation of the scheme (2). Additionally, the length and deviation relative to the optimum are shown.

In turn, it builds only one next-level route with a zero index (level 30), with an estimated integration of city 10 between cities 15 and 37 on tour 3, 15, 37, 30, 16, 3 with the result (length 33948) indicated in Fig. 5 left. Already from this position is also obtained the only route of level 40 with the expected inclusion of city 25 between cities 3 and 15 in tour 3, 15, 10, 37, 30, 16, 3, and with length of 33932 (Fig. 5 in the center). Further, the route of level 50 is also created in a single copy indicating the inclusion of city 12 between cities 15 and 10 in tour 3, 25, 15, 10, 37, 30, 16, 3, and with length of route 33628 (Fig. 5 on the right). And already from that, after the implementation of the specified inclusion, we immediately get the optimal route (Fig. 1 on the right). As a result, all transitions from the initial to the last level pass without a single branching in the process of construction.

Sample of construction 4. Let us take, as the second similar initial tour, a route with a supposed embedding in the two-member tour 3, 16, 3 pairs of cities 22 and 30 between cities 3 and 16. There is also no corresponding route within scheme (1), since embedding from level 1 with city 22 in the case of constructions from level 00 (or with city 30 in case of constructions from level 025) will have an index number of level 1 greater than 0, which is beyond the scope of the scheme. Within the framework of scheme (2), the corresponding full routes exist and coincide, their length is 35935 (Fig. 6, left). The calculation from tour 3, 22, 30, 16, 3 gives a single route of level 20, which implies embedding city 11 between cities 22 and 30 with a full route 34694, corresponding to the central route in Fig. 6. There are already two level 30 routes (Fig. 6 on the right), they suggest embedding cities 32 or 45 between cities 3 and 16 in tour 3, 22, 11, 30, 16, 3. Both routes with the specified embedding lead to an optimal result (Fig. 1 right).

Sample of construction 5. It is also possible to note the possibility of obtaining an optimum according to the scheme (2) among the directions existing within the framework of the scheme (1), but “rejected” by the criterion of stopping the calculations. For example, in the framework of scheme (1), according to this

criterion, constructions are rejected from the direction (06–10–20–30) because of the fourfold construction at each level of the same complete route. Namely, the sequential inclusion in the two-member tour 3, 16, 3 the city 43 (3, 43, 16, 3), then the city 10 (3, 10, 43, 16, 3). After that, city 40 (3, 40, 10, 43, 16, 3) and finally, city 0 (3, 40, 10, 0, 43, 16, 3), each time they build the same full length route 33633 (100, 2% of the optimum of Fig. 7, in the center). In fact, there is a threefold repetition of the length of the route and, according to the criterion, the calculation stops.

Let us present another extended scheme similar to (2), but with three levels combined.

$$(1) \quad \text{level } (0 - 1 - 2)_i - \text{level } 3_0 - \text{level } 4_0 \dots \\ (i=0,1,2,\dots)$$

It provides even greater opportunities for obtaining record-breaking results due to the impressive “re-selection” component of the combined entry level. $(0 - 1 - 2)_i$.

Sample of construction 6. So, for example, thanks to such a scheme, from tour 3, 12, 10, 43, 16, 3 (which corresponds to the construction from the combined level $(0 - 1 - 2)_i$) in our current example of 48 cities, we immediately get the optimal route. This tour is not the only one on the level combined by the scheme (3).

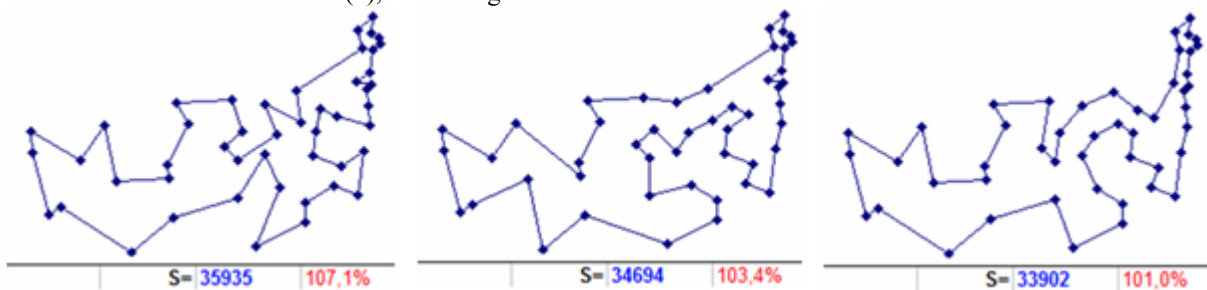


Figure 6. Intermediate routes of the “combined” level $(0-1)_i$ (left), level 20 (center), level 30 (right) as part of the implementation of the scheme (2). Additionally, the length and deviation relative to the optimum are shown.

3.2. Calculation based of the scheme with permissible deterioration of routes

Such a scheme is designed to provide flexibility in achieving record routes. This is a way to improve from a slightly worse calculation position (which is longer than the best route of the current level), which at the next or later levels can lead to a change in the current minimum in the considered direction of calculation. It is significant that this allows to avoid the situation “for a long time, no deviations from the found tour length are observed”. This approach is implemented based on the use of index variation not only as previously at the zero level, but also at subsequent levels of construction. That is, in the general case, the calculation scheme can be as follows [16]:

(2) level 0_{i0} – level 1_{i1} – level 2_{i2} ... ($i_0, i_1, i_2, \dots = 0, 1, 2, \dots$)

The total number of calculations here is much larger relative to the scheme (1). Therefore, from a practical point of view, one should artificially limit the potential excessive growth of the level-by-level calculations. And from this position, it is more expedient to introduce into (4) additional restrictions or to use some hybrid variants from schemes (1) - (3) and (4), for example let us consider the scheme (5), which is defined by the next three paragraphs.

If at the levels of J_0 , where $J = 1, 2, 3, \dots$, only one route is observed (with index 0), then further constructions lead only from the initial $(J + 2)$ -member tour for it, i.e., according to scheme (1).

If there are several routes with index 0, then further constructions will lead not only from the initial $(J + 2)$ -member tours for them, but also from all other $(J + 2)$ -member tours that are initial for the level routes with corresponding non-zero indices, i.e., as in scheme (4).

Moreover, in order to reduce computations, directions of calculations from non-zero indices, for constructions from which the full route with a zero index of the next level coincides with the parent route (that is, in the absence of a decrease in the length), shall not be considered.

To illustrate the scheme with permissible degradations of routes, we turn to the already considered version of the calculation from construction sample 1, where for the first time instead of one, in the direction of 20 (in the direction 00 - 10 - 20), 15 routes appear with a zero index and not improved length of 33961 (Fig. 3 left). This is just the reason to use the scheme (5).

Sample of construction 7. So, we shall start constructing routes of the 3rd level. Embedding in a certain place of the initial tour (3,

30, 16, 46, 3) each of the identified cities corresponding to the level 2 zero index does not change the generated routes of the next level with the zero index (see construction pattern 1, level 30). According to the adopted scheme, we will construct from non-zero level indices. So embedding of city 15 between cities 3 and 30 (corresponding to the route with the last, i.e., the largest index of level 2 with an initial length of 35582) results in two routes of level 30 with a decrease in length to 33614.

Further, from the first of these routes, through the implementation of the supposed embedding of city 33 between cities 3 and 15 (on tour 3, 15, 30, 16, 46, 3), 25 routes of level 40 are obtained. In one of them, the embedding of city 32 between cities 16 and 46 (in tour 3, 33, 15, 30, 16, 46, 3) allows to obtain the optimal route which is several levels earlier than indicated in construction sample 1.

From the second route, an optimal tour is also obtained, but it is built one level further.

Sample of construction 8. A stronger result is obtained when constructing from level 06, where there is (before the start of the intended embedding of city 43 in the two-member tour 3, 16, 3) a full route (Fig. 7, left) 35143 in length. The route of level 10 from the implementation of such embedding is one, with length of 33633 (Fig. 7, in the center) and it involves embedding the city 10 between 3 and 43 (on tour 3, 43, 16, 3). When implementing it, we already get 10 level 20 routes and, according to scheme (5), along with them, it is possible to conduct calculations also from levels 2_i (where $i \neq 0$). As a result of the calculation from level 27 (where the full route 34479 in Fig. 7 is on the right), when the city 22 is embedded between 3 and 10 (in tour 3, 10, 43, 16, 3), we immediately get the optimal route (Fig. 1 on the right).

It can be noted that in a similar way and with the same laboriousness (when building from level 2), optimum is obtained when calculating from the starting level 012 (the full potential route is no longer than 35808), with an estimated only integration of city 10 into the two-member tour 3, 16, 3.

4. Natural innovation in the basic calculation procedure

Having considered the main directions of calculations, and based on the statistics of the calculations carried out earlier [7–9, 16], we can draw some conclusions. For the considered possible directions of calculations in various test examples with dimensions up to 100 and a little more elements, it is always possible to obtain the optimal result with varying degrees of complexity. In this case, the optimum for each of the possible calculation schemes, as a rule, is achieved in many different ways. The least laborious route

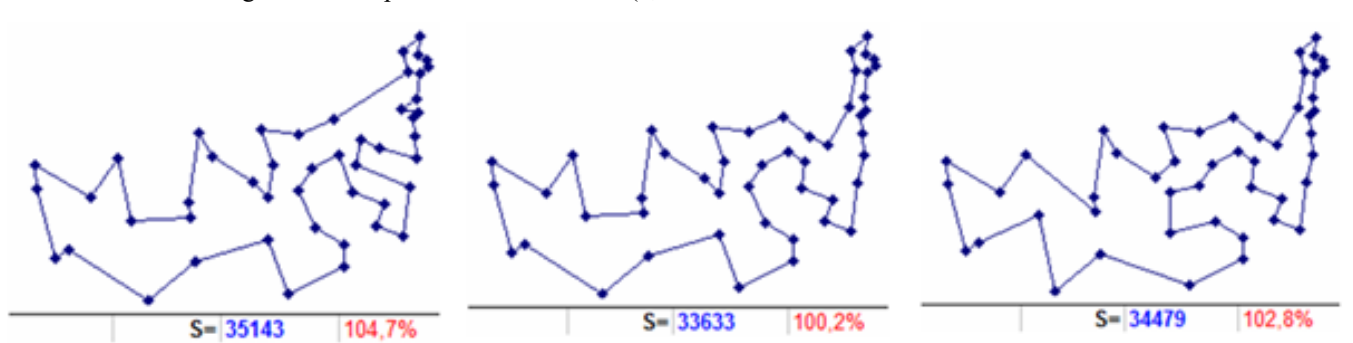


Figure 7. Routes corresponding to the constructions according to the scheme (5): level 06 (left), level 06-10 (center), level 06-10-27 (right). Additionally, the length and deviation relative to the optimum are shown.

(when comparing calculations from level 0 starting tours) was often the one where the optimal route is built at earlier levels. Less often - at later levels, but in the absence of a large number of branches. The results of experiments [7–9, 16] also show the following. The more extensive the base of the starting (initial) level, as is the case with the use of advanced schemes, the higher the probability of an “earlier” approach to optimum in individual calculations from starting tours. The total number of such exits is also growing.

There is a natural opportunity to significantly increase the number of zero level starting tours from the current n-2 options within the framework of the scheme (1). That will also increase the number of corresponding approximate level routes. The essence of such an extension is as follows. For convenience and certainty, the baseline calculation procedure used the directive choice of the first two cities (maximum distance from each other). From this pair, all subsequent constructions were carried out. If calculations would be carried out from each unique pair of source cities, then the number of starting two-member tours increases to $n * (n-1) / 2$, where n is the number of cities. Let us return to current example (48 cities) to clarify application of new starting conditions.

Samples of construction 9. In practice, probably, such a large number (in our new conditions 1128) of initial tours is not necessary. Let us see what happens if we restrict ourselves, for example, to considering only a few starting two-member tours, from which when forming routes of the zero level we get the smallest routes with zero index (i.e., the smallest among all possible routes of levels 00).

In our example, there are only 68 (out of 1128) two-member tours with a minimum route (at level 00) of a length of 33614 (as in Fig. 3 on the right). By constructing according to scheme (1), we will reveal among them those two-member tours, from which the optimal route will be constructed at the nearest of the subsequent levels, and accordingly the number of this level. Thus, directly from all possible three-member tours, being built from each of the 68 two-member tours, the optimal route is not immediately reached. But for the individual subsequent four-member tours the optimal route (Fig. 1 on the right) is already

calculated. That corresponds to the constructions from level 10. The code of these constructions, corresponding to scheme (1), is contained in table 1; it indicates the sequence of obtaining the optimal route (length 33551).

Table 2 contains similar information, but this time from a small sample of the remaining starting two-member tours. In it, the best routes of zero levels (routes of level 00) exceed the minimum length of 33614 (that is, more than the tours from the previous table). The corresponding constructions for each of those two-member tours of scheme (1) also lead to the identification of the optimal route, with the implementation of inclusion from level 10.

The above tabular data suggests that there are no clear preferences when choosing a two-element tour from the available starting conditions in order to obtain a “short” calculation of the optimal route. For example, our initial starting two-member tour (3 16 3) from the most distant cities is with index 75 according to the criterion for the route length of level 00 almost exactly in the middle of the totality of 1,128 two-member tours (among the sequence numbers in ascending order from 532 to 574). From the point of view of building an optimal route (sample of building 1 - 8), it is not very fast relative to others. But the starting two-member tour 12 38 12 with the sequence number 1127 (Table 2 last line) with the length of the 00 level route equal to 35400 (Fig. 8 on the left) is almost perfect from the same point of view (like some others presented in both Tables).

And this is despite the fact that its starting indicators are next to the last ones in proximity to the optimum length of its best route among all 1128 results of levels 00. However, the optimal path is quickly and unequivocally (without a single branching) built from level 03 (basic length 35659 in Fig. 8 in the center). Specifically, first begins the inclusion in the two-member tour of the city 28 (12 28 38 12). As a result, a route with the length of 34603 is built (Fig. 8 on the right). Then, the only possible inclusion of the next level 10 is the insertion of city 32 (12 32 28 38 12) and, as a result, the already known optimal route is obtained (Fig. 1 on the right).

Table 1: Sequential Data

The sequence number of the original tour	The order of the cities for the starting two-member tour	The length of the level 0i tour(at the beginning of the direction for constructing the optimal route)	Order of cities of a three-member tour (level 0i)	Length of a tour of level I0	Order of cities in a four-member tour, directly from which the optimal route is built (level I0)
3	1 30 1	35374 34110	1 12 30 1 1 15 30 1	33628 33628	1 15 12 30 1 — // —
8	1 43 1	35137 35131	1 22 43 1 1 24 43 1	33879 33879	1 22 24 43 1 — // —
22	6 37 6	34947 35613	6 22 37 6 6 24 37 6	33955 33955	6 22 24 37 6 — // —
39	17 37 17	35191	17 24 37 17	33955	17 22 24 37 17
45	22 27 22	34483	22 24 27 22	33955	22 24 37 27 22
53	27 37 27	34704 35769	27 12 37 27 27 24 37 27	34701 33955	27 28 12 37 27 27 22 24 37 27

Table 2: Tour Measurements

Tour Number	Order of starting cities according to the scheme (1) tour	Index among all tours of levels θ_0	Tour length of level θ_0	Tour length of level θ_i (starting to calculate)	Order of three-member tour cities (level θ_i)	Tour length of level I_0
69	3 10 3	1	33633	35877	3 22 10 3	34479
70	3 43 3	1	33633	35254	3 22 43 3	34479
331	24 38 24	33	34151	35539 36239	24 21 38 24 24 40 38 24	34668 34151
332	24 43 24	33	34151	35131 35420 36138 35880	24 1 43 24 24 8 43 24 24 22 43 24 24 40 43 24	33879 33985 33879 34151
333	38 43 38	33	34151	36575	38 22 43 38	34510
394	22 24 22	46	34276	35881 35128 35561 36138 34483 36885 36138	22 1 24 22 22 6 24 22 22 8 24 22 22 17 24 22 22 27 24 22 22 37 24 22 22 43 24 22	33879 33955 33985 33955 33955 33955 33879
639	22 43 22	90	34479	36427 35137 35254 35818 36152 36138 37297	22 0 43 22 22 1 43 22 22 3 43 22 22 9 43 22 22 20 43 22 22 24 43 22 22 38 43 22	33991 33879 34479 33991 34094 33879 34510
854	10 22 10	156	34755	35877 35022	10 3 22 10 10 28 22 10	34479 34094
1127	12 38 12	258	35400	35659	12 28 38 12	34603

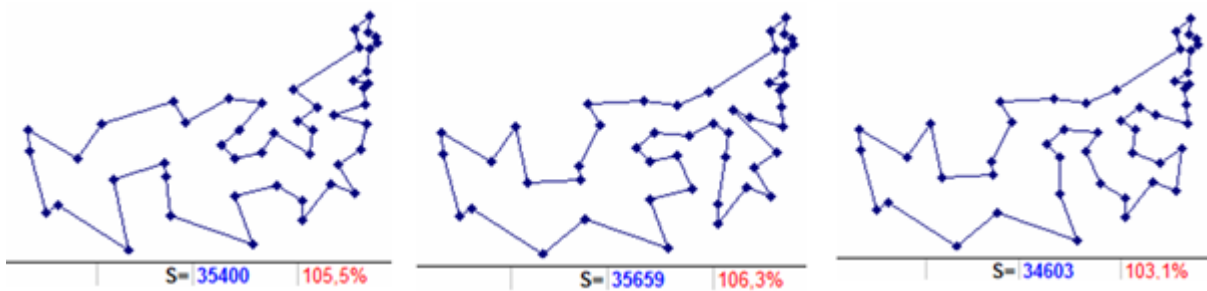


Figure 8. Routes from the starting cities 12 and 38 corresponding to the constructions according to the scheme (1): level 00 (left), level 03 (center), level 03 - 10 (right). Additionally, the length and deviation relative to the optimum are shown.

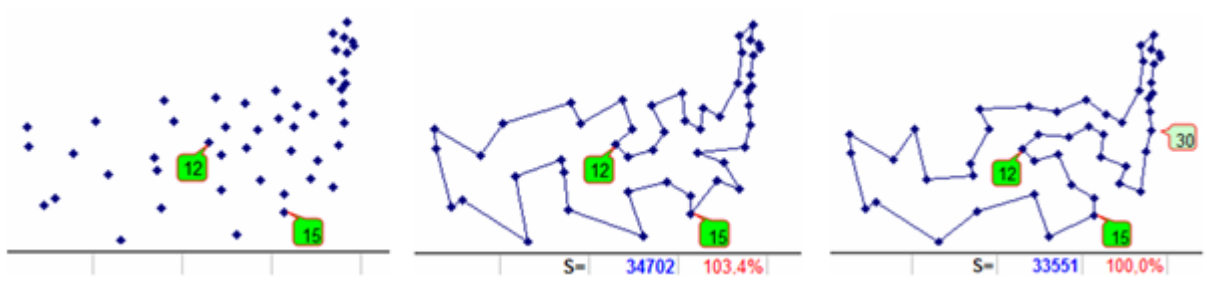


Figure 9. Cities 12 and 15 that make up the starting two-member tour (left). The route of level 00, corresponding to the constructions according to scheme (1), implies the inclusion of city 30 (in the center). Route level 00 - 10 (right), resulting from the inclusion of the city 30 (optimal). Additionally, the length and deviation relative to the optimum are shown.

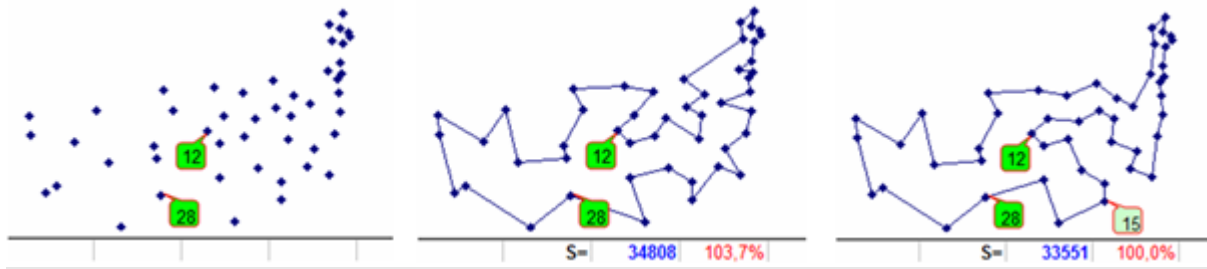


Figure 10. Cities 12 and 28 constituting the starting two-member tour (left). The route level 00, corresponding to the constructions according to scheme (1), implies the inclusion of the city 15 (in the center). Route level 00 - 10 (right), obtained as a result of the inclusion of the city 15 (optimal). Additionally, the length and deviation relative to the optimum are shown.

Samples of construction 10. It is necessary to separately distinguish the group of constructions where the optimal tour is obtained at the earliest stage (from the beginning of the calculations) when implementing the inclusion in the starting two-member tour of the third city, i.e., after calculations at the zero level.

Fig. 9 (left) shows an illustration of the starting conditions for the calculation of a two-member tour between cities 12 and 15. In the center of this figure, the shortest route (length 34702) is being built for those cities which have a zero level route (level 00). It is the only one and involves the inclusion in the two-member tour of the city 30. After the realization of this inclusion, we obtain the optimal route Fig. 9 (right).

Also, when implementing the inclusion of level 00, an optimal route is constructed from a two-member tour between cities 12 and 28. Fig. 10 (left) is an illustration of the starting calculation conditions for this two-member tour. In the center of the figure, the shortest (length 34808) zero level route (level 00) is shown. There are two such routes and, accordingly, two directions of construction according to the scheme (1). The first one, which implies the inclusion of city 2, is a dead end one for reaching the optimum. The second one, which implies the inclusion of the city 15, after its implementation immediately builds the optimal route Fig. 10 (right).

For more commonality, Fig. 11 shows 4 more routes of the zero level from binomial tours, but with a non-zero index (levels $0i$, $i \neq 0$: the first row in the center and to the right, the second row in the center and to the left). Each of the directions corresponding to them, after the implementation of the proper inclusion of the third city in the corresponding route of the two-member tour, also immediately leads to an optimal result (Fig. 11, second row rightmost). This happens again when constructing from the zero level, but only already when implementing the inclusion corresponding to a non-zero index (from the level $0i$, $i \neq 0$).

Unfortunately, in the absence of a general criterion for choosing the “correct” pairs of starting two-member tours, in order to identify all three-member tours, from which the optimal route is immediately constructed, 17296 possible options will have to be considered (in general, $n * (n-1) * (n-2) / 6$, where n is the dimension of the problem).

5. Features of finding the best route in tasks with a large dimension

In the light of the innovations outlined in the previous section, at the very beginning, there emerges a very large number of directions of calculations from each of which it is potentially possible to get the best route.

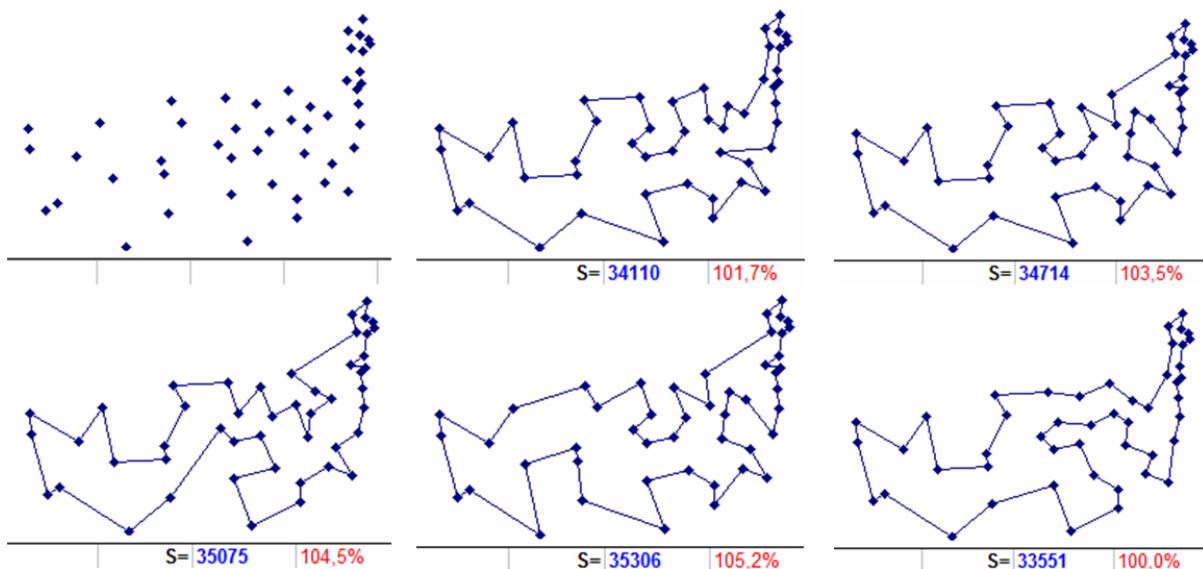


Figure 11. The mutual position of 48 cities (upper left). The optimal route is 33551 long (bottom right). The rest: routes of level $0i$, directly transformed after the inclusion of the corresponding third city in the optimal route. Additionally, the length and deviation relative to the optimum are shown.

In such examples as ours (dimensionality of 48 cities) and even twice as large dimensionality, it is still possible to authorize the definition of not only the best route, but also the youngest level where it is reached (along with a full set of options for building it on it). This was shown in construction samples 10.

For examples with a dimension somewhat greater than 100 elements, such an approach (i.e., the calculation of "breadth") is already quite laborious. Therefore, we first denote a simplified version of the organization of calculations among the set of initial two-member tours. For each unique pair of source cities, determine the minimum route of level 00, and the corresponding city for inclusion in the two-member tour (if there are several, then one of them). Index all $(n * (n-1) / 2)$ obtained routes in order of increasing length (we call these indices global). Limit the number of calculations from starting two-member tours (or groups of tours) only to those of them that have small global indices.

In addition, we will somewhat limit the number of calculations in scheme (5). One of its possible modifications can be formulated as the scheme (6) determined by the next paragraph.

If at the levels of J_0 , where $J = 1,2,3 \dots$, only one route is observed (with index 0), then further constructions will only lead from the initial $(J + 2)$ -member tour, i.e., according to the scheme (1). If there are several routes with index 0, then further constructions will lead not only from the initial $(J + 2)$ -member tours for them, but also from all other $(J + 2)$ -member tours that are initial for the level routes with corresponding non-zero indices, i.e., as in the scheme (4). At the same time, in order to further reduce the computations, as compared with scheme (5), we shall consider only separate directions of calculations from non-zero indices of the current level ("current-baseline"). These are the directions from which a chain of routes with zero indices of subsequent levels with decreasing lengths is built without branching. Moreover, the length at the end of the chain (before branching occurs), even if only one link actually turns out, should be less (or at least not more) than the length of the route with zero index of the "current-base" level.

Let us consider how this is implemented using the example of calculation, for a test with a dimension of 131 elements (Fig. 12, top row to the left) from the ASIC data set of the VLSI TSP collection (file xqf131.tsp) [20]. For convenience, we will use the numbering of the test elements (in this case, vertices or points) defined by a set of Cartesian coordinates, starting from 0 (and not from 1, as in the original source [20]).

Sample of construction 11 (from a two-member tour with the global index of 0). Among the routes of level 00, constructed from all possible combinations of initial two-member tours, the shortest length is 580.23 (Fig. 12 upper row in the center - for simplicity, we shall ignore digits after hundredths) and, accordingly, two of them have a global index of 0. The first one combines the vertices 28 66 and the second one has the vertices 28 62. Both assume that vertex 111 is included in the initial tour. Let us consider the constructions from the first one.

Thus, from the three-member tour (28 111 66 28) a minimum route is built (Fig. 12 upper row to the right) with a length of 579.99. It predetermines 3 variants of the inclusion vertices (25, 26 or 18) corresponding to the zero level index, and therefore along with them, according to scheme (6), other remaining vertices of the possible inclusion should be considered. Among all the directions, the inclusion of the vertex 34 (28 34 111 66 28) seems to be one of the most promising for further development. After it, a diminishing chain of lengths for complete routes is constructed with zero indexes of subsequent levels up to level 5, where there are three routes with zero index length 574.77 (Fig. 12 bottom row to the left) again. Therefore, the condition of the scheme (6) for the final length of the chain is less than the length of the route with the zero index (579.99) of the "current-base" level. Since then there are three inclusions (from zero level indices) into a seven-member tour (28 34 97 111 106 96 66 28), all remaining vertices with nonzero indices should also be considered. At this stage, there is a good potential for the development of directions at the inclusion vertices with zero indices (60 94 and 95). Consider the first of them - the inclusion of vertex 60 (28 34 97 111 106 96 60 66 28), from which a sequence of decreasing routes to level 8 is unambiguously constructed, where 36

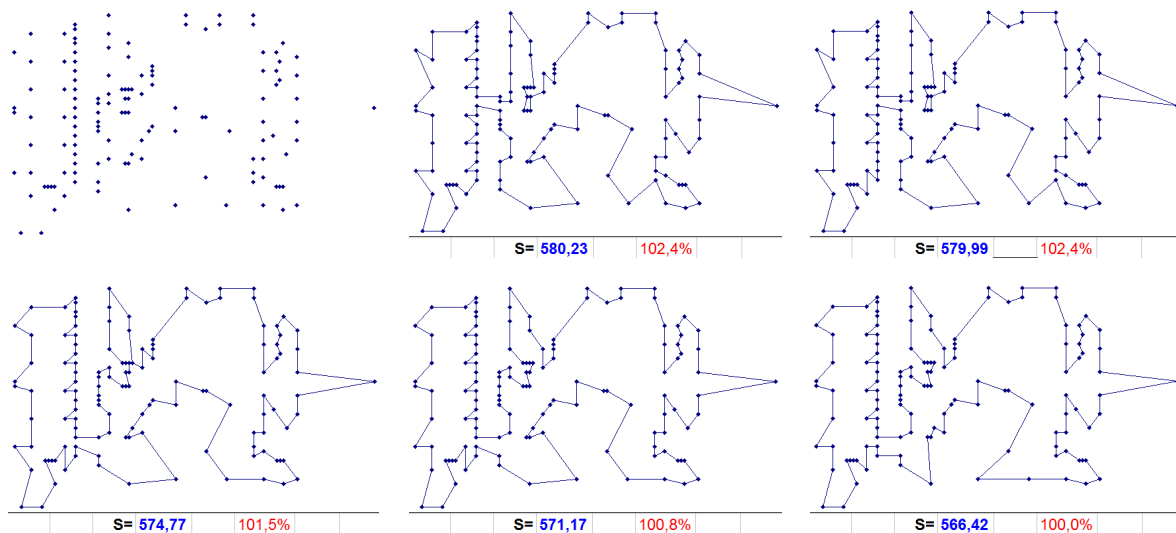


Figure 12. The mutual position of 131 points (top left). Record route length 566,42 (bottom right). The rest: a sequence of key intermediate routes. Additionally, the length and deviation relative to the record are shown.

directions (vertices) characterizing routes with a length of 571.17 with a zero index appear (Fig. 12 bottom row in the center). Any of them does not reduce the length of the route with a zero index on the next level. But from one of them - from the inclusion of vertex 93 in the ten-member tour (28 34 11 93 97 111 106 79 79 60 66 28), 44 more routes are formed with a zero index and with the same length. And here already (when constructing routes of the next level), a record 566.42 length is not improved (Fig. 12 bottom row to the right), from the inclusion of vertices corresponding to non-zero level indices. This happens when vertices 77, 87 (correspondence to index 4) or 13 (correspondence to index 9) are included between vertices 11 93 in tour 28 34 11 93 97 111 106 96 79 60 66 28.

Here is the record tour: 66 62 57 58 59 60 72 71 79 78 82 83 84 85 89 90 94 95 96 103 102 110 109 108 114 118 115 119 116 121 128 127 126 130 125 124 123 112 107 106 105 101 100 99 104 113 117 120 129 122 111 97 88 92 98 93 91 87 86 81 80 77 76 74 67 63 73 52 44 25 26 18 24 16 15 14 13 17 12 4 11 5 0 6 7 1 2 8 9 3 10 23 43 42 41 40 39 22 38 37 36 21 35 34 33 20 32 31 30 19 29 28 27 45 53 54 46 47 48 49 50 51 56 55 61 64 68 65 69 75 70 66.

It is achieved according to the scheme (6), at constructions from level 9 (00 – 163 – 20 – 30 – 40 – 50 – 60 – 70 – 80 – 9i, где i=4, 9), which is two levels earlier than at calculation according to scheme (5) from a two-member tour with the most distant cities in [15]. If the optimal route stated in [20] is calculated with an

accuracy that is used in applied calculations for CNC (conventionally with “computer” accuracy), then we obtain the length 567.20, i.e. Our record route was shorter. In the integer metric used in [20], our route will be one unit longer.

Sample of construction 12. Let us check whether the path of obtaining a record result in certain areas will be shorter if we use scheme (1) for calculations from the same two-member tour with a global index of 0, as in construction 11.

And this direction is found. Let us consider the same two-member tour which unites the vertices of 28 66, but with the vertex of embedding 34, which in the framework of scheme (1) corresponds to the construction from level 084 with a route of 630.94 in length (Fig. 13 on the left).

From it, there are unambiguous constructions of improving routes with zero indices of subsequent levels up to the formation of routes of level 5. Where it appears after being included in the six-membered tour of vertex 11 (28 34 11 13 80 86 66 28), 33 routes with zero index that duplicate the route of level 40 (Fig. 13 in the center).

From it, there are unambiguous constructions of improving routes with zero indices of subsequent levels up to the formation of routes of level 5. Where appear, after being included in the six-membered tour of vertex 11 (28 34 11 13 80 86 66 28), 33 routes with zero index that duplicate the route of level 40 (Fig. 13 in the center).

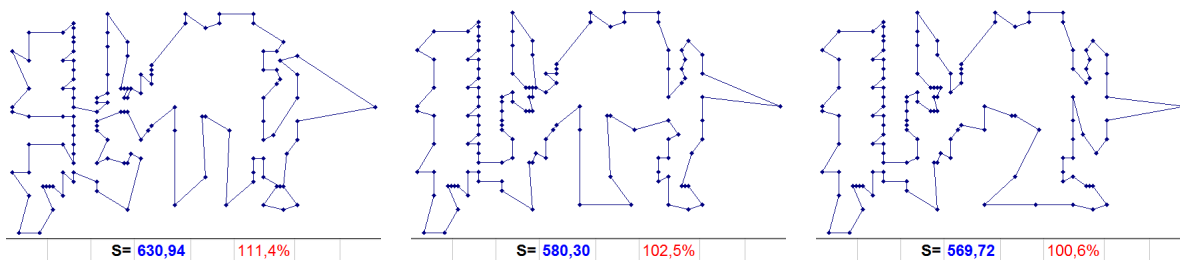


Figure 13. The sequence of key intermediate routes. Additionally, the length and deviation relative to the record are shown.

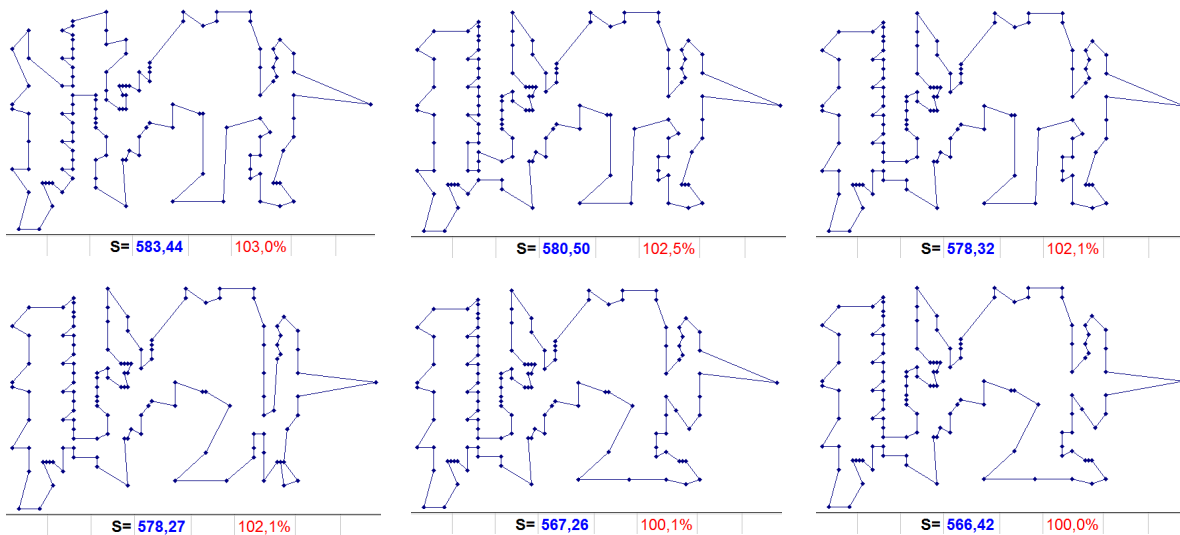


Figure 14. The sequence of key intermediate routes leading to the record. Additionally, the length and deviation relative to the record are shown.

From one of them (as a result of consecutive unambiguous inclusions in the seven-member tour 28 34 11 13 80 86 66 28 vertices 122 106 102), a chain of routes with zero indices of subsequent levels with a minimum length of the resulting route equal to 569.72 is built without branching (Fig. 13 on right). Moreover, there are two such routes and two new directions respectively from the inclusion of vertices 100 and 99 (28 34 11 13 80 86 122 100 106 102 66 28 - the first, and 28 34 11 13 80 86 122 99 106 102 66 28 - the second). Both directions, in order to reach the record route with a length of 566.42 (Fig. 12 bottom row to the right) assume that the 75 vertex is definitely turned on between 28 and 66, and at the next level 100 - alternative inclusions of the vertex 108 (between 102 and 106).

Sample of construction 13. The record route can be calculated at earlier levels by a repeated number of times and also according to a less laborious scheme (1), if not limited to the calculation, only from the global zero index.

So, for example, among the routes built from all possible combinations of the original two-member tours, the route from the two-member 6 6 6 level 00 with a length of 583.44 (Fig. 14 at the top left) has a global index of 16. It alone determines the inclusion of the 87 in the two-member tour.

Then, there is an unambiguous inclusion in the three-member tour of the vertex 54 (6 87 89 54 6), after which 9 routes appear with the index 0 level 2 with a length of 580.50 (Fig. 14 in the center above). Among the corresponding vertices of inclusion, only embedding vertices 4 and 112 reduce the route to the next level. Let us track the first of them (6 4 87 89 54 6).

From it, there is an unambiguous reduction in the length of the routes to the formation of tours of level 5, where 8 routes appear with the index 0, length 578.32 (Fig. 14 at the top right). The corresponding embedding in each of the tours of level 50 again multiplies routes with index 0 of the next level (an average of 8 routes without a change in length). And only embedding already in one of the many tours of level 60 leads to a decrease in the length of the route of level 70 to 578.27 (Fig. 14 at the bottom left).

This reduced route resulted from the inclusion of the vertex 88 in the eight-membered tour (6 0 11 4 87 88 97 89 54 6). The constructed route predetermines three possible embeddings in the tour 6 0 11 4 87 88 97 89 54 6 peaks 98 93 or 91.

Each of these inclusions gives a new reduced route with a length of 567.26 (Fig. 14 in the center below), and, accordingly, 3 initial tours with a unique inclusion of vertex 106. This inclusion in any of the 3 tours immediately gives a record route (Fig. 14 at the bottom right), which repeatedly manifests itself in different directions when building at level 8. Thus, up to a record from the initial two-member tour 6 89 6, it was built on the levels 00 - 10 - 20 - 30 - 40 - 50 - 60 - 70 - 80 in accordance with the scheme (1).

The record route can be obtained even earlier - at level 70. With similar constructions according to scheme (1), from a two-member tour, 89 87 89 with a sequence number (lengths of level 00 routes, arranged in ascending order) are approximately midway between the same numbers for "sample build 11" and "sample build 13".

6. Conclusion

A significant number of two-member tours (proportional to the square of the dimension of the problem), practically from each of which with varying degrees of difficulty can be built a record route, at least close to the optimal result, allow us to draw some preliminary conclusions based on the results of a set of calculations.

It can be argued with a high degree of probability that a significant layer of applied problems associated with the efficient operation of CNC equipment, equivalent in setting to a symmetric traveling salesman problem, can be solved using the method of step-by-step approach to the minimum route.

Moreover, many problems with a dimension of up to 50 elements, close to the present limit of the possibilities of exact methods, can be solved with obtaining an optimal result already at an early stage of calculations. Namely, when calculating at the level of possible combinations of two-member (formation and ranking of routes of zero level) or three-member (formation of routes of level 10) tours.

For problems of greater dimension, especially over 100 elements, it is advisable to reduce the calculations using one of the proposed schemes for organizing calculations (or its modifications). And as experiments show, you can use a more favorable criterion for stopping calculations in the current direction, for example, not threefold (as recommended), but double "repetition of the length of routes". In most cases, this will only lead to a reduction in the directions of calculations leading to a record result, but on the whole, it will accelerate (n - fold) the calculation process.

References

- [1] Nikolay Starostin, Konstantin Mironov: Algorithm Modification of the Level-by-Level Approximation to the Minimum Route, 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 270-275, December 2017.
- [2] N. D. Starostin The method of level-by-level exceptions for optimization of tool movements (in Russian) Collection of materials of the Second International Conference "Intelligent Technologies for Information Processing and Control", Ufa, 2014.
- [3] N. D. Starostin. Improvement and detailing of the method of level-by-level exceptions for optimization of tool movements (in Russian) // Information Technologies and Systems [Electronic resource]: tr. Fourth Intern. sci. Conf., Bannoe, Russia, Feb. 25. - March 1, 2015, pp. 33-44
- [4] I. I. Melamed The traveling salesman's problem. Exact methods (in Russian) II Melamed, SI Sergeev, I. Kh. Sigal // Avtomat. and telemekh. - 1989. - No. 10. - P. 3-29.
- [5] V. V. Borodin Experimental study of the effectiveness of heuristic algorithms for solving the traveling salesman problem (in Russian) VV Borodin, SE Lovetskii, II Melamed, Yu. M. Plotinsky // Avtomat. and Telemekh., 1980, No. 11, 76-84.
- [6] Starostin N. D. Finding the shortest path. Accelerated optimization of tool movements - Saarbrücken: LAP LAMBERT Academic Publishing, 2015. - 89 p.
- [7] N. D. Starostin Variants of calculation of the minimum route of processing (in Russian) Information technologies and systems [Electronic resource]: tr. Fifth International. sci. Conf., Bannoe, Russia, 24 February. - February 28, 2016, pp. 42-48.
- [8] N. D. Starostin Basic variation of the algorithm of level-by-level approximation to the minimal route [Electronic resource, in Russian]: tr. The Sixth Intern. sci. , Bannoe, Russia, March 1 - 5, 2017 P.295-302.
- [9] N. D. Starostin , DV Kurenov Level-by-level approach to the minimum route // International scientific and technical conference "Prom-Engineering" in 2017 was held May 16-19 in St. Petersburg.

- [10] MP-TESTDATA - The TSPLIB Symmetric Traveling Salesman Problem Instances – URL: <http://elib.zib.de/pub/mp-testdata/tsp/tsplib/tsp>. (visited on 08.02.2016).
- [11] Lmatrix Laboratory Featured articles Solutions - URL: <http://lmatrix.ru/news/problems> (reference date: 01/05/2016)
- [12] I. I. Melamed The traveling salesman's problem. Approximate algorithms (in Russian) / II Melamed, SI Sergeev, I. Kh. Sigal // Avtomat. and telemeh. - 1989. - No. 11. - P. 3-26
- [13] Golden B. L., Skiscim C. C. Using simulated annealing to solve routing and location problems // Nav. Res. Log. Quart. 1986. V. 33. № 2. P. 261-279
- [14] National TSP – URL: <http://www.math.uwaterloo.ca/tsp/world/countries.html> (visited on: 05.01.2017)
- [15] Yu. A. Kochetov, Computational possibilities of local search in combinatorial optimization, Zh. Vychisl. Math. and Math. fiz. - 2008 - Vol. 48, No. 5 - P. 788-807.
- [16] An Interview with Vasek Chvatal. – URL: <https://www.cs.rutgers.edu/~mcgrew/Explorer/1.2/#Farmers> (visited on 16.08.2017)
- [17] Kochetov Yu., Mladenovich N., Hansen P. Local Search with Alternating Neighborhoods // Discrete Analysis and Operations Research. Series 2. 2003. Volume 10, No. 1. P.11-43.
- [18] Metaheuristics: From Design to Implementation. El-Ghazali Talbi. John Wiley & Sons, 2009, - 624 p
- [19] R. T. Murzakaev, V. S. Sholov, A. V. Burylov. Application of Metaheuristic Algorithms for minimizing the path of a cutting tool (in Russian) Transactions of PNIPU University on Electrical Engineering, IT, and Control. 2015. №14. C.123-133
- [20] VLSI TSP Collection – URL: <http://www.math.uwaterloo.ca/tsp/vlsi/index.html> (дата обращения: 31.08.2017)
- [21] MP-TESTDATA - The TSPLIB Symmetric Traveling Salesman Problem Instances – URL: <https://people.sc.fsu.edu/~jburkardt/datasets/tsp/tsp.html> (дата обращения: 06.02.2018)
- [22] In Pursuit of the Traveling Salesman : Mathematics at the Limits of Computation. Author(s), Cook, William J. Publication, Princeton, NJ : Princeton University Press, 2011. - 245 p.

Over-The-Air Testing of Automotive Antennas and Wireless Links in The Installed State on The Basis of LTE Downlink Communication Parameters

Philipp Berlt*, Lisa Jäger, Andreas Schwind, Frank Wollenschläger, Christian Bornkessel, Matthias A. Hein

Thuringian Centre of Innovation in Mobility, RF and Microwave Research Laboratory, Technische Universität Ilmenau, 98693 Ilmenau, Germany

ARTICLE INFO

Article history:

Received: 20 December, 2018

Accepted: 28 January, 2019

Online : 20 February, 2019

Keywords:

Antenna measurement

LTE

Over-the-air

ABSTRACT

Modern automobiles have been turning more and more into wireless sensor and communication networks. Accordingly, the number of radio systems is steadily increasing. Due to strict safety requirements, these radio systems need to be tested extensively for functionality and reliability, especially under poor radio channel conditions. Beside the large electrical size of cars at frequencies of mobile communication services, access to the antennas imposes challenges for testing, due to an increasingly high integration of the antennas with frontends and digital signal processing. This paper proposes an over-the-air testing procedure for automotive radio systems on the basis of the wireless communication standard Long Term Evolution (LTE). A method to derive the radiation patterns of automotive antennas from reference signals in the LTE downlink scheme without requiring any access to the analogue RF feed point is proposed. A comparison of the LTE approach with the usual antenna measurement techniques shows good agreement. As a logical step from the antenna towards a complete wireless link, a concept of spatially distributed channel emulation on the basis of Software Defined Radio (SDR) modules is proposed, aiming at the emulation of essential multipath features of the wireless channel. RF measurements of the channel transfer function as well as by over-the-air end-to-end tests prove this approach to be a cost-efficient alternative to commercial channel emulators.

1. Introduction

This paper is an extension of work originally presented at the 2018 IEEE 87th Vehicular Technology Conference [1]. The paper deals with automotive antenna radiation pattern measurements on the one hand, and with end-to-end communication testing on the other hand. Both subjects are closely related, as the antenna performance has an essential impact on the overall system performance. Moreover, wireless signal transmission is strongly affected by the radio channel conditions. As a consequence, end-to-end testing of the entire communication link, including application antennas as well as radio channel conditions, is essential in order to ensure reliable functionality in real applications.

Connected and highly automated driving is a key to future mobility, especially in terms of increasing safety and improving traffic flow. Vehicle-to-vehicle (V2V) or vehicle-to-infrastructure (V2I) communication, generalized as vehicle-to-everything (V2X), has been of great interest in research and development and pushed forward by industry organizations like the 5G Automotive

Association (5GAA), Car to Car Communications Consortium (C2CCC), or the Virtual Drive Test Alliance (VDT Alliance) [2][3][4]. There has been a controversial discussion about optimal technologies for the implementation of V2X networks. Beside WLAN-related networks like IEEE 802.11p, derivatives of the LTE standard are under consideration [5]. This paper focuses on the latter due to its greater variability in terms of cellular and ad-hoc networking, and in terms of its applications in mobile communications, data transmission, and V2X communications. When moving towards higher levels of driving automation, the number of sensor nodes and wireless systems in modern automobiles will increase significantly, e.g., in order to enable cooperative perception or manoeuvring. Satellite-based navigation, automotive radar sensing, but also mobile communication systems will provide altogether the basis for smart cooperative connected cars. Due to the increasing amount of wireless air interfaces, the installation space for antennas is seriously limited and, thus, the future of antennas lies in integrated modules, where radiating elements, frontends, and signal processing units are merged. As a consequence, access to the analogue RF antenna feed cannot be taken for granted any more.

*Corresponding Author: Philipp Berlt, E-Mail: philipp.berlt@tu-ilmenau.de

Consequently, conventional antenna measurement techniques, e.g. far-field measurements or nearfield measurements with subsequent nearfield-to-far-field transformation [6] [7], seem unsuitable for such integrated systems. Hence, alternative test procedures need to be identified and investigated. In this paper, we propose a concept that allows for the derivation of automotive antenna gain patterns in their installed state on the basis of LTE communication parameters, monitored by a commercial communication modem with USB drive. Instead of a wired RF connection to the antenna feed, a serial connection to a laptop computer is required, in order to retrieve information about the signal quality from the modem.

In addition to the radiation patterns of the antennas, the overall performance of automotive wireless communication systems needs to be tested, in order to ensure functionality even under poor mobile channel conditions. Drive tests have been the procedure-of-choice for the verification and validation of automotive systems. However, these tests are expensive and time consuming and, due to changing conditions in a real-world environment, reproducibility cannot be maintained. Therefore, time- and cost-efficient test procedures under reproducible laboratory conditions need to be investigated, leading to the concept of virtual drive tests. One approach for testing mobile devices under realistic radio channel conditions is the electromagnetic wave field synthesis. However, this approach is not feasible for electrically large objects like cars, as the size of the test zone is strictly limited by the available number of illumination antennas and channel emulators [7]. Therefore, in this paper, we follow a cluster-based concept for spatially distributed channel emulation, which has been implemented using commercially available Software Defined Radio (SDR) modules like the Universal Software Radio Peripheral (USRP) [8]. The functionality of the channel emulator has been demonstrated by end-to-end LTE measurements as well as by RF measurements of the emulated transfer function. All measurements were conducted in our multipurpose test facility virtual road – simulation and test area (VISTA).

2. Automotive test facility “Virtual Road Simulation and Test Area”

The unique automotive antenna test facility VISTA was installed at the Technische Universität Ilmenau under the roof of the Thuringian centre of innovation in mobility [9]. It is a multi-purpose test and measurement facility, including over-the-air testing and channel emulation for automotive wireless systems, three-dimensional radiation pattern measurements of antennas in their installed state, and electromagnetic compatibility testing. The facility comprises a shielded semi-anechoic chamber, a spherical nearfield antenna measurement system, a roller dynamometer integrated into an EMC-compatible turntable, and powerful software tools for the digital post-processing of the measured raw data. The walls and the ceiling of the chamber are covered with 60-inches pyramidal absorbers, designed for a lower frequency limit of 70 MHz. The metallic floor may be left blank, e.g. for radiated emission measurements, or covered with absorbers, e.g. for antenna measurements. The antenna measurement system is a spherical nearfield system, consisting of a multi-probe antenna arch for rapid electronically switched elevation scanning, a turntable for mechanical scanning along the azimuth, and a vehicle lift for elevating the car-under-test to the centre of the measurement arch at a height of 2.30 m above ground, as depicted in Figure 1. The left-hand semi-arch is equipped with 111 antennas covering the elevation range from 0° to 20° below horizon in 1°

steps, for the frequency range from 400 MHz to 6000 MHz. The right-hand semi-arch contains 22 antennas from 2.5° to 17.5° below horizon with a spacing of 5°, for the frequency range from 70 MHz to 400 MHz. The resulting far-field radiation patterns are obtained by applying nearfield to far-field transformation algorithms. A more detailed description of VISTA can be found in [10].

3. Derivation of automotive antenna gain patterns from LTE downlink communication parameters

In this section, the procedure to derive gain patterns of automotive antennas in the installed state from LTE downlink parameters is described. In contrast to conventional antenna measurement techniques, the approach benefits from the fact that no RF cable connection to the antenna-under-test is required, rather than a serial connection to the LTE modem, which is interrogated by the connected laptop.

3.1. LTE hardware and transmission parameters and measurement setup

The commercially available communication tester CMW500 (Rohde&Schwarz) was used to establish the LTE link between an emulated base station and the LTE user equipment in VISTA [11], according to the setup outlined in Figure 2. This instrumentation allows for a full control of downlink and uplink parameters in terms of bandwidth, power levels, and resource allocation. Single-input single-output (SISO) as well as multiple-input multiple-output (MIMO) modes are supported for up to two receive and transmit antennas (2x2 MIMO). The antenna measurements described in this paper were conducted in SISO mode. The mobile terminal was a Huawei modem of type E3276s-150 (LTE category 4), connected to a laptop with a universal serial bus (USB). External monopole antennas were installed on the roof of a mid-sized passenger car mock-up and connected to the modem (see Figures 3 and 4). The modem includes one antenna for receive only, and a second antenna for reception and uplink signal transmission. Several power-related parameters, such as signal-to-interference-and-noise ratio (SINR), received signal strength indicator (RSSI), and reference signal received power (RSRP), are provided by the modem.



Figure 1: Photograph of VISTA with a car installed for automotive antenna measurement: Multi-probe antenna arch for the frequency range from 400 MHz up to 6 GHz (Ia) and from 70 MHz to 400 MHz (Ib). II: A customized vehicle lift elevates the car-under-test to the centre of the measurement arch 2.30 m above ground. III: The EMC-compatible turntable is used for azimuthal rotation with an accuracy of 0.1°.

RSSI is a measure of the total power received across the selected bandwidth. Its value depends on interference, noise, and channel traffic. In contrast, RSRP quantifies the average received power of

dedicated reference signals (RS) in the LTE downlink. The reference signals are used for downlink channel estimation and are transmitted with constant power, independent of channel traffic [12]. For this reason, RSRP was selected as the parameter optimally suited for antenna gain measurements.

In order to conduct reliable measurements with the user equipment, the input/output power relation of the UE needs to be calibrated. The modem was characterized in terms of linearity and dynamic range in a wired measurement setup as depicted in Figure 5. Figure 6 displays the results for the I/O relation of the modem.

The downlink transmit power, given in (dBm/15 kHz), was reduced stepwise by 0.25 dB and the received power was detected by the user equipment. The measurements were performed at a carrier frequency of 2655 MHz with 1.4 MHz channel bandwidth, which is the smallest bandwidth supported by the LTE standard. The bandwidth is a measure of the frequency resolution in the RSRP measurement. As expected, the RSRP follows the power

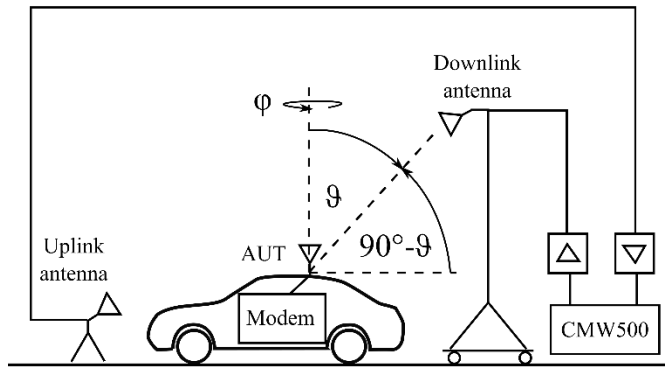


Figure 2: Sketch of the RSRP-based measurement setup. The downlink horn antenna is installed on a height-adjustable antenna mast. Depending on the elevation angle, $90^\circ - \theta$, the height may be set between 1 m and 4 m. The distance between the downlink transmit antenna and the centre of rotation is 7 m. An additional uplink antenna is installed near to the car, in order to provide the feedback channel from the modem to the communication tester. The implemented setup is discussed below and can be seen in Figure 7.

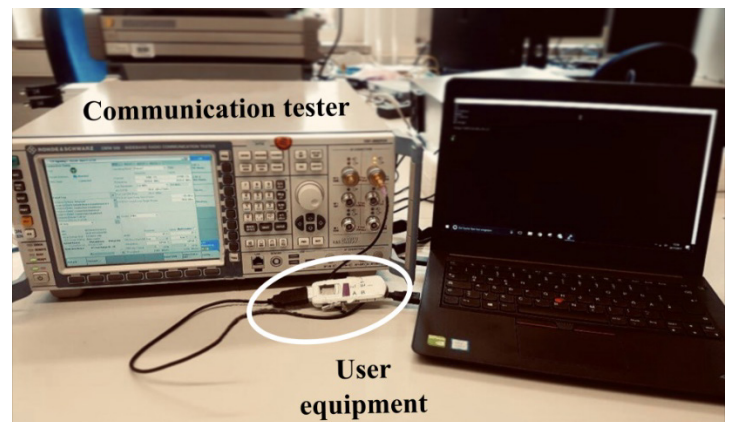


Figure 5: Setup for the measurement of the I/O relation of the LTE modem (right-hand side) with the communication tester (left-hand side). The downlink transmit power is detected by the user equipment.



Figure 3: Photograph of the antenna-under-test, installed in VISTA. The AUT (indicated by the black circle) is connected to the Tx/Rx path of the UE, while the second Rx path is matched with a 50 Ω load.

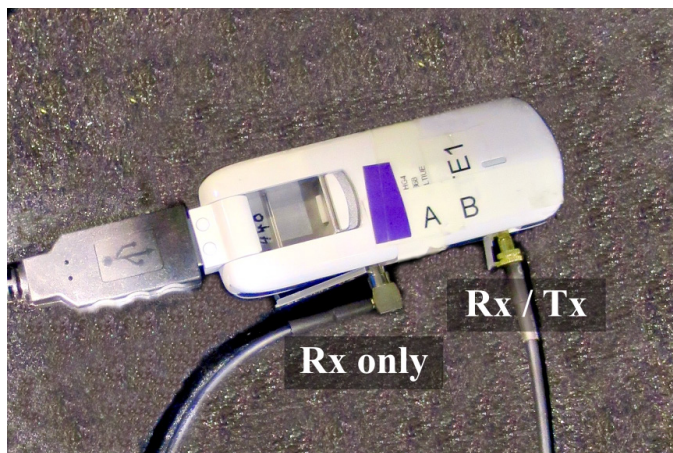


Figure 4: Photograph of the USB drive LTE modem, connected to external monopole antennas on the roof of the mock-up of a mid-size passenger car (see Figure 3). The AUT is connected to the Tx/Rx path of the UE, while the second Rx path is matched with a 50 Ω load.

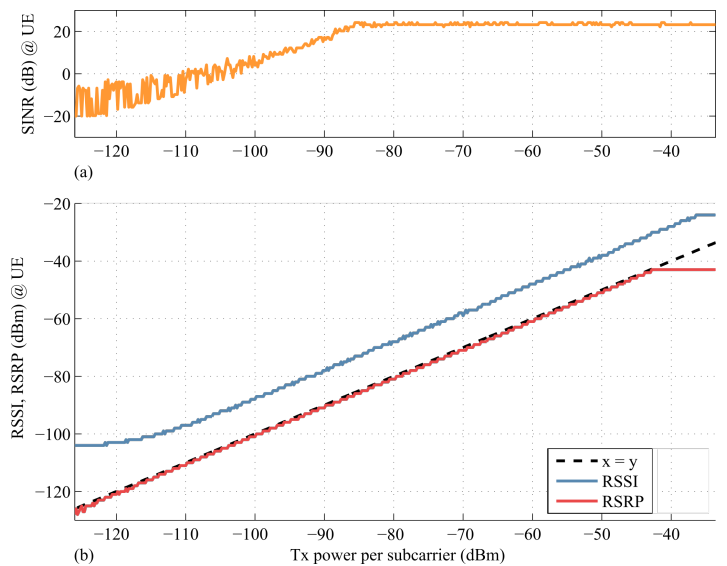


Figure 6: Typical transmission curves measured with the LTE modem. (a): The SINR (orange curve) increases linearly with the input power above the noise floor of about -100 dBm, and reaches a saturation level of 23 dB above -85 dBm. (b): The RSRP (red curve) follows the configured input power linearly for values less than -43 dBm, as indicated by the dashed black curve. The RSSI (blue curve) reaches saturation at approximately -110 dBm input power.

values set by the CMW500 very well, as it is a measure of the average power per RS resource element. Enabling detection of power levels less than -120 dBm, the RSRP offers high sensitivity, whereas the RSSI reaches saturation at -110 dBm input power. The measured data have a quantisation of 1 dB, set by the UE.

3.2. Radiation patterns derived from RSRP measurements

Antenna pattern measurement setup

A LTE link was established in VISTA between the communication tester and the antenna-under-test. The downlink signals were transmitted with a horn antenna (Schwarzbeck BBHA 9120 A) with a gain of 9.3 dBi, which was installed on a height-adjustable antenna mast (see Figure 2), in order to achieve various elevation angles-of-arrival. A second antenna was installed near the car, in order to receive the uplink signals transmitted by the user equipment. The turntable with the car was rotated stepwise, and the RSRP was evaluated as a function of the azimuth angle-of-arrival. The implementation is visualized in Figure 7. The distance between the downlink antenna and the AUT was $R_0 = 7$ m. Taking into account the height of the car and the antenna mast, elevation angles up to 15° were adjusted. The gain of the receive antenna G_{Rx} (AUT), in decibels, can be calculated from the RSRP according to [13]:

$$G_{Rx} = RSRP - P_{Tx} - G_{Tx} + L_{Cbl} + PL \quad (1)$$

P_{Tx} denotes the transmit power in dBm per 15 kHz, G_{Tx} denotes the gain of the horn antenna (downlink antenna) in dBi, L_{Cbl} and PL are cable losses and the free-space path loss in dB, respectively. For given wavelength λ and locations of transmit and receive antennas, the path loss is computed at each turntable angle ψ :

$$PL(\psi) = 20 \cdot \log_{10} \left(\frac{4\pi}{\lambda} \cdot |\vec{R}_z(\psi) \cdot \vec{p}_{Rx,0} - \vec{p}_{Tx}| \right) \quad (2)$$

Chosen the center of rotation of the turntable as the origin of the coordinate system, $R_z(\psi)$ denotes the rotation matrix around the z-axis, p_{Tx} is the position of the transmit antenna, and $p_{Rx,0}$ denotes the initial position of the AUT at turntable position $\psi = 0^\circ$. The location of the receive antenna is derived from phase center calculations, as further described in [14].

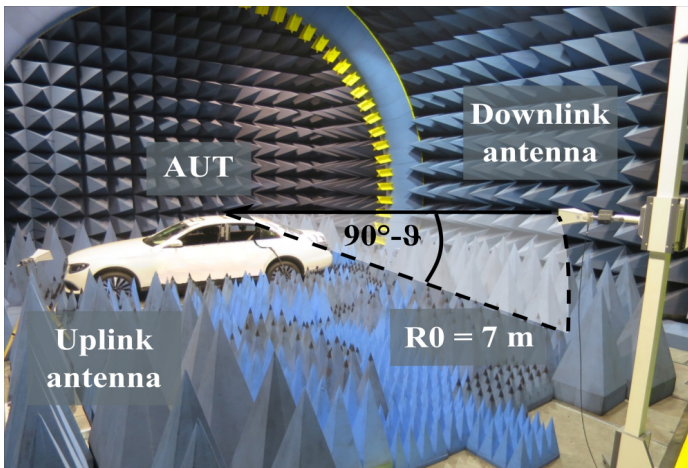


Figure 7: Photograph of the setup for the RSRP and VNA measurements. In order to suppress reflections from the metallic ground, the path between transmit antenna and antenna-under-test was covered with absorbers.

Comparison between RSRP approach and direct power transmission measurements with a network analyser

In order to validate the RSRP measurements, the gain of the AUT was additionally derived from calibrated power transmission measurements with a vector network analyser. The communication tester was replaced by the VNA, and the uplink antenna became obsolete. The positions of the transmit antenna and the AUT remained the same as in the previously described setup. The transmission factor $|S_{21}|_{dB} = 10 \cdot \log_{10}(|S_{21}|^2)$ was recorded at each angular step for a full 360° rotation of the car-under-test and, according to (3), the receive antenna gain was calculated from the measured transmission coefficient.

$$G_{Rx} = |S_{21}|_{dB} - G_{Tx} + L_{Cbl} + PL \quad (3)$$

Figure 8 displays horizontal cuts of the vertically polarised radiation pattern of the AUT at a co-elevation of $\vartheta = 10^\circ$ and at a frequency of $f = 2655$ MHz, derived from both setups. A more detailed comparison, in terms of mean and variance of the error, is depicted in Figure 9 for a range of elevation angles. While the mean value indicates a static offset between the two datasets, the variance of the deviation reflects the similarity between the shapes of the patterns since it ignores a possible offset in the data. In order to express the similarity of two patterns by a single parameter, the cosine similarity can be calculated from (3) in analogy to antenna correlation [15].

$$\rho = \frac{\sum_m \sum_n G_1(J_m, \varphi_n) \times G_2(J_m, \varphi_n) \times \sin(J_m)}{\sqrt{\sum_m \sum_n G_1^2(J_m, \varphi_n) \times \sin(J_m)} \times \sqrt{\sum_m \sum_n G_2^2(J_m, \varphi_n) \times \sin(J_m)}} \quad (4)$$

G_1 and G_2 denote the gain patterns #1 and #2 on a linear scale. The weighting factor $\sin(\vartheta_m)$ compensates for the irregular spacing of the samples across the measurement sphere, given by ϑ_m and φ_n . According to (4), ρ can take values between 0 and 1, where a value of 0 implies orthogonal patterns and a value of 1 implies identical patterns, except for a scaling factor. Both datasets show good agreement regarding the angle-dependent shape of the radiation pattern, indicated by the low variance of the error (≤ 1 dB) and a high cosine similarity above 98%.

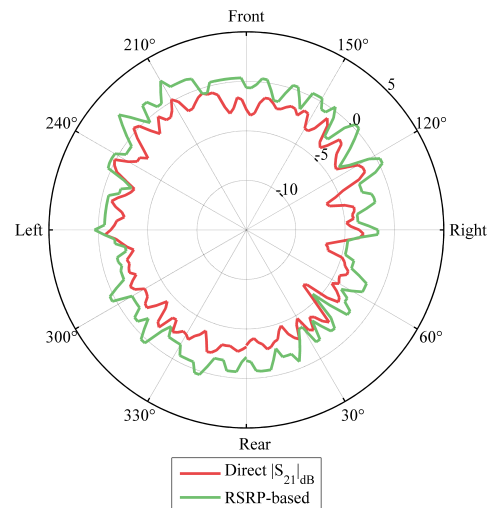


Figure 8: Comparison of the azimuth gain pattern measured via network analyser (red curve, $G_{max} = 0.4$ dBi) and RSRP data (green curve, $G_{max} = 2.4$ dBi) for an elevation angle of 10° .

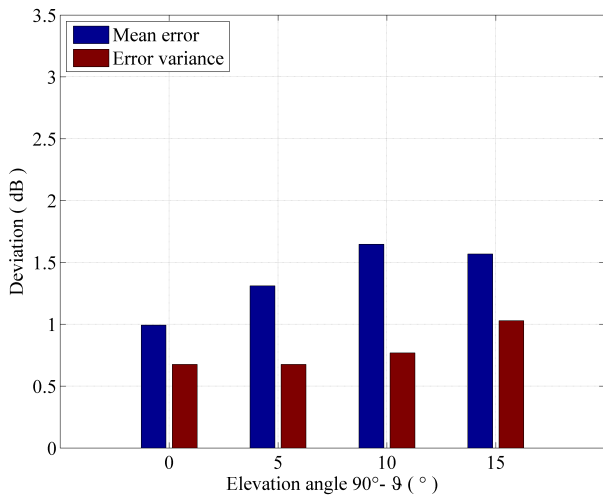


Figure 9: Bar diagram of the mean value (blue) and the variance (red) of the measurement deviation (in dB) between the direct power transmission measurement and the RSRP-based approach.

The RSRP approach tends to yield slightly higher gain values than the direct power transmission approach. Higher signal levels might be explained by measurement uncertainties during calibration and by the modem itself. Path loss and transmit antenna gain from (1) and (3) can be excluded as a reason for uncertainty between both datasets since kind and position of the antennas remained the same in the measurements and the same values were used in data analysis.

Comparison between RSRP approach and far-field pattern obtained from nearfield to far-field transformation

In addition to the RSRP and VNA measurements and the proof of their consistency, the AUT was measured with the spherical nearfield system in VISTA. For this purpose, the car mock-up was elevated, in order to place the antenna close to the phase centre of the measurement arch, as indicated by Figure 1. The far-field radiation pattern was derived from a nearfield to far-field transformation. The three-dimensional radiation pattern for vertical polarisation is depicted in Figure 10.

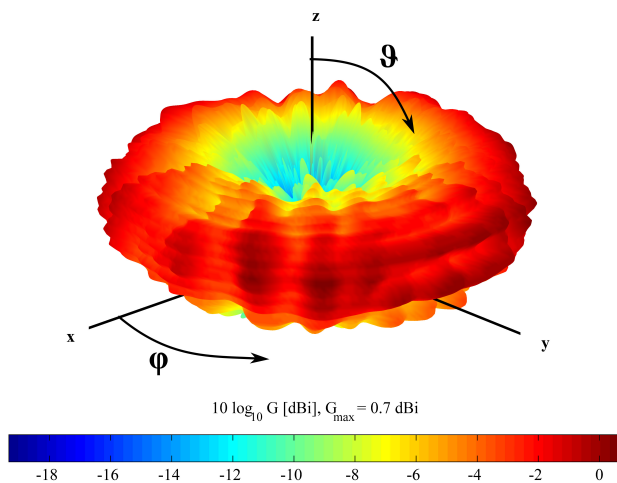


Figure 10: Three-dimensional realised-gain pattern derived from a nearfield to far-field transformation of the AUT for vertical polarisation at a frequency of 2655 MHz. The maximum gain value is 0.7 dBi.

The coordinate system was oriented as usual for automotive antenna measurements: the y-axis points along driving direction, and the z-axis to the zenith. The maximum realised gain was 0.7 dBi.

Since the vehicle was lifted to the centre of the measurement arch, a comparative measurement with the network analyser was conducted, in order to ensure comparability between the two setups with the car located on ground level and lifted, respectively. Due to the limited height of the antenna mast, just a single cut of the pattern for an elevation of 0° could be recorded with the direct power transmission method in the elevated state. Figure 11 shows that the two data sets compare favourably. The comparison between the RSRP and the far-field patterns is displayed in Figure 12, depicting the horizontal cuts of the gain patterns at an elevation of 10° . Figure 13 analyses the deviation between the data sets for the elevation angles considered. The patterns display very similar shapes, as quantified by a cosine similarity of 96 %.

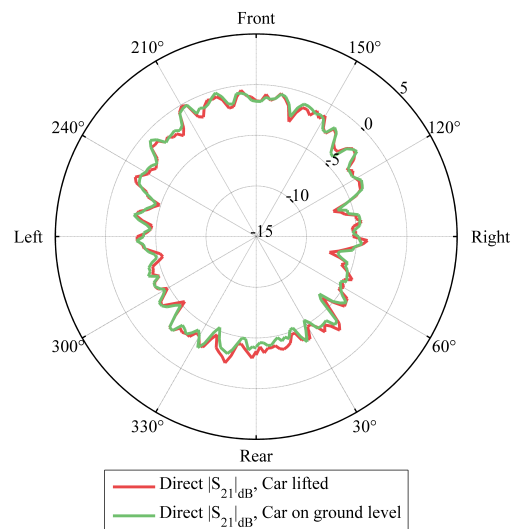


Figure 11: Comparison of the measurement setups in elevated state (red curve) and on ground level (green curve). The curves agree at all angles within 1.1 dB.

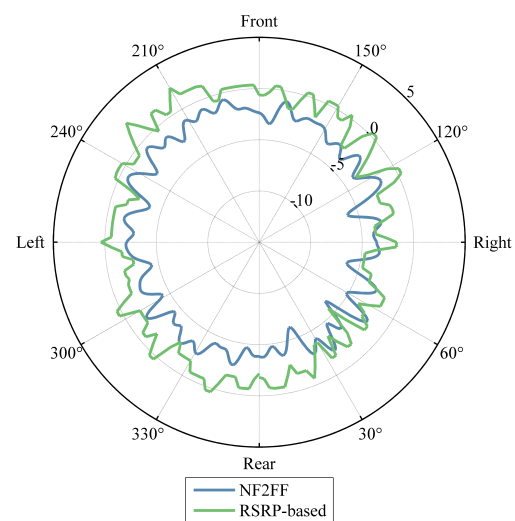


Figure 12: Comparison of the horizontal cuts at an elevation angle 10° of the gain patterns derived from the RSRP approach (green curve, $G_{\max} = 0.4$ dBi) with the dataset obtained from nearfield to far-field transformation (blue curve, $G_{\max} = 2.4$ dBi).

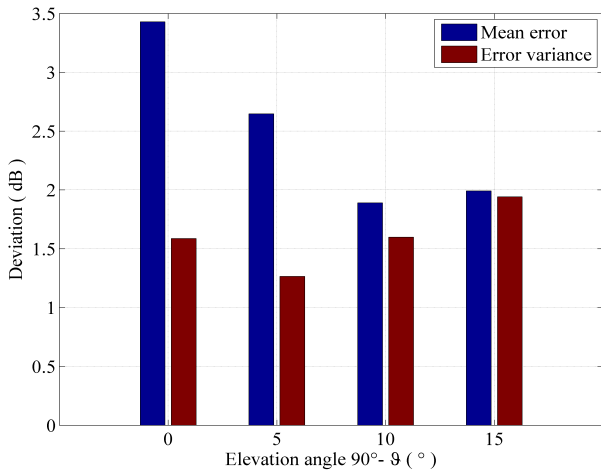


Figure 13: Bar diagram of the mean value (blue) and the variance (red) of the measurement error in dB between the RSRP nearfield pattern measurements and the nearfield to far-field transformation.

Summary of LTE antenna pattern measurements

Three types of measurements were performed, in order to derive horizontal cuts of the gain pattern of an automotive antenna: direct power transmission, nearfield measurement with subsequent nearfield to far-field transformation, and LTE-based RSRP measurement. The patterns derived from the three data sets compared very well in their shape. The small deviations observed between the absolute gain values of the three data sets may have been caused, in addition to the usual measurement uncertainties, by different post-processing of the measured signals and through reception of scattered signals by the user equipment within the car, partially bypassing the external antenna path. The user equipment was utilised as a measurement instrument, in order to evaluate downlink-specific reference signals and thus to obtain information about the received signal strength. It could be shown that the gain of the antenna can be derived from measured RSRP values without requiring access to the RF feed. Knowledge of the antenna gain is sufficient for many applications. The retrieval of phase information, following approaches like [16-18] will be in the focus of future work.

4. Over-the-air end-to-end testing of LTE links in a virtual electromagnetic environment

The overall wireless link performance of a device-under-test (user equipment) is affected by the wireless propagation channel, the antenna patterns involved in downlink and uplink communication, and the digital signal processing at both sides of the link. In order to exploit channel capacity more efficiently, multiple antennas may be installed on transmit and receive side (MIMO).

MIMO systems allow for distinctly higher data rates under favourable radio channel conditions, deploying multiple data streams by spatial multiplexing; in case of poor signal-to-noise ratio, transmission may be more robust against distortions through transmit diversity [19]. This section addresses end-to-end testing of LTE-based automotive wireless communication systems, identifying the realized data throughput as the figure-of-merit, as this describes the quality-of-service experienced by the end user best. The testing strategy follows a deterministic approach, based on the emulation of basic parameters of the wireless channel like propagation delay and Doppler shift. In contrast to stochastic fading models, physical propagation parameters are mapped

deterministically onto a number of N illumination antennas. Furthermore, LTE-specific features like link adaptation and their effect on the link quality are considered.

4.1. Emulation of basic features of a wireless multipath channel

Mobile communications are characterized by multipath propagation. The signal received at the user equipment results from a superposition of many multipath components, including line-of-sight (LOS) and non-line-of-sight (NLOS) components. The latter result from reflections, scattering, refraction, or diffraction, and can be characterized by parameters like propagation delay, Doppler shift, amplitude variation, and phase shift. Additionally, spatial properties of the channel, like angles-of-arrival and angles-of-departure, need to be taken into account, especially when MIMO systems are considered. A time- and cost-efficient test procedure is favoured, reducing the amount of required hardware and software, but still projecting relevant real-world features onto the test scenario. Complexity may be reduced by aggregating several multipath components with similar properties into so-called clusters, either in the spatial domain or by joint evaluation of the angular direction-delay domains [20] [21].

In this study, channel emulation has been applied to the downlink signal by employing a number of Software-Defined Radio (SDR) modules (Universal Software Radio Peripherals, USRP) of type NI USRP 2954-R. For this purpose, an emulator previously developed on the basis of a PXI computer platform, developed by National Instruments, was adapted for the application of radio channel emulation synchronised between several USRP devices [22] [23]. The emulator architecture is divided into parameter generation on a host computer, and digital signal processing on the internal FPGA of the USRP. The USRP is controlled via a PCIe connection to the host PC, which allows for fast exchange of data and control traffic. As shown in Figure 14, each SDR module provides two RF channels, including digital down conversion (DDC) and digital up conversion (DUC), as well as the application of propagation parameters to the baseband input signal. The discrete impulse response of one specific path n of the distributed emulator can then be described by

$$h_n(\tau, t) = a_n \times e^{j(2\pi f_{D,n}t - \varphi_{0,n})} \delta(\tau - \tau_n) \quad (5)$$

with the Dirac delta function $\delta(\tau)$ [24], the path coefficient a_n , the static phase shift $\varphi_{0,n}$, the delay τ_n , and the Doppler shift $f_{D,n}$. Different directions-of-arrival can be emulated by a proper geometrical arrangement of the illumination antennas, placed around the device-under-test at a distance R.

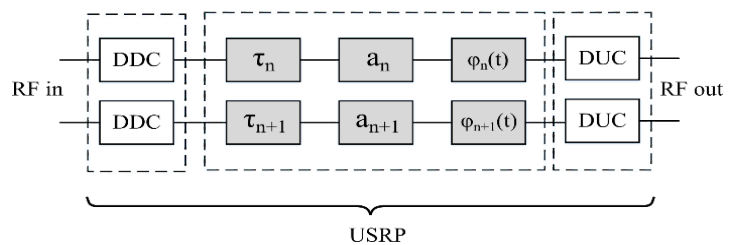


Figure 14: Simplified block diagram of RF signal processing implemented in the USRP. The analogue input signal RF in is converted via digital down conversion (DDC) to the digital baseband, where propagation delay τ_n , attenuation a_n , and a phase shift φ_n are applied. The phase shift may be time-dependent, in order to generate Doppler shifts. The modified RF signal RF out is then generated by digital up conversion (DUC).

According to (6), the resulting emulated channel impulse response is the superposition of time-delayed waves (wave number $k = 2\pi/\lambda$) impinging on the receive antenna, including the effects of antenna patterns and free-space propagation.

$$h(\tau, t) = \sum_{n=1}^N \hat{a}_n h_n(\tau, t) \mathbf{e}_{R_x}(\Omega_{A,n}) \mathbf{e}_{T_x,n}(\Omega_{D,n}) \frac{\lambda}{4\pi R} e^{-jkR} \quad (6)$$

The quantities involved are the complex-valued transmit and receive polarimetric antenna element patterns \mathbf{e}_{T_x} and \mathbf{e}_{R_x} , corresponding to direction-of-departure $\Omega_{D,n}$ and direction-of-arrival $\Omega_{A,n}$, respectively. The proof-of-functionality of the emulator was achieved by measurements of the channel transfer function $H(f)$ for given values of delay, Doppler frequency, and the amplitude ratio of multipath components, with transmitter and receiver connected by cables. The simple, yet very relevant two-path model was chosen as a reference scenario, in order to generate a transfer function with well-known characteristics suitable for comparison with analytical results. The LTE test signal was generated by an arbitrary-waveform generator and distributed to the USRP with a power splitter, as depicted in Figure 15.

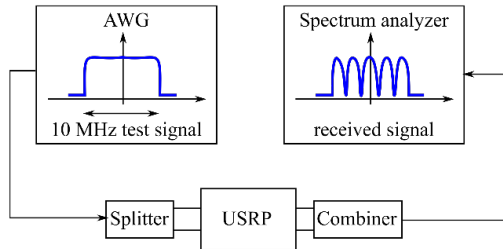


Figure 15: Block diagram of the setup for the verification of the emulated channel transfer function in the frequency domain. An arbitrary-waveform generator (AWG) provides a physical-layer LTE signal with a bandwidth of 10 MHz. The test signal is distributed via a power splitter to the USRP, which sets basic channel parameters. The emulated signal is evaluated with a spectrum analyser.

For $N = 2$ propagation paths, the channel transfer function emulated by the USRP in the frequency domain results from the Fourier transform of (5):

$$H(f, t) = a_1 e^{j(2\pi[f_{D,1}t - \tau_1] - \varphi_1)} + a_2 e^{j(2\pi[f_{D,2}t - \tau_2] - \varphi_2)} \quad (7)$$

In a first step, a static scenario ($f_D = 0$ Hz) was considered, in order to verify the proper emulation of amplitude weights and path delays. In this case, (7) simplifies to a sum of two complex exponentials, which is entirely determined by the propagation time difference $\Delta\tau = \tau_2 - \tau_1$, the phase difference $\Delta\varphi = \varphi_2 - \varphi_1$, and the amplitude ratio $A = a_1/a_2$. Figure 16 shows the results to the maximum values of the measured transfer function for $\Delta\tau = 500$ ns and $A = \{1, 0.5\}$. Measured and theoretical values agree very well. Corresponding to the inverse of $\Delta\tau$, the separation between two maxima of the transfer function was found to be exactly 2 MHz. In case of equally strong paths ($A = 1$), deep fading minima emerge. However, the measured data are affected by noise on the one hand, and the limited frequency resolution of 14.5 kHz on the other hand. For $A = 0.5$, the fading notch amounts to -9.5 dB, and matches the theoretical expectations accurately. After initialisation of the emulator, a random phase shift was observed to occur between the RF frontends of the USRP due to the phase-locked loops, leading to a frequency shift of the transfer function by $\Delta\varphi = 2\pi\Delta f\Delta\tau$. This phase shift was adjusted manually in the emulator settings, since the local oscillators cannot be coupled with the device in use.

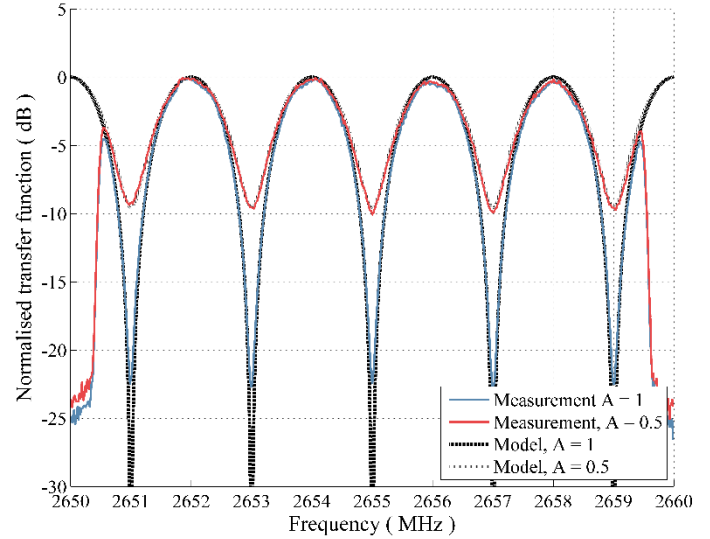


Figure 16: Comparison of emulation results (amplitude ratio of the two paths $A = 1$: blue curve, $A = 0.5$: red curve) and analytical data of the static two-path model (black and grey curves). The decay of the transfer function at the edges reflect the band-limited test signal reaching from 2650.5 MHz to 2659.5 MHz.

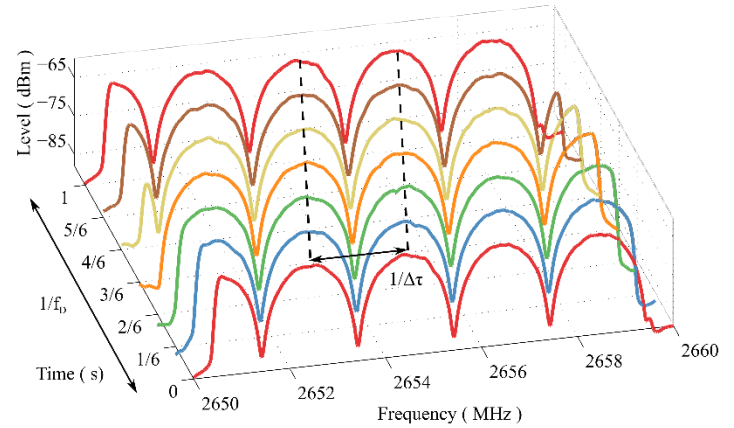


Figure 17: Temporal change of the emulated transfer function, recorded with a spectrum analyser. In case of the two-path model, the shape of the curve remains the same, but is shifted along the frequency axis, until one full Doppler-period is completed.

In a second step, a Doppler shift was applied to one of the two equally strong ($A = 1$) multipath components, whereas the other path remained static. The transfer function becomes time dependent and has to be recorded in temporal steps, whose resolution is limited by the update rate of the analyser. Figure 17 displays the temporal change of the transfer function for a Doppler shift of 1 Hz. After the time interval $\Delta t = 1/f_D = 1$ s, the initial spectrum was recovered, confirming the correct emulation of the Doppler effect. Higher Doppler shifts up to 500 Hz were verified with similar quality.

4.2. Over-the-air throughput measurements in VISTA

Propagation model of a simplified 2 x 2 MIMO scenario

A simplified 2 x 2 MIMO communication scenario, sketched in Figure 18, provided the basis for the OTA emulation experiment. Beside verification of the emulator on the physical layer in a cable connected setup, its operational capability for end-to-end testing was studied in over-the-air throughput measurements in VISTA. The transmission mode was chosen as open-loop MIMO (TM3),

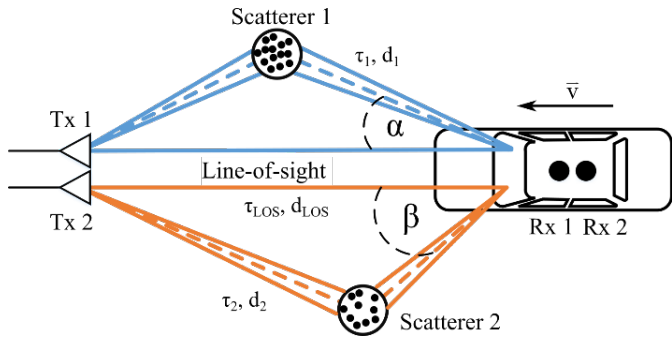


Figure 18: Sketch of the simplified 2 x 2 MIMO scenario with a passenger car driving with velocity v towards a fixed base station. Line-of-sight as well as non-line-of-sight components are considered through different path lengths d_{LOS} , d_1 , and d_2 respectively. Blue colour denotes MIMO data stream 1, orange colour denotes data stream 2.

as this is the mode widely used in practical applications. The delays τ_n were chosen from a real drive test in Ilmenau [25]: $\tau_{LOS} = 500$ ns, $\tau_1 = 959$ ns, and $\tau_2 = 1050$ ns. The remaining channel parameters were derived from geometric considerations. With c_0 being the speed of light and $d_n = c_0 \cdot \tau_n$, the relative path weights result from $a_n = a_{LOS} \cdot \tau_{LOS} / \tau_n$. According to the speed of the vehicle and the angle-of-arrival of the impinging waves, the Doppler shift is given by $f_D = v/\lambda \cdot \cos(\text{AOA})$.

Over-the-air measurement setup

The proposed model provides the basis for the evaluation of the achievable data throughput in a driving scenario under realistic channel conditions, emulated by two USRP modules. The angles-of-arrival were set to $\alpha = 25^\circ$ and $\beta = 40^\circ$. The standard-compliant Huawei LTE modem mentioned in Sec. 3 was used in combination with the communication tester, to establish a LTE link for throughput measurements; Figure 19 and Figure 20 show the block diagram and the implementation of the setup, respectively. In order to control the two USRP modules with one PC, and to synchronize their internal clocks, additional hardware was required, as explained in relation to Figure 21. The four downlink antennas were arranged on a circle around the test object with a radius of 3.95 m, in a way that the designated angles-of-arrival are emulated. The throughput measurements were conducted in LTE band 7 at a centre frequency of 2655 MHz and with a channel bandwidth of 10 MHz.

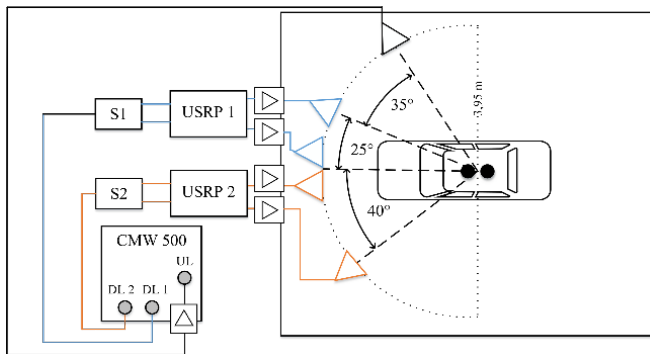


Figure 19: Block diagram of the emulation setup. The communication tester CMW500 provides the LTE protocol stack as well as the downlink RF signals. The downlink signal is distributed to the USRP via power splitters. For the sake of simplicity, the figure is limited to the relevant RF paths. Control signals and conduction for synchronisation are omitted.

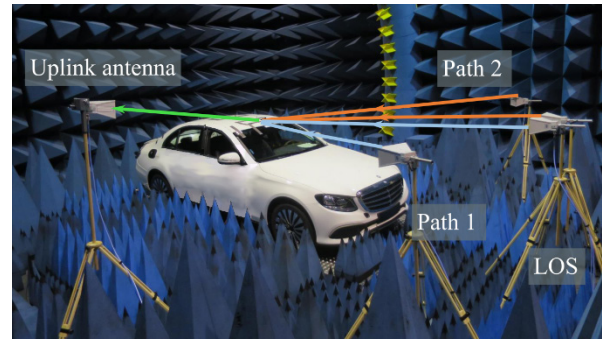


Figure 20: Measurement setup in VISTA for the emulation of the proposed MIMO scenario. The downlink signal paths are coloured blue and orange corresponding to MIMO data streams 1 and 2, respectively (c.f., Figure 18). The uplink path is highlighted in green colour. The downlink antennas are fed by two USRP modules, as indicated in Figure 21.

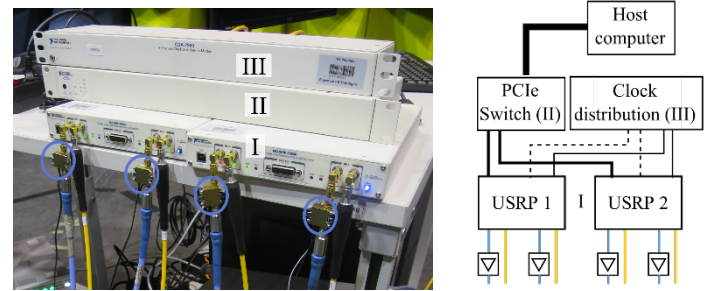


Figure 21: SDR setup for downlink channel emulation. I: The USRP modules perform the digital signal processing. The RF input signal is carried by the yellow cables. After signal processing on the internal FPGA, the signal is amplified (amplifiers indicated by blue circles) and conducted to the illumination antennas via the blue cables. II: PCIe switch for controlling both USRP modules by one host computer. The host computer is connected via a PCIe x8 connection to the switch, whereas the USRP are conducted via x4 connections. III: 10 MHz (right-hand side, solid black line) and 1 PPS (right-hand side, dotted black line) clock distribution in order to synchronize the clock of the both USRP.

Two types of measurement were performed, distinguishing between a reference measurement channel (RMC) mode and the follow-CQI (channel quality indicator) mode. Operating in RMC mode, the modulation and coding scheme is fixed; in this study, a 64 QAM scheme was chosen for the downlink. In contrast, the follow-CQI mode represents a more realistic option as it applies link adaptation, corresponding to the channel quality indicator, which is reported by the user equipment. In both modes of operation, a combination of forward-error-correction (FEC) and automatic retransmission request (ARQ), known as hybrid ARQ (HARQ), was used [26].

Over-the-air Emulation Results

Figure 22 shows the measured throughput values dependent on the received signal level for the static case. The limit of performance of the user equipment, regarding the maximum throughput, is found to be -92 dBm in RMC mode. For input levels below -92 dBm, the throughput dropped drastically. When link adaptation is applied, poor channel conditions are compensated by using a more robust modulation, e.g. 16 QAM or even QPSK instead of 64 QAM, or by reducing the code rate.

In a second step, vehicle motion was emulated, in order to stress the device-under-test when operated at the power limit. Since the LTE downlink uses frequency division access with orthogonal subcarriers (OFDMA), it is expected that Doppler shifts will have a detrimental effect on the performance due to loss of orthogonality (inter-carrier interference). Additionally, synchro-

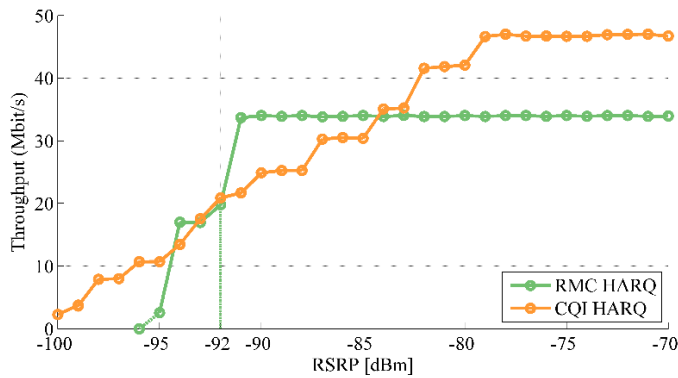


Figure 22: Measured data throughput results versus RSRP. For the reference measurement channel (green curve), the throughput remains constant at 34 Mbit/s until a critical power level of -92 dBm is reached. The data rate drops drastically when the input power level is further reduced. In contrast, when link adaptation is applied (orange curve), the maximum throughput is increased to 47 Mbit/s and the data rate decreases stepwise as channel quality (received power level) decreases.

nisation failures were also found to affect the throughput, when the device-under-test moved faster than 250 km/h [27]. Figure 23 shows the measured data throughput dependent on the emulated vehicle speed. The input power level remained constant at -92 dBm, while various velocities were tested. Confirmed by the cable-connected measurements, data transmission could be sustained for velocities up to 400 km/h. However, as this is not a realistic benchmark for passenger cars, the diagram is limited to a maximum velocity of 300 km/h, corresponding to a maximum Doppler shift of 740 Hz in the considered LTE band. Shallow variations of the measured throughput occur at emulated speeds below 150 km/h, corresponding to a maximum Doppler shift of 370 Hz, in both RMC and CQI modes. Different behaviour is observed for higher speeds: For the reference measurement channel, the throughput decreases strongly until data transmission fails. With link adaptation, the throughput decreases to a minimum of approximately 7.5 Mbit/s, while the link remains operational.

5. Conclusions

Over-the-air testing is an essential concept for the evaluation of automotive wireless systems, because antennas might not be accessible in installed state due to functional integration. Though knowledge of the radiation patterns of the application antennas remains a key for the evaluation of the radio system, the functionality and functional safety of connected cars require more than mere physical-layer testing. Antenna radiation pattern measurements as well as end-to-end communication testing can be applied in virtual electromagnetic environments like the virtual road simulation and test area (VISTA). In this context, this paper presented a holistic testing concept for automotive wireless systems, entirely based on LTE downlink transmission parameters. It was shown that the gain pattern of automotive antennas can be derived from known reference symbols (RSRP) in the LTE downlink scheme, detected by a commercial LTE communication modem. In contrast to conventional measurement techniques, this approach works without access to the antenna feed, rather utilizing the user equipment as the measurement instrument. The gain pattern of a monopole mounted on the roof of a mid-sized passenger car mock-up was derived and compared to data sets from direct power transmission and conventional nearfield measurements with subsequent nearfield-to-far-field transformation. The three sets of measurement were found fully consistent. The shapes of the pattern showed very good agreement,

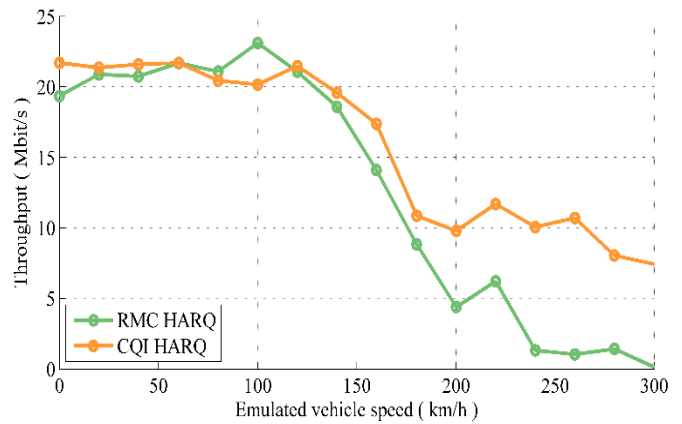


Figure 23: Results for the measured throughput dependent on the emulated vehicle speed. The device was operated at the power limit of -92 dBm. Up to speeds of 150 km/h, the user equipment behaves similarly in both operational modes RMC (green curve) and CQI (orange curve).

as indicated by a low variance of the deviations and a high cosine similarity > 96%. The RSRP concept tends to deliver slightly higher gain values compared to the other two approaches. Altogether, the RSRP-based measurement presents a promising effective alternative to conventional techniques, especially when access to the antenna feed is limited due to the installation place, or the integration with the RF frontend.

The end-to-end wireless link performance of the device-under-test was examined in terms of downlink data throughput. While widely used channel models rely on stochastic fading models, this paper proposed a simplified emulation concept based on mapping physical parameters to a set of illumination antennas arranged around the device-under-test. Compared to expensive commercial channel emulators, software-defined radio modules offer a cost-efficient opportunity to emulate basic channel parameters like path delays, Doppler shifts, and the corresponding parameter spreads. The viability of the SDR-based channel emulator was verified by a cable-connected evaluation of the transfer function with a spectrum analyser for the relevant reference scenario of a two-path model. The comparison of the emulated data with analytically calculated reference values revealed very good agreement. By using two USRP modules, the channel emulator was installed in VISTA, and over-the-air throughput measurements were conducted. In addition to low received power levels, rapid channel fluctuations due to Doppler shifts may interrupt the LTE connection since they cause a loss of orthogonality and induce synchronization failures. The measured data throughput dropped for emulated velocities above 150 km/h, when the device operated at the critical power level of -92 dBm for the static case. Beside physical channel conditions and user equipment specifications, the configuration settings of the base station affect key performance indicators like the downlink throughput. In order to simulate the behaviour of the user equipment in a real-world scenario, the application of link adaptation is advisable.

Future work will address the retrieval of phase information of highly integrated antennas from LTE parameters, without accessing the RF feed of the AUT, as well as full three-dimensional recording of the radiation patterns. For end-to-end testing, the proposed setup employing multiple SDR modules will be extended by additional USRP devices, in order to represent a richer multipath environment and thus approximate realistic conditions more closely.

Acknowledgment

The research has been funded by the Federal State of Thuringia, Germany, and the European Social Fund (ESF) under the grant 2015 FGR 0088. The antenna test range VISTA has been funded by the Federal State of Thuringia, the European Regional Development Fund (ERDF), and the German Research Foundation (DFG, 250758117) with financial support from Fraunhofer IIS. The authors would like to thank Mario Lorenz and Dr. Tobias Nowack for their support during the measurements. We further acknowledge funding for the article processing charge by the German Research Foundation (DFG) and the open access publication fund of the Technische Universität Ilmenau.

References

- [1] P. Berlt, F. Wollenschläger, C. Bornkessel and M. A. Hein, „Reliable derivation of automotive antenna gain patterns from LTE communication parameters“, 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, 2018
- [2] 5G Automotive Association, <http://5gaa.org>; online, 2018
- [3] Car 2 Car Communications Consortium, <http://car-2-car.org>; online, 2018
- [4] Virtual Drive Test Alliance, <http://tu-ilmenau.de/vdt-alliance>; online, 2018
- [5] Z. H. Mir and F. Filali, “LTE and IEEE 802.11p for vehicular networking: a performance evaluation.” *EURASIP J. Wireless Comm. and Networking* 2014 (2014)
- [6] C.A. Balanis, „Antenna Measurements“ In: *Antenna Theory Analysis and Design*. Wiley, 2005.
- [7] J. E. Hansen, Ed., "Spherical Near-field Antenna Measurements", Institution of Engineering and Technology, London, UK, 2008.
- [8] Specification USRP-2954, 10 MHz to 6 GHz Tunable RF Transceiver, National Instruments. Available: <http://ni.com/pdf/manuals/375725c.pdf>; online, 2018
- [9] Thuringian Center of Innovation in Mobility, <http://mobilitaet-thueringen.de>; online, 2018
- [10] M. A. Hein et al., "Emulation of virtual radio environments for realistic end-to-end testing for intelligent traffic systems," 2015 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM), Heidelberg, 2015, pp. 1-4.
- [11] R&S CMW Wideband Radio Communication Tester Specifications, Rohde&Schwarz. Available: <https://tinyurl.com/rscmw500>; online, 2018.
- [12] European Telecommunications Standards Institute, “LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer – Measurements” *European Telecommunications Standards Institute*, ETSI TS 136 214 V15.3.0 (2018-10), p. 8. Available: <http://etsi.org>.
- [13] Harald T. Friis, "A Note on a Simple Transmission Formula", *Proceedings of the I.R.E. and Waves and Electrons*, May, 1946, pp 254–256
- [14] F. Wollenschläger, S. N. Hasnain, C. Bornkessel and M. A. Hein, "Precision Phase Measurements of Automotive Antennas for Localization in Anechoic Chambers", 12th European Conference on Antennas and Propagation (EuCAP 2018), London, UK, 2018.
- [15] A. Stjernman, “Relationship between radiation pattern correlation and scattering matrix of lossless and lossy antennas”, *IEEE Elect. Lett.*, vol. 41, no. 12, pp.678-680, Jun. 2005.
- [16] A. Paulus, J. Knapp and T. F. Eibert, "Phaseless Near-Field Far-Field Transformation Utilizing Combinations of Probe Signals," in *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 10, 2017.
- [17] C.H. Schmidt and Y. Rahmat-Samii, "Phaseless spherical near-field antenna measurements: Concept, algorithm and simulation," 2009 IEEE Antennas and Propagation Society International Symposium, Charleston, SC, 2009, pp. 1-4.
- [18] D. Smith, M. Leach, M. Elsdon and S. J. Foti, "Indirect Holographic Techniques for Determining Antenna Radiation Characteristics and Imaging Aperture Fields," in *IEEE Antennas and Propagation Magazine*, vol. 49, no. 1, pp. 54-67, Feb. 2007.
- [19] European Telecommunications Standards Institute, “LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception” *European Telecommunications Standards Institute*, ETSI TS 136 211 V10.0.0 (2011-01), pp. 56ff. Available: <http://etsi.org>.
- [20] C. Schneider et al., "Directional analysis of multipath propagation in vehicle-2-vehicle channels," 2016 10th European Conference on Antennas and Propagation (EuCAP), Davos, 2016, pp. 1-5.
- [21] D. Shutin, “Clustering wireless channel impulse responses in angular-delay domain”, 2004 IEEE 5th Workshop on Signal Processing Advances in Wireless Communications, 2004, pp. 253-257
- [22] NI, “Real-Time MIMO Channel Emulation on the NI PXIe-5644R,” National Instruments, White Paper, May 2013. [Online]. Available: <http://www.ni.com/example/31556/en/>
- [23] L. Jäger, P. Berlt, C. Bornkessel and M.A. Hein, „Distributed Spatial Channel Emulation for Virtual Drive Testing Based on Multiple Software-Defined Radios“, 2019 13th European Conference on Antennas and Propagation (EuCAP), Krakow, 2019 (accepted)
- [24] F.G. Friedlander and M.S. Joshi, “Introduction to the Theory of Distributions”, Cambridge University Press, 1998
- [25] J. Gedschold et al., “Tracking based Multipath Clustering in Vehicle-to-Infrastructure Channels”, 2018 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, 2018
- [26] E. Dahlmann, S. Parkvall, and J. Sköld, 4G LTE / LTE Advanced for Mobile Broadband. Academic Press, 2011.
- [27] D. Micheli, M. Barazzetta, R. Diamanti, P. Obino, R. Lattanzi, L. Bastianelli, V. M. Primiani, and F. Moglie, “Over-the-Air Tests of High-Speed Moving LTE Users in a Reverberation Chamber,” in *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, May 2018, pp. 4340–4349

Insight into the IEEE 802.1 Qcr Asynchronous Traffic Shaping in Time Sensitive Network

Zifan Zhou^{*}, Michael Stübert Berger, Sarah Renée Ruepp, Ying Yan

Department of Photonics Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:

Received: 20 December, 2018

Accepted: 12 February, 2019

Online: 26 February, 2019

Keywords:

Network Traffic Shaping

Time Sensitive Network

Latency Critical Transmission

ABSTRACT

TSN is an attractive solution for latency-critical frame transmission built upon IEEE 802 architecture. Traffic scheduling and shaping in TSN aim to achieve bounded low latency and zero congestion loss. However, the most widespread solution (i.e. Time-Aware Shaper) requires a network-wide precision clock reference and only targets on cyclical traffic flows. This paper focuses on the performance evaluation of the ATS, which applies shaping algorithm to any flows and requires no clock reference. Simulations are proposed for evaluation and comparison. Metrics including end-to-end delay, buffer usage and frame loss rate are collected to assess the shaping performance. Results show that ATS achieves effective traffic shaping and switching without synchronous mechanisms, while there is an evident trade-off for using these specific algorithms.

1 Introduction

To facilitate the information exchange in the network world, diverse devices and network technologies were introduced over the years, the communication has now become more comprehensive and autonomous. The increase of interconnecting level within the network system cause an explosive growth of network traffics.

The utilization of Internet of Things (IoT) and Cyber Physical System (CPS) in industrial domain bring about the concept of *Industry 4.0*, the industrial networks are now a mixture of field bus system, Ethernet approaches and wireless solutions [1]. In mobile communication networks, centralized baseband processing and cloud service, i.e. Centralized-Radio Access Network (CRAN) and Cloud-Radio Access Network (Cloud-RAN) are introduced to the access network between devices and radio transceiver [2][3], regarding as a typical approach to achieve Fifth Generation (5G) mobile network [4]. Accordingly, distributed data transmission turns into be more integrated, it becomes more challenging to fulfill the stringent transmitting requirements for time-sensitive flows.

Having the development of the network applications in mind, it is a prerequisite condition to build a fundamental transmission network which is capable of providing appropriate services. The ongoing works of Time Sensitive Network (TSN) aim to enhance standard Ethernet to fulfill the need for deterministic, reliable and efficient communication. TSN comprises a

set of standards, as a part of work in the IEEE 802.1 working group, this work originated from the Audio Video Bridging (AVB) standards, it provides services of bridging, network management and building real-time transmission over Ethernet within a LAN or MAN domain, different features are defined in separated standards to ensure the performance from several perspectives.

One of the primary features in TSN is the timing synchronization, where all nodes and end stations within a TSN domain are set by a common timing signal. Relying on the synchronization, general traffic scheduling and shaping in TSN enhances the delivery of frame with high predictability, namely, the time instance when each transmission occurs is guaranteed in the network. However, it raises strict requirement on the precision of global timing mechanism - any timing misalignment possibly imposes failures to the network. Thus high complexity is required for implementation and maintenance. Time-triggered scheduling and shaping also requires a consistent and recurrent egress gate behavior and it has to be synthesized before flows are transmitted from the source, thus the synchronous scheduling and shaping procedure only applies to a subset of traffic flows arrive periodically.

The Asynchronous Traffic Shaping (ATS) project is created by the IEEE 802.1Qcr working group [5]. It is an approach designed without any dependency on network-wide planning, cycle synchronization or time-triggered actions, while the purpose is to provide

^{*}Corresponding Author: Zifan Zhou, Department of Photonics Engineering, Technical University of Denmark, zifz@fotonik.dtu.dk

deterministic and relatively low transmission delay for general time-sensitive flows and has no requirements on the traffic pattern. An Urgency-Based Scheduler (UBS) solution was proposed at an early stage of development [6], which contains two algorithms based on the Rate-Controlled Service Disciplines (RCSDs) [7], besides, a concept of Paternoster scheduling is also included in the 802.1 Qcr web page [5]. At the time of writing, a new ATS algorithm is included in the latest version of the draft standard, all of these algorithms are involved in this paper. The main contributions of this paper are summarized as follows:

- Elaborating the principles of ATS by designing relatively accurate models and measuring the performance in simulation scenario. All models are built in software modeler that describes network topology and functionalities.
- Collecting the average per-hop delay, buffer usage and frame loss rate deriving from simulations. Based on the results, comparisons are done between all ATS approaches, also the models are set with different configurations to optimize scheduling utilization.

2 Related work

To the best of our knowledge, few paper have performance evaluation of ATS algorithms through software simulation. The synchronous scheduling has been mentioned in many works, most of them indicate that it is essential to apply scheduling to provide time-sensitive services. On the other hand, existing researches on ATS accomplish the theoretical analysis on the features of ATS, in this section, a few works that are relevant with the analysis, measurement and modeling of TSN scheduling are included.

2.1 Researching on traffic scheduling in TSN

Some works in the literature elaborate the requirements and implementation of the real-time scheduled traffic in TSN networks. For instance, references [8] and [9] give examples of applying Ethernet and scheduled TSN to in-vehicle and wireless communication systems, emphasizing the importance of scheduling. The evaluation in [8] proves that Ethernet is able to transport the traffics mixing of different vehicle functions but scheduling is necessary in the overload situations. In [9], results show that it is difficult for conventional Ethernet to fulfill the jitter requirements of Common Public Radio Interface (CPRI), while this problem could be solved by implementing enhanced scheduled traffic.

Multiple time-sensitive flows usually co-exist in the same network, taking into account the mutual interference among these flows, G. Alderisi et al proposed a temporal isolation of flows that refer to the same traffic

class[10]. The work comprises of simulations of scheduled traffic in Audio Video Bridging (AVB) network, the traffic scheduling is driven by strict priority and off-line configuration. The results show that the temporal isolation between Scheduled Traffic (ST) and other traffic classes guarantees low and predictable latency for ST class.

The procedure of configuring synchronous scheduling is considered to be time-consuming, in [11], a graphical network modeling tool was designed to automate synthesis of gate control list in TSN scheduling, it is able to convert user-defined flow, topology and Quality of Service (QoS) to the constraints for synthesis. The tool applies object-oriented modeling, logic programming and Satisfiability Modulo Theories (SMT) to achieve automation of synthesis, it also simplifies the procedure for configuration. In [12] and [13], two performance analysis of real-time Ethernet are introduced. In [12], the evaluation of AVB standards are carried out in a simulation environment, and it shows the interfering flows have limited influence on the latency of AVB flows and the latency of the flows are constrained by size of payload. And in [13], an experimental setup is proposed to analyze the latency and jitter of synchronous traffic scheduling in TSN. The results in this work also indicates that the latency and jitter of scheduled traffic are independent from unscheduled traffics, besides, it is worth mentioning that the network stack software of end station has a strong effect on the behavior of critical periodic traffics in such experimental environment.

A prototype real-time Ethernet switch is proposed in [14], the switch provides real-time communication based on a time-triggered schedule. The switch supports frame transmission with a network-coherent time line and online administration control, and it enforces isolation of three different traffic classes so as to prevent any interference from non-time-sensitive traffic. A hardware/software co-design concept of Ethernet controller is presented in [15], the controller is partitioned into communication and application components, dedicated modules are allocated to critical transmission to fulfill timing requirement and reduce the load of microcontroller. The results show that the hardware extension of Ethernet controller significantly reduces the working load of software communication stacks, especially with a mixture of scheduled and non-scheduled traffics. With the controller, the jitter of time-scheduled transmission can be improved sharply.

2.2 Relevant work on ATS

One of the most significant metrics to evaluate TSN networks is the worst-case delay, the calculation of ATS delay bounds in [6] does not account for accumulative burstiness of the same traffic class. E. Mohammadpour et al. proposed a performance evaluation of ATS and Credit Based Shaper (CBS) in [16]. Firstly, the delay calculation included in [6] is extended in this paper, in regard to generic features of TSN. Moreover, backlog bounds of buffers are given based on network calculus.

A relatively stringent end-to-end latency bound of TSN is computed instead of adding up the bounds calculated at every switch on the path. The work increases the tightness of upper bound of end-to-end delay in TSN network and benchmarks a theoretical analysis for such a network.

In order to achieve the flexibility of ATS by aggregating flows and assign separate priority level at each hop, the synthesis of ATS becomes a more complicated process, in terms of forwarding flows to queues and assigning priority levels to queues. In [17], Johannes and Soheil present a SMT based solution along with a topology rank, cluster based heuristic of this method. The work demonstrate that with the SMT method, it is feasible to find an existed solution of synthesis and the Topology Rank Solver (TRS) heuristic reduces the computational effort to achieve the method significantly.

3 Asynchronous Shaping

3.1 Modules and Architecture

As depicted in Figure1, the switch with asynchronous shaping implements an independent clock that does not synchronize with other switches. For a given queue that supports asynchronous shaping, flows sent out from the queue are shaped by a bonded shaper, which calculates eligibility time and assigns the time to frames, the time are then used for traffic regulation by the transmission selection algorithm, a frame is eligible for transmission if the assigned eligibility time is less than or equal to the current time. The flow shaping actions are implemented through an open/closed gate control instance attaching to the queue: the gate for the specific queue will be opened when the frame in that queue is eligible to be transmitted. The algorithms used for calculating eligibility time is described in the following sections.

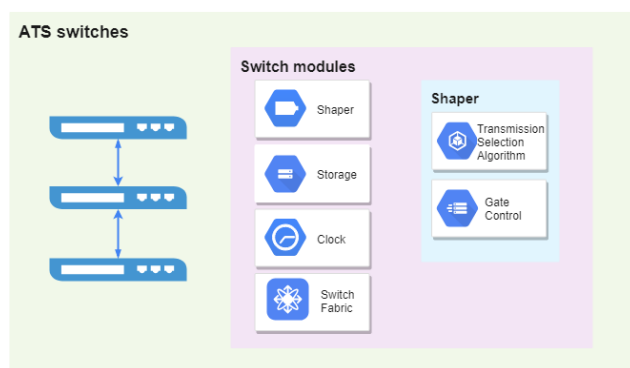


Figure 1: Architecture of ATS switch

Transmission latency usually comprises of link propagation delay and intermediate devices delay, a desirable queuing discipline should reduce effectively the storage delay in the devices. In the UBS proposal, a queuing hierarchy is introduced to the ATS pipeline, as Figure2 shows, the queuing framework contains: (1) per-flow shaped queues, which are classified ac-

ording to the identification of the frame, e.g. flow ID, traffic class and flow destination address (2) Shared queues, which merge frames with the same internal priority level and egress port but are transmitted from different shaped queues, in shared queue, frames are transmitted based on the First Come First Serve (FCFS) principle.

Queuing schemes for input frames are defined as [6]: **QAR1**: frames from different transmitters are not allowed to be stored in the same shaped queue. **QAR2**: frames from the same transmitter but not belong to the same priority in the transmitter are not allowed to be stored in the same shaped queue. **QAR3**: frames from the same transmitter with the same priority in the transmitter, but not belong to the same priority in the receiver are not allowed to be stored in the same shaped queue.

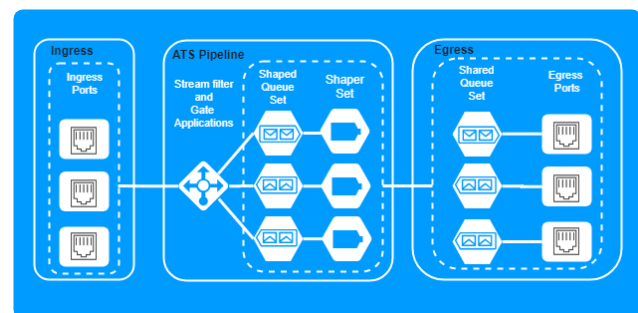


Figure 2: Architecture of ATS switch

The main purposes of implementing these queuing schemes are to enable flexible configuration among different flows, and to fulfill network services from various network domains and administrators.

According to the queuing schemes, the minimum number of shaped queues is limited by the number of ports in device. An n -port node needs at least $n - 1$ mandatory queues to fulfill QAR1 scheme.

Basically, scheme QAR2 and QAR3 achieve the separation of flows on a priority base, which enable frames with higher priority can bypass the lower-priority frames, to ensure the transmission delay of high-priority flows will not be affected by interfering flows. The isolation of queuing brings benefit to prevent the propagation of malicious flows, assuring that the ordinary flows will not get influence, it also enables flexible operations according to administrative requirements e.g. flow blocking or transmitter blocking. Considering asynchronous shaping, due to the classification of frame queuing, the shaper is able to conduct more granular operations based on larger scale of in-queue frame state. Thus ATS could lessen the queuing time of time-sensitive frames and achieve fast forwarding.

3.2 UBS algorithms

In order to achieve asynchronous shaping and keep low storage delay, an interleaved scheduling algorithm is introduced in UBS. Two approaches deriving respec-

tively from frame-by-frame leaky bucket algorithm and token-based leaky bucket algorithm [7] are introduced: Length-Rate Quotient (LRQ) and Token Bucket Emulation (TBE). Both algorithms enable the shaper with a constraint on the rate of input and output flow as:

$$l_i(d) \leq b + d \cdot r \quad (1)$$

Where l_i denotes the accumulative amount of transmitted bits, as a function of the time; b is the size of burstiness; d and r are the time duration and data rate, the constrain can be regarded as a benchmark for mixed traffic network without interaction among different flows.

LRQ and TBE are both designed for asynchronous shaping in TSN, however differ in the shaping concept. The principle of LRQ is to shape the traffic flow with a stable transmitting/leaking rate, regardless of the incoming flow pattern, it is able to convert bursty flow or flows with any pattern to stable, constant and distributed output flows. On the other hand, in the TBE method, the shaper controls the traffic flow with an average rate while allows a certain level of burst, namely, as long as sufficient number of "token" exists in the "bucket", a transmission can therefore get started immediately.

Instead of scheduling synchronously on timing basis, each asynchronous shaper keeps a local eligibility time to indicate when next frame is allowed to be transmitted. For LRQ algorithm, the eligibility time is calculated as the quotient between the size of the previously transmitted frame and the reserved link rate of the particular class, as shown in the pseudo code:

Algorithm 1 LRQ algorithm Pseudo code

```

1: /* Initialization */
2: for i in (0 : I) do
3:   flow[i].timestamp = 0
4: end for
5: /* Shaping */
6: while true do
7:   if queue[i].size > 0 then
8:     f = queue[i].head
9:     L = f.length
10:    i = f.index
11:    ti = flow[i].timestamp
12:   end if
13:   if tnow ≥ ti then
14:     output f from queue
15:     ti = tnow + (L/ri)
16:   end if
17: end while

```

The shaper updates the per-flow state every time a transmission is finished. Consequently, the LRQ shaper forces a time vacancy between frames and closes the gate for the shaped queue until next frame gets eligible for transmission, so that it keeps a stable average output rate. For TBE, the eligibility time is

calculated as the time it needs to accumulate enough "tokens", as shown in the pseudo code:

Algorithm 2 TBE algorithm Pseudo code

```

1: /* Initialization */
2: for i in (0 : I) do
3:   flow[i].timestamp = 0
4:   flow[i].token = Burstsize
5: end for
6: /* Shaping */
7: while true do
8:   if queue[i].size > 0 then
9:     f = queue[i].head
10:    L = f.length
11:    i = f.index
12:    ti = flow[i].timestamp
13:    Ki = flow[i].token
14:   end if
15:   if Ki + (tnow - ti) * flow[i].bitrate ≥ L then
16:     output f from queue
17:     ti = tnow
18:     Ki = min(Burstsize, Ki + (tnow - ti) * flow[i].bitrate)
19:     - L
20:     flow[i].timestamp = ti
21:     flow[i].token = Ki
22:   end if
23: end while

```

In principle, TBE algorithm could increase the utilization of network resources than LRQ, especially in the case when the network is lightly loaded, since in TBE algorithm, the spacing time between two adjacent frames is not added every times, unless the token level of the flow is less than the pending frame.

Accordingly, the state of output gate relies on current number of "token" in the per-flow "bucket": if the length of pending frame exceeds the current amount of "token", the shaper has to shut down the gate until the number of "token" increases with time and accumulates to an enough amount. Therefore, the TBE shaper allows a limited extent of bursty flow when the number of token is sufficient.

3.3 ATS algorithm

In the recently proposed draft of ATS standard[5], a new shaping approach is included. Basically, the approach is also derived from Leak Bucket algorithms, including the concept of token bucket which is used to constrain the output rate of flows, preventing bursty flows spreading along the path. A local system clock function determines the selectability time per frame, which is the time when the frame is queued and available for transmission selection. All frames that reach their selectability time are selected for transmission in ascending order of the assigned eligibility times. Any frame may experience an additional, non-negative processing delay between its arrival time and its selectability time. This delay may vary per frame, thus

there is a delay variation over a sequence of frames. The pseudo code of ATS algorithm is shown below:

Algorithm 3 ATS algorithm Pseudo code

```

1: /* Initialization */
2:   $T_{eligibility} = 0$ 
3:   $T_{bucketFull} = 0$ 
4:   $T_{groupEligibility} = 0$ 
5:   $T_{bucketEmpty} = -(burstSize/rate)$ 
6: /* Frame Processing */
7:   $D_{lengthRecover} = frame.length/rate$ 
8:   $D_{emptyToFull} = burstsize/rate$ 
9:   $T_{shaperEligibility} = T_{bucketEmpty} + D_{lengthRecover}$ 
10:  $T_{bucketFull} = T_{bucketEmpty} + D_{emptyToFull}$ 
     $T_{eligibility} = \max(T_{arrival},$ 
11:          $T_{groupEligibility},$ 
             $T_{shaperEligibility})$ 
12: /* Shaping */
13: if  $T_{eligibility} \leq (T_{arrival} + MaxTime/1.0e9)$  then
14:      $T_{groupEligibility} = T_{shaperEligibility}$ 
         $T_{bucketEmpty} = (T_{eligibility} < T_{bucketFull}) ?$ 
15:      $T_{shaperEligibility} :$ 
         $T_{shaperEligibility} + T_{eligibility} - T_{bucketFull} ;$ 
16:      $AssignAndProcessd(frame, T_{eligibility})$ 
17: else
18:      $Discard(frame);$ 
19: end if

```

The *bucket full time* is the time instant when the bucket is full with tokens, the size of bucket is equivalent to the burst size, on the contrary, *bucket empty time* is the time when there are no tokens existing in the bucket. The initial *bucket empty time* should be at least *empty to full duration* before the initial *bucket full time*. Basically, the *empty to full duration* is the duration needed to fill up the bucket with tokens from empty to full by the committed information rate. The *length recovery duration* denotes the duration that the tokens are accumulated by a number equaling to the length of the frame.

Considering a single shaper, the *shaper eligibility time* is the time when the number of tokens in the bucket is more or equal to the *frame size*. Taking into account a group of shapers within the same shaper class, the *group eligibility time* means the most recent *eligibility time* from the previous frame processed by the shaper in the same class. *Max residence time* is a parameter used to limit the time a frame residing in one node, a frame is valid only within the *Max residence time*.

As the code indicates, the calculation of eligibility time of the frame strongly depends on the size of last transmitted frame and the arrival time of itself. Different with the TBE algorithm, the eligibility time is not directly calculated from number of tokens in the bucket, instead, the bucket full time and bucket empty time are considered. The ATS algorithm also allows

a certain scope of bursty flows, while for oversized flow bulk, the shaper will still limit the amount of outputting flow to avoid accumulating bigger flow bulk in the downstream node.

3.4 Paternoster queuing and scheduling

Paternoster algorithm is developed based on a cyclically scheduling approach, Cyclic Queuing and Forwarding (CQF) [18], it provides deterministic and bounded delay but removes the dependence on synchronous timing. The principle of Paternoster is to implement four cyclic egress queues per class of service per port, each node and end station has local timing, and the time is counted in the unit of epoch duration τ . Four terminologies: *prior*, *current*, *next* and *last* are used to describe all epochs and cyclic queues, Table 1 illustrates the mechanism:

TABLE I: Timing and queuing in Paternoster

Epoch \ Queue	Queue0	Queue1	Queue2	Queue3
Epoch0	prior	current	next	last
Epoch1	last	prior	current	next
Epoch2	next	last	prior	current
Epoch3	current	next	last	prior
Epoch4	prior	current	next	last

Every epoch has an associated *current* queue, all incoming frames will be directed to the same *current* queue during one epoch unless the queue gets full, the following frames arrive at the same epoch are forwarded to the *next* and *last* queue as far as next epoch starts. Frames will be dropped if the volume of frames exceed reserved storage in all three queues. At the egress ports, only the *current* queue works as outbound queue per epoch, which means only the *current* queue is allowed to transmit and receive frames simultaneously.

The length of epoch of each traffic class remains its consistency within the defined network. Higher-priority flows are assigned with a shorter epoch to ensure less delay bound. Transmission of best-effort flows only fills the remaining bandwidth left from reserved flows. The best-effort frames will be dropped if the anticipated transmission time is beyond the current epoch. In principle, the length of τ should be configured to long enough for all reserved transmission and at least one best-effort frame with maximum size.

Compared with synchronous scheduling, Paternoster sacrifices some of the delay predictability but removes the synchronous timing signaling. Meanwhile, it reduces the lower bound of delay and distributes received frames to four queues, which provides similar scheduling performance with synchronized schedule and simplifies the implementation of synchronization. From the perspective of buffer usage, the division of queuing in Paternoster offers more available storage resources, thus guarantee a lower frame loss rate compared with conventional CQF.

The end-to-end queuing delay of Paternoster is independent of the network topology and interfering traffics, the primary factor that bounds the delay is the duration of cycle epoch τ . The best case of end-to-end delay occurs when frames are forwarded from and to *current* queues in all relays, accordingly, the waiting time in queues is negligible. The minimum end-to-end delay depends on the number of hops (h) and processing time. The worst case caused by the situation where all three queues - current, next and last are assigned fully with frames. Thus per-hop queuing delay increases to:

$$d_{p_hop} \leq (Q - 1) \cdot \tau \quad (2)$$

Where Q denotes the total number of queues, then end-to-end queuing delay becomes

$$d_{p_ETE} \leq (Q - 1) \cdot \tau \cdot h \quad (3)$$

Therefore, in Paternoster scheduling, frames are distributed to egress queues in a more sparse manner, and cut-through transmission is also feasible when the *current* queue receives and transmits frames at the same time, which cannot be done with CQF. These features of Paternoster guarantee more accurate per-flow state to the shaper and enable fast forwarding without having synchronous timing signaling.

4 Modeling

In this section, models of ATS approaches in a simulation environment are proposed. We used Riverbed modeler for designing models and running simulations, it is a discrete-event simulation tool providing performance evaluation for internet technology applications.

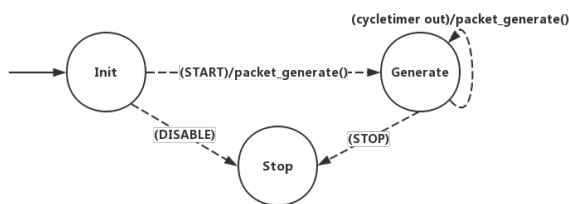


Figure 3: Process domain of traffic generator in Riverbed modeler

In the modeler, behavior of all modules in nodes are defined by process state machine, a process domain is usually consisted of multiple states and the execution of system kernel transits from one state to another as responding to events have occurred, such as expiration of timers and frames arrival. Actions and functions are included inside states. One state could have several transitions corresponding to different transiting conditions or events. An example process domain of traffic generator is given in Figure 3.

The modeler supports a multi-layer process hierarchy, a root process could create its own child processes,

multiple child processes are allowed to coexist at the same time, the first generation child processes may then in turn create new processes, which would be referred to as second generation. A tree structure of process relationships are given in Figure 4. The simulation kernel provides communication mechanisms that allow memory sharing among root and child processes, which is fitting for building the pipeline of queuing and shaping schemes, in this work, processes in the simulator emulate different modules in the switch, the internal forwarding of frames is achieved by passing the memory among processes.

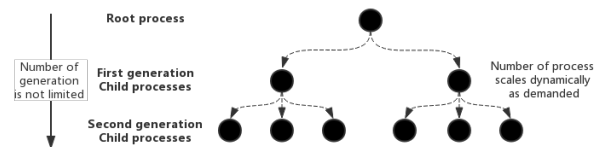


Figure 4: Process Hierarchy

4.1 Modeling UBS

Based on the root-child process hierarchy, the ingress modules and functionality in UBS are done in the root process, which parses the frames and forwards to the shared queue group, which are a series of child processes created by the root, representing the per-flow shaper associated with each shaped queue, the benefits of applying this hierarchy for UBS is that each bridge and end station, from the root process point of view, has the direct access to every shaper, which means it is able to monitor and evaluate real-time state of all per-flow shaping, for instance, it could terminate a child process of one shaper when there are no more queued frames, and generates a new process for the shaper once new frames arrive at the same class of flow.

Figure 5 depicts the root process in UBS. In process model, green circles represent forced states and red circles for unforced states: unforced states allow a pause between enter and exit executives of the states during execution, and the process remains suspended only until invocation causes it to progress into the exit executives of its current state; On the contrary, forced state does not allow the process to wait during the execution, thus forced state starts, finishes both enter and exit executives immediately.

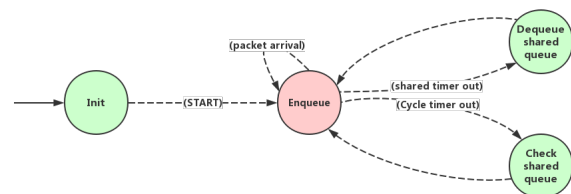


Figure 5: Process model of UBS

The root process of UBS includes two kinds of in-

terruptions for all the states:

- Packet arrival: external interrupt when new frames arrive at the ingress port, frames are then parsed and recognized by an assigned identification tag (e.g. VLAN, source and destination address) in header field and will be forwarded to the corresponding child process.
- Timer out: internal interrupt scheduled by the root process itself, two timers are used here: a shared timer is set to simulate the elapsed time for the whole transmission procedure of each frame, from the first bit to the last bit leave the egress port; another cycle timer is set with a higher frequency than the shared timer, it is used for a periodic check on shared queues, since in the modeler, a process could not detect the frames being passed from child process to root process, namely, the timer is used to check periodically if there are any frames waiting in the shared queue to be transmitted to egress ports.

Two types of child processes, implementing LRQ and TBE algorithms respectively, are presented in this paper. In the simulation, root and child processes both have the same access to the shared memory, frames received by the root process will be allocated to the specific memory block and invoke the corresponding child process, afterwards the child process extracts the frames from the given memory address, calculates the eligibility time and forwards them to the corresponding shaped queue.

For LRQ algorithm, the child processes first need to get the size of head-of-queue frame, and sets up a timer lasts for a duration equals to the quotient between frame size and reserved link rate. A draft of state machine for LRQ child process is shown in Figure 6.

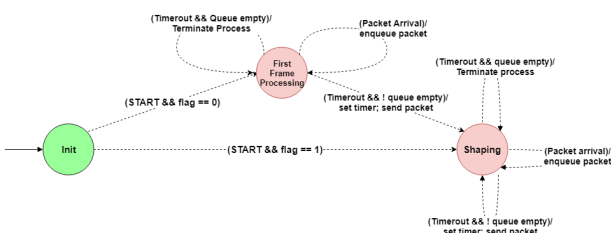


Figure 6: Child process model of LRQ

A flag is used in the child process to indicate whether the shaper is being occupied, its value decides the next state of transition from *initial* state when child process receives a new frame. In the case when the last frame in queue has been completed with transmission, the child process destroys itself to release the system resources and also to inform the root process the current vacant state of a specific shaper. A timer is used in child process to implement the time interval between frames as defined in LRQ algorithm, thus, a frame is eligible to be transmitted when it becomes head of the queue and the timer set after the transmission of previous frame expires.

TBE is achieved by the other type of child process, similar to LRQ child process, TBE also needs the acknowledgement of frame size to calculate the eligibility time for transmission in the shapers, however, it is not necessary to delay the transmission of each frame with this shaping algorithm, the state of the shaper highly depends on the token level, the state transitions are like in Figure 7.

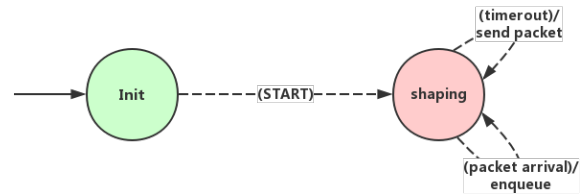


Figure 7: Child process model of TBE

In *shaping* state, the token level is kept by a variable b_i , the level increases with a constant rate of each shaper, the system kernel updates the state of variable t_i with current time, when an eligible frame is transmitted from the shaper, two possible comparison results between token level and frame size are considered in *shaping* state:

$b_i \geq size$: forward head-of-queue frame to shared queue directly

$b_i < size$: start a timer with duration of $(size - b_i)/r_i$ once the timer expires the frame is forwarded

The process model of ATS use the same architecture as the TBE model, as introduced in this section, the model contains two major parts: root process and child process, the former covers the operations such as forwarding frames to shaped queue and extracting frames from shared queue, the latter implements the ATS algorithm inside each shaper. Shared memory block between root and child processes enables the internal frame transfer inside one node.

4.2 Modeling Paternoster

In the model of Paternoster, the epoch updating is done by setting a cyclic timer each with duration τ . The queue indexes of each service class are represented by four fixed numbers, so that *prior*, *current*, *next* and *last* queues are allocated with corresponding numbers at different epochs, the number of queue groups is a configurable option, in this work, two groups of queues numbered from 0 to 3 and from 4 to 7 stand for two service classes, 8 shows the state transitions in the Paternoster model.

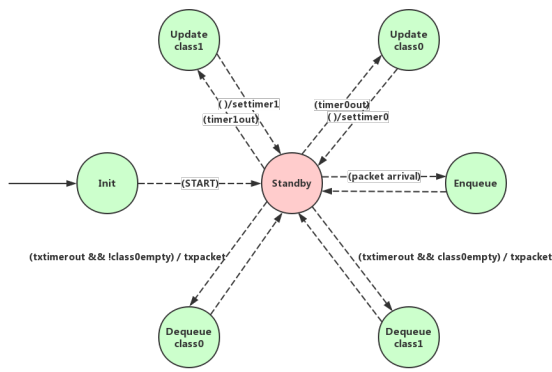


Figure 8: Process model of Paternoster

Timer 0 and 1 are set for class 0 and class 1 epochs updating, every time before transmit frames from *current* queue the shaper has to make sure *prior* queue is drained, otherwise it is supposed to dequeue frames from *prior*, the remaining frames need to be dropped if the amount exceeds a specific threshold, if not, the frames in *prior* are allowed to be transmitted. *Threshold* variables are defined in the model and used in epoch updating states *Update class1* and *Update class0*. Additionally, if no transmission happens during the epoch updating time and frames exist in the *current* queue, the shaper has to launch new transmission after the update is finished. Thus the functions of epoch update contains:

- Update index of *prior*, *current*, *next* and *last* queues
- Update reserved bandwidth
- Check the amount of remaining frames in *prior*, if exceeds the threshold then drop all left frames, if not, transmit all frames
- Check any undergoing transmission exists, if not, start transmitting from *current* queue, else, wait for the ongoing transmission to be finished

According to Stream Reservation Protocol (SRP), the bandwidth reservation is done in terms of allow a certain amount of frames counting in bit during a time period, the amount equals to epoch duration (τ) multiply reserved data rate (r_i), it is represented by *rsv_remaining* variable in the model, an amount equals to the frame size is subtracted from the reserved bandwidth when a frame is transmitted. Inside the *Enqueue* state, the model enqueues frames to *current*, *next* and *last* queues successively, and drops the frames when all three queues are full with reserved bandwidth.

Dequeue state contains the procedure of getting frames from queues and transmission selection based on priority classes, class with higher priority, *class0* in this project, is checked before other classes, an example of checking sequence is: *prior* queue of *class0* \rightarrow *current* queue of *class0* \rightarrow *prior* queue of *class1* \rightarrow *current* queue of *class1* \rightarrow ... when a frame is extracted

from queue, the process starts a timer stands for the transmission time of the frame from egress port, since the modeler is driven by discrete events.

5 Simulations and Results

To evaluate the asynchronous shaping algorithms, in this section, the LRQ, TBE and Paternoster models proposed in the last section are used for carrying out simulations in Riverbed simulator, the ATS algorithm proposed in the standard draft is not given in this work. A simple topology where only one flow exists, as depicted in Figure 9, is tested to evaluate the behavior.

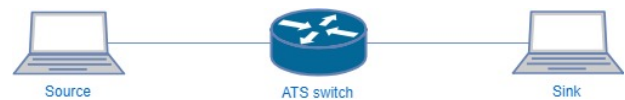


Figure 9: Simulation Scenario

Simulation parameters are given in Table. II. In this paper, the main concern is to simulate different working environments for the ATS switch, thus all the values are taken based on simulation requirements instead of real-world use cases.

TABLE II: Simulation Parameters

Parameters	Value
Frame size	720 – 1328 bytes
Reserved bandwidth	50 Mbps
Link rate	1 Gbps
Bandwidth of input flow	32 – 1928 Mbps
Paternoster epoch length	0.01, 0.005, 0.0025 second

The simulation results are based on our previous proposed paper [19]. In Table. III, the average frame loss rate comparison among Paternoster with different epoch length and UBS are given. Since all devices have limited storage space and each traffic class is assigned with dedicated bandwidth, excess frames over the bandwidth limit will be dropped. Independent to the algorithm, the input flows to UBS follow the leaky bucket constraint, thus LRQ and TBE algorithms have similar frame-loss performance as shown in the first row.

TABLE III: Frame loss rate (in percentage) comparison

Algorithms	Input Flow(Mbps)										
	32	48	64	80	96	112	128	144	160	176	192
TBE,LRQ	0	0	0.26	0.56	1.7	3.27	9.23	12	17.16	53.6	70.2
Paternoster A($\tau = 0.01s$)	0	0	0.33	1.6	2.87	6.81	10.34	17.72	23.97	44.32	77.76
Paternoster B($\tau = 0.005s$)	0	0	1.51	2.85	3.51	6.33	12.13	18.83	25.26	44.21	75.92
Paternoster C($\tau = 0.0025s$)	0	0.05	3.15	4.46	5.85	7.65	16.96	19.31	26.52	52.62	78.51

TABLE IV: Average number of queued frames comparison

Algorithms	Input Flow(Mbps)										
	32	48	64	80	96	112	128	144	160	176	192
LRQ	0.83	3.75	54	522	187	80	39	26	17	8.65	3.03
TBE	0.08	0.51	36	320	129	76	37	31	19	0.04	0.04
Paternoster A($\tau = 0.01s$)	0.02	1.5	24	63	96	115	101	110	110	111	110
Paternoster B($\tau = 0.005s$)	0.12	2.21	36	45	50	50	40.4	67.4	67.5	67.5	67.4
Paternoster C($\tau = 0.0025s$)	0.7	2.3	10.8	15.1	26.2	22.3	19.5	19.2	19.8	20.25	22.7

TABLE V: Average per-hop delay (millisecond) comparison

Algorithms	Input Flow(Mbps)										
	32	48	64	80	96	112	128	144	160	176	192
LRQ	0.35	0.89	28	86	25	10	7.9	3.7	3.4	1.75	0.73
TBE	0.0003	0.0015	4.8	87	39	19.1	7.7	4.5	4	0.0058	0.03
Paternoster A($\tau = 0.01s$)	0.00071	0.066	2.8	19.7	15.5	12.2	17.1	15	15	19.5	17.87
Paternoster B($\tau = 0.005s$)	0.0257	0.37	7.9	5.23	5.65	7.6	8.5	8.5	8.1	8.7	9.8
Paternoster C($\tau = 0.0025s$)	0.063	0.724	4.6	5.35	3.65	3.95	3.14	3.65	3.23	4.5	5.8

As traffic intensity keeps increasing, the loss rate also constantly increase for all scheduling algorithms. The comparison shows that under the same input intensity, UBS has relatively lower loss rate, while the rate of Paternoster is related to the epoch length (τ): shorter length means less storage space per epoch thus causing higher loss rate, the reserved bandwidth of each flow is calculated as $: 3 \cdot \tau \cdot \text{datarate}$.

Table IV shows the storage usage of the switch during simulation. Regarding the switch, time for transmission of one frame depends on frame size, data rate also arrival and departure time of adjacent frames in LRQ algorithm. From the results: when the bandwidth of input flow is less than reserved level (input intensity = 32, 48Mbps), UBS/LRQ has the most queued frames on average while Paternoster A with the longest epoch length has the least. As outlined above, the storage of frames in Paternoster schedulers are related with epoch length. Compared with Paternoster B and C, A is able to accommodate and forward more frames during one epoch, thus A has the least queued frames. Since UBS/LRQ algorithm enforces a pending time after each transmission, the forwarding efficiency is lower than others.

On the other hand, when the bandwidth of input flow is equal or greater than reserved level (input intensity $\geq 50Mbps$), Paternoster C has the least number of queued frames: according to Table III, C discards most frames among all schedulers under the same condition, moreover, it iterates more epoch update during the entire running time, which means more forwarding operations are executed.

Finally, table V lists the average delay measurement comparison. The variation of the delay statistics conforms with that of average number of queued frames. Paternoster A and UBS/TBE have the shortest delay when the input intensity is less than reserved, because Paternoster with longer epoch length enables more forwarding operations, also less frames are dropped due to bandwidth limitation, while Paternoster C performs faster operations when the input overflows. The average delays of all Paternoster schedulers keep increasing with input intensity.

The average delay of LRQ and TBE increases with the input traffics before overload, however, because of the linearly increasing feature of the Leaky Bucket Constraint, LRQ and TBE schedulers allow more transmissions of frames with smaller size under overload environment, thus the per-hop delay decreases sharply with the increase of input intensity.

6 Conclusion

Currently, the standard *802.1Qcr* for Asynchronous Traffic Shaping (ATS) is still not finalized. Asynchronous shaping aims at providing low congestion loss and deterministic performance while not using time synchronization in TSN, the objective of this paper is to test the performance of ATS in a series of simulations. With models built in Riverbed modeler, it

is able to simulate the work of ATS and collect results for analyses.

The evolution of ATS starts with priority-based traditional Ethernet, where traffic flows are sorted by assigned priority, upon which frames are selected for transmission. Then approach like CBS is introduced to shape egress flows to prevent congestion caused by bursty flows. Also new traffic class like AVB is created, providing services with limited level of deterministic to specific flows.

In TT Ethernet and TSN, the notion of global time synchronization is integrated in all nodes and end stations, scheduling approach is then able to carry out operations on time-triggered schedule and offer determinism for time-sensitive flows. Schedulers like TAS and CQF are both set up on the time base, the performance requires high precision on time synchronization, which is possible to be affected by failures.

ATS presents scheduling without global notion of timing, while still provide service with high-level of determinism. From results collected in the simulations, UBS with LRQ algorithm performs weakly in lightly loaded networks, in terms of delay and bandwidth utilization. While TBE algorithm allows a certain level of bursty transmission during idle period also limits the egress flow with leaky bucket constraints when network is rather full or overfull.

Deriving from CQF, Paternoster scheduling increases the number of cyclic queues for each traffics class, providing bounded transmission delay for all flows. Knowing from simulation results, the trade-off between frame loss probability and delay in Paternoster depends on the length of epoch, short duration could be set for flows taking less bandwidth than reserved value, also assigning shorter epoch to higher priority flows results in lower transmission delay.

7 Acknowledgement

This work has been partially supported by the "Fronthaul (FH) and time sensitive network (TSN) technologies for Cloud Radio Access Network (C-RAN)" project funded by Eurostars Funding.

References

- [1] Vitturi, Stefano, Paulo Pedreiras, Julian Proenza, and Thilo Sauter. "Guest editorial special section on communication in automation." *IEEE Transactions on Industrial Informatics* 12, no. 5 (2016): 1817-1821.
- [2] Wu, Jun, Zhifeng Zhang, Yu Hong, and Yonggang Wen. "Cloud radio access network (C-RAN): a primer." *IEEE Network* 29, no. 1 (2015): 35-41.
- [3] Checko, Aleksandra, Henrik L. Christiansen, Ying Yan, Lara Scolari, Georgios Kardaras, Michael S. Berger, and Lars Dittmann. "Cloud RAN for mobile networks—A technology overview." *IEEE Communications surveys & tutorials* 17, no. 1 (2015): 405-426.
- [4] Chih-Lin, I., Corbett Rowell, Shuangfeng Han, Zhikun Xu, Gang Li, and Zhengang Pan. "Toward green and soft: a 5G perspective." *IEEE Communications Magazine* 52, no. 2 (2014): 66-73.

- [5] IEEE Std. '802.1Qcr IEEE Standard for Local and metropolitan area networks - Bridges and Bridged Networks Amendment: Asynchronous Traffic Shaping', 2018. [Online]. Available: <https://1.ieee802.org/tsn/802-1qcr/>. [Accessed: 06- Nov-2018].
- [6] Specht, Johannes, and Soheil Samii. "Urgency-based scheduler for time-sensitive switched ethernet networks." In *Real-Time Systems (ECRTS), 2016 28th Euromicro Conference on*, pp. 75-85. IEEE, 2016.
- [7] Zhang, Hui, and Domenico Ferrari. "Rate-controlled service disciplines." *Journal of high speed networks* 3, no. 4 (1994): 389-412.
- [8] Lim, Hyung-Taek, Lars Völker, and Daniel Herrscher. "Challenges in a future IP/Ethernet-based in-car network for real-time applications." In *Proceedings of the 48th Design Automation Conference*, pp. 7-12. ACM, 2011.
- [9] Wan, Tao, and Peter Ashwood-Smith. "A performance study of CPRI over Ethernet with IEEE 802.1 Qbu and 802.1 Qbv enhancements." In *Global Communications Conference (GLOBECOM), 2015 IEEE*, pp. 1-6. IEEE, 2015.
- [10] Alderisi, Giuliana, Gaetano Patti, and Lucia Lo Bello. "Introducing support for scheduled traffic over IEEE audio video bridging networks." In *Emerging Technologies & Factory Automation (ETFA), 2013 IEEE 18th Conference on*, pp. 1-9. IEEE, 2013.
- [11] Farzaneh, Morteza Hashemi, Stefan Kugele, and Alois Knoll. "A graphical modeling tool supporting automated schedule synthesis for time-sensitive networking." In *Emerging Technologies and Factory Automation (ETFA), 2017 22nd IEEE International Conference on*, pp. 1-8. IEEE, 2017.
- [12] Lim, Hyung-Taek, Daniel Herrscher, Martin Johannes Waltl, and Firas Chaari. "Performance analysis of the IEEE 802.1 ethernet audio/video bridging standard." In *Proceedings of the 5th International ICST Conference on Simulation Tools and Techniques*, pp. 27-36. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2012.
- [13] Farzaneh, Morteza Hashemi, and Alois Knoll. "Time-sensitive networking (TSN): An experimental setup." In *Vehicular Networking Conference (VNC), 2017 IEEE*, pp. 23-26. IEEE, 2017.
- [14] Santos, Rui, Ricardo Marau, Alexandre Vieira, Paulo Pedreiras, Arnaldo Oliveira, and Luis Almeida. "A synthesizable ethernet switch with enhanced real-time features." In *Industrial Electronics, 2009. IECON'09. 35th Annual Conference of IEEE*, pp. 2817-2824. IEEE, 2009.
- [15] Groß, Friedrich, Till Steinbach, Franz Korf, Thomas C. Schmidt, and Bernd Schwarz. "A hardware/software co-design approach for ethernet controllers to support time-triggered traffic in the upcoming IEEE TSN standards." In *Consumer Electronics-Berlin (ICCE-Berlin), 2014 IEEE Fourth International Conference on*, pp. 9-13. IEEE, 2014.
- [16] Mohammadpour, Ehsan, Eleni Stai, Maaz Mohiuddin, and Jean-Yves Le Boudec. "Latency and Backlog Bounds in Time-Sensitive Networking with Credit Based Shapers and Asynchronous Traffic Shaping." In *2018 30th International Teletraffic Congress (ITC 30)*, vol. 2, pp. 1-6. IEEE, 2018.
- [17] Specht, Johannes, and Soheil Samii. "Synthesis of Queue and Priority Assignment for Asynchronous Traffic Shaping in Switched Ethernet." In *Real-Time Systems Symposium (RTSS), 2017 IEEE*, pp. 178-187. IEEE, 2017.
- [18] IEEE Std. '802.1Qch IEEE Standard for Local and metropolitan area networks - Bridges and Bridged Networks Amendment: Cyclic queuing and forwarding', 2016. [Online]. Available: <https://1.ieee802.org/tsn/802-1qch/>. [Accessed: 06- Nov-2018].
- [19] Zhou, Zifan, Ying Yan, Michael Berger, and Sarah Ruepp. "Analysis and modeling of asynchronous traffic shaping in time sensitive networks." In *2018 14th IEEE International Workshop on Factory Communication Systems (WFCS)*, pp. 1-4. IEEE, 2018.

A Study on the Efficiency of Hybrid Models in Forecasting Precipitations and Water Inflow Albania Case Study

Eralda Gjika^{1*}, Aurora Ferrja¹ and Arbesa Kamberi²

¹Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Tirana, 1001, Albania

²Albanian Power Corporate, Head of Production Analysis and Programming Sector, 1001, Albania

ARTICLE INFO

Article history:

Received: 25 December, 2018

Accepted: 13 February, 2019

Online : 28 February, 2019

Keywords:

Time series

Automatic forecasting

Exponential smoothing

Neural network

Regression

Hybrid model

Energy

ABSTRACT

Climatic changes have a significant impact on many real life processes. Climacteric position of Albania makes precipitations and water inflows in HPP the main variables influencing the amount of electric energy produced in the country. Taking into account their volatility it has considerably increased the need of using hybrid models to improve the quality of predictions. After a detailed analysis of the time series components, we develop a group of hybrid models and propose modifications to increase the accuracy in prediction. Among the contributions of this work is the challenge to choose between hybrid models presented earlier in literature and the modified version according to the nature of data. The final decision on the most accurate model is made based on many goodness of fit test. This study suggest that an accurate selection of the forecasting techniques increases significantly the quality of forecast on precipitations and water inflow data.

1. Introduction

Forecasting the energy production is essential for maintaining and raising the performance of a country and their presence in the regional market and beyond. Many forecasting techniques have been introduced and developed for different situations of energy production. Among these forecasting techniques, the combination of time series models, Artificial Neural Network (ANN) and Optimization Techniques have proven to be highly effective by providing satisfactory predictions for different indicators such as economic, financial, energy, demography etc. One of the earlier work on combinations of time series models was done by Bates and Granger [1] and since then it has expanded very rapidly as the necessity to obtain qualitative and accurate forecasts. ARIMA (Autoregressive Integrated Moving Average) models presented by Box and Jenkins [2] are remarkable in the literature of forecasts for their ability to build accurate predicting models for a wide range of time series data.

Combinations of some methods in order to improve forecasting quality is a good idea because they can handle at the same time the patterns (trend, seasonality) of a time series but it is not always easy for forecasters to select the best model among those proposed. Each time series is of different nature and the external factors effects vary from one situation to another. To select the most suitable model for forecasting purposes also requires extensive

experience in predictions and time series nature as well as qualitative experience. In many scientific research, it is accepted the fact that a single technique has no better predictor quality than a combination of some techniques. Also, there are many empirical findings that suggest combining forecasting techniques to improve the forecast. One of the most well-known competitions which gives contribution on the quality of forecasting by combining different techniques is the M-competitions [3-6]. Not always the principle "the more, the better" is right, so it is also important to discuss and determine the number of potential models that can be combined for prediction. Similar discussions have been observed in recent years by many authors [7-18].

One of the disadvantages of the ARIMA model is that they have the difficulties in detecting and considering the nonlinear pattern of the data and ANN, on the other hand, have difficulties to consider the linear nature of time series. Combining ARIMA models and ANN in most cases increase the forecasting performance since both can specifically deal with linear and nonlinear patterns of the time series and together they can simultaneously consider these two qualities and offer an accurate forecast. In his work Zhang [19] presents a state-of-the-art survey of ANN applications in forecasting. The research directions of ANN for forecasting purposes became very popular in recent years. Many authors [15, 18, 20, 21] in their work have proposed a combination of ARIMA and ANN to increase the forecast

*Eralda Gjika, E-Mail: eralda.dhamo@fshn.edu.al

www.astesj.com

<https://dx.doi.org/10.25046/aj040129>

performance of a time series. Other authors [10, 22, 23] have proposed combinations of ARIMA models and optimization techniques, such as PSO (Particle Swarm Optimization), to increase the forecasting accuracy of the time series.

This study is an effort to construct suitable predictive models for two important hydrological variables which highly affect the electricity production in the country. The water inflow and precipitation are the main hydrological variables which affect the energy production in our country. There have been previous attempts to build appropriate models for predicting these indicators. In their work Gjika and Ferrja [8,9] have studied the water inflow time series in three HPP built in Drin river in Albania (Fierza, Koman and Vau-Dejes). SARIMA (Seasonal Autoregressive Integrated Moving Average) and ETS (Error-Trend- Seasonality) models were tested on the three HPP and according to the minimum value of error measure and graphical test. The most accurate model for water inflow in Fierza HPP was SARIMA with seasonality 12 and ETS with multiplicative errors and seasonality for the two other HPP (Koman and Vau-Dejes).

Hybrid models which have been proposed previously in the literature are concentrated on forecasting electricity demand time series based on time series of different nature such as economic (GDP, electricity price), demographic (total population) and in some cases the average temperature, CO2 emission etc. [24-28].

The hydrological forecast has proved to be a challenge considering the unstable nature of these data. The novelty of this paper lies in the fact that we analyze hybrid models known in the literature of forecasting and propose modifications in order to fit the hydrological nature of the data.

This study is organized as follows. In the 2-nd section, we review ARIMA, ETS, ANN and LSSVM modeling approaches to time series forecasting; Section 3 presents the baseline scenario forecasting and the goodness of fit test used to evaluate accuracy of the models. In section 4 we present results and discussions for the efficiency of the evaluated forecasted models. Section 5 contains the concluding remarks and further work.

2. Forecasting Models

Classical time series models such as simple linear models (linear regression), Exponential Smoothing (ES) methods, Autoregressive Integrated Moving Average models (ARIMA) and their modifications (SARIMA, GARCH, etc.) are easy to apply on many statistical software’s and that is one of the reasons these models are widely used in time series modeling. But unfortunately regarding the volatility of the time series there is a necessity to modify the existing algorithms and models to obtain more accurate predictions.

In the first approach of this work, we have analyzed the possibility to combine classical time series models which take into consideration different components of the time series. More precisely, we have worked with: ARIMA model which considers in particular the linear behavior of the time series; the ETS model, which takes into account particularly the seasonality nature of the time series and the ANN model which considers in particular the non-linear behavior of the time series. By combining these models

into a multilinear regression model with evaluated weights in terms of the impact they have in time series, we pretend to achieve an accurate prediction for the hydrological time series.

In the second approach we have used the multiple linear regression model to estimate the values for the observed period and then we use a classic (SARIMA or ETS) model to the fitted value to predict the values in the upcoming months.

In the third approach, we have used an automatic algorithm to build hybrid models to the observed data and obtain their forecasted values for the next months.

2.1. ARIMA model

The classic ARIMA model can deal with trend and adding a seasonal term it may capture the behavior along the seasonal part of the time series. Based on Box and Jenkins model [2], the seasonal autoregressive integrated moving average model is given by equation (1):

$$\Phi_p(B^s)\phi(B)\nabla_s^D\nabla^d X_t = \alpha + \Theta_Q(B^s)\theta(B)\omega_t \tag{1}$$

where, s is the seasonal lag, ϕ is the coefficient for AR process, Φ the coefficient for seasonal AR process, θ coefficient for MA process, Θ coefficient for seasonal MA process, B is the backward shift operator, $\nabla_s^D = (1 - B^s)^D$ and $\nabla^d = (1 - B)^d$, ω_t is an uncorrelated random variable with mean zero and constant variance.

2.2. ETS model

The triplet (E,T,S) refers to the three components: error, trend and seasonality. We choose this model because it gives weight to the three components of a time series and because the water inflows in hydropower plant are highly affected by precipitations which also have seasonal nature.

The classic exponential smoothing method proposed by Holt [30] assigned weights to observations based on the time of registration. The older the observation the lower the impact in forecast. The Holt-Winters method takes into consideration trend as well as seasonality of the time series. A state space framework for automatic forecasting using exponential smoothing methods was presented by Hyndman et.al [31] and Taylor [32]. Twelve of the exponential smoothing methods are written as follow:

$$l_t = \alpha P_t + (1 - \alpha)Q_t \tag{2}$$

$$b_t = \beta R_t + (\phi - \beta)b_{t-1} \tag{3}$$

$$s_t = \gamma T_t + (1 - \gamma)s_{t-m} \tag{4}$$

where, l_t at time t, s_t denotes the slope

P_t, Q_t, R_t vary according to which of the cells the method belongs regarding the combination of the trend and seasonal component, α, β, γ and ϕ are constants. Additive and multiplicative methods give the same point forecasts but different forecast intervals. To fit an ARIMA and ETS models in R there are many forecasting packages. In our study we have used the forecast v8.3 package managed by Hyndman [33-35].

2.3. ANN model

In the literature of forecasting it is widely used the fact that ANNs are flexible computing frameworks for modeling a range of nonlinear problems [19]. Although there are some heuristic rules for the selection of the activation function it is not clear whether different activation functions have major effects on the performance of the networks. The single hidden layer feed forward network is widely used for time series modeling and forecasting. The model is characterized by a network of three layers of simple processing units connected by acyclic links. The relationship between the output (y_t) and the inputs ($y_{t-1}, y_{t-2}, \dots, y_{t-p}$) has the following mathematical representation:

$$y_t = \alpha_0 + \sum_{j=1}^q \alpha_j g \left(\beta_{0j} + \sum_{i=1}^p \beta_{ij} y_{t-i} \right) + \varepsilon_t \quad (5)$$

where, $\alpha_j (j = 0, 1, 2, \dots, q)$ and $\beta_{ij} (i = 0, 1, 2, \dots, p; j = 1, 2, \dots, q)$ are the model parameters often called the connection weights; p is the number of input nodes and q is the number of hidden nodes. The logistic function is often used as the hidden layer transfer function, $f(x) = (1 + \exp(-x))^{-1}$.

Hence, the ANN model of (5) in fact performs a nonlinear functional mapping from the past observations ($y_{t-1}, y_{t-2}, \dots, y_{t-p}$) to the future value y_t , i.e.,

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-p}, w) + \varepsilon_t,$$

where, w is a vector of all parameters and f is a function determined by the network structure and connection weights.

Both ANN and ARIMA models usually require a large sample in order to achieve a successful forecasting model. It is always advisable to undertake a subjective analysis of the data when choosing among the proposed forecasting models.

2.4. Linear Least Squares Regression model

Linear least squares regression (LSSVM) is a process which approximates linearly a set of data points $\{(x_i, y_i), x_i \in R^p, y_i \in R \text{ for } i = 1 \dots n\}$, where x is the input vector, y is the expected output and n is the number of data. Fundamentally, SVR is linear regression in the feature space. The goal of SVR is to find a function f(x) that deviates not more than ε from the targets y for all the training data, and at the same time,

is as flat as possible. LSSVM have been developed to find the optimally of non-linear regression function $y(x) = \beta^T \varphi(x) + \beta_0$.

The optimization problem of LSSVM for regression function is given:

$$\min \phi(\beta, \varepsilon) = \frac{1}{2} \beta^T \beta + \frac{\gamma}{2} \sum_{i=1}^n \varepsilon_i^2 \quad (6)$$

subject to:

$$y(x) = \beta^T \varphi(x) + \beta_0 + \varepsilon_i, \quad i = 1, \dots, n \quad (7)$$

LSSVM use a fitting function is $y(x) = \sum_{i=1}^n \alpha_i K(x_i, x) + \beta_0$, where α_i, β_0 are the solution of the linear system and $K(x_i, x)$ is a Kernel function. The most popular Kernel function is Radial Basis Function [36].

3. Baseline scenario forecasting

Time series models and neural network models are widely used in modeling of time series for prediction purposes. Many studies have shown their performance as separate forecasting models and combined with each other. Interesting results on different nature of time series are presented by Wang L. et al. [29]; Tseng et al [37]; Sheta et al. [38]; Barba et al. [39]. In the field of energy forecasting the studies are focused on choosing one of the models among some of them as has been discussed by many authors [8-9, 11-18, 40-43].

Zhang in his work [20] propose a combination of ARIMA and ANN model with the aim to increase the accuracy of the forecasting dealing both with linear and nonlinear patterns. He present a methodology which first fits an ARIMA model to the real data and then an ANN model to the residuals of the first model dealing this way the nonlinear pattern of the time series. In his study, he show that this methodology increases the accuracy of the forecast on time series data. In their work Khashei and Bijari [44] proposed a hybrid model which first used the ARIMA model to fit the real-time series and then the ANN model to obtain the final forecast. The forecast model seems to have a better performance on the same data set used before by Zhang [20]. Latter, Wang L. et al [29] have presented improvements of Zhang methodology. In 2017, Khairalla et al. [45] proposed a hybrid methodology, using additive multilinear regression methods on forecasting techniques taken as independent variables. They came out with the recommendation that using this hybrid scheme will improve the accuracy of the forecast, especially in the exchange rate time series.

The main strength of SARIMA and ETS is the capability of dealing with linear and seasonal patterns, and combined with the ANN capability of dealing with nonlinear pattern of a time series they are a good combination to offer a potential forecast model for the precipitation and water inflow time series which may be used later to predict the electricity production of the country.

3.1. Hybrid Methodology Forecast

In the first approach of our work, we propose a hybrid methodology of forecasting models using multiple linear regression method. More precisely, we have used the Least Square Support Vector Machines (LSSVM) [46,47] to the forecasting models (SARIMA, ETS, ANN) with the main goal to assign to each of the models the appropriate weight in final forecast.

Our first approach follows this steps:

Step 1. Fit a forecast model (SARIMA, ETS, and ANN) to the observed data.

Let X_t denote the observation at time t , which will serve as the dependent variable in multiple linear regression model and as independent variables we will use the estimated values obtained from each forecasting model independently. So, X_t^{ARIMA} in our case is the estimated time series obtained from a SARIMA model, X_t^{ETS} is the estimated time series obtained from an ETS model and X_t^{ANN} is the estimated time series obtained from an ANN model.

Step 2. Use the fitted values from the models in step 1 as independent variables to the multiple linear regression model and estimate the weights for each model based on LSSVM (Least Square Support Vector Machines).

Then, the estimated values from the additive multiple linear regression model will be:

$$X_t = \beta_0 + \beta_1 X_t^{ARIMA} + \beta_2 X_t^{ETS} + \beta_3 X_t^{ANN} + \varepsilon \quad (8)$$

with the constraint that $\sum_{i=0}^3 \beta_i = 1$.

Step 3. Use the multiple linear regression model fitted in Step 2 to obtain the final forecast. Use as input values the forecasted values from each single model (SARIMA, ETS, ANN).

So, after evaluating the coefficients of the model through the LSSVM procedure [46] we obtain the equation which will serve as the final forecasting model of our time series. The forecasted time series from each technique (denoted by $X_t^{(Model,F)}$, where Model={SARIMA, ETS, ANN} and F stands for Forecast for a given period) serve as input variables in the multilinear regression model:

$$\hat{X}_t^{(F)} = \beta_0 + \beta_1 X_t^{(ARIMA,F)} + \beta_2 X_t^{(ETS,F)} + \beta_3 X_t^{(ANN,F)} \quad (9)$$

This procedure will be followed for the time series of precipitations and water inflow in Fierza HPP. At this step we have considered two hybrid models (Hybrid 1 and Hybrid 2) with two and three forecasting models respectively.

3.2. Improved Hybrid Methodology Forecast

In our second approach we have used the fitted values from the “best” hybrid model (Hybrid 1 or Hybrid 2) obtained in the first approach as “real” observations and we have fitted a SARIMA and ETS model. Then, after the evaluation of the two models (SARIMA, ETS) we obtain the final forecast for the next period.

3.3. Automatic Hybrid Forecast

To compare the forecasting models we have chosen in our third approach an automatic forecasting time series package in R (*forecast* and *forecastHybrid* v8.3) managed by Rob J. Hyndman (2018). Forecasts generated from *auto.arima()*, *ets()*, *thetam()*, *nnetar()*, *stlm()*, and *tbats()* can be combined with equal weights, weights based on in-sample errors, or CV weights. The *forecastHybrid* package includes the ARIMA, ETS and ANN model along with other forecasting techniques. The results are obtained by optimizing the prediction features of the model based on minimizing error. The automatic methodology was applied to the water inflow and precipitation time series and a list of 12 models (single and combined) is obtained.

3.4. Model Performance measures

In both cases (non-automatic procedure and the automatic procedure) the accuracy of the model is evaluated based on some performance measures: the Mean Error (ME), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), symmetric MAPE (sMAPE) and Root Mean Square Error (RMSE). The evaluation of the model performance is made based on the lower value of these accuracy measures [48,49]. The selection of the “best” model between all proposed was affected also on subjective indicators observed in the behavior of the time series such as seasonality and production requirements from OSHEE.

$$MAE = \frac{1}{n} \sum_{t=1}^n |\hat{X}_t - X_t| \quad (10)$$

$$ME = \frac{1}{n} \sum_{t=1}^n (\hat{X}_t - X_t) \quad (11)$$

$$MAPE = \left(\frac{1}{n} \sum_{t=1}^n \frac{|\hat{X}_t - X_t|}{|X_t|} \right) \cdot 100\% \quad (12)$$

$$sMAPE = \left(\frac{1}{n} \sum_{t=1}^n \frac{2|\hat{X}_t - X_t|}{|X_t| + |\hat{X}_t|} \right) \quad (13)$$

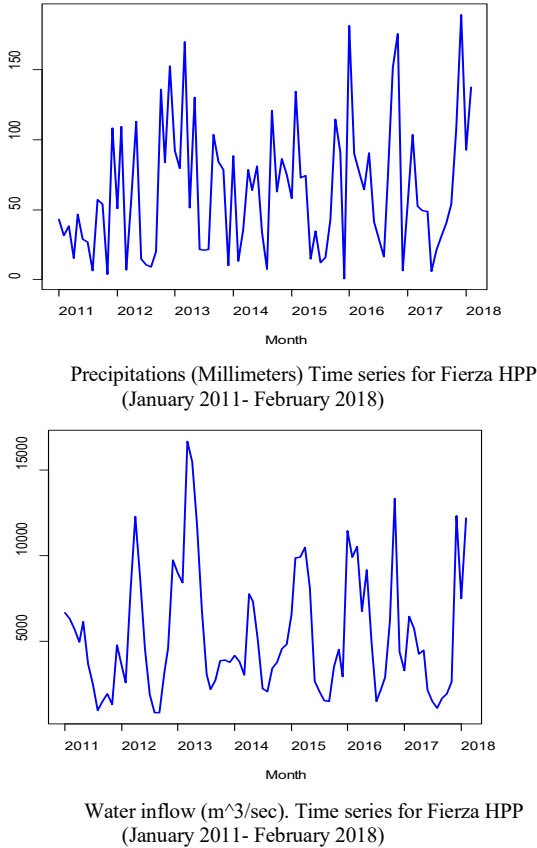
$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (\hat{X}_t - X_t)^2} \quad (14)$$

where, X_t denote the observation at time t and \hat{X}_t denote the estimated time series .

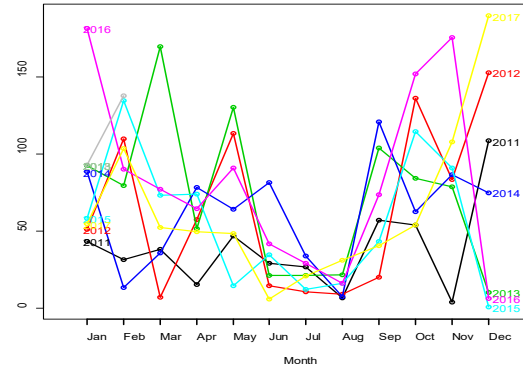
4. Empirical results and discussion

Since Albania is part of the subtropical belt and included in the Mediterranean climate zone (with short winters and hot-dry summers) the production of electrical energy is mainly based in the level of precipitations (millimeters) and water inflow (m3/sec). According to the Kesh – Gen procedure the year is divided into four energetic periods, which are October-February, March, April-May, June-September. Fierza is the oldest and most important hydropower plant built on the river Drin and thus it has a stronger impact on energy production compared to other HPPs in the country. Also, it has the highest height (or otherwise Hash) which directly determines the output power.

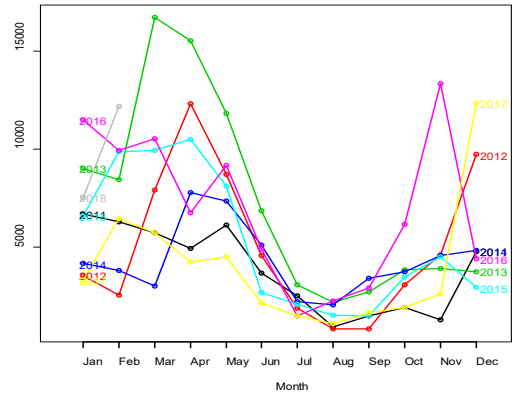
The data are collected every day from January 2011 to February 2018. Considering the fact that the necessity for energy production is long-term we have considered monthly data of precipitations and water inflow. Figure 1 shows the behavior of these time series. Being a country with a Mediterranean climate, it is not a surprise that the water inflow in the HPP is affected by the melting of the snow in mountains but we should not forget that the main impact on water inflow are precipitations.



A preliminary pre-processing of the time series of precipitation and water inflow in Fierza show that the trend is not a critical component of the time series. We observe monthly seasonality as well as a climacteric season which plays an important role in the model selection procedure of the forecasting techniques. In figure 1 the trend is “hidden” in the whole time series but if we observe carefully it is present within the season of the time series. Figure 2 shows the presence of trend and seasonality in each time series: precipitation and water inflow for Fierza HPP.



Precipitation (Millimeters), Seasonal plot for Fierza HPP (January 2011- February 2018)



Water inflow (m³/sec). Seasonal plot for Fierza HPP (January 2011- February 2018)

The seasonal plot of water inflow is more stable compared to that of the precipitation which seems to be mostly affected by the climacteric changes.

4.1. Results for Hybrid Forecast

Using the *forecast* package offered in R we have fitted separately the SARIMA, ETS and ANN model for the precipitation and water inflow time series. The corresponding models and parameters are presented in Table 1.

Table 1. The parameters of the SARIMA, ETS and ANN model for the precipitation and water inflow time series

	Precipitation	Water in flow
SARIMA	ARIMA(0,0,0)(1,0,0)[12]	ARIMA(1,0,0)(1,0,0)[12]
Coefficients	sar1=0.1685	ar1=0.6177, sar1= 0.2266
ETS	(M,Ad,M)	(M,N,M)
	alpha=1e-04 beta=1e-04 gamma=1e-04 phi=0.9443	alpha=0.1966 gamma = 1e-04
ANN	NNAR(1,1,2)[12]	NNAR(1,1,2)[12]

For the ANN model in both time series there were an average of 20 networks, each of which is a 2-2-1 network with 9 weights.

We have tested two multiple linear regression models corresponding with two and three parameters. Based on our non-automatic hybrid model of combining the forecasting models (ARIMA, ETS, ANN) in one multiple linear regression model we have obtained the following results:

Hybrid Model 1: (with two models)

$$Y = 0.04079269 + 0.48340735X^{ANN} + 0.47579996X^{ETS}$$

Hybrid Model 2: (with three models)

$$Y = 0.01847692 - 0.59771354X^{ARIMA} + 1.14123748X^{ANN} + 0.43799914X^{ETS}$$

The computed accuracy measures for each hybrid model are given in Table 2. Analyzing these values we observe that the *Hybrid model 2* has a lower value of the errors and s.d. of the errors compared to other models. This is a good sign which shows the importance of each technique on the prediction of precipitation and water inflow time series.

Table 2. Comparison of fitted models for precipitation and water inflow time series in Fierza HPP

Model	Precipitation in Fierza HPP			
	RMSE	MAPE	SMAPE	SD
ARIMA	46.722	2.513	0.396	46.996
ANN	40.292	1.992	0.359	40.529
ETS	41.342	2.392	0.299	41.585
Hybrid 1	38.505	2.07	0.322	38.73
Hybrid 2	37.379	1.813	0.269	37.599
Model	Water Inflow in Fierza HPP			
	RMSE	MAPE	SMAPE	SD
ARIMA	2860.68	0.545	0.246	2877.46
ANN	2439.6	0.46	0.203	2453.9
ETS	3551.65	0.455	0.194	3572.48
Hybrid 1	2393.31	0.44	0.193	2407.35
Hybrid 2	2376.64	0.43	0.19	2390.58

We have used MAPE as a popular measure for forecast accuracy and the calculated value for Precipitation Hybrid 2 model is 1.813% which is the lower value between the proposed models; and for Water in-flow Hybrid 2 model has again the lower value compared to other models, 0.43% error.

A view of the real-time series, fitted and forecasted values from the hybrid models (Hybrid 1, Hybrid 2) for precipitation and water inflow time series data in Fierza are shown in Figure 3.a and Figure 3.b respectively. From the graphical representations, we may observe that the multi-linear regression model (Hybrid 2) offers a good approximation to the real-time series data compared to Hybrid 1 model.

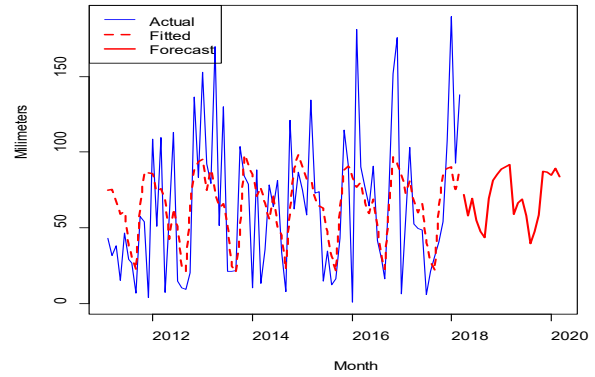


Figure 3.a Hybrid model 1, time series of precipitation real, fitted and forecast

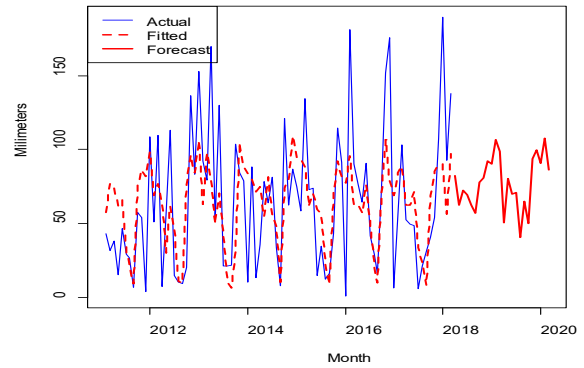


Figure 3.b Hybrid model 2. Time series of precipitation real, fitted and forecast

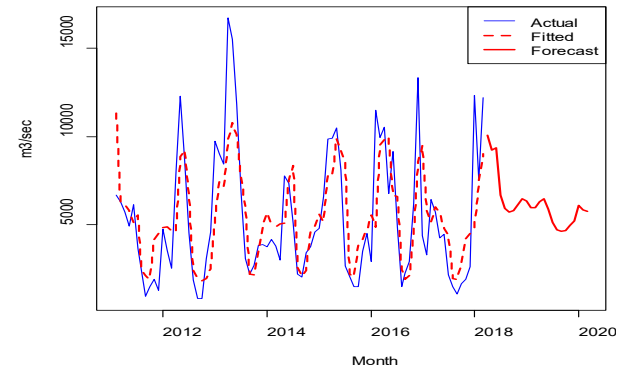


Figure 4.a Hybrid model 1. Time series of water inflow real, fitted and forecast

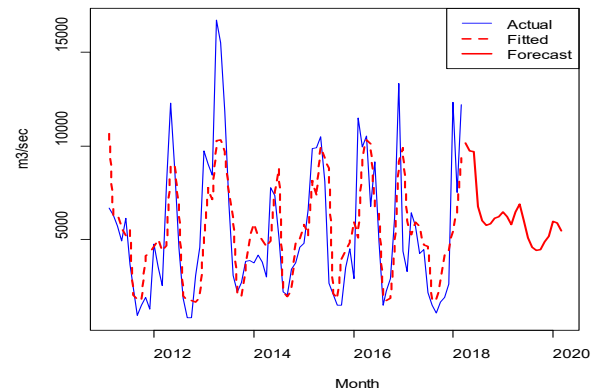


Figure 4.b Hybrid model 2. Time series of water inflow real, fitted and forecast

4.2. Results for Improved Hybrid 2 Forecast

A second approach was considered in order to achieve better predictions on the water inflow times series. Goodness of fit test for the first approach show that Hybrid 2 model performed better than the Hybrid 1 model. So, in the second approach we use the fitted values from Hybrid 2 model and build a SARIMA model to the fitted values. The results show that a seasonal model with seasonality 12 gives the lower value of the accuracy measures (the characteristics of the fitted model are: ARIMA(0,0,2)(0,1,1)[12], Coefficients: ma1=0.7579, ma2=0.3682, sma1=-0.7012; MAPE=23%). An ETS model was also fitted to the Hybrid 2 values and the best model among the proposed was ETS(M,N,M); Smoothing parameters: alpha = 0.0277, gamma = 1e-04; MAPE=25%.

From the values of the accuracy measures and graphic tests among Hybrid 2 model and the improved Hybrid 2 model, it is noted that the improved model has the best qualities to be used as a predictive model.

For the precipitation time series the SARIMA and ETS model built on the Hybrid 2 fitted values were: SARIMA with seasonality 12 and characteristics SARIMA(0,0,0)(2,1,0)[12], Coefficients: sar1=-0.5886, sar2=-0.3055; MAPE=16.8% .

The ETS(A,N,A) model has the characteristics: Smoothing parameters: alpha = 1e-04, gamma = 1e-04; MAPE=15.9%).

4.3. Results for Results for Automatic Hybrid Forecast

Using the forecastHybrid v8.3 package in R [34,35] we have obtained the following results among the possible combinations of forecasting models offered by this package. We may notice from the empirical results of accuracy measures (shown in Table 3) that the hybrid model of ANN and STLM (Seasonal and Trend decomposition using Loess) perform better than other models [50]. It has the lower value of RMSE and MAE as well. The ANN model detects the nonlinear behavior of the time series and is therefore important to the model, as well as the seasonal behavior of the model which is detected by STLM. In both time series (precipitation and water inflow in Fierza HPP) the combination of ANN and STLM gives the lower value of accuracy measures.

Table 3. Comparison of models computed from automatic package in R for precipitation and water inflow time series in Fierza HPP

ANN-TBATS	3.927	40.306	31.23	124.465	2322.08	1644.9
ARIMA-STLM	0.072	43.978	33.57	-122.38	2810.57	1960.62
ARIMA-ANN-STLM	-2.46	40.458	31.42	-103.83	2385.47	1689.96
ARIMA-ANN-TBATS	1.581	41.886	32.68	-19.784	2404.52	1711.96
ANN-STLM-TBATS	1.248	39.038	29.51	82.482	2325.47	1593.13

The closest values of errors after the combined ANN-STLM model are those of the hybrid model ANN-STLM-TBATS which is very close to the Hybrid 2 model proposed at the beginning. Since the differences in value are sufficiently small they can be considered irrelevant, and therefore the two models can serve to provide forecasts in the future.

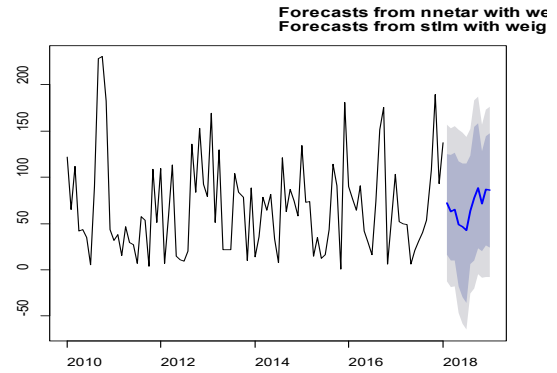


Figure 5.a Precipitation, fitted and forecast from automatic hybrid model in R

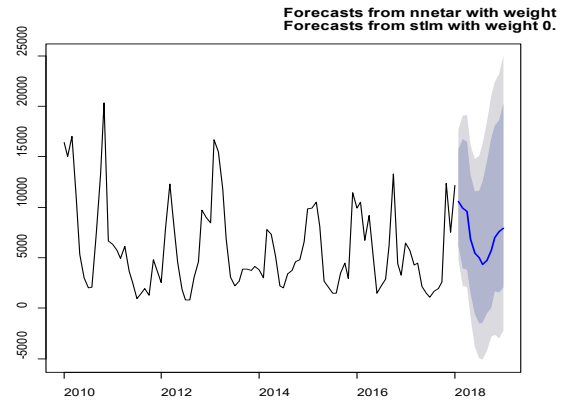


Figure 5.b Water inflow real, fitted and forecast from automatic hybrid model in R

Figure 5 shows for precipitation and water inflow time series the real values, fitted and forecasted values obtained from automatic hybrid model (ANN+STLM) in R. Due to the apparent stationarity and seasonal behavior the water inflow time series has good qualities to be modelled.

Figure 6 shows the water inflow: real observations, improved Hybrid 2 based on SARIMA, improved Hybrid model 2 based on ETS, the automatic (ANN+STLM) forecast from the forecastHybrid v8.3 package and the Hybrid 2 forecast.

Model	Precipitation in Fierza HPP			Water Inflow in Fierza HPP		
	ME	RMSE	MAE	ME	RMSE	MAE
ARIMA	-0.19	49.705	39.58	-170.85	3093.26	2268.99
ETS	-3.98	43.496	32.12	-143.48	3318.93	2084.24
ANN	0.007	41.306	32.67	-0.466	2459.17	1857.74
ARIMA-ANN	-1.61	43.019	34.09	-155.39	2506.95	1839.74
ARIMA-TBATS	6.243	47.632	36.01	61.031	2888.54	1984.21
ANN-STLM	-2.07	38.59	29.43	1.683	2291.94	1626.79
ARIMA-ANN-STLM-TBATS	0.145	40.149	30.9	-15.618	2387.21	1643.96

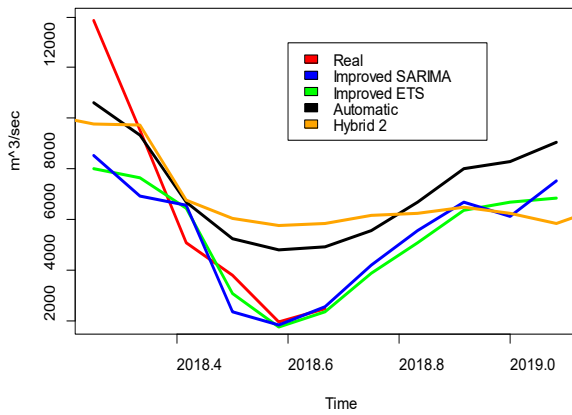


Figure 6. Water inflow: real observations, forecasted values from the improved models (SARIMA and ETS on Hybrid 2 fitted values), the automatic forecast in R and Hybrid 2 forecast (period: April 2018-April 2019)

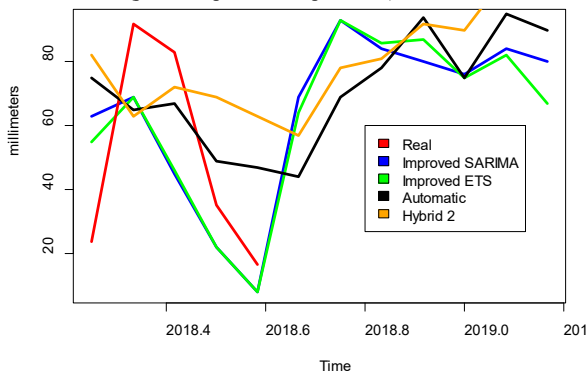


Figure 7. Precipitations: real observations, forecasted values from the improved models (SARIMA and ETS on Hybrid 2 fitted values), automatic forecast in R and the Hybrid 2 forecasted values (period: April 2018-April 2019)

Figure 6 and Figure 7 show a graphical comparison of forecasting results from the Hybrid 2 model, improved Hybrid 2 model and automatic forecast in R. Due to the geographical position and Mediterranean climate of Albania the precipitations are mainly present in two periods: October-February and March. In the case of precipitation time series, it is noticed that the “best” approximation to real observations are obtained from the improved Hybrid 2 model (both SARIMA and ETS show satisfied approximations to real observation compared to other models). And, for the precipitation time series again the improved Hybrid model show a satisfactory approximation to real data. It is known that precipitations are influenced by temperatures and changes in global warming so, the variations between the real values and the forecasted from the improved Hybrid 2 (SARIMA and ETS) are acceptable and within the confidence levels.

5. Conclusions

In this study two main indicators of energy production were analyzed, precipitation and water inflow recorded for every month in the period January 2011- February 2018 for the largest HPP in the country (Fierza HPP) which has the main impact on electricity production. After a detailed analysis of the characteristics of the time series: trend, seasonality, and randomness we have considered hybrid models in order to obtain an accurate prediction for the upcoming months. In this work we present three methodologies: Hybrid models based on LSSVM method;

improved Hybrid models with SARIMA and ETS forecasting models and automatic hybrid models proposed by forecastv8.3 package in R. The water inflow time series is the most regular; therefore, it is easier to achieve a qualitative forecasting method compared to the precipitation time series which show irregular patterns and has therefore a lower accuracy level.

The challenge of this work was to show that not all the proposed methodology on forecasting are effective because they depend on the nature of the time series. Especially, for the hydrological time series which are affected from various unstable factors it is necessary to work on many techniques and combinations to achieve the best accuracy forecast model. The improved hybrid model proposed in this study was considerably more effective compared to the models proposed earlier in the literature.

Conflict of Interest

The authors declare no conflict of interest.

References

- [1] Bates, J.M.; Granger, C.W.J. The combination of forecasts, *Oper. Res. Q.* 1969, 20, 451–468.
- [2] Box, G. E. P.; Jenkins, G. *Time Series Analysis: Forecasting and Control*, Oakland, CA: Holden-Day, 1976.
- [3] Chatfield, C., What is the ‘best’ method of forecasting? *J. Appl. Statist.* 1988, 15, 19–39;
- [4] Makridakis, S.; Anderson, A.; Carbone, R.; Fildes, R.; Hibon, M.; Lewandowski, R.; Newton, J.; Parzen, E.; Winkler, R.; The accuracy of extrapolation (time series) methods: results of a forecasting competition, *J. Forecasting*, 1982, 1, 111–153.
- [5] Makridakis, S. Why combining works? *International Journal of Forecasting*, 1989, 5, 601–603.
- [6] Makridakis, S.; Chatfield, C.; Hibon, M.; Lawrence, M.; Millers, T.; Ord, K.; Simmons, L.F. The M-2 competition: a real-life judgmentally based forecasting study, *International Journal of Forecasting*, 1993, 9, 5–29. https://en.wikipedia.org/wiki/Makridakis_Competitions
- [7] Yu, L.; Wang, Sh.; Lai, K. K.; Nakamori, Y. Time series forecasting with multiple candidate models: selecting or combining, *Journal of Systems Science and Complexity*, 2005, Vol. 18 No. 1.
- [8] Gjika, E.; Ferrja, A.; Efficiency of Time Series Models on Predicting Water Inflow (Case Study, Drin river in Albania), *Proceedings of the 3rd Virtual Multidisciplinary Conference, Slovakia, December, 7- 11, 2015*, Publisher: EDIS - Publishing Institution of the University of Zilina, <http://doi.org/10.18638/quaesti.2015.3.1.225>
- [9] Simoni, A.; Dhama (Gjika), E. Evolutionary Algorithm PSO and Holt-Winters method applied in hydro Power Plants Optimization. *Proceedings of the STATISTICS, PROBABILITY & NUMERICAL ANALYSIS 2015 METHODS AND APPLICATIONS Conference, Albania, December 5-6, 2015.* ISSN 2305-882X. <https://sites.google.com/a/fshn.edu.al/fshn/home/botim-special>
- [10] Simoni, A.; Dhama (Gjika), E. Forecasting the maximum power in hydropower plant using PSO, *Proceedings of the 6th INTERNATIONAL CONFERENCE Information Systems and Technology Innovations: inducing modern business solutions, Tirana, Albania, June 5-6, 2015.* ISBN: 978-9928-05-199-8.
- [11] Suhartono, S.P.; Prastyo, D.D.; Wijayanti, D.G.P.; Juliyanto, Hybrid model for forecasting time series with trend, seasonal and calendar variation patterns, *IOP Publishing, IOP Conf. Series: Journal of Physics: Conf. Series* 890, 2017, 012160, <https://doi.org/10.1088/1742-6596/890/1/012160>
- [12] Ozozen, A.; Kayakutlu, G.; Ketterer, M.; Kayalica, O. A Combined Seasonal ARIMA and ANN Model for Improved Results in Electricity Spot Price Forecasting: Case Study in Turkey, *Proceedings of Portland International Conference Management of Engineering and Technology (PICMET)*, 2016, 2681–2690.
- [13] Wulansari, R.E.; Setiawan, Suhartono, A Comparison of Forecasting Performance of Seasonal ARIMAX and Hybrid Seasonal ARIMAX-ANN of Surabaya’s Currency Circulation Data, *International Journal of Management and Applied Science*, 2016, Vol.2, Issue-10, ISSN: 2394-7926

- [14] Wangsoh, N.; Watthayu, W.; Sukawat, D. A Hybrid Climate Model for Rainfall Forecasting based on Combination of Self-Organizing Map and Analog Method, *Sains Malaysiana*, 2017, 46, 12, 2541–2547, <http://dx.doi.org/10.17576/jsm-2017-4612-32>
- [15] Khandelwal, I.; Adhikari, R.; Ghanshyam, V. Time Series Forecasting using Hybrid ARIMA and ANN Models based on DWT Decomposition, *Proceedings of International Conference on Intelligent Computing, Communication & Convergence (ICCC-2015)*, India, 2015, *Procedia Computer Science*, 48, 173-179, <http://doi.org/10.1016/j.procs.2015.04.167>
- [16] Hamzaçebi, C. Primary energy sources planning based on demand forecasting: The case of Turkey, *Journal of Energy in Southern Africa*, 2016, Vol.27, No.1
- [17] Szolgayová, E.P.; Danačová, M.; Komorníková, M.; Szolgay, J. Hybrid Forecasting of Daily River Discharges Considering Autoregressive Heteroscedasticity, *Slovak Journal of Civil Engineering*, 2017, 25, 2, 39-48, <http://doi.org/10.1515/sjce-2017-0011>
- [18] Du, Y.; Cai, Y.; Chen, M.; Xu, W.; Yuan, H.; LI, T. A Novel Divide-and-Conquer Model for CPI Prediction Using ARIMA, Gray Model and BPNN, *Proceeding of 2nd International Conference on Information Technology and Quantitative Management (ITQM 2014)*, Moscow, Russia, 2014, *Procedia Computer Science*, 31, ISBN: 978-1-63266-899-8.
- [19] Zhang, G.; Patuwo, B. E.; Hu, M.Y. Forecasting with artificial neural networks: The state of the art, *International Journal of Forecasting*, 1998, 14, 35–62
- [20] Zhang, G. P. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing*, 2003, 50, 159–175.
- [21] Qi, M.; Zhang, G.P. Tend Time-Series Modeling and Forecasting with Neutral Networks, *IEEE Transactions on Neural Networks*, 2008, 19, 5. <http://doi.org/10.1109/TNN.2007.912308>
- [22] Talari, S.; Shafie-khah, M.; Osório, G.J.; Wang, F.; Heidari, A.; Catalão, J.P.S. Price Forecasting of Electricity Markets in the Presence of a High Penetration of Wind Power Generators. *Sustainability*, 2017, 9, 2065. <http://doi.org/doi:10.3390/su9112065>
- [23] Papaioannou, G.P.; Dikaiakos, C.; Dramountanis, A.; Papaioannou, P.G. Analysis and Modeling for Short- to Medium-Term Load Forecasting Using a Hybrid Manifold Learning Principal Component Model and Comparison with Classical Statistical Models (SARIMAX, Exponential Smoothing) and Artificial Intelligence Models (ANN, SVM): The Case of Greek Electricity Market. *Energies* 2016, 9, 635. doi: 10.3390/en9080635
- [24] Hasanov, F.J.; Hunt, L.C.; Mikayilov, C.I. Modeling and Forecasting Electricity Demand in Azerbaijan Using Cointegration Techniques. *Energies* 2016, 9, 1045. doi: 10.3390/en9121045
- [25] Liang, Y.; Niu, D.; Cao, Y.; Hong, W.-C. Analysis and Modeling for China's Electricity Demand Forecasting Using a Hybrid Method Based on Multiple Regression and Extreme Learning Machine: A View from Carbon Emission. *Energies* 2016, 9, 941. doi: 10.3390/en9110941
- [26] Candelieri, A. Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection. *Water* 2017, 9, 224. doi: 10.3390/w9030224
- [27] Ma, X.; Liu, D. Comparative Study of Hybrid Models Based on a Series of Optimization Algorithms and Their Application in Energy System Forecasting. *Energies* 2016, 9, 640. doi: 10.3390/en9080640.
- [28] Wang, L.; Zou, H.; Su, J.; Li, L.; Chaudhry, S. An ARIMA ANN hybrid model for time series forecasting, *Wiley-Syst. Res. and Behav. Sci.*, 2013, 30, 3, 244–259. <https://doi.org/10.1002/sres.2179>
- [29] Holt, C.C. Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages. *ONR Memorandum*, Vol. 52, 1957, Carnegie Institute of Technology, Pittsburgh. Available from the Engineering Library, University of Texas, Austin
- [30] Hyndman, R. J.; Khandakar, Y. Automatic Time Series Forecasting: The forecast Package for R, *Journal of Statistical Software*, 2008, 27, 3. <http://doi.org/10.18637/jss.v027.i03>
- [31] Hyndman, R. J.; Koehler A. B.; Snyder R.D.; Grose S. A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting*, 2002, 18(3), 439–454.
- [32] Taylor, J. W. Exponential smoothing with a damped multiplicative trend, *International Journal of Forecasting*, 2003, 19, 715–725. <https://cran.r-project.org/web/packages/forecastHybrid/forecastHybrid.pdf>
- [33] <https://robjhyndman.com/hyndsight/forecast83/>
- [34] Cristianini, N.; Shawe-Taylor, J. *An Introduction To Support Vector Machines and Other Kernel Based Learning Methods*. Cambridge, Cambridge University Press. New York, USA, 2000, ISBN:0-521-78019-5
- [35] Tseng, K.-C.; Kwon O.; Tjung L. Time series and neural network forecast of daily stock prices, *Journal of Investment Management and Financial Innovations*, 2012, 9 (1). ISSN 1810-4967 (print), 1812-9358 (online)
- [36] Sheta, A.; Ahmed, E.; Faris, S.H. A Comparison between Regression, Artificial Neural Networks and Support Vector Machines for Predicting Stock Market Index, *International Journal of Advanced Research in Artificial Intelligence*, 2015, 4 (7), 55-63. <http://doi.org/10.14569/IJARAI.2015.040710>
- [37] Barba, L.; Rodríguez, N.; Montt, C. Smoothing Strategies Combined with ARIMA and Neural Networks to Improve the Forecasting of Traffic Accidents, *Hindawi Publishing Corporation, The Scientific World Journal*, 2014, Article ID 152375, 12 pages <http://dx.doi.org/10.1155/2014/152375>
- [38] Salas, J.D. Analysis and Modeling of Hydrological Time Series. In: Maidment, D.R., Ed., *Handbook of Hydrology*, McGraw-Hill, New York, 1993, 19.1-19.72.
- [39] Noakes, D.J.; McLeod, A.I.; Hipel, K.W. Forecasting monthly riverflow time series, *International Journal of Forecasting*, 1985, 179-190, North Holland
- [40] Collischonn, W.; Tucci, C.E.M.; Clarke, R.; Chou, S.C.; Guilhon, L.G.; Cataldi, M.; Allasia, D. Medium-range reservoir inflow predictions based on quantitative precipitation forecasts. *Journal of Hydrology*. 2007, 344. 112-122. <http://doi.org/10.1016/j.jhydrol.2007.06.025>
- [41] Sulandari, W.; Subanar S.; Suhartono S.; Utami H.. Forecasting electricity load demand using hybrid exponential smoothing-artificial neural network model, *International Journal of Advances in Intelligent Informatics*, 2016, 2(3), 131-139, ISSN: 2442-6571, <https://doi.org/10.26555/ijain.v2i3.69>
- [42] Khashei, M.; Bijari, M. A novel hybridization of Artificial Neural Networks and ARIMA models for time series forecasting, *Appl. Soft Comput.*, 2011, 11(2), 2664–2675. <https://doi.org/10.1016/j.asoc.2010.10.015>
- [43] Khairalla, M.; Xu-Ning; Nashat, T.; AL-Jallad. Hybrid Forecasting Scheme for Financial Time-Series Data using Neural Network and Statistical Methods, *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2017, 8(9). <http://dx.doi.org/10.14569/IJACSA.2017.080945>
- [44] Suykens, J.A.K.; Vandewalle, J. Least squares support vector machine classifiers, *Neural Processing Letters*, 1999, 9 (3), 293-300. <https://doi.org/10.1023/A:1018628609742>
- [45] Wang, H.; Hu, D. Comparison Of SVM And LSSVM For Regression. *Proceedings of International Conference on Neural Networks and Brain, Beijing*, 2005, 1: 279–283
- [46] Hyndman, R. J.; Koehler, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 2006, 22(4), 679-688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- [47] Hyndman, R. J. Measuring forecast accuracy, 2014, <https://pdfs.semanticscholar.org/af71/3d815a7caba8dff7248ecea05a5956b2a487.pdf>
- [48] Cleveland, R. B.; Cleveland, W. S.; McRae, J. E.; Terpenning, I. J. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 1990, 6(1), 3–73.

Talk Show's Business Intelligence on Television by Using Social Media Data in Indonesia

Eris Riso, Abba Suganda Girsang*

Department of Computer Science, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia

ARTICLE INFO

Article history:

Received: 20 December, 2018

Accepted: 09 February, 2019

Online : 28 February, 2019

Keywords:

Business Intelligence

Data Warehouse

Sentiment Analysis

Word2vec

Naïve Bayes

Random Forest

ABSTRACT

Knowing how and types of talk shows discussed in social media is significant to all stakeholders in a talk show's program. There are many messages that can be found in social media that need to be noticed so the messages from the user could reach the viewer. Social media provides promising as well as challenging data to business intelligence. By using data warehouse and business intelligence, the data from social media can be proceeded to monitor and understand adoption and sentiment. In this research, data warehouse and business intelligence will be designed and implemented using data from social medias; Facebook, Twitter, Instagram, and Youtube which are expected to be the source to do monitoring to social media contents that are shared by each stakeholder from the talk shows that are going to be analyzed. As for the sentiment analysis, word2vec, Naïve Bayes and Random Forest will be applied for the qualification methods. As a result, this monitoring is used to give understanding of how information spread and the sentiment which underlies the spreading messages could give critical perception to the management on how their product recognized and valued in society.

1. Introduction

Based on the data of Television Program Quality Index in 2016 Period 5 which is released by Broadcasting Commission of Indonesia (Komisi Penyiaran Indonesia-KPI) that talk show program get scale index 3, 48 out of 4 in according to KPI's standard. This index is higher than news program that gets 3,44, comedy at 3,27, variety show at 3,06, film and tv series at 2,75 and infotainment at 2,71 [1]. Talk show program as a product or commodity is possible to compete with other products. As a pop-culture product, talk show program must be marketable [2]. Through social media [Facebook, Twitter, Instagram, Youtube], communication occurs in real time with a wider coverage area. This program recognized as a place to share ideas and information as well as entertainment in order to give valuable knowledge to the viewer.

Business starts to see technology as an effective mechanism to interact more with their client and customer [3]. The content, time and frequency of conversations in social media about a product between customers are beyond the control of the manager. In that, managers need to learn how to form consumer's

discussion in a way which is in line with the mission and target of the organization [4]. The key to success in taking the advantage of consumer involvement is in the understanding of the consumer itself, such as; dissatisfaction, wishes or even their unconscious feeling [5].

At least there are two implications why a company directly involves in social media. First is the need to evaluate social objects and the second one is related to how far a content should be spread so the content could be managed as best as it could, therefore a content management system is needed to identify and manage the content in social media [6]. The Monitoring system is also needed in order to see viewer sentiment, so it could give confidence to the continuity of the talk show program, at least in public perception. With data monitoring, every company that broadcast talk show program could also see the competitiveness level so they could quickly formulate the strategy to face the competition. To date, TV programs are rated by the quantitative survey, thus qualitative survey is also important to be implemented to determine which programs are qualified and worth to be watched by the viewer. This will tell how popular a talk show program in each TV station is and what aspect needs to be improved based on the responses of the TV viewers.

*Abba Suganda Girsang, Email: agirsang@binus.edu

The high amount of information that users continually provides to the Internet requires methods that sort this information and stores it in a way that can be easily accessed and processed [7]. In its original forms, the data contains noise and unstructured that cause difficulty to extract the information in real time. By using data warehouse and business intelligence, the data from social media can be proceeded to monitor and understand adoption and product sentiment. Understanding how the information spread and the sentiment which underlies the messages will give critical perspective to the management about how their product recognized [8]. The application of business intelligence as a tool to monitor could provide reliable information to make effective decision as well as helping in evaluating consumer values and the desire to gain profit [9].

Sentiment analysis is an important tool for the extraction of information about the human emotional state [10]. Since the internet has become an excellent source of consumer review, the area of sentiment analysis has seen a large increase in academic interest [11]. Researches and the use of social media data for sentiment analysis or business intelligence have been done by some people, either for business necessity or else. Machine learning technique can be used to conclude sentiments on social media data that shows the perspective and experience of medicines and cosmetics consumer. The result of this study suggests considering spam comments, comparing the performance of different machine learning sentiment classification, temporary analysis to detect the decreasing tendency or the decline of their sentiment or of a specific product [12]. Decision Support System base on sentiment analysis able to efficiently support companies and enterprises in managing promotional and marketing campaigns on multiple social media channels [13]. Survey analysis shows that people are willing to be involved in the creation of educational policy and their opinion on social media could be directed to the creation of effective educational policy. However, the twitter classification [positive, negative and neutral] has not done in this research [14].

Other research has shown that social media data could be used to form a data warehouse to analyze data in social media such as; likes, comments and sentiment analysis that will be used as the object to design business intelligence to observe the company performance based on the data from their social media. This research will also state that the classification accuracy could be improved by using the algorithm and machine learning like naïve Bayes classification and in the future, Facebook emoticon could also be analyzed to get more comprehensive information [9].

Other research is the analysis of social media content by using some methods of data mining, thus recent information could be collected to support decision making that could give benefit to either company or personal. The objective of this research is to create business intelligence dashboard to observe the performance of news channels that is posted to social media accounts like Facebook and Twitter. This study also suggests wording qualification using Word2vec that could enhance the accuracy of text qualification process [15,16].

In this research a data warehouse and business intelligence software will be designed and implemented by using the data from social media such as; Facebook, Twitter, Instagram, and Youtube which are expected as the sources to do the monitoring towards the contents of those social media which are shared by each stakeholder of talk show programs that are going to be analysed.

As for the sentiment analysis, word2vec and Naïve Bayes and Random Forest algorithms will be applied for the qualification methods. So, this research is also expected to give perception of the rating of a talk show program qualitatively.

2. Research Method

Generally, steps in this research are collecting data, sentiment analysis, design of data warehouse, and design of business intelligence. Figure 1 is the proposed model which consists of data collection process, sentiment analysis process, and data warehouse process. The detail of these steps can be seen in the next subsection.



Figure 1: Proposed Model

2.1. Data Collection

This part consists of how the data collected from social media for each official account of the talk shows. Chosen social media for this research are Facebook, Twitter, Instagram, and Youtube. The data is taken by using web scraping method with Node.js and Python programming language. Collected data will be saved in a file with comma separated value (csv) form. Each web scraping process for each social media account will produce 2 files in CSV format, which is one file containing the admin post and one file containing the user post. Each file with the same type will be combined into one data set file that is used for two purposes. The first is aimed for sentiment analysis process and the second is to be saved in database staging for the necessity of making data warehouse and business intelligence. In this research, will use two years of data, from March 2018 to August 2018.

2.2. Sentiment Analysis

The sentiment analysis will be carried out on this stage. Pre-processing is done in each data set such as tokenization which divides sentences into words or tokens. Then delete mentions, hashtags, and URL. The next is the stopwords process which consists of non-semantic words like article, preposition, conjunction, and pronoun. Pronouns like “he”, “they”, and “I” only have less or no information about sentiments.

After getting the data set from the pre-processing process, it will be continued to the classification process. It is using word2vec to produce word embeddings and Naïve Bayes (NB)

and Random Forest as the machine learning’s text classification method. This classification will produce a vector for each word from the data set, then these vectors represent how we use words to describe the relation such as the word “man” to “woman”, as “king” to “queen”. Representation of words within a sentence can define the characteristics of the words and word embedding of sentences can determine the word characteristics and the context [17].

Talk Show	Facebook	Twitter	Instagram	Youtube
Hitam Putih	-	@HitamPutihT7	officialhitamputihtrans7	TRANS7 OFFICIAL
Ini Talk Show	-	@Ini_Talkshow	initalkshowofficial	Ini Talk Show
Kick Andy	-	@KickAndyShow	kickandyshow	-
Rumah Uya	-	@RumahUya_Trans7	rumahuya_trans7	TRANS7 OFFICIAL
Indonesia Lawyer Club	ILCtv1	@ILC_tvOnenews	indonesialawyersclub	Indonesia Lawyer Club tvOne
Sarah Sechan	-	@SarahSechan_NET	sarahsechan_net	SarahSechanNet
Economic Challenges	ecmetrotv	@EC_MetroTV	-	-
Pagi-pagi	-	@PagiPagi_NET	pagipagi_netv	Pagi Pagi NET

Figure 2: Social Media Accounts of Talk Show

Word2vec method treats each word equally because the purpose of word2vec is to calculate word embeddings. In word embedding, not all words represent the meaning of a particular sentence. To obtain the feature vector for each review, first, we learn the vector representation of words and then average all vectors of the words in each review. It uses 10-fold cross validation for Naïve Bayes and Out of Bag (OOB) error for Random Forest to fit and validate data training.

2.3. Data Warehouse Process

In designing data warehouse, the Kimbal method will be used where there are 9 stages that must be passed, such as: choosing the process, choosing the grain, identifying and confirming the dimensions, choosing the facts, storing pre-calculation in the fact table, rounding out the dimension tables, choosing the duration of the database, tracking slowly changing dimensions, and deciding the physical design [18].

When the data has been included into the staging database, then the ETL process will be carried out, such as calculating the number of comments, posts, likes, sentiments, and others. Data from the ETL process is then stored in the data warehouse.

Figure 3 is the scheme which is going to be used to create a data warehouse and business intelligence which explained that there are 2 fact tables and 6 dimensional tables. The Admin Posts Fact table is a collection of admin data posts history which is done by the admin for each social media account and has foreign key to each dimensional table. Similarly, to Fact User Comment table is the user history comments towards the admin posts for each social media’s talk show and has a foreign key to each dimensional table. Meanwhile for each dimensional table is a detailed summary of admin post data and user comments, as well as dimensional time table and dimensional talk show table to complete the needs of the data warehouse manufacture.

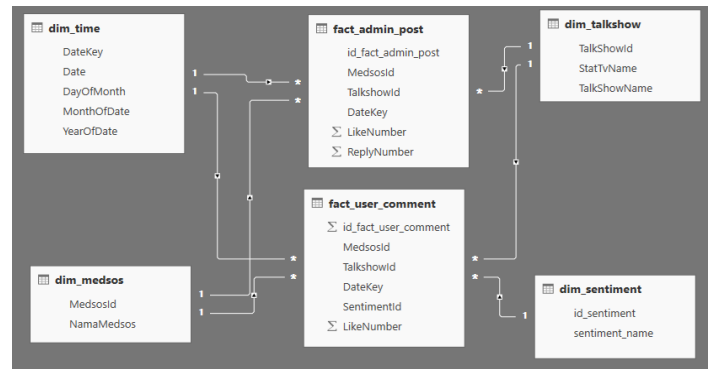


Figure 3: Star Scheme Data Warehouse

2.4. Business Intelligence System

The result of this process is a web-based business intelligence system that can be accessed by the users on the client side. Business intelligence design method that will be used in this research is business intelligence design recommended by Carlo Vercellis. He was written in his book that in developing business intelligence of a company, there are four main phases or stages that must be carried out; analysis, design, planning, implementation, and control [19].

3. Results

Based on the result of the web scrapping that has been done to the data collection process in this research, each talk show’s data from each social media’s account on period of July 2018 – August 2018 as follows:

Talk Show	Facebook		Twitter		Instagram		Youtube	
	Admin	User	Admin	User	Admin	User	Admin	User
Hitam Putih	1	-	350	635	266	5,182	171	59,947
Ini Talk Show	-	-	610	1,388	215	5,940	68	15,889
Kick Andy	-	-	261	568	114	443	-	-
Rumah Uya	-	-	243	2,035	34	888	277	21,991
Indonesia Lawyer Club	32	2,929	343	4,730	27	859	91	153,057
Sarah Sechan	-	-	11	-	54	683	157	4,264
Economic Challenges	18	26	-	-	-	-	-	-
Pagi-pagi	-	-	2	13	11	163	95	489

Figure 4: Admin Post and User Comment

Figure 4 shows that “Pagi-pagi” talk show has the least admin posts with only have 2 posts in Twitter account and “Pagi-Pagi” talk show has the least user comments with only 13 comments in Twitter account. Meanwhile, for the highest admin post is “Ini Talk Show” with 610 posts in Instagram account and the highest user comment is “Indonesia Lawyer Club” talk show with 153,057 comments in Youtube channel.

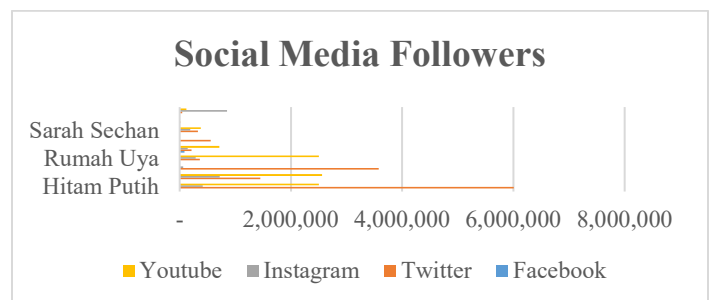


Figure 5: Social Media Followers

On Figure 5 shows the social media followers' data for each social media's account of the talk show. The most talk show's follower is "Hitam Putih" about 6.000.000 followers on social media's twitter, meanwhile, the least talk show's follower is "Economic Challenges" which only has 84 followers on Facebook account

3.1. Sentiment Analysis

In this research, the training corpus for Word2vec contains the comment data for each social's account (Twitter, Facebook, Instagram, and Youtube) from each talk show's programs in Bahasa. The data that has been collected, is the data without label so it needs label to be processed as training data. For the labelling process is used Naïve Bayes algorithm until it gets sentiment positive label, negative, and neutral for each of user's comments. We trained our corpus with 90% of user's comments and each user's comments from each talk show social media account as a testing data. For each of the tests that has been done, 10-fold cross validation is implemented on Naïve Bayes and Out of Bag error on Random Forest.

Talk Show	Facebook			Twitter			Instagram			Youtube		
	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3
Hitam Putih	-	-	-	59.24%	59.18%	59.35%	59.01%	59.08%	59.27%	59.14%	59.06%	59.31%
Ini Talk Show	-	-	-	59.28%	59.23%	59.01%	59.08%	59.16%	59.25%	59.08%	59.12%	59.02%
Kick Andy	-	-	-	59.01%	58.94%	59.24%	59.16%	59.14%	59.22%	-	-	-
Ramah Uya	-	-	-	59.12%	59.25%	59.14%	59.26%	59.23%	59.27%	59.17%	59.20%	59.20%
Indonesia Lawyer Club	59.11%	59.07%	59.08%	59.26%	59.11%	59.28%	59.14%	59.13%	59.21%	58.84%	59.14%	59.01%
Sarah Sechan	-	-	-	59.26%	59.00%	59.28%	59.16%	59.08%	59.29%	59.08%	59.25%	59.17%
Economic Challenges	59.18%	58.73%	59.35%	-	-	-	-	-	-	-	-	-
Pagi-pagi	-	-	-	59.04%	59.11%	59.41%	59.17%	59.20%	59.02%	59.39%	58.97%	59.36%

Figure 6: Naïve Bayes Algorithm Test Result using 10-Fold Cross-Validation

Talk Show	Facebook			Twitter			Instagram			Youtube		
	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3
Hitam Putih	-	-	-	67.84%	67.95%	68.01%	67.93%	67.84%	67.91%	67.78%	67.94%	67.84%
Ini Talk Show	-	-	-	67.76%	67.95%	68.02%	67.89%	67.83%	67.85%	67.97%	68.01%	67.78%
Kick Andy	-	-	-	67.83%	67.88%	67.91%	67.90%	67.79%	67.74%	-	-	-
Ramah Uya	-	-	-	67.92%	67.96%	67.76%	67.87%	67.89%	67.78%	67.89%	67.81%	67.64%
Indonesia Lawyer Club	67.73%	67.92%	67.85%	67.72%	67.74%	67.81%	67.90%	67.74%	67.87%	67.78%	67.89%	67.78%
Sarah Sechan	-	-	-	67.91%	68.01%	67.88%	67.85%	67.81%	67.95%	67.72%	68.03%	67.84%
Economic Challenges	67.91%	67.89%	67.68%	-	-	-	-	-	-	-	-	-
Pagi-pagi	-	-	-	67.78%	67.74%	67.76%	67.85%	67.77%	67.89%	67.84%	67.87%	67.83%

Figure 7: Random Forest Algorithm Test Result with Out-of-Bag error

Figure 6 is the summary of Naïve Bayes Algorithm test result as the text classification with cross-validation which the experiments have been conducted 3 times. As the whole test gets the result that the 2nd test gets the highest accuracy score of the overall test which is 59.41%. Figure 7 is the summary of Random Forest Algorithm test result as the text classification with Out of Bag error which the experiments have also been conducted 3 times for each social media's accounts. The highest accuracy score of the overall test is 68,03%. From the 2nd test result of classification algorithm, therefore Random Forest Algorithm test has the highest accuracy score compared to Naïve Bayes.

Figure 8 is the summary graphic of Sentiment Analysis Prediction which can be concluded based on Sentiment Analysis Prediction towards user's comments that have been done accumulatively. Therefore, the highest negative percentage prediction of the talk show is Indonesia Lawyers Club which is 22.16% and for the highest positive percentage prediction is also own by Indonesia Lawyers Club which is 34.17%. Meanwhile,

the second position for negative percentage prediction is Hitam Putih which is 9,30% and for positive percentage prediction is 29.47%.

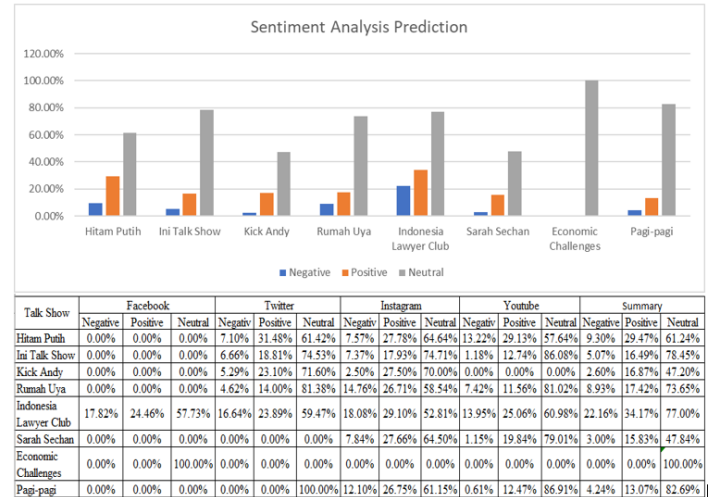


Figure 8: Summary of Sentiment Analysis Prediction

3.2 Business Intelligence System Development

In order to make easier on monitoring data, Business Intelligence Dashboard is created for admin posts data and for user comments use Microsoft Power BI. Business Intelligence System on admin's side gives facilitation which can be used as monitoring tools, such as the admin post's amount in a certain time scale that is needed. Also, it can be seen how many "like" have been accepted and how many "comments" which have been given by the users towards each admin's posts on each its social medias.

In this monitoring system, it is also displayed sentiment analysis towards each social media such as Facebook, Twitter, Instagram, and Youtube, which are owned by each talk show in order to sell their talk show's product. Besides that, it is also displayed some graphics and which post data that has the most comment and comment displays from each user's comment.

On user's side, this monitoring system displays data which relates with user's comment such as the most active users for each social media's account, sentiment prediction for each user's comment, and which admin's post that is popular for each talk show's users displayed statistically.

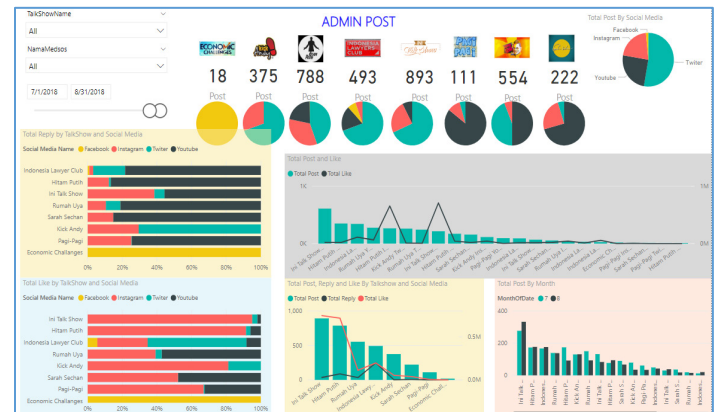


Figure 9: Admin Post Dashboard

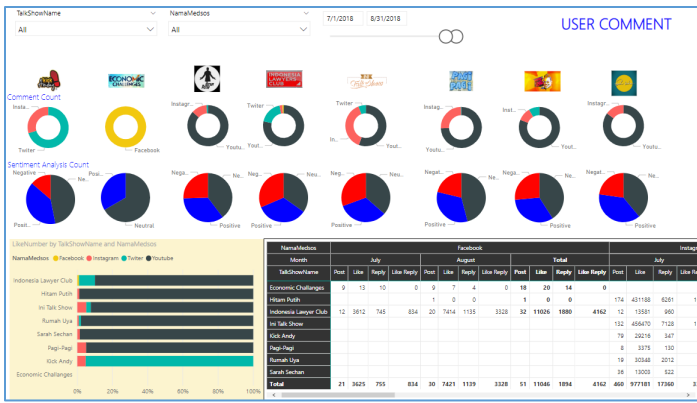


Figure 10: User Comment Dashboard

4. Discussion

In the previous research, sentiment analysis used Word2vec which was more focused on review data with the standard language, so that preprocessing process could be done maximally. In contrast to the sentiment analysis process, it used user’s comments data on social media for a talk show which comment did not use standard language. As the result, each of users could give their comments with their own language. Therefore, in the stemming process, there were a lot of words which could not be changed into basic words thus the test scores’ results were not maximal.

In business intelligence system that has been made, the data is categorized into 2 categories. They are admin posts’ data and user comments’ data. For admin’s posts category, is made into several pages of data in the form of tables and graphics to display admin’s posts, the posts’ amount, the popularity of each posts, like, share, total comment, the sentiment’s average which is categorized based on talk show’s programs and social media’s accounts. Meanwhile, for users’ comments category is made into several pages which contain tables and graphics to display users’ comments data, total comment, like, share, the sentiments’ average is also categorized into talk shows’ programs and social media’s accounts.

#	KPI’s Survey 2017	Social Media				
		Followers	Admin Post	User Comment	Like Count	Sentiment Analysis
1	Hitam Putih	Hitam Putih	Ini Talk Show	Indonesia Lawyer Club	Ini Talk Show	Indonesia Lawyer Club
2	Indonesia Lawyer Club	Ini Talk Show	Hitam Putih	Hitam Putih	Hitam Putih	Hitam Putih
3	DR OZ Indonesia	Kick Andy	Ramah Uya	Ramah Uya	Indonesia Lawyer Club	Ramah Uya
4	Satu Meja	Ramah Uya	Indonesia Lawyer Club	Ini Talk Show	Ramah Uya	Kick Andy
5	Talk to I News	Indonesia Lawyer Club	Kick Andy	Sarah Sechan	Kick Andy	Ini Talk Show

Figure 11: Top Five By Number of Viewers and Social Media Activities

Figure 11 shows the five highest data based on the amount of audience which was taken from Broadcasting Commission of Indonesia (Komisi Penyiaran Indonesia-KPI) 2nd period in the year of 2017 (18) and the number of user’s activities on social media which based on the result of KPI’s survey, the “Hitam Putih” talk show was the most watched talk show. Meanwhile, based on social media data concluded that “Hitam Putih” has the highest score and by this result, we can confirm the KPI’s survey result.

5. Conclusion

Generally, Word2vec can be used for sentiment analysis’ process in Bahasa, especially in talk shows’ programs on

television. However, the result of the sentiment analysis did not show positive sentiment or negative sentiment towards its talk show, but it was more likely used as a media to express or comment the opinion freely on social media and it was not meant to give “likes” towards a talk show’s program.

Business Intelligence Monitoring System that has been created, it can automatically calculate data from admin’s side or user’s side which is sourced from the data warehouse. The data such as admin’s posts’ amount, users’ comments, like, share, dislike, who made the comments, who has the most often comments, what are the comments for each of the posts. All of these can be easily displayed in this monitoring system without immediately opening one by one each of the social media’s accounts. Therefore, through this monitoring system, the performance can be measured for each of talk show’s programs, so that it can be rated from admin’s posts and user’s comments which talk show’s program is popular in social media.

For the next research, particularly for sentiment analysis process, it needs to be developed a stemming method and stop word for nonstandard words especially in Bahasa in order to get basic word precisely from each of user’s comments words. Business Intelligence Monitoring System which has been created, it will be better if the data can be processed fast in real time with the use of existing big data technology.

Acknowledgment

The researchers would like to thank for magister programs at bina nusantara university who have supported this research.

References

- [1] Komisi Penyiaran Indonesia, “Survei Indeks Kualitas Program Siaran Televisi Periode 5 tahun 2016,” 2016.
- [2] B. M. Timberg, *Television Talk: A History of The Tv Talk Show*. The University Of Texas Press, 2002.
- [3] A. Abdallah, N. P. Rana, Y. K. Dwivedi, and R. Algharabat, “Social media in marketing : A review and analysis of the existing literature,” *Telemat. Informatics*, vol. 34, no. 7, pp. 1177–1190, 2017. <https://doi.org/10.1016/j.tele.2017.05.008>
- [4] W. G. Mangold and D. J. Faulds, “Social media: The new hybrid element of the promotion mix,” *Business Horizons*, vol. 52, no. 4, pp. 357–365, 2009. <https://doi.org/10.1016/j.bushor.2009.03.002>
- [5] M. H. Saragih and A. S. Girsang, “Sentiment Analysis of Customer Engagement on Social Media in Transport Online,” in 2017 International Conference on Sustainable Information Engineering and Technology (SIET), 2017. <https://doi.org/10.1109/SIET.2017.8304103>
- [6] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, “Social media? Get serious! Understanding the functional building blocks of social media,” *Business Horizons*, vol. 54, no. 3, pp. 241–251, 2011. <https://doi.org/10.1016/j.bushor.2011.01.005>
- [7] J. A. Morente-Molinera, G. Kou, C. Pang, F. J. Cabrerizo, and E. Herrera-Viedma, “An automatic procedure to create fuzzy ontologies from users’ opinions using sentiment analysis procedures and multi-granular fuzzy linguistic modelling methods,” *Information Sciences*, vol. 476, pp. 222–238, 2019. <https://doi.org/10.1016/j.ins.2018.10.022>
- [8] Y. Lu, F. Wang, and R. Maciejewski, “Business Intelligence from Social Media: A Study from the VAST Box Office Challenge,” *Comput. Graph. Appl. IEEE*, vol. 34 no 5, pp. 58-69, 2014. <https://doi.org/10.1109/MCG.2014.61>
- [9] M. Yulianto, A. S. Girsang, and R. Y. Rumagit, “Business Intelligence for Social Media Interaction In The Travel Industry In Indonesia,” *J. Intell. Stud. Bus.*, vol. 8, no. 2, pp. 72–79, 2018.
- [10] H. H. Do, P. W. C. Prasad, A. Maag, and A. Alsadoon, “Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review,” *Expert Syst.*

Appl., vol. 118, pp. 272–299, 2019.
<https://doi.org/10.1016/j.eswa.2018.10.003>

- [11] Y. Fang, H. Tan, and J. Zhang, “Multi-strategy sentiment analysis of consumer reviews based on semantic fuzziness,” *IEEE Access*, vol. 6, no. c, pp. 20625–20631, 2018. <https://doi.org/10.1109/ACCESS.2018.2820025>
- [12] H. Isah, P. Trundle, and D. Neagu, “Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis,” in 2014 14th UK Workshop on Computational Intelligence (UKCI), 2014. <https://doi.org/10.1109/UKCI.2014.6930158>
- [13] P. Ducange, M. Fazzolari, M. Petrocchi, and M. Vecchio, “An effective Decision Support System for social media listening based on cross-source sentiment analysis models,” *Eng. Appl. Artif. Intell.*, vol. 78, no. October 2018, pp. 71–85, 2019. <https://doi.org/10.1016/j.engappai.2018.10.014>
- [14] M. S. Omar, A. Njeru, S. Paracha, M. Wannous, and S. Yi, “Mining Tweets for Education Reforms,” in 2017 International Conference on Applied System Innovation (ICASI), 2017. <https://doi.org/10.1109/ICASI.2017.7988441>
- [15] P. F. Kurnia and Suharjito, “Business Intelligence Model to Analyze Social Media Information,” *Procedia Comput. Sci.*, vol. 135, no. September, pp. 5–14, 2018. <https://doi.org/10.1016/j.procs.2018.08.144>
- [16] P. F. Kurnia, “Perancangan dan implementasi bisnis intelligence pada sistem social media monitoring and analysis (studi kasus di pt. dynamo media network),” Bina Nusantara University, 2017.
- [17] M. A.- Amin, M. S. Islam, and S. Das Uzzal, “Sentiment Analysis of Bengali Comments With Word2Vec and Sentiment Information of Words,” in 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2017. <https://doi.org/10.1109/ECACE.2017.7912903>
- [18] R. Kimball and M. Ross, *The Kimball Group Reader: Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence*. 2010.
- [19] C. Vercellis, *Business Intelligence, Data Mining and Optimization for Decision Making*. John Wiley & Sons, Ltd, 2009.

Application Layer Security Authentication Protocols for the Internet of Things: A Survey

Shruthi Narayanaswamy, Anitha Vijaya Kumar*

Department of ECE, Dayananda Sagar College of Engineering, Bangalore 560078, Karnataka, India

ARTICLE INFO

Article history:

Received: 15 December, 2018

Accepted: 04 February, 2019

Online : 28 February, 2019

Keywords:

IoT

Wireless Sensor Networks (WSNs)

Application Layer

Security Authentication Protocol

Vulnerabilities

ABSTRACT

Network security challenges due to nearly limitless internet connectivity, platform limitations, ubiquitous nodal mobility and huge data transactions is burgeoning by the day and the need for transcend Internet of Things (IoT) based cloud security authentication protocols is on an exponential rise. Even though many secure classic layered security mechanisms are available for implementation, they cannot be applied on IoT devices because of the huge energy that they consume. The essence of the paper is an attempt to revisit the existing IoT based security authentication protocols operating in the Application Layer (AL), AL being the end user's actual service provider. This gateway to the outside world definitely demands stringent and safe data handling and processing. The main objective of the paper is to highlight the positives of the AL protocols and also take a note of the drawbacks in terms of security and defensive measures. The author intends to support the users with information sufficient enough to decide on the type of protocol based on the application. The paper helps the future researchers to have a comparative analysis of each AL protocol's performance and further work on effective improvised defensive measures to tackle the threat-prone IoT environment even better. The paper discusses the architecture implementations, security provisions as well the pros and cons of certain avowed AL protocols currently being used in an IoT environment. Furthermore, the vulnerabilities and possible open issues currently encountered in the AL contribute valuably to the paper since they unravel the path to future research opportunities for secure interconnection of communicating devices.

1. Introduction

IoT envisages millions of communicating nodes with sensing, actuating and processing capabilities actively connected to the Internet and the number of physical objects eyeing to get connected to the Internet is booming to an unprecedented rate. IoT environments require their sensory nodes to sense continuously and communicate with the environment; needless to say the polling method of data collection fails.

Almost all layers of the protocol stack are vulnerable to security threats and attacks. Layered security protections have to be introduced to combat unique physical security concerns of IoT [1]. It is vital to bring in the security mechanisms of existing IoT based protocols, analyze existing open research security issues, and evolve with better security mechanisms for existing protocols and a step ahead to innovation of many more ingenious IoT based protocols. Significant obstacles in IoT security involve

Application, System, Communication Network and Infrastructure security [2]. Also, IoT still does not have global policies and standards to standardize application development, interaction and implementation. Hence, best security practices and standards requirements must evolve to enhance data integrity.

The Application Layer on the top of the protocol stack is the most open ended of all of the layers providing the widest attacker surface to hackers and hence is more vulnerable to network threats when compared to the other layers of the stack. All application dependent high level functions operate from this layer. The primary focus of this paper is on Application Layer security, prominent security authentication protocols of the layer and their security implementations. Even though breakthrough researches have made their way into the world of IoT security, each day the network threats and vulnerabilities are not failing to create network troubles. The ever growing jargon of vulnerabilities motivate the author to discuss the existing defensive measures offered by the prominent AL protocols and further provoke

*Anitha Vijaya Kumar, Email: anithavijaya@gmail.com

www.astesj.com

<https://dx.doi.org/10.25046/aj040131>

researchers to evolve with techniques and measures superior to the existing ones. In fact, each drawback tabulated under vulnerabilities encourages us to work on solutions for the same.

The paper is organized as follows. Section II describes few IoT architectures encompassing the essentials of the IoT system namely heterogeneous physical objects, sensors and actuators, data storage and handling and smart network technologies. The section also projects a pictorial representation of an IoT protocol stack. Section III recalls the existing IoT based security authentication protocols based on three classification criteria. Section IV deals with vulnerabilities and security issues, specifically in the Application layer and highlights few contributions attempting for better security mechanisms. Section V focuses on research challenges and required enhancements for the IoT based security authentication protocols followed by the conclusion of the survey in Section VI.

2. Architectural Support

The authors of [3] provide references to four proposed architectures, one of them being the five-layered generic architecture which many IoT implementations relate to. Lack of standardization and common IoT designs encourage researchers to dwell more into generic architectures whereas an efficient standardization would probably drive researchers to fix common security issues much more effectively. The architectures mentioned are: (a) Three layered architecture (b) Middleware based architecture (c) Service oriented architecture (SOA) (d) Five layered architecture.

There is no single consensus over the choice of IoT architectures. The Five layered architecture in which the AL provides an interface to the Business Layer for high level analysis of data. Data accession control mechanisms are mainly handled in this layer. These reasons are quite a reason for network engineers and designers to settle down for the Five- layered architecture comprising of the following layers [4].

- (a) Business Layer at the top constitutes the financial and service benefits yielded from the Application layer provided data.
- (b) Application Layer defines the various applications in which IoT can be deployed.
- (c) Processing Layer is the middleware layer which stores, analyzes and processes to accomplish Service Management.
- (d) Transport Layer is responsible for mutual data transfer between Processing and Perception layers using different communication networks.
- (e) Perception Layer is responsible for sensing and information gathering from the IoT environment.

The content of this paper is a primitive contribution to the implementation of an IoT based security authentication protocol in the Application layer as shown in the stack diagram above in Figure 1. The protocol flow would definitely prove to be more performance oriented than the regular IP flow in terms of security features and protocol efficiency.

IoT protocols like Constrained Application Protocol (CoAP), Datagram Transport Layer Security (DTLS), User Datagram Protocol (UDP) and IPv6 over Low Power Wireless Personal

Area Networks (6LoWPAN) are designed for optimized IP access and smaller data overhead of few tens of bytes in a network of constrained devices.

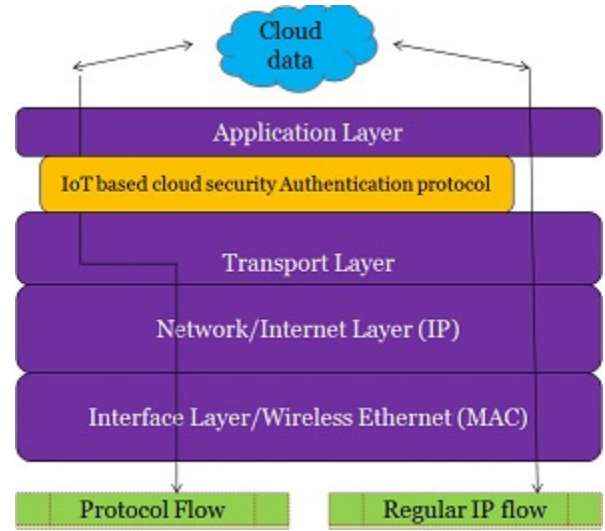


Figure 1. Proposed Architecture Diagram of Protocol Stack

3. Existing IoT based Protocols

The following three approaches define a fine way to categorize the IoT based protocols, the details of which are put down in the table below.

- (a) Based on the layer to which the protocol belongs - Application Layer protocols, Network Layer protocols, Data Link Layer protocols
- (b) Based on key distribution schemes - Symmetric Key, Asymmetric Key
- (c) Based on the nature of IoT application - Application protocols, Service Discovery, Infrastructure protocols, Influential protocols.

3.1. Based on the Layer

Messaging among various subsystems of the IoT environment is enabled by the session layer or transport layer protocols [5] like Message Queue Telemetry transport (MQTT), Secure MQTT (SMQTT), Data Distribution Service (DDS), Advanced Message Queuing Protocol (AMQP), Extensible Messaging and Presence Protocol (XMPP), CoAP, HTTP, Embedded Binary HTTP (EBHTTP), Lean Transport Protocol (LTP), Simple Network Management Protocol (SNMP), IPfix, DNS, Network Time Protocol(NTP), Secure Shell Protocol (SSH), Device Language Message Specification (DLMS/COSEM), Distributed Network Protocol (DNP), MODBUS. All these protocols are built on either TCP or UDP. However, the protocol stack standardized by Institute of Electrical and Electronics Engineers (IEEE) and Internet Engineering Task Force (IETF) shows the Application layer as the topmost in the stack [6]. The above list of protocols may belong to Session layer, Transport layer or Application layers.

3.1.1. MQTT and versions

MQTT, led by Dr Andy Stanford-Clark has been standardized by OASIS in 2013 [7] and now is an open standard. MQTT is an IBM’s event-driven, lightweight many-to-many communication, publish-subscribe based protocol developed on TCP. MQTT is message-oriented, every message is published to an address, called “topic”. Every client subscribed to a topic receives every message published to the topic. An intermediate broker distributes messages from publishers to the respectively demanding client machines. Some of the brokers used are Mosquitto, Really Small Message Broker (RSMB), MQTT.js, HiveMQ, RabbitMQ and VerneMQ. Jose Luis Espinosa-Aranda et.al has proposed a tiny open-source MQTT broker for flexible and secure IoT deployments in [8].

The message format shown in Figure is usually expressed as 2-byte fixed header, a variable length header and payload, out of which the fixed header is mandatory whereas the other two are optional.

Fixed Header field (minimum 2 bytes)				Packet Length (1 to 4 bytes)	Variable Length Header (size depends on the type of message)	Variable Length Payload (Payload refers to the data sent)
Control Header (1 byte)						
Message Type	DUP flag	QoS Level	Retain			
4bits	1 bit	2 bits	1 bit			

Figure 2. MQTT Message Format

Message types include CONNECT, CONNACK, PUBLISH, PUBACK, PUBCOMP, SUBSCRIBE, SUBACK, UNSUBSCRIBE and many more. The DUP flag when set conveys to the receiver of already having received the data and indicates duplication. The QoS field indicates the delivery assurance assisted by three modes/profiles namely (a) Fire and forget/ At most once/ QoS0 (b) Acknowledged delivery/ At least once/ QoS1 (c) Assured delivery/ Exactly once/ QoS2.

MQTT is TCP/IP based and designed for constrained devices and low-bandwidth, high-latency networks, best suited as communications bus for live data. MQTT is therefore, an ideal messaging protocol for IoT and M2M communications. MQTT ensures reliability by providing three QoS levels. Semantic data extraction is supported by MQTT protocol and is one of the best suited paradigms for IoT [9], especially on battery-run devices. In fact, MQTT outperforms CoAP in managing higher traffic, lower latency, higher throughput, optimal memory, low power operation and CPU usage [10].

The MQTT design is suitable to operate in secure networks and has no security mechanisms imposed. Security in MQTT is based on SSL/TLS encryption, a relative standard for authentication in an IoT environment. A matter of concern is that SSL/TLS is quite expensive to be used for a constrained IoT environment. SMQTT is secure MQTT in which a message is encrypted and delivered to multiple nodes which is suitable for IoT applications. This broadcast encryption feature dependent algorithm of SMQTT has 4 stages of operation namely setup,

encryption, publish and decryption. MQTT- SN v1.2, formerly known as MQTT-S is a dedicated MQTT version for Sensor Networks handling embedded devices on non-TCP/IP networks, such as Zigbee. MQTT-SN too is a publish/subscribe messaging protocol operating beyond the reach of TCP/IP infrastructure i.e. UDP based for Sensor and Actuator solutions. MQTT-SN envisages power constraint oriented communication with a UDP platform and adds broker support to index topic names unlike MQTT. Secure versions SMQTT and SMQTT-SN have been augmented to MQTT and MQTT-SN respectively based on an attribute-based Key/Cipher Text Policy using Elliptic Curve Cryptography (ECC). The authors of [11] have explained the possible solutions in MQTT systems to implement different protection levels from varied network threats; however, ECC always has played a good choice for MQTT implementations. MQTT is finding its way into many domains [12] like Healthcare, Energy and Utilities, Industry and Irrigation systems, Social Networking and many IoT based applications.

The protocol has a low overhead in spite of operating on TCP when compared to other TCP based Application layer protocols [13]. MQTT can carry only a maximum of 256 MB of data and is hence suitable for expensive, unreliable networks. MQTT also experiences lower delays; uses bandwidth and battery moderately and hence well preferred in lower delay message delivery applications. The limitations of MQTT include limited security, broker overloading and hence message expiry, message ordering challenge and no message priority principle. The authors of [14] have experimentally compared the protocol efficiencies of CoAP, MQTT and WebSocket and revealed the mediocre performance of MQTT with QoS1 in terms of protocol efficiency. Results show CoAP to be the best, followed by WebSocket and MQTT with QoS0.

3.1.2. CoAP

CoAP is the brainchild of CoRE (Constrained Resource Environments) IETF group and enables web applications on smart objects [15]. CoAP is a one-to-one protocol best suited for a partially event based state transfer model and is built on UDP to provide a reliable low weight mechanism. CoAP provides a request and response communications model and supports end-to-end communication at the application layer between constrained IoT devices and other Internet devices. It works similar to HTTP in order to benefit from existing web-based technologies using the same methods (GET, PUT, POST, and DELETE) as HTTP, but with an additional ability for resource discovery and observation [16].

A standard interface called Representational State Transfer (REST) is the standard interface used between client and the servers in CoAP. CoAP operates with a 2-layer convention of Request/Response and Transaction/Messaging. The Request/Response layer manages the REST operation and the Messaging layer ensures reliable UDP communication with the help of exponential backoff.

The four messages of CoAP are Confirmable (CON), Non-Confirmable (NON), Acknowledge (ACK) and Reset (RESET).

A typical CoAP message is 10 to 20 bytes, the first fixed part is a 4-byte header, a token, Options and Payload fields which are not mandatory. The details of the CoAP message format are depicted in Figure 2. “Ver” is 2 bit unsigned integer for the version number of CoAP, followed by another 2 bit unsigned integer “T” to indicate the type of messages (CON, NON, ACK or RESET). The 4 bit “TKL” field represents the token. The 8 bit “Code” splits as a 3 bit class and 5 bit detail for MSBs and LSBs respectively. “Message ID” is used for matching responses and message duplication indication.

Ver	T	TKL	Code	Message ID
2 bits	2 bits	4 bits	8 bits	16 bits
Token (optional) – 0 to 8 bytes				
Options (optional)				
Payload (optional)				

Figure 3. CoAP Message Format

CoAP promises to fulfil low overhead, asynchronous message exchanges, URI and content-type support, simple parsing, enhanced reliability due to data reduction, reduced latency in low-power lossy wireless networks and multicast communication support in IoT based resource constrained environments. Other important features include Resource Observation, Block-wise resource Transport, Resource Discovery and easy interaction with HTTP [17,18].

Comparing MQTT with CoAP in terms of overhead, CoAP allures with its appreciable low overhead. However, due to the dearth of TCP retransmission mechanisms, packet losses tend to be on the higher side. CoAP races over MQTT with lesser traffic generation in the case of small-sized messages. CoAP outruns MQTT with lower delays but only when packet loss rate is high. On the contrary, for lower packet loss rates, MQTT delivery rates are comparatively quicker.

The authors of [19] propose an adaptive (Retransmission Time Out) RTO method rather than a fixed RTO, which consists of a Smooth Round-trip Time getting multiplied by a constant parameter (K) to reduce energy consumption of nodes by nearly 8% and improve reliability enhanced packet delivery ratio (PDR) of MQTT-SN and CoAP protocols. A comparative study by the authors of [20] for a Smartphone application between CoAP and MQTT showed that CoAP’s bandwidth usage is lesser than that consumed by MQTT. The authors of [21] have provided recorded results demonstrating the better side of CoAP in terms of energy usage and transmission time. Other metrics have also been analysed such as discarded publication message ratio, retransmitted message ratio and duplicated message ratio in support of the adaptive RTO method. The authors here state that a maximum achieved PDR is better with CoAP than what is achieved with MQTT-SN. CoAP does offer a basic congestion control mechanism for unicast messages, better congestion control mechanisms are required to only handle the gradually increasing traffic of multicast communications.

CoAP secure communication was earlier based upon IPSec [22]. Since CoAP is built on UDP, SSL/TLS cannot provide

security but can be achieved with DTLS. CoAP backed up by DTLS is unitedly termed as Secured CoAP (CoAPs). DTLS with enhanced features of TLS to deal with UDP communications of CoAP succeeds in targeting Confidentiality, Integrity, Non-Repudiation and Data Protection against Replay Attacks. CoAP provides inbuilt support for content negotiation and discovery thereby allowing devices to probe each other to find ways of exchanging data. The introduction of raw public keys with compressed DTLS in CoAP reduces message size and hence energy savings, avoidance of 6LoWPAN fragmentation at the link layer for larger datagram sizes and reduced burden on constrained devices during DTLS handshake. The suitability of DTLS and IPSec for CoAP security implementations is questionable in spite of their usage in many IoT based applications. Secured CoAP or CoAPs by implementing the three modes of DTLS namely:

- (a) RawPublicKey mode – An asymmetric raw public key pair is generated by the manufacturer and installed on the device. However, devices may have one or more raw public keys.
- (b) PreSharedKey – This mode is based on a list of pre-shared keys. Each key in turn includes many communicating nodes. The communication process to a new node includes a DTLS session start using the pre-shared key, the system selecting an appropriate key based on the destination nodes.
- (c) Certificate – The devices operating in this mode use a root trust anchor - validated X.509 certificate with an asymmetric key pair.

Although security implementations like DTLS for CoAP is a necessity, one should also be aware of the fact that quite a significant overhead will be added in constrained environments, thereby challenging the limitations on memory and/or bandwidth. A sensible solution would be to dislodge from unused nodes and make the protocol lighter and re-introduce them only when required. The suitability of DTLS and IPSec for CoAP security implementations is questionable in spite of their usage in many IoT based applications. However, DTLS does not support multicast communications since it lacks group key management. IPSec faces Network Address Translation (NAT), Port Address Translation (PAT) and multicast communication issues. Both IPSec and DTLS have an incompetent QoS, Access Control and network trust and rely upon out-of-the-box extra protocols like Extensible Authentication Protocol (EAP) and Internet Key Exchange (IKE).

On the contrary, CoAP is also known for its high latency, poor packet delivery and inability to be used for complex data types [23]. MQTT and MQTT-SN are quite prevalent than CoAP and find applications in the area of social networks, Vehicle to Vehicle communication (V2V) and sensor networks.

In spite of the constantly evolving upgrades of CoAP, cost, power efficiency, supreme data security, network robustness and application deployment, gullibility of a CoAP based system still

remains a challenge. In fact, the generic CoAP can no longer be used with increased number of transmissions and network congestion. The authors of [24] mention and discuss about an advanced CoAP termed as CoAP Congestion Control Mechanism (CoCoA). Comparative study is made based upon parameters like latency, throughput and re-transmission. The authors in [25] have carried out a CoCoA analysis and have implemented a 4-state-Strong CoCoA adaptation that uses a 4-state estimator for variable backoffs. Results signify an improvement in throughput and goodput even in highly lossy networks. CoCoA+ is add on to the CoAP and CoCoA and the drafters of this mechanism in [26] prove the upper hand performance of CoCoA+ with many use cases in a variety of network topologies. On the contrary, one of the authors in their works uplifts the degradations of CoCoA+ when compared to generic CoAP [27]. The results indicate that CoCoA+ can perform significantly worse than default CoAP, especially with burst traffic and in networks with few clients as a result of an improper selection of the retransmission timeouts (RTOs).

3.1.3. XMPP

XMPP originally coined as “Jabber” is a well demonstrated IETF protocol which provides both asynchronous (publish/subscribe) and synchronous (request/response) messaging supports. This TCP based, instant messaging standard protocol supports a variety of authentication patterns via the Simple Authentication and Security Layer (SASL – RFC4422). XMPP was designed for near real-time communications and therefore it supports small message imprint and low latency message exchanges [28] and is used in multi-party chatting, voice and video calling. XMPP was extended to IoT applications because of its eXtensible Markup Language (XML) feature, addressing, security and scalability features.

In terms of security, SASL provides a set of authentication methods from which the client can choose the best fit. SASL uses Base64 coding to hide recognizable information. While SASL is responsible for authentication, TLS looks after channel encryption for XMPP. XMPP is fulfilling the needs of IoT cloud providers in terms of message management and security. However, XMPP lacks native advanced security features to address security requirements of emerging federation-enabled IoT cloud scenarios [29]. The authors of [30] provide a security mechanism for XMPP based communication in sensor networks as well, but at the cost of extra overhead.

The overhead of XMPP too remains a concern to be used in IoT sometimes and requires a makeover in preferably the architecture. The cons are additional overhead due to gratuitous tags, increased power consumption due to complex computation and not many QoS options. In an attempt to unify XMPP with IoT, the authors of [31] have proposed a solution to unify sensors and actuators with Internet by omitting application protocol gateways and protocol translators. XMPP has been evolving from a simple Instant Messaging (IM) system to Cloud Computing.

3.1.4. DDS

DDS is a data-centric, PKI based certificate authentication protocol based on a brokerless, publish/subscribe architecture and hence more reliable with impressive QoS and suited for M2M as well as IoT. Object Management Group (OMG)’s DDS uses multicasting and also supports token mechanism catalysed by RSA and DSA algorithms. DDS uses a device-to-device relational data model to transfer data directly to the device using bus communication. DDS architecture is 2 layered as Data-Subscribe Publish-Subscribe (DCPS) and Data-Local Reconstruction Layer (DLRL). DCPS delivers data to subscribers. DLRL is an optional interface to DCPS. DDS is a standards-based QoS-enabled data centric middleware platform that enables applications to communicate by publishing information they have and subscribing to information they need in a timely manner [32]. DDS offers detailed QoS control, multicast, configurable reliability and pervasive redundancy [33] and resolves data distribution and management challenges [34].

3.1.5. AMQP

AMQP is a message-centric standard which is based on the publish/subscribe architecture similar to MQTT and runs on TCP. AMQP is an open standard used to send large number of messages per second [35] when compared to other RESTful services.

Exchanges and message queues constitute the AMQP broker and exchange information between each other according to pre-defined protocols. The exchanges route messages to appropriate queues. Queues store the received information and deliver to appropriate subscribers when required.

The key capabilities of AMQP are its ability to connect across technologies, organizations and time domains and hence, AMQP finds applications based on control plane or server-based analysis functions. AMQP is not very suitable for constrained environments and real-time applications. It does not support automation discovery too. However, AMQP is well interoperable in multiple environments. AMQP 1.0 is now approved as an International standard and has become an OASIS standard too.

3.1.6. EBHTTP

EBHTTP is a space-efficient, binary formatted, stateless encoding of the standard HTTP/1.1 protocol. It is used to transfer smaller messages in a constrained environment [37].

3.1.7. LTP

LTP allows constrained nodes/devices to exchange web service messages. The authors of [38] penned this versatile, lightweight Web service transport protocol in 2010. LTP allows the transparent exchange of Web Service messages between all kinds of resource-constrained devices and server or PC class systems.

3.2. Based on Key Distribution Schemes

IoT security solutions may either rely upon asymmetric key schemes or pre distributed symmetric keys. Each of the categories

Table 1: Summary of the Application Layer Protocols

Protocol	Architecture	Transport	QoS	Security	Areas of Application	Advantages	Limitations
MQTT	Asynchronous	TCP	Yes	SSL	Healthcare, Energy & Utilities, Industry & Irrigation, Social networking, IoT based applications	Low overhead, delays and power consumption, high latency, better than CoAP in traffic management, higher throughput, optimal memory and CPU usage	Moderate bandwidth and battery usage compared to CoAP
CoAP	Synchronous	UDP	Yes	DTLS	Live data communication, sensor networks, IoT based applications	URI & Content-type support, enhanced reliability, reduced latency, single parsing, multicasting, reduced bandwidth usage, good PDR	Packet losses due to TCP retransmissions, high cost, network robustness, application deployment gullibility
DDS	Asynchronous	TCP/UDP	Yes	SSL DTSL	M2M and IoT based applications, air traffic and vehicle control systems, industrial automation systems	Excellent QoS control, configurable reliability, pervasive redundancy, multicasting	Limited scalability, resiliency in data delivery, network heterogeneity
EBHTTP	Asynchronous	UDP	No	SSL	Applications involving transfer of smaller messages in constrained, hypermedia information systems	Resource discovery due to RESTful design, simplicity in design, extensibility of HTTP to suit highly constrained networks	No support for fragmentation, must follow HTTP caching behaviour
LTP	Synchronous	TCP/UDP	Yes	SSL	Web service message exchanges	Standard -compliant to Web services, combines with microfiber to give SOAP messages, header compression, message fragmentation	High implementation and maintenance cost
XMPP	Synch/Asynchronous	TCP	No	SSL	Voice & Video calls, chatting & message exchange applications	Good to use if application is already built and running with XML	High power consumption due to complex computations, additional overhead, no QoS and not suitable for M2M
AMQP	Asynchronous	TCP	Yes	SSL	Applications based on control plane & server-based analysis functions	Can connect across technologies, organizations and time domains, store-and-forward strategy for good reliability	Not suitable for constrained real-time applications, no support for automation discovery

has its own pros and cons. However, researchers use both of these and optimizing the existing solutions of Asymmetric and Symmetric Key schemes continues to be the area of prominence.

3.2.1. Asymmetric Key Schemes (AKS)

Asymmetric algorithms are commonly deployed in conventional internet. But, AKSs quote comparatively higher computation cost and consume higher energy for operation. Such

schemes are widely implemented for IoT based applications since they offer high resilience against node capture attacks, have low memory requirements for keying materials, few message exchanges and high scalability for large networks. Asymmetric approaches can be Public key encryption key based transport wherein a public key (secret code) is the authentication link to information sharing parties. This approach is more vulnerable to man-in-the-middle attacks. The key establishment techniques

may range from simple traditional mechanisms like raw public key encryption, certificate based encryption and identity based encryption to higher levels of complicated X.509 based implementations. Security certificates are expensive and hence require few hardware and software improvements in design. The need for optimization and cryptographic hardware accelerators arises. Identity Based Schemes (IBSs) provides a well-known identity which acts as the public key. A trusted party called the Public Key Generator (PKG) generates the private key of each entity. Even though certificates are eliminated here, IBSs are prone to key-escrow attacks. IBSs based implementations like RSA or ElGamal type IBE, IBAKA, TinyIBE are already being used in many applications.

3.2.2. Symmetric Key Schemes (SKS)

There is a demand for high memory space for keying materials, lower scalability for wider networks and vulnerability against node capture attacks. Therefore, SKS cannot be considered as the default protocol for IoT. Diffie-Hellman (DH) protocol is expensive and not suitable for constrained environments, a variant of DH called the Elliptic Curve DH (ECDH) protocol helps. ECDH is based on Elliptic Curve Cryptography (ECC) and has a relatively smaller key size than with the RSA algorithm. A Digital Signature Algorithm ECDH-EDSA algorithm [39] is an effective key agreement protocol too. The HIP-DEX algorithm generates an ECDH encrypted session key between two entities after 4 messages and uses the least number of cryptographic primitives. Many IoT based works rely upon HIP-DEX. The authors of [40] have reused ECDH boosted with the session resumption mechanism. The authors of [41] came up with a combination of ECDH-IBE, IBE uses an ECC primitive. However, it still demands 2 bilinear pairings and 3 scalar point multiplications each time a session key is bootstrapped. As an attempt to eliminate pairing, TinyIBE [42] was proposed in which the session key between two nodes is retrieved after just two messages.

3.3. Based on nature of IoT Application

The authors of [43] have attempted to categorize IoT protocols based on the nature of IoT applications and their core functionalities. The categories are the Application protocols, Service Discovery protocols, Infrastructure protocols and Influential protocols. The Application protocols include DDS, CoAP, AMQP, MQTT, MQTT-SN, XMPP and HTTP REST.

The Service Discovery protocols are listed as Multicast DNS (mDNS) and DNS Service Discovery (DNS-SD). mDNS is well suited for Internet based embedded devices since its working is unaffected by infrastructure failure. mDNS can execute the unicast DNS server operation. The working operation of mDNS can be analyzed as follows. mDNS sends IP multicast messages at once to all nodes in its local domain inquiring by NAME for the preferred client node. When the target node receives its NAME, it will send a response to the calling mDNS along with

its IP address. Devices which receive the response message will update their NAME and IP address in their respective local cache.

DNS-SD is similar to mDNS with respect to the fact that it too does not require additional manual configuration or administration. In fact, the client machines use mDNS and pair required services to constitute the DNS-SD. mDNS finds the required services by host name and pairs their IP addresses with them. Since mDNS are DNS-SD are configuration independent protocols, they are suited for IoT based implementations wherein smart devices can join or quit the platform without affecting the entire system operation. On the contrary, these protocols demand caching DNS entries for constrained devices and cache handling and timing operations can be challenging enough to consider other protocols instead.

The Infrastructure protocols are actually the whole sum set of Network, Link and Physical layers which are Long Term Evolution – Advanced (LTE-A), EPCglobal, IEEE 802.15.4, Z-Wave, 6LoWPAN, IPv4, IPv6 and Routing Protocol for Low Power and Lossy Networks (RPL). LTE-A promises reasonable service costs and scalability as far as cellular solutions matter. Architectural essentials include the Core Network (CN) dealing with packet flows and device control and Radio Access Network (RAN) for radio access. Base stations typically called evolved nodes and represented as eNBs connect each other through the X2 interface. RAN and CN connect through S1 interface. And finally, other mobile devices connect through the gateway. LTE-A uses the Orthogonal Frequency Division Multiple Access (OFDMA) to partition bandwidth into smaller bands called Physical Resource Blocks (PRBs). Problems of QoS compromise and network congestion come along with LTE-A protocol, solutions however exist to lower contention in network.

EPCglobal manages Electronic Product Code (EPC) and RFID technologies and standards. Its architecture supports good interoperability, reliability and scalability. The entire RFID based tag system works on two components – the tag and tag reader. A chip in the tag is the storage element which has an object's unique identity. This chip communicates with the tag reader with an antenna using radio waves. The tag reader passes over the unique identity/tag number to a computer application named Object Naming Service (ONS) which further interacts with the IoT applications.

IEEE 802.15.4 finds a reasonable place in the choices for IoT, M2M and WSNs due to its reliability in communication, low cost, power consumption and data rate, high throughput and interoperability but at the cost of poor QoS. The protocol uses three Direct Sequence Spread Spectrum (DSSS) modulation technique.

Z-Wave is a low power protocol preferred for low distance data transmissions typically of few meters and hence finds applications in home appliances, light control, access control, wearable technology etc. The architecture comprises of the

controller and slave nodes, controller maintains a table for updating and hence monitoring routing strategies of the topology.

6LoWPAN supports IPv6 with mapping services, provides fragmentation and header compression (headers typically compressed to two bytes [44]). Link layer forwarding for multi hop delivery and IPv6 overhead reduction are added features of 6LoWPAN.

RPL supports multipoint-to-point, point-to-multipoint and point-to-point communications. The essence of RPL is a directed acyclic graph with a single root node called Destination Oriented Directed Acyclic Graph (DODAG) responsible for routing. The RPL routers work in either Storing mode or Non- Storing mode. In the former, destination IPv6 addresses direct the downward routing whereas, in the latter, IP source routing come into picture.

There is the Influential protocol category that includes IEEE 1888.3, IPSec and IEEE 1905.1. It is evident that IoT environments have many underlying technologies and interoperability is essential and this category of protocols aims for the same. In fact, IEEE 1905.1 standard was designed for heterogeneous technologies and convergent digital networks.

4. AL Security – Vulnerabilities and Issues

IoT Security Systems Engineering is constantly evolving with state-of-the-art security approaches to counter the exponentially growing “headless” security threats. Defining and designing a protective architecture is definitely a security requirement at the system or architecture level. However, we restrict our discussion to protocol based security authentication, especially at the Application layer. Achieving end-to-end security triggers network challenges due to the discrepancy between the high demand for security standards and the available envisioned constrained hardware. Unprotected protocols (without security based implementations) are often vulnerable to various network attacks, eavesdropping, spoofing etc. Having SSL/TLS, IPSec, DTLS or any other security mechanism still does not assure the protocol of flawless security. In fact, IPSec faces Network Address Translation (NAT), Port Address Translation (PAT) and multicast communication issues. DTLS does not support multicast communications since it lacks group key management. Both IPSec and DTLS have an incompetent QoS, Access Control and network trust and rely upon out-of-the-box extra protocols like Extensible Authentication Protocol (EAP) and Internet Key Exchange (IKE). SSL/TLS is expensive to be used in constrained devices.

Vulnerabilities are the weaknesses of a system due to poor design which allow the network to be hacked illegally. An attacker may bank upon improperly maintained network access and permissions, buffer overflow, cross site scripting, error configurations, data tampering and poor data authentication mechanisms. The authors of [45] provide a classification for security threats in the Application layer. They are Privacy Leak, DoS Attack, Malicious Code and Social Engineering.

Another major setback to AL security has ever since been the distribution of keys among devices [46]. Few solutions like vendor based access control and virtual networks have helped, but not been a major breakthrough to handle key distribution issues very effectively. General security measures and counter attacks can be put up as Data Security, Authentication, Trust Management, Risk Assessment and Intrusion Detection. However, below is a tabulation of the possible vulnerabilities and challenges threatening the Application Layer.

5. Research Challenges and Proposals

Research still prevails to minimize potential threats and probable network attacks. The authors of [68] had proposed a DTLS improvement to send multiple CoAP messages in a multicast group using a common group key. Large buffers are required at the receiver end to hold data for retransmission due to inadequate timers in DTLS and code size required to support DTLS. Stateless compression of DTLS headers help to reduce overhead [69]. There are few DTLS header compression implementations, one of them being the usage of LOWPAN_IPHC 6LoWPAN [70]. The authors of [71] proposed the RESTful DTLS handshake to confront the fragmentation limitation. Larger messages are transferred in blocks. To mitigate costs of DTLS operations common security gateways are mapped between TLS and DTLS as well as between CoAP and HTTP. Mutual authentication using DTLS not using ECC too was a proposal of an end-to-end architecture that used specialized trusted-platform modules (TPM) that supports RSA cryptography. Public-key and Digital Certificates support involve computational complexity.

Works attempting to optimize this complexity [72] have come up with certification pre-validation and session resumption to eliminate the need for additional handshake. The authors of [73] have proposed an optimized DTLS integration within CoAP with minimum ROM usage and ECC technique. The proposal highlights block wise message transfer and message reordering. Newer assembly routines which use registers more effectively have been added to minimize the number of memory operations and reduce RAM and ROM occupancy. The authors have used an ECC library of their own which are based on TinyECC and Relic libraries further reducing complex operation execution time.

The authors of [74] provide another breakthrough CoAP based Communication Architecture for sensor and actuator networks (CASAN) to reduce device constraints, reduce intelligence (software complexity, code execution) at the minimum needed in constrained nodes and transfer it to a more competent device which acts as gateway between the sensors and Internet. The basic idea is to have a “REST” level communication by providing a RESTful interface for all sensors and simplify application programming tasks. CoAP based communication proves to be in the best list if we are able to minimize the number of intermediate servers and yet provide secure data delivery over large distances. The authors of [75] have proposed a scalable, flexible embedded CoAP solution for Web applications or the

Table 2: Application Layer Threats and Vulnerabilities

Vulnerability/Challenges		Problem Description	Solutions Proposed
Attacks	DoS	Deceiving node to breach defensive system	<ul style="list-style-type: none"> ➤ Dynamic threat anticipation ASTM [47, 48] – Adaptive learning technique with changing internal parameters ➤ Risk transfer mechanism based security systems [49] ➤ Support for Software Defined Networks (SDNs) architectures [50]
	Sphear Phishing	Luring emails for adversary gains	
	Sniffing	Introduction of a sniffer application into the system	
	Overwhelm	Undue consumption of energy by nodes and bandwidth	
Insecure web interface & Data Privacy		Log and keys leakage at IoT end-node, illegitimate malicious nodes feeding contaminating data and/or accessing critical information (Malicious Code Injection due to end user hacking techniques)	<ul style="list-style-type: none"> ➤ Preference Based Privacy ➤ Protection Method - Third party evaluation, report to service provider and appropriate security level based sensed preferences [51]
Insecure mobile interface & Cloud Interface		Unsecured apps, no Device Lockout, In-Cloud data leakage, Cross site scripting, poorly configured SSL/TSL	<ul style="list-style-type: none"> ➤ Stronger passwords ➤ Testing the interface against the vulnerabilities of software tools (SQLi and XSS) ➤ Using https along with firewalls [52]
Insecure Remote Security Configuration		Fails to implement security measures @ interfaces, IoT end -node, end-device, end-gateway, no security logging, lack of granular permission model, lack of add-on password security options, lack of comprehensive security management	<ul style="list-style-type: none"> ➤ Remote safe configuration ➤ Scalable security enhancement system of the SMC model for distributed resources – SMC [53] ➤ Simplified security management of network security teams
Insecure Software/Firmware		Threats to system from pirated softwares, malware installations, unencrypted update files, inability to receive timely security patches	<ul style="list-style-type: none"> ➤ Encryption with validation ➤ Anti-virus, anti-adware, firewalls, Real Time Intrusion Detection Systems (IDS) [54] ➤ Security patches ➤ Code with languages such as JSON, XML, SQL and XSS needs to be tested carefully
Insufficient Authentication/Authorization		Lack of multifactor authentication, unsecure password recovery mechanism, Account Enumeration, lack of Role based access, No account Lockout	<ul style="list-style-type: none"> ➤ Cross-layer authentication and authorization ➤ Sensitive information isolation/Data leakage protection ➤ Administrator/Identity Manager authentication ➤ Effective Key coordinate sharing, frequent key coordinate updates [55, 56] ➤ Identity Authentication and Capability based Access Control (IACAC) [57] ➤ Strong Encryption schemes ➤ Cryptographic Hash functions & Feature Extraction – [58] ➤ Decentralized control of authentication using user-dependent security context [59]

Risk Assessment/Trust Management	Lack of convenient tools for real time risk expectancy, threat detection and security reporting, absence of global and standard trust policies	<ul style="list-style-type: none"> ➤ Security quantified in terms of incident and asset loss – CCM [60] ➤ Mutual trust for inter-system security [61] ➤ Agent-based and weight-based trust models
Lack of Protocol Standardization &	Lack of global standards and policies guiding development of security protocols, failure of existing policies to provide 100% protection from threats	<ul style="list-style-type: none"> ➤ Smart Object Lifecycle Architecture for Constrained Environments (SOLACE) [62]
Existing protocols coping with newer & stronger threats	Network bottlenecks are still prevalent in existing security protocols which are only relatively successful (like CoAP) [63]	<ul style="list-style-type: none"> ➤ TLS/DTLS and HTTP/CoAP mapping ➤ Mirror Proxy (MP) and Resource Directory ➤ TLS-DTLS tunnel and message filtration using 6LBR 64-67]

Web of Things (WoT) in general, integrating the browsers and Web clients without intermediate gateways and proxies. Authors of [76] discuss upcoming CoAP options to enhance security in CoAP by highlighting a granular per message based security scheme.

There are a few proposed approaches to few CoAP research challenges discussed below: (i) Group key management mechanisms may be applied externally to CoAP or integrated within the DTLS handshake. (ii) Security gateways can offer intrusion detection and attack tolerance mechanisms [77, 78, and 79]. (iii) Online certification validation can be improved with a foundational idea like the one discussed in [80] about Online Certification Status Protocol (OCSP). Another related work is based on OCSP stapling using TLS Certificate Status Request extension defined in RFC 6066 [81]. (iv) Optimized hardware design to handle computational complexity and cost imposed due to ECC implementations. (v) Support for varying heterogeneous Convergent Networks considering the possible compatibility and performance issues. (vi) Combination of communication paradigms such as open cloud resource access and single hop long range rather than the former multi hop short range communications.

6. Conclusion

IEEE, IETF and International Telecommunication Union (ITU) have provided several standards and security mechanisms in order to cater to the demands of the uprising IoT. However, a designer is free to rise up with an entirely new authentication protocol or bring about modifications in the existing chain of protocols.

Working in an IoT environment involves IoT devices operating in a wireless environment which are constrained in terms of battery life, processing power, and memory which invites a number of networking challenges. Each of the graded and regulated protocols complement each other and work in coordination for the very cause of IoT security in spite of the fact that they behave differently at different layers in their individual operation. Object security is of primary concern rather than the

layer security, be it transport or the application layer. Security mechanisms have to be incorporated or embedded within the protocol itself.

However, the primary objective of this study was to lay the foundation to a state-of-the-art security authentication protocol in the Application layer for IoT applications. The paper can help understand the concerns, issues and progress of research ideas to secure the IoT protocols of the Application layer. The review of the most widely used protocols in terms of operation and security indicates that none of them actually are the best. Each protocol has its own pros and cons, but a wise trade-off between protocol parameters has to be made which is purely application dependent. As far as Application Layer security is concerned, it must be rated the topmost priority parameter which needs to be taken care of since it is the most vulnerable layer to the user world of cyber attacks. The end point of this paper leads to the beginning of research ideas to defend the AL, be it improvisations in Trust Management, Key Management Strategies, Intrusion Detection systems, Encryption schemes and many more to follow. Further research intends to provide a broader contribution to improvised key management strategies for AL security generic to any domain. However, choice of protocol would be application and domain dependent and performance may be evaluated experimentally for each of the predominant AL protocols.

Conflict of Interest

The authors declare no conflict of interest.

References

- [1] Cloud Security Alliance (CSA), April 2015, 'Security Guidance for Early Adopters of the Internet of Things'.
- [2] Keoh, S., Kumar, S., and Tschofenig, H. (2014), 'Securing the Internet of Things: A Standardization Perspective', Internet of Things Journal, IEEE, Vol. 1, No. 3, pp. 265-275.
- [3] Lavinia, Nastase 2017, 'Security in the internet of Things: A Survey on Application Layer Protocols', 2017 21st International Conference on Control Systems and Computer Science, IEEE.
- [4] Pallavi, Sethi & Smruti, Sarangi 2017, 'Internet of Things: Architectures, Protocols, and Applications', Journal of Electrical and Computer Engineering Volume, Article ID 9324035, 25 pages, <https://doi.org/10.1155/2017/9324035>.

- [5] Tara, Salman & Prof. Raj, Jain 2015, 'Networking Protocols and Standards for Internet of Things', Washington University in St.Louis, Recent Advances in Networking.
- [6] J, Granjal, E, Monteiro & J, Silva 2010, 'Security for the internet of things: A survey of existing protocols and open research issues', IEEE Communications Surveys Tutorials, vol. 17, no. 3, pp. 1294-1312.
- [7] D. Locke 2010, 'MQ telemetry transport (MQTT) v3. 1 protocol specification', IBM Developer Works Technical Library.
- [8] Jose, Espinosa 2015, 'A tiny open-source MQTT broker for flexible and secure IoT deployments', Communications and Network Security (CNS), IEEE.
- [9] Satyavrat, Wagle 2016, 'Semantic Data Extraction over MQTT for IoT-centric Wireless Sensor Networks', International Conference on Internet of Things and Applications (IOTA) Maharashtra Institute of Technology, Pune, India 22 Jan - 24 Jan, 2016.
- [10] Muneer, Yassein, Mohammed, Shatnawi, & Du'a, Al-Zoubi 2016, 'Application Layer Protocols for the Internet of Things', IEEE International Conference on Internet of Things and Pervasive Systems, At 22-24 September 2016, Agadir, Morocco.
- [11] Giovanni, Perrone, Massimo, Vecchio, Riccardo, Pecori & Raffaele, Giuffreda 2017, 'The Day After Mirai: A Survey on MQTT Security Solutions After the Largest Cyber-attack Carried Out through an Army of IoT Devices', Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security (IoTBSDS 2017), pages 246-253.
- [12] Dipa, Soni & Ashwin, Makwana 2017, 'A Survey on MQTT: A Protocol of Internet of Things (IoT)', International Conference on telecommunication, Power Analysis and Computing Techniques (ICTPACT-2017).
- [13] Dinesh, Thangavel, Xiaoping, Ma, Alvin, Valera, Hwee-Xian, Tan, Colin, Keng-Yan Tan 2014, 'Performance Evaluation of MQTT and CoAP via a Common Middleware', IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 21-24 April 2014, pp. 1-6.
- [14] Stefan, Mijovic, Erion, Shehu & Chiara, Buratti 2016, 'Comparing Application Layer Protocols for the Internet of Things via Experimentation', 2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI).
- [15] Z, Shelby, K, Hartke & C, Bormann 2013, 'Constrained application protocol (CoAP), draft-ietf-core-coap-18', IETF, 2013.
- [16] Angelo, Castellani, Mattia, Gheda, Nicola, Bui, Michele, Rossi & Michele, Zorzi, 'Web Services for the Internet of Things through CoAP and EXI', IEEE International Conference on Communications Workshops (ICC), 5-9 June 2011, pp. 1-6.
- [17] C. Bormann, A. P. Castellani, Z. Shelby, "CoAP: An application protocol for billions of tiny Internet nodes", *IEEE Internet Comput.*, vol. 16, no. 2, pp. 62-67, Mar./Apr. 2012.
- [18] C. Lerche, K. Hartke, M. Kovatsch, "Industry adoption of the Internet of Things: A constrained application protocol survey", *Proc. IEEE 17th Conf. ETFA*, pp. 1-6, 2012.
- [19] Davis, Ernesto, Calveras, Anna & Demirkol, Ilker 2013, 'Improving packet delivery performance of publish/subscribe protocols in wireless sensor networks', Vol. 13, No. 1. Multidisciplinary Digital Publishing Institute, 2013, pp. 648-680.
- [20] De Caro, N, Colitti, W, Steenhaut, K, Mangino, G & Reali, G 2013, 'Comparison of two lightweight protocols for smartphone-based sensing', Communications and Vehicular Technology in the Benelux (SCVT), 2013 IEEE 20th Symposium On, pp. 1-6.
- [21] Colitti, W, Steenhaut, K, De Caro, N, Buta, B and Dobrota, V 2011, 'Evaluation of constrained application protocol for wireless sensor networks', Local & Metropolitan Area Networks (LANMAN), 2011 18th IEEE Workshop On, 2011, pp. 1-6.
- [22] Thamer, Alghamdi, Aboubaker, Lasebae & Mahdi, Aiash 2013, 'Security Analysis of the Constrained Application Protocol in the Internet of Things', Second International Conference on Future Generation Communication Technologies (FGCT), London, UK.
- [23] Talaminos-Barroso, A, Estudillo-Valderrama, M. A., Roa, L. M.rrrr, Reina-Tosina, J., & Ortega-Ruiz, F 2016, 'A Machine-to-Machine protocol benchmark for eHealth applications-Use case: Respiratory rehabilitation', Computer Methods and Programs in Biomedicine, 129, 1-11, 2016.
- [24] Ananya, Pramanik, Ashish, Luhach, Isha, Batra & Upasana, Singh 2017, 'A Systematic Survey on Congestion Mechanisms of CoAP Based Internet of Things', Advanced Informatics for Computing Research, July 2017, pp 306-317.
- [25] Rahul, Bhalerao, Sridhar, Subramanian & Joseph, Pasquale 2016, 'An analysis and improvement of congestion control in the CoAP Internet-of-Things protocol, Consumer Communications & Networking Conference (CCNC), Jan 2016.
- [26] August, Betzler, Carles, Gomez, Ilker, Demirkol, Josep, Paradells 2015, 'CoCoA+: An advanced congestion control mechanism for CoAP', Elsevier, October 2015.
- [27] Emilio, Ancillotti & Raffaele, Bruno 2017, 'Comparison of CoAP and CoCoA+ Congestion Control Mechanisms for Different IoT Application Scenarios', IEEE Symposium on Computers and Communications (ISCC), Heraklion, Greece.
- [28] Sven, Bendel, Thomas, Pringer, Daniel, Schuster, Alexander, Schill, Ralf, Ackermann & Michael, Ameling 2013, 'A Service Infrastructure for the Internet of Things based on XMPP, IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), 18-22 March 2013, pp. 385-388.
- [29] Antonio Celesti, Maria Fazio, Massimo Villari, 'Enabling Secure XMPP Communications in Federated IoT Clouds Through XEP 0027 and SAML/SASL SSO', Sensors (Basel), 2017 February 17(2): 301, doi: 10.3390/s17020301.
- [30] Longhua Guo, Jun Wu, Zhengmin Xia, Jianhua Li, 'Proposed Security Mechanism for XMPP-Based Communications of ISO/IEC/IEEE 21451 Sensor Networks', IEEE SENSORS JOURNAL, VOL. 15, NO. 5, MAY 2015.
- [31] Michael, Kirsche, Ronny, Klauk 2012, 'Unify to bridge gaps: Bringing XMPP into the Internet of Things', Pervasive Computing and Communications Workshops (PERCOM Workshops), March 2012.
- [32] Douglas, Schmidt, Angelo, Corsaro & Hans, Hag 2008, 'Addressing the Challenges of Tactical Information Management in Net-Centric Systems with DDS, <http://citeseerx.ist.psu.edu>.
- [33] Gururaj, Kulkarni, Manoor, S, Mitragotri, P V 2017, 'Enabling Technologies, Protocols, and Applications: A Detailed Survey on IOT', International Journal of Advance Research, Ideas and Innovations in Technology, Vol 3, Issue 2.
- [34] Douglas, Schmidt & Hans, Hag 2008, 'Addressing the challenges of mission-critical information management in next-generation net-centric pub/sub systems with OpenSplice DDS', Parallel and Distributed Processing, IEEE International Symposium, April 2008.
- [35] Joel, Fernandes, Ivo, Lopes, Joel, J P, Rodrigues, C & Sana, Ullah 2013, 'Performance Evaluation of RESTful Web Services and AMQP Protocol', Fifth International Conference on Ubiquitous and Future Networks (ICUFN), 2-5 July 2013, pp. 810-815.
- [36] Xi Chen, Constrained Application Protocol for Internet of Things, <http://www.cse.wustl.edu>.
- [37] Nils, Glombitza, Dennis, Pfisterer & Stefan, Fischer 2010, 'LTP: An Efficient Web Service Transport Protocol for Resource Constrained Devices Sensor Mesh and Ad Hoc Communications and Networks (SECON)', 2010 7th Annual IEEE Communications Society Conference, June 2010.
- [38] Kim Thuat, Nguyen, Maryline, Laurent & Nouha, Oualha 2015, 'Survey on secure communication protocols for the Internet of Things', Elsevier, ScienceDirect, February 2015, Pages 17-31.
- [39] Moskowitz, R & Jokela, P 2013, 'Host Identity Protocol version 2 (HIPv2), Draft-Internet, 2013.
- [40] De Meulenaer, G et al. 2008, 'On the energy cost of communication and cryptography in wireless sensor network', IEEE International Conference on Wireless and Mobile Computing, Network & Communication, 2008.
- [41] Yang, L, Ding, C & Wu, M 2013, 'Establishing Authenticated Pairwise Key using Pairing-based Cryptography for Sensor Network', 8th Chinacom, 2013.
- [42] Szczechowiak, P & Collier, M 2009, 'TinyIBE: identity-based encryption for heterogeneous sensor networks', 5th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2009.
- [43] Al-Fuqaha, Guizani, Mohammadi, M, Aledhari, M & Ayyash, M 2015, 'Internet of things: A survey on enabling technologies, protocols and applications, IEEE Communications Surveys Tutorials, vol. PP, no. 99, 2015.
- [44] J. W. Hui, D. E. Culler, "Extending IP to low-power wireless personal area networks", *IEEE Internet Comput.*, vol. 12, no. 4, pp. 37-45, Jul./Aug. 2008.
- [45] Weizhe, Zhang, Baosheng, Qu 2013, 'Security Architecture of the Internet of Things Oriented to Perceptual Layer', International Journal on Computer, Consumer and Control (IJ3C), Vol. 2, No.2 (2013).
- [46] I. Ishaq et al., "IETF standardization in the field of the Internet of Things (IoT): A survey", *J. Sens. Actuator Netw.*, vol. 2, pp. 235-287, 2013.
- [47] Abie H., and Balasingham I., "Adaptive security and trust management for autonomic message-oriented middleware", IEEE 6th Int. Conference on Mobile Ad hoc and Sensor Systems (MASS'09), pp. 810-817, 2009.

- [48] Sathish Alampalayam Kumar, Tyler Vealey, Harshit Srivastava, "Security in Internet of Things: Challenges, Solutions and Future Directions", 2016 49th Hawaii International Conference on System Sciences, IEEE.
- [49] Vipindev Adat ; Amrita Dahiya ; B. B. Gupta, "Economic incentive based solution against distributed denial of service attacks for IoT customers", 2018 IEEE International Conference on Consumer Electronics (ICCE), March 2018.
- [50] Akhunzada, A., Gani, A., Anuar, N. B., Abdelaziz, A., Khan, M. K., Hayat, A., and Khan, S. U. (2016). Secure and dependable software defined networks. *Journal of Network and Computer Applications*, 61, 199–221.
- [51] Tao and Peiran, "Preference-based Privacy Protection Mechanism for the Internet of Things", *International Symposium on Information Science and Engineering*, (ISISE), pp. 531 - 534 2010
- [52] OWASP, Top IoT Vulnerabilities, 2016. URL https://www.owasp.org/index.php/Top_IoT_Vulnerabilities
- [53] Pierre de Leusse., Panos Periorellis., Theo Dimitrakos., and Srijiith K., Nair., "Self-Managed Security Cell, a security model for the Internet of Things and Services", *First International Conference on Advances in Future Internet*, pp. 47 – 52, 2009.
- [54] S. Raza, L. Wallgren, and T. Voigt, "SVELTE: Real-time intrusion detection in the Internet of Things," *Ad Hoc Netw.*, vol. 11, no. 8, pp. 2661–2674, Nov. 2013.
- [55] Q. Wen, X. Dong, and R. Zhang, "Application of dynamic variable cipher security certificate in internet of things," in *Int'l Conference on Cloud Computing and Intelligent Systems (CCIS)*, 1062-1066, 2012.
- [56] Tasneem Yousuf, Rwan Mahmoud, Fadi Aloul, Imran Zuakernan, "Internet of Things (IoT) Security: Current Status, Challenges and Countermeasures", *International Journal for Information Security Research (IJISR)*, Volume 5, Issue 4, December 2015.
- [57] P. N. Mahalle, B. Anggorojati, N. R. Prasad, and R. Prasad, "Identity authentication and capability based access control (iacac) for the internet of things," *J. of Cyber Security and Mobility*, vol. 1, 309-348, 2013.
- [58] G. Zhao, X. Si, J. Wang, X. Long, and T. Hu, "A novel mutual authentication scheme for Internet of Things," in *Int'l Conference on Modelling, Identification and Control (ICMIC)*, 563-566, 2011.
- [59] X. Duan and X. Wang, "Authentication handover and privacy protection in 5G hetnets using software-defined networking," *Communications Magazine*, vol. 53, 28-35, 2015.
- [60] Weiß, S., Weissmann, O., and Dressler, F., "A comprehensive and comparative metric for information security", In *Proceedings of IFIP International Conference on Telecommunication Systems, Modeling and Analysis (ICTSM'05)*, pp. 1-10, 2005.
- [61] Y. Xie and D. Wang, "An Item-Level Access Control Framework for Inter-System Security in the Internet of Things," in *Applied Mechanics and Materials*, 1430-1432, 2014.
- [62] IETF SOLACE Info Page. Available online: <https://www.ietf.org/mailman/listinfo/solace> (accessed on 27 December 2012).
- [63] Minhaj Ahmad Khan , Khaled Salah, "IoT security: Review, blockchain solutions, and open challenges", *Future Generation Computer Systems*, 2017.
- [64] M. Brachmann, S.L. Keoh, O.G. Morchon, S.S. Kumar, End-to-end transport security in the IP-based Internet of Things, in: 2012 21st International Conference on Computer Communications and Networks, ICCCN, 2012, pp. 1–5. <http://dx.doi.org/10.1109/ICCCN.2012.6289292>.
- [65] J. Granjal, E. Monteiro, J.S. Silva, Application-layer security for the WoT: extending CoAP to support end-to-end message security for internet-integrated sensing applications, in: *International Conference on Wired/Wireless Internet Communication*, Springer Berlin Heidelberg, 2013, pp. 140–153.
- [66] M. Sethi, J. Arkko, A. Kernen, End-to-end security for sleepy smart object networks, in: 37th Annual IEEE Conference on Local Computer Networks - Workshops, 2012, pp. 964–972. <http://dx.doi.org/10.1109/LCNW.2012.6424089>.
- [67] M. Brachmann, O. Garcia-Morchon, S.-L. Keoh, S.S. Kumar, Security considerations around end-to-end security in the IP-based Internet of Things, in: 2012 Workshop on Smart Object Security, in Conjunction with IETF83, 2012, pp. 1–3.
- [68] Keoh, S, Kumar, S, Garcia-Morchon, O & Dijk, E 2014, 'DTLS-Based Multicast Security for Low-Power and Lossy Networks (LLNs)', 2014.
- [69] Hartke, K 2014, ' Practical Issues With Datagram Transport Layer Security in Constrained Environments', Issues-01, 2014.
- [70] Shahid, R, Daniele, T & Voigt, T 2012, '6LoWPAN compressed DTLS for COAP', *Proc. 8th IEEE Int. Conf. DCOSS*, 2012, pp. 287–289.
- [71] Keoh, S, Kumar, S & Shelby, Z 2013, 'Profiling of DTLS for CoAP-Based IoT Applications', draft-keoh-dice-dtls-profile-iot-00, 2013.
- [72] Hummen, R, Ziegeldorf, J, Shafagh, H, Raza, S & Wehrle, K 2013, 'towards viable certificate-based authentication for the Internet of things', *Proc. 2nd ACM Workshop Hot Topics Wireless Netw. Security Privacy*, 2013, pp. 37–42.
- [73] Angelo, Caposelle, Valerio, Cervo, Gianluca, Cicco & Chiara, Petrioli 2015, 'Security as a CoAP resource: An optimized DTLS implementation for the IoT', *IEEE ICC 2015 SAC – Internet of Things*.
- [74] Pierre, David & Thomas, No'el 2015, 'CASAN: A New Communication Architecture for Sensors Based on CoAP', *Proceedings of 2015 IEEE 12th International Conference on Networking, Sensing and Control*, Howard Civil Service International House, Taipei, Taiwan, April 9-11, 2015.
- [75] Miguel, Castro, Antonio, Jara & Antonio, Skarmeta 2014, 'Enabling end-to-end CoAP-based communications for the Web of Things', Elsevier, October 2014.
- [76] Granjal, J, Monteiro, E & Sá Silva, J 2013, 'Application-layer security for the WoT: Extending CoAP to support end-to-end message security for Internet-integrated sensing applications', *Wired/Wireless Internet Communication*. Berlin, Germany: Springer-Verlag, 2013, pp. 140–153.
- [77] Butun, E, Morgera, S D & Sankar, R 2014, 'A survey of intrusion detection systems in wireless sensor networks', *IEEE Commun. Surveys Tutorials* vol. 16, no. 1, pp. 266–282.
- [78] Young, M & Boutaba, E 2011, ' Overcoming adversaries in sensor networks: A survey of theoretical models and algorithmic approaches for tolerating malicious interference', *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 617–641, 2011.
- [79] Abduvaliyev, A, Pathan, A, Jianying, Z, Roman, R & Wong, W C 2013, 'On the vital areas of intrusion detection systems in wireless sensor networks', *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1223–1237, 2013.
- [80] Myers, M, Ankney, R, Malpani, A, Galperin, S & Adams, C 1999, 'X. 509 Internet Public Key Infrastructure Online Certificate Status Protocol-OCSP, RFC 2560, 1999.
- [81] Eastlake, D 2011, 'Transport Layer Security (TLS) Extensions: Extension Definitions, RFC 6066.

Managing and Optimizing Quality of Service in 5G Environments Across the Complete SLA Lifecycle

Evgenia Kapassa^{*1}, Marios Touloupou¹, Panagiotis Stavrianos¹, Georgios Xylouris², Dimosthenis Kyriazis¹

¹University of Piraeus, Department of Digital Systems, Piraeus, Greece

²Institute of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece

ARTICLE INFO

Article history:

Received: 20 December, 2018

Accepted: 15 February, 2019

Online : 28 February, 2019

Keywords:

Quality of service provisioning

5G networks

Service level agreements

Cost-efficient monitoring

Network service recommendation

ABSTRACT

The 5G is the fifth generation of mobile broadband, cellular technologies, and networks that promises a major change in mobility by evolving connected business realities. In such an emerging environment, reliable Service Level Agreements (SLA) and anticipation of breaches of Service Level Objectives (SLO) become compulsory. Thus, guaranteeing the required service quality, while also ensuring efficient recourse allocation becomes a challenge. In addition, 5G networks are expected to provide diverse Quality of Service (QoS) guarantees for a wide range of services, applications and users with a variety of requirements. However, there is an increased difficulty in translating user-friendly business terms into resource-specific monitoring attributes that can be used to manage resources in the 5G core network. To address these gaps, an SLA management framework, enabling QoS provisioning is introduced. The aforementioned framework will be supported by an adaptive monitoring algorithm, which removes the static time interval used in the monitoring system, in order to provide highly accurate information in real time, without the produce of unnecessary traffic to the network. The proposed architecture also incorporates a recommendation mechanism to determine the significance of various QoS parameters in order to ensure that relevant QoS metrics are included in the SLAs, using enriched metadata information from a Network Function Virtualization (NFV) Catalogue.

1. Introduction

The explosive growth in mobile data, forces network operators to transform their networks [1]. The emerging 5G in combination with Software Defined Networks (SDN) aims to comply with different quality of service (QoS) requirements in different application scenarios [2]. Nevertheless, promoting delay-limited QoS over emerging 5G wireless networks with limited resources for bandwidth-intensive and time-sensitive scenarios, presents many new challenges not arised in 4G networks [3,4]. Therefore, it becomes crucial to adopt high performance Virtual Network Functions (VNF) and Network Services (NS) that require a plethora of system resources [5,6]. However, this kind of provisioning in a virtualized environment is a challenging task. Various virtual services in fields like Augmented Reality (AR), Virtual Reality (VR), or autonomous guided vehicles (AGVs), require corresponding Service Level Agreements (SLAs) that capture different levels of guarantees as QoS requirements, such as performance and availability. As stated by many Telco's and cloud service providers, they are tempted by the promise of QoS

such as high speed/performance, high reliability, low latency, increased capacity, availability and connectivity, as well as dynamic bandwidth allocation from 5G RANs to core networks [7]. As a result, the responsibility to provision the necessary infrastructure recourses and QoS assurances, lies with the 5G network operators and network service providers, in order to ensure this kind of demands. Consequently, there are challenges to address this topic which include a pro-active SLA management framework, an efficient and high adaptive monitoring system on top of the virtualized 5G infrastructure, while also support the provisioning of appropriate SLAs, in terms of business aspects and recourses allocation. One of the most significant challenges, is the role of an SLA Management Framework on top of the 5G Core network [8]. It should be pointed, that managing a virtualized network attempts to preserve acceptable quality, and at the same time push the customers to negotiate with the service providers precise QoS thresholds. Considering that the customer-related QoS requirements are indicated in an SLA, the thing that is vital for the service providers, refers to the assessment of the necessary resources, for each NS that is going to be deployed.

*Evgenia Kapassa, University of Piraeus, +30 2104142746, ekapassa@unipi.gr

To this end, monitoring the infrastructure but also the running services, should be considered as the mediator for supporting QoS provisioning. In 5G networks, where QoS assurance through SLA enforcement is a crucial process, data monitoring is required to evaluate the health of the network [9]. Additionally, service provisioning and assurance tools in NFV should monitor a large number of end points [10]. Keeping that in mind, a frequency analyzer tool could be the key, where real time data and low running costs are essential in this kind of networks. Thus, monitoring and evaluation of monitoring data are considered as important aspects to track the overall progress and reliability of a running service. However, decisions on adaptive and pro-active tracking should be made without the possible absence of information that could negatively impact strategic decisions [11]. Such solutions should also be able to adapt the virtual service during runtime, while at the same time maintain all health and performance historical data as the service evolves [13].

At this point, questions such as how the QoS parameters would be described, or where the SLAs would be stored, and how the service providers could manage all this information arise. Furthermore, the efficient description of the QoS parameters comprises an issue of paramount importance, along with the storage and the management of the SLA descriptors from the service providers. Primarily, for fulfilling this demand in 5G infrastructures, the concept of the NFV Catalogues is introduced. Here comes the vision of the NFV Catalogues, going beyond a plain data store to the promotion of the service offerings and the facilitation of the commercial activity and fluent interaction among the different business stakeholders. The above-mentioned concept not only enables the main storage of the 5G infrastructure, but also, promotes the management and the exploitation of this information. As a result, recommendations mechanisms, or even predicting tools can be developed, providing the most suitable SLAs for the selected by the customers network service [12].

Considering the aforementioned set of challenges, in this paper, we are trying to portray an “SLA-Oriented Framework” which targets the profitable provisioning of QoS guarantees, in 5G environments. The proposed approach is based on mapping the high-level customers’ requirements to low-level recourse attributes, as shortly described in [14,15]. The proposed Framework adopts an Artificial Neural Network (ANN) approach, making it easily adaptable to different software components like VNFs and NSs. It is widely known that ANN have many benefits, but the main reason for selecting them is their fault tolerance. This kind of advantage is very important in the SDN environment, where networks are scaled across multiple machines and multiple servers. Additionally, we also introduce an extra set of mechanisms for suggesting the most important QoS parameters, for endorsing the most reasonable and profitable SLAs for diverse services and customers. Furthermore, we also introduce an adaptable “Monitoring Framework, in order to supervise the established SLAs. The latter is based on a scheduling algorithm that provides decision logic on probe level as initially described in [17]. On top of them, an enhanced metadata “NFV-Catalogue” is adopted for valuable management of the NSs and the corresponding SLAs.

The remaining of the paper is organized as follows. Section 2 presents the related work and motivation of this work. Section 3

introduces the overall “SLA-Oriented Framework”, while in Section 4, 5 and 6 we describe in detail the three major architectural components. Section 7 states the evaluation of the proposed approach in a real 5G testbed, through an end-to-end scenario. Finally, in Section 8, we close up with ideas for future experiments and current study capabilities

2. Background and Motivation

2.1. SLAs & QoS assurance

It is doubtless that Active and advanced work on SLA management for the cloud an 5G infrastructure has been carried out. Therefore, there are many researches for solutions that handle QoS parameters and monitor in an efficient way the guaranteed SLOs. A framework is discussed in [16] to utilize QoS into grid applications. In this paper a performance model is used to calculate the response time and the pricing model to determine the cost of conducting a job. It should be noted that a baseline for our work, is presented in [19], where the architecture of an SLA management system is described. Another intriguing work is described in [20], which demonstrates the importance of the implementation of ANNs in the service- oriented field. In this approach, the ANN's main objective is to set technical goals in terms of quality goals for the design attributes of web service systems. Moreover, the LoM2HiS framework is presented in [21]. In this case, the authors provide a paradigm that implements the reverse process of the one we propose, where low- level specifications are translated into high- level requirements that are used in cloud SLAs. More details concerning the mapping of requirements can be found in [22,23]. Moving forward, to the next generation of networks, the 5G Network Slice Broker [24] is an innovative network element that builds on the capacity broker function block, for advanced RAN sharing considered by 3GPP. It maps incoming SLA requirements to physical resources in connection with network slice requests, having as a result to get a “slice” of the relevant elements of the Radio Access Network (RAN) [25].

In addition, a lot of work has been done to provide guaranteed QoS for enhanced user experience. There was a migration from QoS management at the user equipment level, to QoS management at the network level during the evolution of the QoS management mechanism in 3GPP networks, a shift which maintained also in 5G networks [26]. The QoS level provided in 5G systems should meet the requirements of future Internet industry and go beyond what can today be accomplished with any wireless communication technology. Essential QoS requirements in currently studied 5G systems include: a) maximum acceptable end-to-end latency (delay) less than 5ms and b) reliability around 10-9% or 99.999% [27, 28].

It is worth mentioning, that delay- limited QoS requirements are relatively difficult to guarantee, due to highly differing wireless channels. Alternatively, the statistical delay- limited QoS supply theory has been initially proposed and demonstrated to be a useful method for characterizing and implementing the delay- limited QoS guarantee for wireless real- time traffic [29]. In addition, several works have proposed solutions in the framework of the QoS scheduling [30], including multi-QoS scheduling as investigated in [31,32]. Instead, authors in [33,34] contended that existing QoS mechanisms do not endorse the implementation of specific policies for a group of network users.

It should be pointed out, that a preliminary description of our work can be found in [14,15,18], where the authors presented an approach for mapping the high-level end-user requirements the low-level policy parameters, and at the same time proposed a mechanism for suggesting the most important QoS parameters to the Service/Infrastructure Provider, in order to achieve better QoS assurance. In the present paper, we are going beyond an isolated SLA management framework, by trying to support business guarantees in the overall lifecycle of NSs. To do so, we co-operate the proposed “SLA Framework” with an “NFV-Catalogue”, providing recommendations based on benchmarking results, as well as with an advanced “Monitoring Framework” for efficient SLA violations detection.

2.2. Efficient Monitoring

The need for network monitoring is a key enabler for efficient network management. Several works address this domain. As stated in [35], OpenTM was presented, where the integrated features provided in the OpenFlow switches are used to directly and accurately measure the low overhead traffic matrix. OpenTM also utilizes the routing knowledge acquired from the OpenFlow controller to intelligently select the switches from which flow statistics can be obtained, decreasing the load on switching elements. Furthermore, the authors in [36] presented Flow Sense, a push-based approach to performance monitoring in flow-based networks, where they let the network inform regarding performance changes, rather than query for metrics information on demand. The key point is that control messages sent by switches to the controller contain information that allows performance estimation. Moreover, another well-presented monitoring framework for SDNs has been introduced in [37]. The authors propose a software-defined traffic measurement architecture that distinguishes the data plane from the control plane, namely OpenSketch. OpenSketch offers a simple three-stage pipeline (hash, filter and count) in the data plane that can be enforced with commodity switch components and support many measuring tasks. In the control plane, OpenSketch provides a measurement library that automatically configures the pipeline and allocates resources for different measurement tasks. What is more, in [38], authors have presented an extension of Prometheus.io, a monitoring framework implemented and integrated within SONATA project [39]. In short, the SONATA monitoring framework gathers and processes data from many sources, enabling the developer to activate measurements and thresholds to capture generic or service-specific behaviors. In addition, the developer can define rules based on metrics collected from one or more VNFs in one or more NFVs to receive runtime notifications. Furthermore, authors in [40] proposed an approach for comprehensive and detailed monitoring of 5G mobile networks characterized by software using an IoT-based system. The corresponding monitoring framework is designed to collect any type of data either in text or in numerical form in a cloud database. Thus, by using this knowledge, SDN controllers make the decisions on network reconfiguration according to current conditions. A great work was published also in [41]. In this article, a Software Defined Monitoring (SDM) was proposed, while it highlights how SDMs can be used to solve the current limitations in legacy monitoring systems. The proposed approach is able to monitor both virtualized and physical network environments in an economical and efficient way. Initially, the authors’ proposed SDM architecture was used only to monitor 5G backhaul network. Last but not least, an automatic monitoring management for 5G mobile networks was

proposed in [42]. In particular, a 5G-oriented architecture was proposed to integrate SDN and NFV technologies to monitor and control the entire service life cycle taking into account network control plane information. This architecture automatically manages network resources to orchestrate network monitoring services, which are developed in their solution as VNF monitoring. VNF monitoring is described by the collection of information from different and diverse sources, such as network (i.e. physical or virtual) infrastructure data, network management services and user- to- network communication

2.3. NFV-Catalogues & Recommendations

The conception of the NFV Catalogues for the coherent storage of the exchanged entities of the 5G infrastructure originates from several applications into the new era of virtualization. The first attempt of delivering a distributed storage component with functional business and service layers for the VNFs/NS operators was from the [43]. The main view of this approach was to provide a digital marketplace that collects VNFs / NS to operate on commodity cloud infrastructures. On top of that, the continuous and real-time network information of the available VNFs/NS is exchanged between the several layers of the NFV Catalogue based on a set of RESTful APIs for a functional service. Though, the first concrete approach was presented from in the framework of the T-NOVA project [44]. The authors presented the storage of the machine-readable descriptions of the VNFs/NSs and was covered solely from the NFV Catalogues as an integrated component. In parallel, the NFV Catalogues were fully aligned the functional components of the T-NOVA infrastructure, responsible for the charge of registering all business relationships and exposure of the related information for the billing component. Although, this approach presents a distributed storage approach with the several QoS/QoE metrics being disperse in the ecosystem, introducing latency in the diverse functionalities. The next milestone in the evolution of the NFV Catalogues was set one more time by the SONATA project. This approach was predominantly based on the efficient storage of the generalized package formats of the NFV landscape and their functional information. The followed approach was strictly correlated with the specification of the ETSI MANO for the diverse employed NFV Catalogues in the infrastructure [45]. What was beyond the specification comprised the introduction of the engagement of the information on the instantiated VNFs/NSs in accordance to the SONATA Service Platform policy updates.

In parallel with the vast development of the Web, the advent of the rapid growth of the available information was obvious to its users. Recommender Systems (RSs) pioneered the web with the aim of incorporating social information and at the same time delivering meaningful suggestions to the end user. While the research field of RSs has been skyrocketed in diverse domains, there is a gradual, yet slow, interest of the application of the RSs in the 5G ecosystems, through their pinpointing in network management applications [46-48]. Through the introduction of the virtualized era, telecom networks generate massive amounts of monitoring data consisting of observations on network faults, configuration, accounting, performance, and security. E- stream also included a predictive, automated network management recommendation with surprising results due to the ever-increasing complexity of the networks, correlated with particular business level constraints [49]. Through the exploitation of the profound streams of data and the efficient application of techniques in dimension reduction, E-stream was based on recommending

actions with four different aligned factors, namely the context, audience, existing responses, and validation. The main factor of these recommendations was the applicable trading-policies for actuating the recommendation module and propelling the necessary actions to get the best price for each VNF be carried out. Yet, in the 5G telecom systems, there is a paramount absence of RSs in utilizing implicitly the plethora of the QoS/QoE metrics of the multiple diverse components, such as monitoring systems, policy and SLA modules, etc. Thus, RSs techniques and methods comprise a subset of tools that need to be examined thoroughly.

2.4. The challenges of 5G & SLAs

An SLA is a contract between a service provider and its end users that documents what services the service provider can provide and establishes the performance standards that the service provider is required to meet [50]. Since each component potentially impacts the overall behavior of the SDN, any high-level target specified for the service (e.g. performance, availability, security) potentially impacts all low-level components.

SLAs establish customer expectations regarding the service provider's performance and quality. In recent years, SLAs have set expectations for the performance of a service provider, set down penalties for missing targets and, in certain cases, rewards for exceeding those targets [21]. With SDN, network devices (routers and switches) can be managed using OpenFlow [51] and a separate set of application programming interfaces (APIs) can handle virtual network overlays. Once the network is virtualized, the SDN controller can set up network devices as fast as it can deploy new VMs. For instance, customers can take a VM image, deploy it to hardware, spin it, apply it via OpenStack Compute and set up the network around it in a short time via OpenStack Neutron (i.e. Orchestrated SDN form) [52]. In this scenario, SLAs can cover the time and cost of deploying new compute resources, as well as their related network resources [53].

Finally, the SLA lifecycle is an important part of the provision of services, in particular in SDN and 5G networks. The SLAs lifecycle in the 5G domain is managed by the accompanying 5G service platforms and is slightly different from the traditional ones as described in [54]. SLA management is a dynamic process comprising four key stages: a) Architecture, b) Engagement, c) Operations and d) Termination as presented in Figure 1. The overall lifecycle is 5G-enabled, due to the fact that is fully aligned with 5G principles and is running in parallel with Network Service Lifecycle [55].

The first phase starts with the selection of a NS and the requirements definition by the developer. Typically, the Operator is the one responsible to examine those, take into consideration important business needs and implement SLA Templates, as initial offers to the NS customers. During the engagement phase, the selection of different NS results from business aspects, which are the basis for different QoS constraints, which can also be defined as requirements of the agreement.

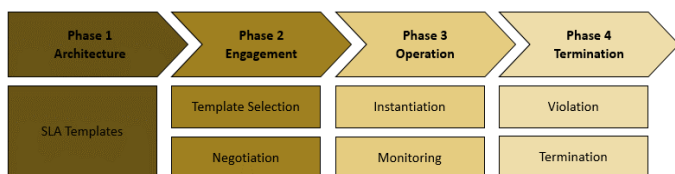


Figure 1 : SLAs Lifecycle in 5G.

The preference of an operator / network service provider depends on the desired network service (NS), its characteristics, budget constraints and so on. QoS expectations also drive end users to negotiate with their operators / service providers precise QoS levels. An SLA is created to describe the agreed QoS parameters after a successful negotiation process. After the successful NS selection and the corresponding SLA negotiation process, the operation phase takes place. This phase comprises the actual deployment of the NS, the population of the respective service with running data, the establishment of communication channels and additional operational activities. Moreover, the operation phase monitors the agreement with real-time data, for the purpose of avoiding or managing unexpected violations. Finally, termination phase deals with the end of the relationship between operator/service-provider and NS customer, including the end of the legal relationship. In general, the latter will continue for a few years after termination in accordance with mandatory laws and legislation. This last phase includes the evaluation of alternatives, settlement and termination commitments, export of data, customer care and diligence, and deletion of data. All the above should be considered either the Network Service was terminated, or the SLA was violated. More details for the aforementioned processes and how they are managed in the proposed architecture will be discussed in Section 3 of this article.

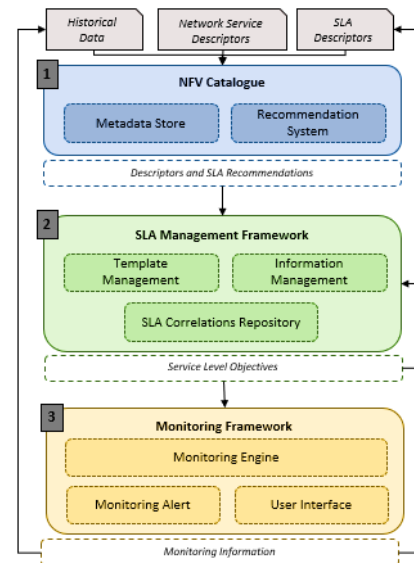


Figure 2: Overall framework architecture.

3. Proposed SLA-Oriented Framework

Considering the challenges presented in Section 2, the developed framework allows to manage the whole lifecycle of SLAs, from the template formulation to the violation detection. Figure 2 depicts the overall architecture of the developed framework. Taking into consideration the inter-connection between a distributed set of users and resources, the implementation of an “SLA-Oriented Framework” aims at governing this interaction. Due to the fact that all running service affects the performance of the service platform, any business parameter (i.e. high-level requirements) included in the SLA, would be linked to recourse demands – encapsulated in the respective policies (i.e. low-level requirements). Initially, a set of NSs, SLAs and historical

monitoring data among with their metadata, are stored into the “NFV-Catalogue”. This preliminary information is used from the “Recommendation System”, to provide feedback toward the “SLA Management Framework” for an optimized SLA Template formulation. As soon as this information is available, the SLA Framework should be considered in two phases: a) the SLA Template Management and b) the SLA Information Management. During the first phase, an optimized SLA Template is prepared, by mapping the high-level expressed by the customer into low-level resource attributes needed by the service provider. On the other hand, during the second phase, where the NS instantiation takes place, our proposed framework oversees the obtain monitoring data from the “Monitoring Framework”, which lies on the bottom of the proposed architecture. The implemented “Monitoring Framework” provides support for the QoS management, while it provides adaptable monitoring feedback based on the infrastructure needs. The following sub-sections explain in detail each component.

4. NFV Catalogue

4.1. Metadata Store

As an initial stage, the “NFV-Catalogue”, which is depicted in Figure 2, is positioned to address storage and management necessities of diverse stakeholders’ (i.e. NS developers, NS providers, customers etc.). The main view of the “NFV-Catalogue”, is to provide a repository for persistent storing of the developed VNFs/NS and their corresponding SLAs, attaching them with additional metadata, which are exploited to leverage its functionalities and interfaces for storing, searching and retrieving. Moreover, additional information, like metric significance outcomes and information related to policies and QoS need to be also stored as metadata of the corresponding VNFs. Thus, the “NFV-Catalogue” is deemed to be a multi-faceted data storage, addressing various stakeholder needs while also forming the main and centralized data storage of the 5G ecosystem. The functionality of the “NFV-Catalogue” is predominantly based on the metadata for NS Descriptors and SLA Descriptors. Prior to the attachment of the metadata, through a RESTful API of the “NFV-Catalogue”, the inspection of the validity of the document structure is a critical and necessary step. Since the documents are specified in machine-readable formats, the review of the format contributes to the eliminations of flaws in the “NFV-Catalogue”. Moreover, the attachment of metadata provides the ability of defining uniquely the individual stored machine-readable objects inside the data storage as depicted in Figure 3. Moreover, “NFV-Catalogue”

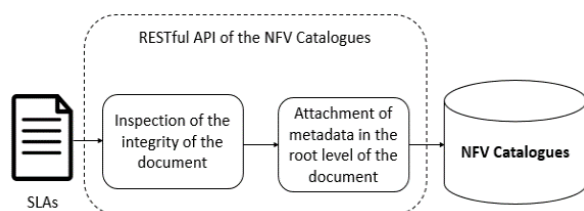


Figure 3: SLAs and metadata storage.

is aligned with the principle of the persistent storage by extending this type of information with valuable fields for successful data integration, accuracy in the format of the document, confirmed time of creation, etc. In this way, it enables

the development of enhanced operations for Creating, Retrieving, Updating and Deleting (CRUD) SLA Descriptors inside the “NFV-Catalogue”, while reassures the correct data format of the stored documents (e.g. SLA Templates).

Going beyond the conventional data storage, the presented “NFV-Catalogue” provide intelligent functionalities in a 5G environment. Since the types of information vary, one of the necessities that is satisfied by the Catalogues is the full-text search capabilities in structure-agnostic documents. Since the schema of the diverse documents (i.e. NS descriptors, SLA descriptors) is variable, the “NFV-Catalogue” provides searching capabilities without the necessity of indexes. Thus, it provides seamless retrieval abilities in deep-hierarchical machine-readable document structures. Furthermore, besides from the plain NoSQL document store for the diverse descriptors, “NFV-Catalogue” provides a scalable file system for hosting the artifact files, required for the instantiation lifecycle of the VNFs/NSs.

4.2. Recommendation System

After the successful storage of the NS Descriptors among with their SLAs, we consider a “Recommendation System”, which will optimize the SLA Templates creation, by providing important knowledge to the Operator/NS-Provider. Thus, through the plethora of the stored SLAs and the historic data of each end-user, it is undisputedly valuable to tap into them and provide optimum combinations of the available QoS parameters through recommendations. The basic principle of these recommendations is that paramount dependencies presented between the user-to-item activity. The aim of the “Recommendation System” is not to provide only specific and optimum combinations of QoS but also to allow users to profit from them. Thus, the recommendations comprise samples from the relevant actions that were followed in several signed SLAs from similar end-users. In the proposed framework, user-based Collaborative Filtering (CF) was used in order to detect similar users and promote recommendations in these terms [56]. The user is deemed that explicitly rated a combination of SLAs, along with the included QoS parameters. Moreover, what is of paramount importance is the metric of computing the correlations between the end users in the “Recommendation System”. More specifically, the Pearson correlation metric was the most appropriate for the best trade-off in the equation of quality-number of predictions [57]. Although, the aforementioned filtering suffers from two severe issues, the Data Sparsity and the Cold Start [58]. The former, comprises the phenomenon that end users provide a small amount of ratings, contributing to memory complexity and inevitability in training the “Recommendation System”. The latter refers to the difficulty in bootstrapping the “Recommendation System” for newly-introduced users and items. Despite this fact, the CF, in the proposed “SLA-Oriented Framework”, eliminating these challenges by introducing “trust relationships” of end-users. The main idea is the provision of the trust metric for each individual user-to-user and, on top of that, contributes to the predictions of the conventional CF, as depicted in Figure 4. The trust metric relevance of the users is denoted from the received implicit rating of each VNF/NS. With the instantiation/upload of a VNF/NS, the “NFV-Catalogue” automatically receives implicit rating of the respective entity. Thus, the selections can provide the relevance of the diverse users of the platform.

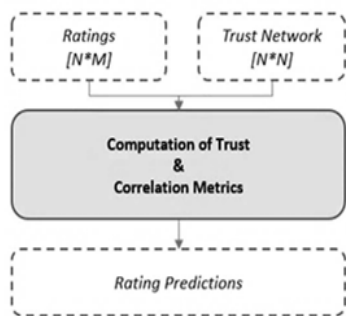


Figure 4: Proposition of trust metrics

5. SLA Management Framework

In the second stage of the proposed “SLA-Oriented Framework”, the SLA Templates are generated, as an initial offer to the end-users. After the successful NS instantiation, along with the corresponding SLA, the signed Agreement starts to be monitored so it can fulfill the signed SLOs to the end-user. The current stage is splitted into two sub-phases, a) SLA Template Management, which takes place prior the NS deployment, and b) the Information Management, which takes place during the NS deployment. The internal architecture and workflow inside the “SLA Management Framework” it is depicted in Figure 5.

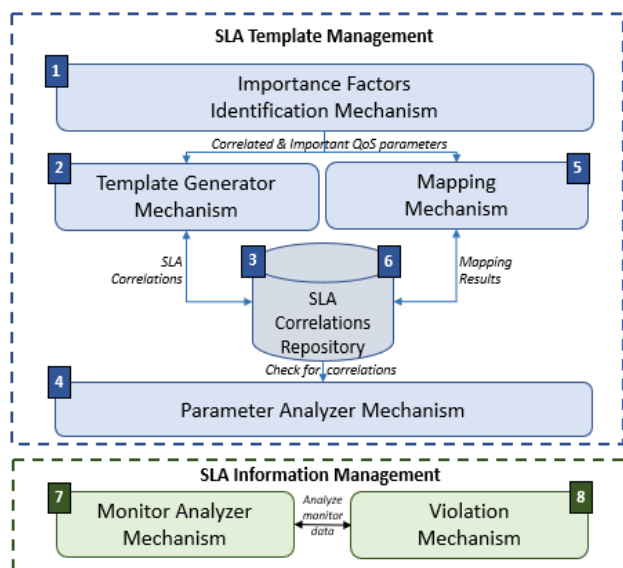


Figure 5: SLA management framework internal architecture

5.1. SLA Correlation Repository

Given the big amount of data generated and analyzed into the SLA Management Framework, an internal repository has been introduced, namely “SLA Correlations Repository”, in order to store and manage all the necessary correlations, between end-users, network services, templates, agreements, violations, as well as recourse mapping results. In particular, the correlations between the high-level and the low-level requirements are stored for future analysis. In addition, it keeps track of all the correlations between the generated templates and the linked network services. At the same time agreements information are also located in the repository, along with the end-user’s authentication details, as well as the violations records.

5.2. SLA Template Management

The SLA Template Management is the first phase of the SLA Management component while it is in charge of receiving the desired business guarantees (i.e. high-level QoS parameters) from various parties (e.g. NS provider, customer), and formulate an initially SLA Template. It is also responsible for mapping the high-level business guarantees to low level resource attributes, so they are able to be included in the Template, and afterwards monitored through the instantiated Agreement. The SLA Template Management consists of four mechanisms which are going to be further described in the following sub-sections.

Importance Factors Identification Mechanism

In the case of NSs, a challenge arises given that many different entities are setting their requirements for the overall service. Those entities may have specific preferences for resource attributes and potentially additional parameters that can be monitored (e.g. number of sessions) and thus be included into an SLA. To address the aforementioned challenge, we would need to develop a mechanism which would analyze monitoring data and performance information in order to identify dependencies between a VNF’s metrics but also realize how these dependencies affect the overall performance of a specific NS. The latter would be reflected to the so called “importance” factors, while it could give feedback to the Operator, about what it is of crucial importance and need to be included in the SLA Template. This is in fact an on-line learning process that updates and dynamically evolves. To this end, the corresponding mechanism, estimates and defines the importance of various QoS parameters, b) predefined policies, c) historical monitoring data and d) recommended QoS metrics. The latter it is depicted in Figure 6.

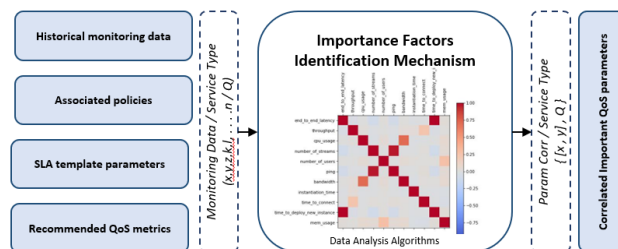


Figure 6: Importance factors identification mechanism

As one understands, the color of the correlogram is white when there is no correlation between the two variables (the correlation value is equal to 0 or close to 0). When the color is deep black it means that our mechanism has calculated a sufficiently large positive correlation, while the deepest grey color indicates there is a large negative correlation.

It should be mentioned that policies are considered out of the proposed framework’s scope, and that’s why it is assumed that are able to be provided by an external Policy Management Framework. In order to correlate the QoS parameters and define the important ones, analysis libraries were used to analyze the performance measurements of the chained VNFs in a NS. Starting an offline learning process with the gathered data, the mechanism calculates and stores the dependencies in an internal NoSQL database. Then, providing REST APIs as well as a GUI, the end-

user can request and get those dependencies of his/her developed NS, and as a result getting an inside knowledge of the NS's performance behavior. Thus, the "Importance Factors Identification Mechanism" can produce essential weight factors and classify parameter dependencies, to suggest and include relevant QoS parameters in the SLA templates. [59-61].

Template Generator Mechanism

As soon as the important QoS parameters are recognized, the "Template Generator Mechanism" takes action, which initially produces the SLA Templates requested by the service provider, and then it is responsible to establish the final Agreement. The "Template Generator Mechanism" as shown in Figure 7, can acquire a set of policies for a clearly defined NS and also historical data of the service provider through the "Monitoring Framework" (i.e. NS performance data, preferences of resource parameters) [62].

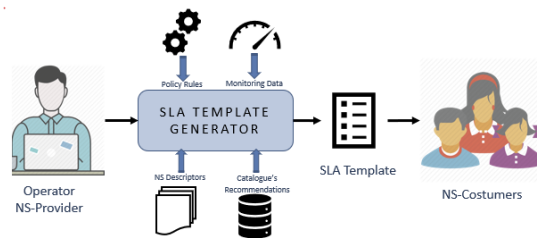


Figure 7: SLA template generator mechanism

In addition, the "Template Generator Mechanism" accesses the "NFV-Catalogue" to retrieve also the NS Descriptor for the corresponding NS, as well as QoS parameters recommendations from the Catalogue's RS. Finally, an also important input to the "Template Generator Mechanism", is the weight factors of the aforementioned recommended QoS parameters, obtained by the "Importance Factors Identification Mechanism". After gathering all the above-mentioned input, the "Template Generator Mechanism", triggers the analysis while it correlates the input in a way that it can formulate an SLA Template with some initial guarantees.

Mapping Mechanism

As soon as the initial Template is formulated, there is essential need to decompose the business guarantees to recourse attributes that can be monitored accordingly by the "Monitoring Framework". In other words, The SLA "Mapping Mechanism" (MM) is the component responsible to translate the high-level requirements described by the end-user into low-level metrics required by the service provider, and vice-versa. More specifically, the MM obtains a set of policy rules from an external Policy Management Framework, a set of low-level and high-level requirements described from the service provider but also from the customer. As a result, the produced output of the MM (i.e. output layer of the ANN) are explicit SLA business metrics, as depicted in Figure 8.

The MM is based on unsupervised learning, using an Artificial Neural Network (ANN) [63]. ANNs can be used to solve this translation problem by mapping service-specific SLOs to resource attributes directly. As they embody a black box approach, ANNs are ideal to be used in an environment where information is not

easily transmitted from one entity to another. In addition, ANNs need no knowledge of the inner structure of the NSs [64]. However, it should be noted that they need a representative execution dataset, in order to detect complex, linear or non-linear dependencies.

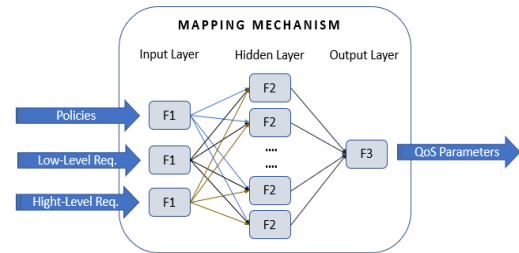


Figure 8: ANN based mapping mechanism

Parameter Analyzer Mechanism

The "SLA Parameter Analyzer" it is capable of deciding whether or not the MM should occur. The decision is made after the component searches into the "SLA Correlations Repository" and check whether the input parameters already correlate with the mapping results calculated and stored.

5.3. SLA Information Management

Once the selected NS has been successfully instantiated, the SLA Information Management should take action by continuous monitoring the service instance. This process aims to optimize the SLA formulation, manage the guaranteed terms in respect with the infrastructure conditions, while also check for any SLA violations. SLA Information Management consists of two mechanisms which are going to be presented explicitly in the next sub-sections. The first is called "SLA Monitor Analyzer Mechanism" while the second "SLA Violation Mechanism".

Monitor Analyzer Mechanism

Starting with the "SLA Monitor Analyzer Mechanism" the component acquires both historical monitoring data for the deployed NS and biases for resource parameters. Sequentially, it should decide whether there is any existed contrast between the mapping results and the runtime monitoring data. Specifically, the "SLA Monitor Analyzer Mechanism", examines the QoS parameters from the "SLA Correlations Repository", with the gathered monitoring data, while it afterwards measures the delta between their values. The MM will obtain the monitoring feedback as an additional dataset and re-train the ANN, in case the delta is greater than "0".

Violation Mechanism

As a final step, in the second stage of the proposed "SLA-Oriented Framework", we consider the "SLA Violation Mechanism". The corresponding mechanism is responsible to ensure that the newly deployed NS would not violate the corresponding SLA upon instantiation, while it successfully fulfills the signed business needs. Nevertheless, the mechanism is also responsible to identify any violations occurred and take the necessary actions (e.g. scale in, scale out). Specifically, upon receipt of the measured metrics through the Monitoring Framework, the "SLA Violation Mechanism" starts an ongoing

process of re-adoption. In particular, the mechanism considers the mapping results stored in the “SLA Correlations Repository” and compare them with the real-time monitoring information. In anticipation of future SLA violation threats, the mechanism readjusts the low-level recourse parameters described for the SLA and push them back to the “Monitoring Framework”. In this way, both the SLA is fulfilled, while at the same time the infrastructure does not waste recourses, but only in times of real need. Although, in case a violation is not prevented in time, an alert is sent from the “Monitoring Alert Mechanism”. Upon receipt, the “SLA Violation Mechanism” calculates the overall value of the specific metric and takes decision whether the SLA is violated or not. In case of an SLA violation, the customer is informed by an e-mail, SMS, or even a live push notification.

6. Monitoring Framework

In the third, and final, stage of the proposed “SLA-Oriented Framework”, lies the “Monitoring Framework”, where, its internal architecture is depicted in Figure 9. The proposed monitoring framework has adopted the SONATA Monitoring Framework [65], and then adapted accordingly in order to support the whole NS lifecycle in respect of the associated SLAs. The corresponding monitoring framework is consisted of: a) the “Monitoring Engine”, which collects monitoring data provided by the NSs based on the signed SLA, b) the “Alert Manager”, which is responsible to produce alert messages when a violation of a SLA rule is occurred, and c) a “User Interface”, used for visualization of the collected monitoring data, while also visualization of each individual rule specified by the “SLA Management Framework”.

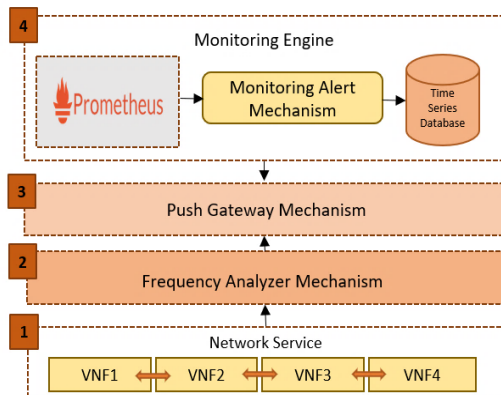


Figure 9: Monitoring Framework: Internal Architecture

6.1. Monitoring Engine

To begin with, a key mechanism of the third stage is the “Monitoring Engine”, on the bottom of the proposed “SLA-Oriented Framework”. We should point out that, the automation of monitoring an SLA, is a difficult task that demands precise specifications and an adaptable mechanism that collects the right measures and models. At the same time, evaluation of an SLA should occur in specific time frames or when some remarkable events happen. In 5G/SDN ecosystem, where chained VNFs in form of NSs are implemented and deployed on top of a service platform, it becomes essential to create a “Monitoring Engine” which is able to manage a variety of specifications and monitor accordingly the recourses of the virtualized infrastructure. While an SLA is already attached and instantiated through the NS, it is

assumed that the desired guarantee terms have been granted to the customer. On the instantiation phase of the service, monitoring rules for the specific NS instance are automatically generated and pushed to the Monitoring Framework, through the “SLA Violation Mechanism”. The “Monitoring Engine”, which is based on a stable version of Prometheus Monitoring System [66] will undertake the collection of monitoring data from the running services. Prometheus scraps metrics for short- lived jobs either directly or through an intermediate push gateway. It stores all scraped samples locally and executes rules on these data to either add new time series from available data or generate alerts. The benefit of using Prometheus as “Monitoring Engine” is the fact that it was designed for reliability, and the ability to allow quick problems diagnosis.

Push Gateway Mechanism

As it is previously mentioned, Prometheus, can scrap monitoring data exported from the running service instances through the push gateway. To be more specific, the Push Gateway, is a subcomponent of Prometheus, acting like an intermediary service, allowing to forward the monitored data from the “Monitoring Engine” towards the “Monitoring Alert Mechanism”, and thus publish them to the external components (i.e. SLA Management Framework, NFV-Catalogue). In our case, the usefulness of the “Push Gateway Mechanism” arrives during the NS scaling up [67]. Scaling up, means that a new VNF is about to start, relieving the service instance when it is actual needed. In this case, the new VNF identifies the “Push Gateway” in terms of authentication and push the monitoring data towards a recognized and reliable host. As a result of the approach, is the fact that the “Monitoring Engine” does not need to know and identify the NS instances, but vice-versa.

Frequency Analyzer Mechanism

After the previous discussion around pushing monitoring data through the “Push Gateway”, a challenge arises, in terms of how often and which data are promoted to the engine. Having this in mind, an important parameter of define the above-mentioned challenge is the time interval used to evaluate the resource metrics and guaranteed SLOs (e.g. every two seconds or every two minutes). Although, too frequent pushes may affect negatively the overall system performance, whereas too infrequent pushes may cause heavy SLA violations, due to lack of monitoring metrics towards the “SLA Management Framework” [68]. To this end, an enhancement to Prometheus Monitoring Framework is introduced, namely the “Frequency Analyzer Mechanism”, which is based to an adaptive monitoring algorithm. Thus, the “Frequency Analyzer Mechanism” acts as a middle agent between an active connection of the NS instance, the “Push Gateway” and the “Monitoring Alert Mechanism” [17]. Its purpose is to provide highly accurate information about the network’s health, while at the same time avoid the production of unnecessary traffic in the network. It aims at adapting during runtime the monitoring time intervals in order to ensure that the data collected and transmitted to the SLA Management Framework, are fruitful and not all raw data. Moreover, it should be noted that the algorithm achieves significant reduction in resource consumption and also reduces the number of SLA violations, due to the pro-active nature of the mechanism.

Time Series Database

Finally, a key component of the Monitoring Framework, is also an internal database. A “Time Series Database” is used for storing and identifying the monitored information, by a metric name and a set of key-value pairs. Following the approach of a time series database, the advantage of having operators for calculating useful information of the monitoring data, is given to the “Monitoring Engine”.

6.2. Monitoring Alert Mechanism

The previously discussed monitoring outcomes are going to be published to the external components (i.e. SLA Management Framework, NFV-Catalogue) through the “Monitoring Alert Mechanism”. During our research we realized that a message queue system (MQ) was the most appropriate solution for the intercommunication of the aforementioned components. Therefore, RabbitMQ was integrated to the Monitoring Framework, as the message broker for asynchronous messaging [69]. Through the implemented “Monitoring Alert Mechanism”, SLA monitoring rules and SLA violations are produced as alerts. Thus, the message (i.e. alert) is pushed to the “Monitoring Alert Mechanism”, and all the RabbitMQ consumers are receiving the message for further actions. It should be noted that, one of the consumers is the “SLA Violation Mechanism”. This mechanism, acts as an intermediate component between the “Monitoring Engine” and the end-user of the NS instance.

6.3. User Interface

For the visualization of the gathered monitoring data, Grafana is used as an open platform for visualize and beautify monitoring data analytics [70]. Grafana, features an advanced chart query editor that lets the user to quickly browse the metric space, add features, change operating parameters, and more.

7. Evaluation

In order to evaluate the performance of the proposed framework in terms of efficiency and ease of use, our approach was included in the innovative 5G infrastructure environment of the 5GTANGO Service Platform [71]. 5GTANGO project is an EU funded Innovation Action, that enables the flexible programmability of 5G networks with a modular Service Platform so it can bridge the gap between business needs and network operational management systems [72]. The 5GTANGO Service Platform offers the service and functional orchestration features, along with all the supplementary and supporting tools required, like the proposed “NFV-Catalogue”, “SLA Management Framework” as well as the “Monitoring Framework”.

7.1. Emulation Environment

During the evaluation of the proposed framework a challenge arised, as many of the mechanisms (i.e. Recommendation System, Importance Weight Factors Mechanism, Mapping Mechanism) need apriory behavior knowledge, in order to be able to deal with unknown VNFs/NSs and train their models properly. Moreover, this becomes even more necessary in the emerging DevOps environments, where new versions of NSs are directly deployed in production (i.e. working environment), and therefore no up-to-date monitoring data is available for the updated services. To deal with this challenge, we adopted the OSM supported VIM emulator [77],

www.astesj.com

in order to run on top of it the 5GTANGO VNF/NS benchmarking framework, to automatically execute performance benchmarks of NFV network services and functions [78, 79]. The benchmarking tool automatically gathers performance information about a service, prior to its deployment without requiring dedicated testbeds, resulting to an offline profiling of the service, and the collection of benchmarking data, so they can used as a starting point of the service modeling.

7.2. Working Environment

As it is previously mentioned, the proposed “SLA-Oriented Framework” is implemented inside the 5GTANGOs’ SP where the installation guide can be found in [80]. For evaluation purposes, we used the NCSR Demokritos’ testbed in order to setup the SP and thus our proposed framework along it. NCSR Demokritos’ testbed is the main node of the 5GTANGO infrastructure in Athens, providing the following infrastructure components: a) WAN network, b) Access network, c) datacenter (computing resources for NFVI realization), and d) end user devices and services. To be more specific, as Network Function Virtualization Infrastructure (NFVI) Queens OpenStack multi node deployment with provider networks configuration is used [82]. Also, service chaining is taken care by a Service Function Chaining (SFC) agent that interfaces with the Service Platform in order to fix the chaining between the Network Service components. For the Wide Area Network (WAN) part, the networking is managed by a WAN Infrastructure Manager (WIM) implemented by a Virtual Tenant Network (VTN) running on top of OpenDayLight (version Oxygene) [83]. Finally, the current study used a processing environment which consists of the following elements:

- One Dell R210 used as the Fuel jump host
 - 1xIntel(R) Xeon(R) CPU X3430 @ 2.40GHz
 - 4GB RAM
 - 1TB HDD
- One Dell T5500 functioning as the Controller
 - 2xIntel(R) Xeon(R) CPU X5550@2.67GHz
 - 16GB RAM
 - 1.8 TB HDD
- Three Dell R610 utilized as Compute Nodes
 - 2xIntel(R) Xeon(R) CPU E5620@2.40GHz
 - 64GB RAM
 - 1.8 TB HDD
- One Dell R310 used as NFVI-PoP
 - 1xIntel(R) Xeon(R) CPU X3450@2.67GHz
 - 16GB RAM
 - 465GB HDD

More details about the testbed’s topology, hardware/software availability and network recourses can be found in [73]. It should be also noted that with regards to the source code availability, it is

currently partially in open-source format, since not all functionalities have finalized yet.

7.3. Experimental Results

In this case study, different stakeholders take place to the overall workflow, in order to provide the necessary inputs. In this study, we consider a) Thomas – the Service Developer, b) Sally – the Commercial Offer Designer (COD), responsible for defining SLAs supporting the whole business, c) Bob – the network engineer of the Service Provider in charge of defining run-time policies, d) Robert – the engineer, in charge of providing monitoring data and e) End-user Customers, such as Brian. The deployment of a network service instance in a service provider’s infrastructure comes with the definition of some requirements. This will ensure the performance estimation and the QoS requested by the customer. Quality of service is introduced in the SLA with the definition of SLOs, along with policies for managing the infrastructure accordingly, and thus enforces the respective SLA.

In order to perform a complete testing and evaluation of our approach, an elastic proxy network service is used as depicted in Figure 10, which is consisted of two chained VNFs: a) a HAProxy VNF, configured as a load balancer and b) a Squid VNF configured as a proxy server.

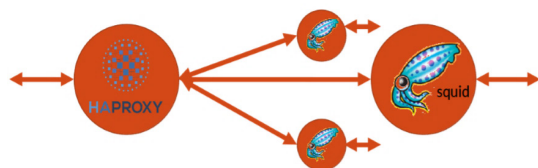


Figure 10: Elastic proxy network service

The end-users use the ingress interface of the HAProxy as proxy IP, and sequentially the HAProxy forwards the incoming requests to one of the Squids in its backend pool. To begin with, the NS-Descriptor, as developed by Thomas, the service developer, is onboarded to the proposed “NFV-Catalogue”. The onboarding is made through a Rest API (i.e. Create operation), as mentioned in Section 3.1. Then, the Sally, the Commercial Offer Designer, is responsible to define the customers’ facing characteristics of the service, in particular the SLAs. The SLA Generator requires four (4) parameters for a successful generation of the SLA template. The most critical one is the selection of the NS (i.e. the haproxy-squid NS) that is going to be correlated with the newly created SLA template. At the same time an SLA name among with a valid future expiration date and at least one SLO are needed. Once this information is gathered, the SLA Generator mechanism triggers the “Importance Weight Factors Mechanism” and the “Recommendation System”. At this moment, the first one will feed the “Template Generator Mechanism” with feedback on “relevant” QoS parameters that need to be included in the template, and at the same time, Catalogue’s “Recommendation System” will also provide recommendations based on previous created templates and relevant costumers’ preferences. It is important to point out that, computing this kind of recommendations and similarities on behalf of Sally, enhance the generation process, by giving the chance to the Commercial Offer Designer to include in the template the most appropriate QoS parameters, and minimize the negotiation between Brian (i.e. the

customer) and the service provider. Tables 1 presents high level parameters included in previous customers’ SLAs, along with their calculated and monitored values, while Table 2 depicts the similarity between Sally and previous costumers of the haproxy-squid service. It should be mentioned that the values presented in Table 1, are real business requirements, as gathered from Communication Pilot use case of the 5GTANGO project [76].

Table 1: Previous customers low-level metrics

	Customer A	Customer B	Customer C
Availability	0,99.99 %	0.95 %	0.9 %
Jitter	10 ms	25 ms	50 ms
Packet Loss	0.1 %	0.05 %	0.5 %

Table 2: Sally’s similarity with previous customers

	Sally	Customer A	Customer B	Customer C
Sally	1	82 %	50 %	30 %

Based on the above results, the “Recommendation System” of the “NFV-Catalogue” and the “Importance Weight Factors Mechanism”, Sally is recommended to include the following guarantees to the SLA Template, as presented in Table 3.

Table 3: QoS recommendations for Sally

	Availability	Jitter	Packet Loss
Sally	99 %	15 ms	0.1

Afterwards, the generation of the SLA template, as an initial offer to the NS-costomers is triggered. The outcome is an SLA Descriptor that will be onboarded to the NFV-Catalogue, and it is based on WS-Agreement specification [74-75]. The main SLA Templates building blocks of the reference model include the root element, the SLA template and the service elements, as depicted in Figure 11.

Afterwards, Brian, the end-user customer, browses through the available NSs, select the haproxy-squid NS and instantiate it, in order to be deployed in the SP. During the instantiation process the customer triggers automatically the one-shot negotiation process through the “Mapping Mechanism”, by selecting the previously generated SLA Template for the specified NS. Based on the business requirements of the customer, the SLA can be accepted, or a new negotiation process can be initiated. Once the SLA Manager has collected all the relevant datasets via the “Parameter Analyzer”, checks if there is already a combination in the “SLA Correlations Repository”, between the latter and the already existing mapping results, in order to decide whether the process of the “Mapping Mechanism” should be triggered or not. In case there is not a correlation yet, the operator’s low-level requirements, the costumer’s high-level business needs and the policies, are mapped in order to produce the actual QoS parameters that can finally be included in the SLA. The objective is to forecast the performance and the quality that is required, to be agreed and signed in the final SLA. Alternatively, if there is an already a

combination between the input dataset (i.e. requirements obtained from the operator and the customer) and the stored mapping results, the SLA Manager bypasses the mapping process and dynamically creates the final Agreement. In order to investigate this mapping, we needed to predict the performance of the Network Service on top of the infrastructure.

```

{
  "name": "silver-template-example",
  "vendor": "UPRC",
  "version": "2.0",
  "author": "Evgenia Kapassa, Marios Touloupou",
  "description": "This is a Gold SLA Template for Haproxy-Squid Service",
  "sla_template": {
    "template_name": "Gold",
    "offer_date": "2019-02-04T11:35:10Z",
    "expiration_date": "2020-02-04T11:35:10Z",
    "provider_name": "Telefonica",
    "template_initiator": "Evgenia Kapassa",
    "service": {
      "ns_uid": "0e69ccfd-d9ba-4439-99b8-cd4f2a059457",
      "ns_name": "ns-squid-haproxy",
      "ns_vendor": "eu.5gtango",
      "ns_version": "0.2",
      "guaranteeTerms": [
        {
          "guaranteeID": "g1",
          "guarantee_name": "Availability",
          "guarantee_threshold": "99%",
          "guarantee_operator": "greater",
          "guarantee_unit": "%",
          "guarantee_period": "Daily",
          "guarantee_definition": "",
          "guarantee_service_level": "50sec/24h",
          "target_slo": [
            {
              "target_kpi": "Downtime",
              "target_value": "50s",
              "target_operator": "less",
              "target_duration": "10s",
              "target_period": "24h",
              "target_service_level": "Downtime less 50s"
            },
            {
              "target_kpi": "Jitter",
              "target_value": "15 ms",
              "target_operator": "less",
              "target_duration": "10s",
              "target_period": "",
              "target_service_level": "Jitter less than 15 ms"
            },
            {
              "target_kpi": "Packet Loss",
              "target_value": "0.1%",
              "target_operator": "%",
              "target_duration": "",
              "target_period": "",
              "target_service_level": "0.1% Packet Loss of the total packets sent"
            }
          ]
        }
      ]
    }
  }
}

```

Figure 11: SLA descriptor example

The output is categorized between simple mapping results and complex ones. A simple mapping result maps “end-to-end”, from low-level to high-level. For instance, mapping the low-level metric “downtime” to high level SLA parameter “availability”. Complex mapping results include predefined formulations to calculate specific SLA parameters using low level resource metrics. Table 4 presents an example of a complex mapping result.

Table 4: Complex mapping result example

Low-Level Metric	SLA Parameter	Mapping Formulation
downtime, uptime	Availability (A)	$A = 1 - \frac{downtime}{uptime}$

In order to investigate the performance of the “Mapping Mechanism”, the emulation environment provided a data set consisted of 360 data points. Of these, 50% was used in order to train the network model, for the haproxy-squid NS, as depicted in Figure 12, resulting into 180 data points. For validation purposes another 20% was used during training, meaning 72 data points. Finally, the overall network capability measured against the

remaining 30% of the data set, (on which the model was not trained), simulating a real test situation.

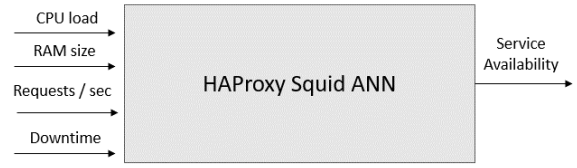


Figure 12: ANN model for HAProxy-Squid Network Service

Then, the remaining 30% was used to check the system’s reliability in accurately predicting the NS’s QoS levels in the Deployment Environment. In order to check the absolute differences between the ANN’s prediction and the actual monitoring observation, the Mean Absolute Error (MAE) was used, as shown in (1), where ‘n’ is the number of data points, y_j represents the observed values and \hat{y}_j the predicted values. The MAE result is depicted in Table 5.

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \quad (1)$$

Table 5: Complex mapping result example

ANN Model	Neuron per Layer	Mean Absolute Error
HAProxy-Squid NS Availability	4 – 3 - 1	2.75 %

The aforementioned ANNs are feed-forward back-propagation networks, trained with the Levenberg-Marquardt algorithm [81]. The criterion for performance was the Mean Square Error (MSE) in the training set, while it was trained for 100 periods, for a training time of 1 minute. Putting all this together, we have the general formula for calculating the MSE in (2).

$$MSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2 \quad (2)$$

It should also be stated that, if the customer wishes to change the recourse parameters that the HAProxy-Squid service can handle, the 5GTANGO platform, which incorporates the service presented in this paper, offers the capability of re-triggering the “Mapping Mechanism” and retrain the model.

Next, the instantiation of the NS takes place. To this end, the deployment aims to enable execution of the NS according to the QoS requirements, while at the same time appropriate monitoring, allows the measurement of QoS parameters at both service and infrastructure levels targeting events of resource provisioning estimation and decision making. For this reason, the Monitoring Framework access the Point of Presence (PoP) that the NS is deployed into and gathers monitoring information for the haproxy-backend-downtime, jitter and packet loss, in order to measure its availability. At this point, the “Frequency Analyzer Mechanism” takes place, by adjusting the monitoring time intervals during runtime to ensure that the data collected and transmitted to the SLA management framework are meaningful. At first, data are collected and compared with linear increase of time intervals, until they reach an initial time threshold. In our case the collected monitored values

of the HAProxy-Squid service downtime were below the certain threshold, indicating that the network has changed towards a better state. Therefore, after the service had pushed the monitored data, the “Frequency Analyzer Mechanism” multiply the time the change occurred, with β , in order to increase the data transmission interval. As a consequence, the new timeout is higher, and the probes will collect data with a linear increase over a longer time period, without wasting recourses.

As depicted in Figure 13, the X axis shows a linear increase in data collection time from the samples. The latter begins in the first second and increases linearly until it reaches the fifth second. while a change in the metric value occurs in the 5th sec. The “Frequency Analyzer Mechanism” commands the probes to send their data to the “Push Gateway”, and at the same time increases the timeout by the current time* β (i.e. current time is the time when the significant change occurred).

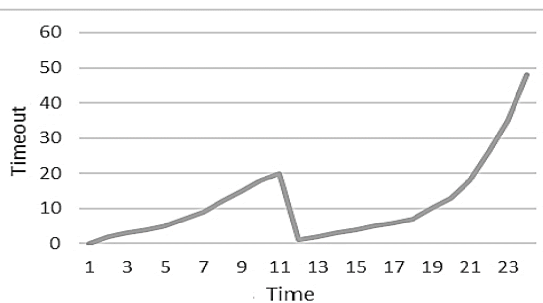


Figure 13: Time intervals adjustments

In our scenario, the equation returned the value of 18 seconds, meaning that the data collection will continue after a linear increase in time until the new timeout which set to 18th second is reached. Moreover, the monitoring process of the proposed approach was tested firstly using the standard Prometheus framework, and then by enhancing it with the aforementioned “Frequency Analyzer Mechanism”. Figure 13 depicts the difference regarding the network workload (i.e. throughput in terms of data requests towards the push gateway per second). Finally, when the network service has completed its lifecycle, Brian, is responsible for terminating it. The termination process, as well as all the aforementioned procedures, are taking place in a user-friendly way, through a unified Portal.

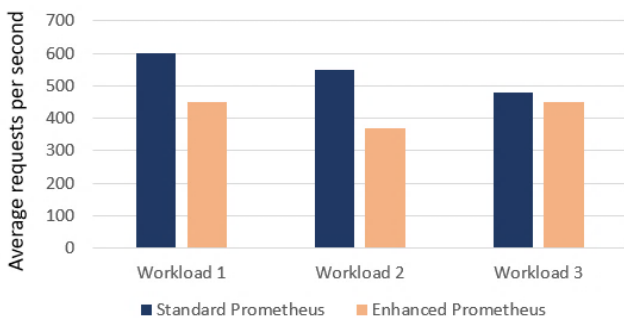


Figure 13: Network workload with and without using the “Frequency Analyzer Mechanism”

8. Conclusions

Based on the aforementioned evaluation, we presented a SLA Management Framework that is used to map high- level business

parameters to low- level attributes of resources. This framework is integrated in the 5GTANGO Service Platform for the autonomous management of SLAs. We considered a generic approach that is based on ANNs in order to efficiently be used as a mediator for the network provider and the end-user. We considered an ANN based approach that can be used as a mediator between the end-user and the provider. Furthermore, we have introduced a mechanism to determine the importance of various QoS parameters so that the “relevant” ones be included in the SLAs for better QoS assurance. The proposed monitoring framework goes a step forward from the traditional implementation, by preventing any unnecessary traffic to the network, and also by providing real-time and high accurate information for better QoS assurance. Last but not least, the authors have presented a beyond a plain data storage to an enriched information – driven repository, the so-called “NFV-Catalogue”. The aggregation of the stored information allows the mechanism to apply recommender system techniques, build on QoS predictions and SLA recommendation systems. Regarding the three major components of the proposed architecture (i.e. the NFV-Catalogue, SLA Management Framework and Monitoring Framework), we conclude to the following, based on the captured experimental results.

All things considered, provisioning of resources in a virtualized 5G infrastructure is a challenging task, that still needs a lot of investigation. Therefore, we plan to extend the framework in order to enable Quality of Experience (QoE) enforcement. This kind of enforcement could be done by adopting the infrastructure recourses accordingly during runtime, considering parameters based on Catalogue’s recommendations as well as monitoring feedback. Moreover, we tend to enhance the SLA violations management, by providing violations prediction models, in order to prevent day zero violations. Additionally, the currently proposed framework is able to monitor and manage business guarantees in a single` 5G environment. Therefore, we envision to manage a 5G network with multiple domains, enabling higher level of integration, and at the same time adapt the proposed architecture from a wide range of verticals, enabling higher level of abstraction.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has been partially supported by the 5GTANGO project, funded by the European Commission under Grant number H2020ICT-2016-2 761493 through the Horizon 2020 and 5G-PPP programs (<http://5gtango.eu>).

References

- [1] F. Hu, Opportunities in 5G Networks: A Research and Development Perspective, CRC Press, 2016.
- [2] A. Osseiran et al., "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," in IEEE Communications Magazine, 52(5), 26-35, 2014. doi: 10.1109/MCOM.2014.6815890
- [3] H. Su, X. Zhang, "Cross-Layer Based Opportunistic MAC Protocols for QoS Provisionings Over Cognitive Radio Wireless Networks," in IEEE Journal on Selected Areas in Communications, 26(1), 118-129, 2008. doi: 10.1109/JSAC.2008.080111
- [4] D. Wu, R. Negi, "Effective capacity: a wireless link model for support of quality of service," in IEEE Transactions on Wireless Communications, 2003. doi: 10.1109/TWC.2003.814353

- [5] C. Liang, F. R. Yu and X. Zhang, "Information-centric network function virtualization over 5g mobile wireless networks," in *IEEE Network*, 29 (3), 68-74, 2015. doi: 10.1109/MNET.2015.7113228
- [6] Q. Zhu, X. Zhang, "Game-theory based power and spectrum virtualization for maximizing spectrum efficiency over mobile cloud-computing wireless networks," in 49th Annual Conference on Information Sciences and Systems (CISS), 2015. doi: 10.1109/CISS.2015.7086818
- [7] Samsung Developers, "5G Requirements", Online: <https://developer.samsung.com/tech-insights/5G/5g-requirements>
- [8] Huawei Technologies Co. Ltd, "5G Network Architecture: A High-Level Perspective", Online: <https://www.huawei.com/minisite/hwmbbf16/insights/5G-Network-Architecture-Whitepaper-en.pdf>
- [9] E. Casalicchio, V. Cardellini, G. Interino, M. Palmirani, "Research challenges in legal-rule and QoS-aware cloud service brokerage", in *Future Generation Computer Systems*, 78(1), 211-223, 2018. doi: <https://doi.org/10.1016/j.future.2016.11.025>.
- [10] Anuta Networks, "Top 6 Challenges for Service Assurance in NFV", Online: <https://www.anutanetworks.com/top-6-challenges-for-service-assurance-in-nfv/>
- [11] A. Tabebordbar, A. Beheshti, "Adaptive Rule Monitoring System" in 1st International Workshop on Software Engineering for Cognitive Services, 2018. doi: 978-1-4503-5740-1
- [12] L.B. López, J. M. Vidal, L.G. Villalba, "An Approach to Data Analysis in 5G Networks", in *Entropy*, 19(2), 74, 2017. doi: 10.3390/e19020074
- [13] Centina, "Service Assurance Critical to SDN/NFV Success", Online: <http://www.centinasystems.com/service-assurance-critical-sdn-nfv-success/>
- [14] E. Kapassa, M. Touloupou, A. Mavrogiorgou, D. Kyriazis, "5G & SLAs: Automated proposition and management of agreements towards QoS enforcement" in 21st Conference on Innovation in Clouds, Internet and Networks and Workshops, 2018. doi: 10.1109/ICIN.2018.8401587
- [15] E. Kapassa, M. Touloupou and D. Kyriazis, "SLAs in 5G: A Complete Framework Facilitating VNF- and NS- Tailored SLAs Management," in 32nd International Conference on Advanced Information Networking and Applications Workshops, 2018. doi: 10.1109/WAINA.2018.00130
- [16] S. Benkner and G. Engelbrecht, "A Generic QoS Infrastructure for Grid Web Services," in Advanced Int'l Conference on Telecommunications and Int'l Conference on Internet and Web Applications and Services, 2006. doi: 10.1109/AICT-ICIW.2006.16
- [17] M. Touloupou, E. Kapassa, A. Kiourtis and D. Kyriazis, "Cheapo: An algorithm for runtime adaption of time intervals applied in 5G networks," in Fifth International Conference on Software Defined Systems. doi: 10.1109/SDS.2018.8370420
- [18] N. Sfondrini, G. Motta, L. You, "Service level agreement (SLA) in Public Cloud environments: A Survey on the current enterprises adoption," in 5th International Conference on Information Science and Technology, 2015. doi: 10.1109/ICIST.2015.7288964
- [19] Xi Zhang, Jia Tang, Hsiao-Hwa Chen, Song Ci and M. Guizani, "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks," in *IEEE Communications Magazine*, 44(1), 100-106, 2006. doi: 10.1109/MCOM.2006.1580939
- [20] L. Zhu, X. Liu, "Technical Target Setting in QFD for Web Service Systems Using an Artificial Neural Network," in *IEEE Transactions on Services Computing*, 3(4), 338-352, 2010. doi: 10.1109/TSC.2010.45
- [21] V. C. Emeakaroha, I. Brandic, M. Maurer, S. Dustdar, "Low level Metrics to High level SLAs - LoM2HiS framework: Bridging the gap between monitored metrics and SLA parameters in cloud environments," in International Conference on High Performance Computing & Simulation, 2010. doi: 10.1109/HPCS.2010.5547150
- [22] T. Cucinotta et al., "Virtualised e-Learning with real-time guarantees on the IRMOS platform," 2010 IEEE International Conference on Service-Oriented Computing and Applications, 2010. doi: 10.1109/SOCA.2010.5707166
- [23] G. Kousiouris, D. Kyriazis, S. Gogouvitis, A. Menychtas, K. Konstanteli, T. Varvarigou, "Translation of application-level terms to resource-level attributes across the Cloud stack layers," in IEEE Symposium on Computers and Communications, 2011. doi: 10.1109/ISCC.2011.5984009
- [24] K. Samdanis, X. Costa-Perez, V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," in *IEEE Communications Magazine*, 54(7), 32-39, 2016. doi: 10.1109/MCOM.2016.7514161
- [25] X. Foukas, G. Patounas, A. Elmokashfi, M. K. Marina, "Network Slicing in 5G: Survey and Challenges," in *IEEE Communications Magazine*, 55(5), 94-100, 2017. doi: 10.1109/MCOM.2017.1600951
- [26] V. Tikhvinskiy, G. Bochechka, "Prospects and QoS Requirements in 5G Networks", in *Journal of Telecommunications and Information Technology*, 1, 23-26, 2015. doi:
- [27] ETSI, "ETSI TR 102 889-2 V1.1.1, Technical Report", Online: https://www.etsi.org/deliver/etsi_tr/102800_102899/10288902/01.01.01_60/tr_10288902v010101p.pdf
- [28] G.C. Madueño, C. Stefanović, P. Popovski, "Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling" in *IEEE Wireless Communications Letters*, 3(4), 429-432, 2014. doi: 10.1109/LWC.2014.2326674
- [29] J. Tang X. Zhang, "Quality-of-Service Driven Power and Rate Adaptation over Wireless Links," in *IEEE Transactions on Wireless Communications*, 6(8), 3058-3068, 2007. doi: 10.1109/TWC.2007.051075
- [30] A. Asadi, V. Mancuso, "A Survey on Opportunistic Scheduling in Wireless Communications," in *IEEE Communications Surveys & Tutorials*, 15(4), 1671-1688, 2013. doi: 10.1109/SURV.2013.011413.00082
- [31] T. Guo, R. Arnott, "Active LTE RAN Sharing with Partial Resource Reservation," in IEEE 78th Vehicular Technology Conference), 2013. doi: 10.1109/VTCFall.2013.6692075
- [32] K. Hammad, A. Moubayed, S. L. Primak and A. Shami, "QoS-Aware Energy and Jitter-Efficient Downlink Predictive Scheduler for Heterogeneous Traffic LTE Networks," in *IEEE Transactions on Mobile Computing*, 17(6), 1411-1428, 2018. doi: 10.1109/TMC.2017.2771353
- [33] J. Pérez-Romero, O. Sallent, R. Ferrús, R. Agustí, "On the configuration of radio resource management in a sliced RAN," in *IEEE/IFIP Network Operations and Management Symposium*, 2018. doi: 10.1109/NOMS.2018.8406280
- [34] I. da Silva et al., "Impact of network slicing on 5G Radio Access Networks," in *European Conference on Networks and Communications*, 2016. doi: 10.1109/EuCNC.2016.7561023
- [35] A. Tootoonchian, M. Ghobadi, Y. Ganjali, OpenTM: Traffic Matrix Estimator for OpenFlow Networks, Springer, 2010
- [36] C. Yu, Lumezanu, Y. Zhang, V. Singh, G. Jiang, H.V. Madhyastha, FlowSense: Monitoring Network Utilization with Zero Measurement Cost, Springer, 2013.
- [37] M. Yu, J. Yu, L. Rui, M. Rui, Software Defined Traffic Measurement with OpenSketch, USENIX Association, 2013.
- [38] P. Trakadas et al., "Scalable monitoring for multiple virtualized infrastructures for 5G services" in The International Symposium on Advances in Software Defined Networking and Network Functions Virtualization, 2018. doi: <http://hdl.handle.net/1854/LU-8569066>
- [39] SONATA Project Consortium, "SONATA NFV: Agile Service Development and Orchestration in 5G Virtualized Networks", Online: <http://www.sonatanfv.eu/>
- [40] T. Maksymyuk, S. Dumych, M. Brych, D. Satria, M. Jo, "An IoT Based Monitoring Framework for Software Defined 5G Mobile Networks" in *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, 2017. doi: 10.1145/3022227.3022331
- [41] M. Liyanage et al., "Software Defined Monitoring (SDM) for 5G mobile backhaul networks," in *IEEE International Symposium on Local and Metropolitan Area Networks*, 2017. doi: 10.1109/LANMAN.2017.7972144
- [42] A. H. Celdrán, M.G. Pérez, F. J. García Clemente, G. M. Pérez, "Automatic monitoring management for 5G mobile networks" in *Procedia Computer Science*, 110, 328-335, 2017. <https://doi.org/10.1016/j.procs.2017.06.102>.
- [43] N. Nikaen. et al., "Network store: Exploring slicing in future 5g networks" in *Proceedings of the 10th International Workshop on Mobility in the Evolving Internet Architecture*, 2015. doi: 10.1145/2795381.2795390
- [44] T-NOVA Project Consortium, "T-NOVA: Network Functions As-a-Service Over Virtualized Infrastructures", Online: <http://www.t-nova.eu/>
- [45] M. Ersue, "ETSI NFV Management and Orchestration - An Overview", Online: <https://www.ietf.org/proceedings/88/slides/slides-88-opsawg-6.pdf>
- [46] R.E. Núñez-Valdéz, et al., "Implicit feedback techniques on recommender systems applied to electronic books" in *Computers in Human Behavior*, 28(4), 1186-1193, 2012. <https://doi.org/10.1016/j.chb.2012.02.001>.
- [47] K. Oku, R. Kotera, K. Sumiya, Geographical Recommender System Based on Interaction Between Map Operation and Category Selection, *ACM*, 2010.
- [48] J. Serrano-Guerrero, E. Herrera-Viedma, J.A. Olivás, A. Cerezo, F.P. Romero, "A google wave-based fuzzy recommender system to disseminate information in University Digital Libraries 2.0" in *Information Sciences*, 181(9), 1503-1516, 2011. doi: <https://doi.org/10.1016/j.ins.2011.01.012>.
- [49] F. Zaman, G. Hogan, S. V. Der Meer, J. Keeney, S. Robitzsch, G. Muntean, "A recommender system architecture for predictive telecom network management," in *IEEE Communications Magazine*, 53(1), 286-293, 2015. doi: 10.1109/MCOM.2015.7010547
- [50] Y. Koh, R. Knauerhase, P. Brett, M. Bowman, Z. Wen, C. Pu, "An Analysis of Performance Interference Effects in Virtual Environments," in *IEEE International Symposium on Performance Analysis of Systems & Software*, 2007. doi: 10.1109/ISPASS.2007.363750
- [51] ONF, "Software-Defined Networking (SDN) Definition", Online: <https://www.opennetworking.org/sdn-definition/>
- [52] OpenStack Project, "OpenStack Networking ("Neutron")", Online: <https://wiki.openstack.org/wiki/Neutron>

- [53] B. Hoff, "IT service-level agreements and SDN: Assuring virtualization performance", Online: <https://searchnetworking.techtarg.com/tip/IT-service-level-agreements-and-SDN-Assuring-virtualization-performance>
- [54] SLA-Ready, "Cloud SLA lifecycle", Online: <http://www.sla-ready.eu/cloud-sla-lifecycle>
- [55] S. Van Rossem et al., "A network service development kit supporting the end-to-end lifecycle of NFV-based telecom services," in IEEE Conference on Network Function Virtualization and Software Defined Networks, 2017. doi: 10.1109/NFV-SDN.2017.8169859
- [56] C. Saluja, "Collaborative Filtering based Recommendation Systems exemplified", Online: <https://towardsdatascience.com/collaborative-filtering-based-recommendation-systems-exemplified-ecbffe1c20b1>
- [57] J. Jin, S. Zhang, L. Li, T. Zou, "A Novel System Decomposition Method Based on Pearson Correlation and Graph Theory" in IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS), 2018. doi: 10.1109/DDCLS.2018.8515967
- [58] X. He, L. Liao, H. Zhang, L. Nie, T. Chua, "Neural Collaborative Filtering", in Proceedings of the 26th International Conference on World Wide Web, 2017. doi: 10.1145/3038912.3052569
- [59] Z. u. Rehman, F. K. Hussain, O. K. Hussain, "Towards Multi-criteria Cloud Service Selection," in Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2011. doi: 10.1109/IMIS.2011.99
- [60] Z. ur Rehman, O. K. Hussain, S. Parvin, F. K. Hussain, "A Framework for User Feedback Based Cloud Service Monitoring," in Sixth International Conference on Complex, Intelligent, and Software Intensive Systems, 2012. doi: 10.1109/CISIS.2012.157
- [61] T. Halabi, M. Bellaiche, "Evaluation and selection of Cloud security services based on Multi-Criteria Analysis MCA," in International Conference on Computing, Networking and Communications, 2017. doi: 10.1109/ICCNC.2017.7876216
- [62] B. K. Tripathy, A. G. Sathy, P. Bera, M. A. Rahman, "A Novel Secure and Efficient Policy Management Framework for Software Defined Network," in IEEE 40th Annual Computer Software and Applications Conference, 2016. doi: 10.1109/COMPSAC.2016.31
- [63] S. Hussain, R. Atallah, A. Kamsin, J. Hazarika, "Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study" in Silhavy R. (eds) Cybernetics and Algorithms in Intelligent Systems, Springer, 2019. doi: https://doi.org/10.1007/978-3-319-91192-2_21
- [64] G. Kousiouris, D. Kyriazis, S. Gogouvitis, A. Menychtas, K. Konstanteli, T. Varvarigou, "Translation of application-level terms to resource-level attributes across the Cloud stack layers," in IEEE Symposium on Computers and Communications, 2011. doi: 10.1109/ISCC.2011.5984009
- [65] SONATA Project Consortium, "D4.3 Service Platform First Operational Release and Documentation", Online: <http://www.sonata-nfv.eu/sites/default/files/sonata/public/content-files/deliverables/SONATA%20D4.3%20Service%20platform%20operation%20release%20and%20documentation.pdf>
- [66] Prometheus Authors, "Prometheus: From metrics to insight", Online: <https://prometheus.io/>
- [67] S. Dutta, T. Taleb, A. Ksentini, "QoE-aware elasticity support in cloud-native 5G systems," in IEEE International Conference on Communications (ICC), 2016. doi: 10.1109/ICC.2016.7511377
- [68] C. Vincent et al., "Towards autonomic detection of SLA violations in Cloud infrastructures", in Future Generation Computer Systems, 28(7), 1017 – 1029, 2012. doi: <https://doi.org/10.1016/j.future.2011.08.018>
- [69] Pivotal Software, "RabbitMQ", Online: <https://www.rabbitmq.com/>
- [70] Grafana Labs, "Grafana: The open platform for beautiful analytics and monitoring", Online: <https://grafana.com/>
- [71] C. Parada et al., "5Gtango: A Beyond-Mano Service Platform," in European Conference on Networks and Communications, 2018. doi: 10.1109/EuCNC.2018.8443232
- [72] 5GTANGO Project Consortium, "5GTANGO: 5G Development and Validation Platform for global Industry-specific Network Services and Apps", Online: <https://5gtango.eu/>
- [73] 5GTANGO Consortium, "D6.1 Infrastructures, Continuous integration approach", 2017, Online: <https://5gtango.eu/project-outcomes/deliverables/38-d6-1.html>
- [74] A. Andrieux et al., "Web Services Agreement Specification (WS Agreement)", Grid Resource Allocation Agreement Protocol (GRAAP) WG, 2011, Online: <https://www.ogf.org/documents/GFD.107.pdf>
- [75] R. Kabert, G. Katsaros, T. Wang, "A RESTful implementation of the WS-agreement specification", in Proceedings of the Second International Workshop on RESTful Design, 2011. doi: 10.1145/1967428.1967444
- [76] SONATA Project Consortium, "Real Time Communications", Online: <https://5gtango.eu/index.php/about-5g-tango/47-real-time-communications>
- [77] M. Peuster, H. Karl, S. v. Rossem, "MeDICINE: Rapid Prototyping of Production-Ready Network Services in Multi-PoP Environments" in IEEE Conference on Network Function Virtualization and Software Defined Networks, 2016. doi: 10.1109/NFV-SDN.2016.7919490
- [78] M. Peuster, H. Karl, "Profile Your Chains, Not Functions: Automated Network Service Profiling in DevOps Environments" in IEEE Conference on Network Function Virtualization and Software Defined Networks, 2017. Doi: 10.1109/NFV-SDN.2017.8169826
- [79] M. Peuster, H. Karl, "Understand Your Chains: Towards Performance Profile-based Network Service Management" in Fifth IEEE European Workshop on Software Defined Networks, 2016. doi: 10.1109/EWSDN.2016.9
- [80] 5GTANGO Project Consortium, "5GTANGO Service Platform Installation Guide" Online: https://sonata-nfv.github.io/component_installation
- [81] G. Kousiouris et al., "Distributed Interactive Real-time Multimedia Applications: A Sampling and Analysis Framework" in 1st International Workshop on Analysis Tools and Methodologies for Embedded and Real-time Systems, 2010. doi: <https://eprints.soton.ac.uk/id/eprint/272323>
- [82] OpenStack Project, "OpenStack Queens Expands Support for GPUs, Containers to Meet Edge, HA, AI Workload Demands", Online: <https://www.openstack.org/software/queens/>
- [83] OpenDaylight Project, "OpenDaylight", Online: <https://www.opendaylight.org>