ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1225-1232 (2020)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Interactive Virtual Rehabilitation for Aphasic Arabic-Speaking Patients

Sherif H. ElGohary[*], Aya Lithy, Shefaa Khamis, Aya Ali, Aya Alaa el-din, Hager Abd El-Azim

Biomedical Engineering and Systems Department, Faculty of Engineering, Cairo University, Giza, 12613, Egypt

ARTICLE INFO

Article history:
Received: 04 August, 2020
Accepted: 19 September, 2020
Online: 24 October, 2020

Keywords:
Aphasia
Arabic
Convolutional neural networks
Rehabilitation
Signal processing
Speech and language disorders

ABSTRACT

Objective: Individuals with aphasia often experience significant problems in their daily lives and social participation. Technologies that address speech and language disorders deficit in merging between therapist's major role and reinforcing the training between sessions at home. It also lacks the Arabic language attention; however, current systems are typically expensive and lack amusement. Moreover, cumulative feedback for both patient and therapist incapacitates the whole home rehabilitation process. This project sought to address these issues by developing an interactive rehabilitation-based system for people with aphasia. Methods: A virtual reality (VR) environment is created providing real-life situations with task specific training of comprehension in addition to a virtual speech-language pathologist (SLP) representing lip motions for correct pronunciation of target words. A speech recognition convolutional neural networks (CNN) algorithm based on signal processing is created and trained on ten isolated Arabic words. We tempted log-spectrograms and Mel-frequency cepstral coefficients (MFCC) as feature extractors for the CNN model which is then integrated for accurate evaluation of input speech from the aphasic patient providing a real-time feedback resulting in measuring speech improvement and sends it to the SLP through the network via a website platform. Results: Our speech recognition assessment algorithm results in a recognition accuracy of 95.2 % using Log-Spectrograms feature extraction method and 92.6 % using MFCC. Significance & Conclusion: We hypothesize that this interactive VR therapy combined with speech function training will result in faster word retrieval and improve language ability of patients with aphasia and that our outcomes contribute to the development of a home-based language and speech therapy.

1. Introduction

Aphasia is the most common language disorder caused by traumatic brain injury or most commonly a stroke. Stroke is the 3rd leading cause of long-term disabilities in the world [1]. According to the world stroke organization, there are over 13 million emerging strokes each year [2] and to the stroke association in the United Kingdom, around a third of stroke survivors experience some level of aphasia [3]. Aphasia is a disorder that results from damage to portions of the brain that are responsible for language production and comprehension which impairs the expression and understanding of language and speech as well as reading and writing [4].

The most common two types of aphasia are Broca's and Wernicke's. Broca's aphasia is associated with the damage to the frontal lobe of the brain in which patients can understand speech and know what they want to say but they cannot express it and is also referred to as non-fluent aphasia because of the halting and effortful speech quality [5]. On the contrary, people with Wernicke's aphasia suffer from damage in the temporal lobe of the brain and have difficulties in language comprehension while producing speech itself is not much affected, therefore, it is referred to as fluent aphasia [5]. Patients mostly suffer from significant problems in their daily lives and social participation which leads to depression, isolation, embarrassment and preventing the expression of everyday needs [6], and therefore, greatly reduce quality of life.

Usually, traditional rehabilitation therapy aims at restoring or improving the impaired functions. It helps patients retrieve their

[*]Corresponding Author: Sherif H. ElGohary, Biomedical Engineering and Systems Department, Faculty of Engineering, Cairo University, Giza, 12613, Egypt, (+20)1020031012 & Email: sh.elgohary@eng1.cu.edu.eg

https://dx.doi.org/10.25046/aj0505148

ability to speak and express as the speech-language pathologists (SLPs) focus on motor production of speech sounds. It depends on the SLP showing the patient some cards or photos of an object while pronouncing its name with a very accurate and relatively slow lip motion so that the patient could mimic it and train several times. But, most often, patients feel embarrassed from their pathologists in addition to that the treatment progress is relatively slow [7].

A range of mobile phone applications have widely spread out for people with stroke, aphasia, brain injury, or dementia. Some of them focus on language and conversation training in a very primitive way and most of the rest are used as communication tools to help patients communicate with other people in form of icons being pressed to say what they need in the time; however, there is still a huge lack in technological solutions that focus on the rehabilitation process itself especially that dealing with Arabic language.

Arabic language is the 6th most used language based on the number of native speakers. Nearly 250 million people use Arabic as their first language and it is the second language for around four times that number [8]. There are three types of the Arabic language: Classical Arabic which is the most formal type, Modern Standard Arabic with some simplifications on the classical and is used in writings and formal speeches, and Colloquial (dialectal) Arabic. Each country or region within a country has its own dialect. Colloquial Arabic is considered the most important for a language rehabilitation process as it is the language used in daily conversations and informal writings.

For the available rehabilitation digital solutions, we found a focus on virtual therapists to validate the tele-rehabilitation delivery by using pre-recorded voices or scripts like in Web-ORLA developed by Cherney et al. [9] and Aphasia Scripts developed by Van et al. [10] which are more reading treatment for people with aphasia so, they are too complicated for people with severe comprehension or reading impairment. Another solution called EVA project developed at City University of London [11] is a multi-user online virtual world headed towards speech training through navigating different environments and conversation; however, it is represented in a fantasy and non-immersive world.

Moreover, virtual reality (VR) has been explored and proved to be an effective rehabilitation method in many physical disabilities' rehabilitations [12], [13] and in other communication disorders such as autism [14], [15] and stuttering [16]. Although this potential has not yet been fully directed to aphasia disorder, there has been some work that investigated the effectiveness of VR in aphasia rehabilitation [17], [18] and proved it to be effective as patients were more attracted and concentrated.

In this paper, we propose a new VR-based rehabilitation approach for Arabic-speaking aphasic patients, which provides them the flexibility and convenience of getting therapy at their homes, with an access to real-life simulation, speech, comprehension, and categorization training in addition to a virtual SLP representing lip motions for correct articulation of words. The boost of an Artificial Intelligence (AI) speech recognition algorithm based on isolated words independent of speakers is integrated for accurate evaluation of input speech from the aphasic patient providing a real-time feedback resulting in measuring speech and comprehension improvement. And taking the Internet of Things (IOT) advantages into account for automatic data transfer over the network to send feedback to the therapist via a website platform for remote following up.

2. System Description

The flow block diagram of the proposed system is depicted in Figure 1. We have created a game-like virtual rehabilitation environment in which we simulate the traditional therapy approach. Therefore, the environment is divided into two significant parts, which can be customized separately depending on the patient's case. The first one is the virtual SLP clinic representing the vocalization learning part based on the very accurate lip movement with several repetitions. Secondly, is the training part which focuses on comprehension problems in which we provide real-life situations with interactive and categorization tasks.

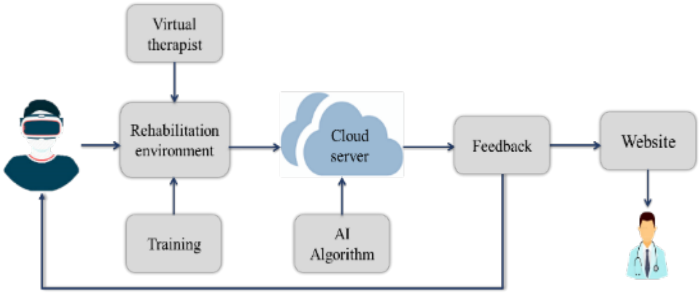


Figure 1: System flow block diagram.

All the manipulations within the game that are in the form of input voice from the patient are continuously recorded and inputted to an algorithm for measuring and calculating speech production and comprehension improvement. A feedback report is sent to the real therapist via a website platform for following up and also an on-screen feedback will be appearing in the game for encouraging the patient. A cloud server connects the whole cycle of sending and receiving data.

3. Methodology

3.1. Rehabilitation Environment

We designed the virtual environment simulating the real rehabilitation pattern but in the form of a game-like scenario to be more interactive and fun. Unity3D (version 2019.3.13f1) game engine was used as the main development platform with C# scripting language to control the various assets. The game activities were divided into two main parts; the virtual therapist or SLP clinic and the training part.

3.1.1. Virtual Therapist

Based on our interviews with different SLPs and their instructions, lip motions are considered the most important part in speech rehabilitation generally and in aphasic cases specially, including synchronization of articulators' movement [19] and also according to a pilot study about the effects of silent visuomotor cueing on verbal execution [20]. Thus, we started by designing a realistic room in unity (shown in Figure 2), representing the clinic's room.

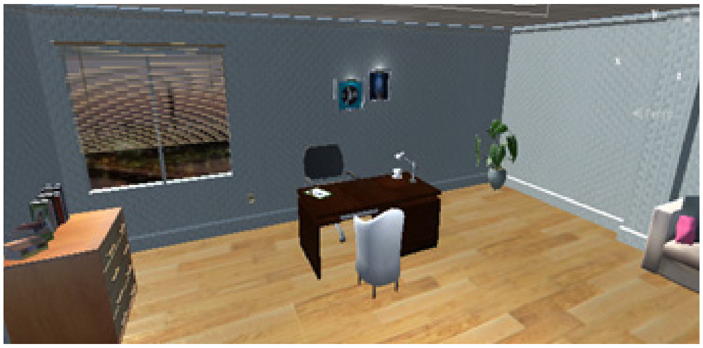


Figure 2: Virtual therapy room.

We created a virtual SLP represented as a humanoid avatar character created using Daz3D studio using Genesis2 characters for rendering the very accurate lip motions of correct pronunciation of words. The character is then exported from Daz3D studio as an FBX object format with its animation and imported into the clinic's scene in Unity as shown in Figure 3. This part is responsible for encouraging the patient to learn words and practice pronunciation regularly.

Patients interact using a screen UI button to make the avatar start saying the targeted word and after listening and observing target speech patterns for reference, another button is used for patient voice input via the device's microphone. Then, a useful screen feedback appears to correct speech deficits and encourage improvement in speech legibility.

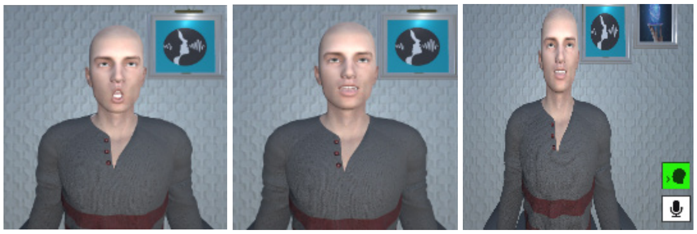


Figure 3: Avatar character's lip motions after importing in Unity and the UI screen buttons.

3.1.2. *Training Part*

For the training part, we designed a city-like environment containing a variety of locations of daily living situations, for example: street, living room and kitchen; sample scenes are shown in Figure 4. Designing real-life scenes was considered so that the patient feels as if he/she is in the real world with free movement and scene navigation. This part focuses mainly on comprehension measurement and how to deal with things in real life with its denotations through carrying out some assigned tasks either categorization and matching or voice input ones.

a) Navigation Exercises

In all scenes in this part, the player can navigate freely in the area and if he/she selects an object that belongs to the specific category of training in the time, the name of this object is pronounced so he can train by listening and learning vocabulary of objects while wandering in the place.

b) Comprehension Tasks

In order to work on patient's ability to comprehend objects' meanings in an interactive manner, one of the categorization-based tasks that is assigned to him while moving is that a food bar is placed in the screen corner (see Figure 5), decreased over time and at a moment, an icon appears that guides the patient to click by mouse on any food around and so, the bar gets refilled and he can keep proceeding.

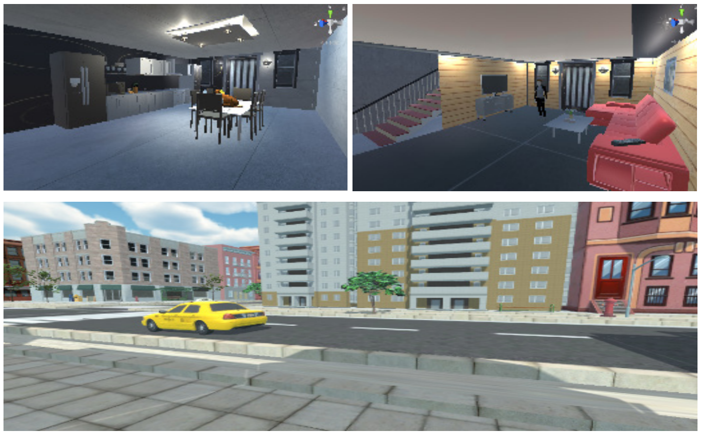


Figure 4: Sample scenes from the training environment.

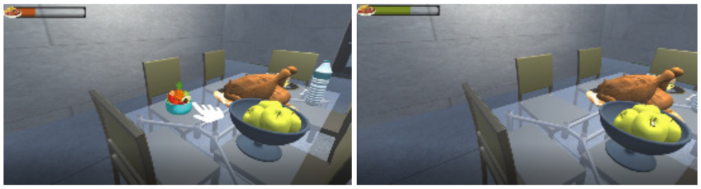


Figure 5: Interactive task based on words categories.

3.2. *Algorithm Generation*

For the AI algorithm used in voice data processing, we created a convolutional neural network (CNN) model using TensorFlow module with python programming language to train the data of targeted words in Arabic language and then testing the patient's speech of those words to assess the patient's improvement. The following parts illustrate the steps followed in order to build the algorithm.

3.2.1. *Data Acquisition*

The undergoing dataset consists of collected audio data of frequent speech of some targeted words in colloquial Egyptian Arabic language. The words were selected based on different major categories used in speech therapy.

We collected the dataset consisting of 12,804 utterances of 10 short words; each word repeated 5 times by the same person; by participants of both genders; about 199 females and 58 males with ages greater than 18. The mother tongue of all participants is Egyptian Arabic.

3.2.2. *Pre-processing*

Data preprocessing is a data mining technique that is used to transform the raw data into a useful and efficient format. For the purpose of data preparation for the system, we applied several preprocessing steps as shown in Figure 6.

a) Noise Removal

The Noise reduction effect dramatically reduces background and broadband noise with a minimal reduction in the signal quality.

We used fre:ac – a free audio converter desktop application for background noise reduction which implicitly uses recurrent neural networks models for noise reduction (RNNoise).

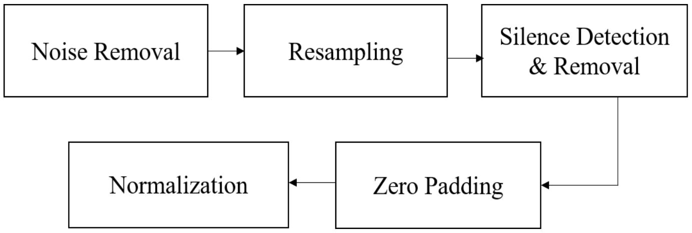


Figure 6: The pipeline of preprocessing steps.

b) Resampling

Resampling is usually done to interface two systems which have different sampling rates. The sampling rates of our collected data were different. So, we resampled all audio signals to 16000 Hz.

c) Silence Detection and Removal

Silence detection and removal help to maintain an acceptable end-to-end delay for the audio signal. We applied silence detection algorithm by segmenting the audio file into chunks of constant size equals to 0.01 s and comparing each chunk's amplitude in decibels with the silence threshold which equals to – 45 dB

d) Zero Padding

Because of the different lengths of the collected audio files, we unified the length of all signals to be one second for each command in the dataset. We applied zero padding at the end of each record that was less than one second.

e) Normalization

In order to overcome the problem that the speech is loud in some portions and quiet in others. Having this variance in volume can hinder transcription, so we had to normalize the audios.

The method that was followed, is used for standard amplitude normalization by scaling the whole audio to the maximum amplitude.

3.2.3. Data Augmentation

We performed this step to generate more data from the available dataset and increase the diversity of it to make our model invariant to perturbations and enhance its ability to generalize. We augmented the dataset randomly by several techniques such as adding a random factor of white noise between 0.004 to 0.009, shifting the starting point of the audio, and then padded it to its original length, increasing signal's amplitude (loudness) by a gain of 5-10 dB and applying random cropping by a mask of random silence of the time that is chosen to be saved and the remaining parts of the audio will reset to zero. After applying data augmentation, the dataset size became 77,040 utterances.

3.2.4. Features Extraction

The most common two approaches in speech recognition are to convert raw audio signals to spectrograms or to extract features using Mel-frequency Cepstral Coefficients (MFCC) [21-23]. As proven that convolutional neural networks work well with image recognition tasks [24], in this paper we followed the two approaches and compared results to choose the best approach that fits our system.

a) Spectrogram Generation

First, we converted the prepared uniform raw data to its log-spectrograms; which are two-dimensional matrices, as a representation of the audio signal in frequency domain and used it as an input to the CNN. The audio signal is used in time domain and broken up into chunks, each chunk is a small frame size of the speech signal data. It is between 0.02-0.04 s and is assumed to be a stationary signal. If the chunk size is more than 0.04 s, the signal will act as a non-stationary signal and if the chunk size is less than 0.02 s, it won't have enough samples to extract features from it. Then Fast Fourier Transform (FFT) is applied for each chunk to calculate the magnitude of the frequency spectrum. Then the spectrum is rotated by 90 degrees to make the frequency in the vertical axis and the amplitude is represented by mapping it to color. These spectrums have some imaginary values, so, a log for the spectrum is applied to visualize it, then these spectrums are added side by side to form the spectrogram. Some spectrograms are visualized in Figure 7.

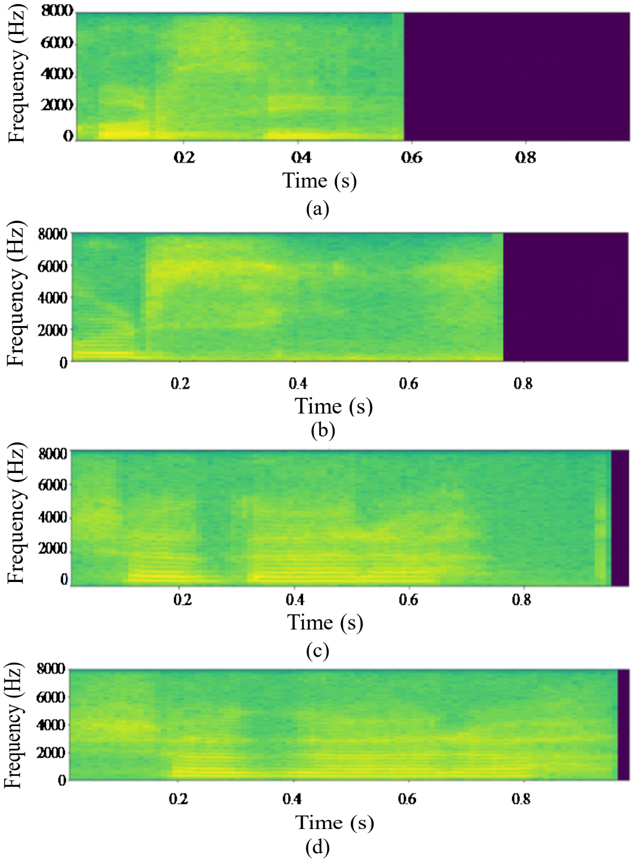


Figure 7: Spectrogram representation of two words in Arabic, each recorded from two different persons: (a), (b) for "chair" and (c), (d) for "tree". Sample rate = 16000 Hz, then frequencies in the range [0-8000] according to Nyquist theory.

b) MFCCs Generation

The MFCCs help to understand the speech and try to determine how sound comes out. To create the MFCC, we used the audio signal in the time domain and frame blocking into frames. Then the Hamming windowing Framing process produces discontinuity frames. Fast Fourier transform (FFT) is applied for each frame to calculate the magnitude of the frequency spectrum. The magnitude spectrum is warped according to Mel scale. Then the magnitude spectrum is segmented into a number of critical bands by means of a Mel filter bank. Then the logs of the powers at each of the Mel frequencies were taken.

The logarithm of the filter bank outputs are called the Mel spectrum. After that, we applied the discrete cosine transform (DCT) of the list of the Mel spectrum to convert it into the time domain. The MFCCs are the amplitudes of the resulting list of the Mel spectrum as shown in Figure 8.

3.2.5 Classification

After extracting the features from the audios, the input became images. These images were used as input to the CNN model. We built our model to classify between 10 words in Arabic language. Thus, we knew if the target word is said or not, and improvement measurements are to be followed.

a) Model Architecture

Our model represents the CNN architecture with three convolutional and pooling layers and one dense layer. Additionally, we add batch normalization to increase stability and dropout to avoid overfitting.

For the hyperparameters of each convolutional layer, we used 32 filters each with a window size of 3×3 with 1 stride. We used max-pooling layers each of 2×2 window cell size with 2 strides after each convolutional layer for down sampling the input by halve. For the dropout hyperparameter, the probability of training a given node in a layer is 0.2.

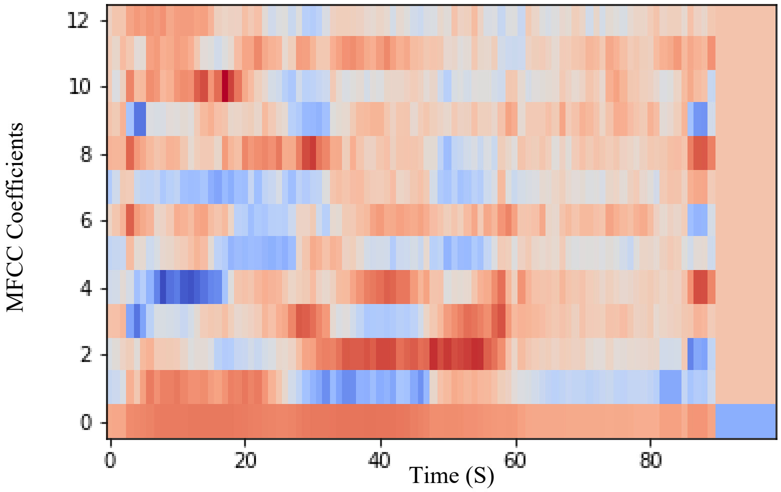
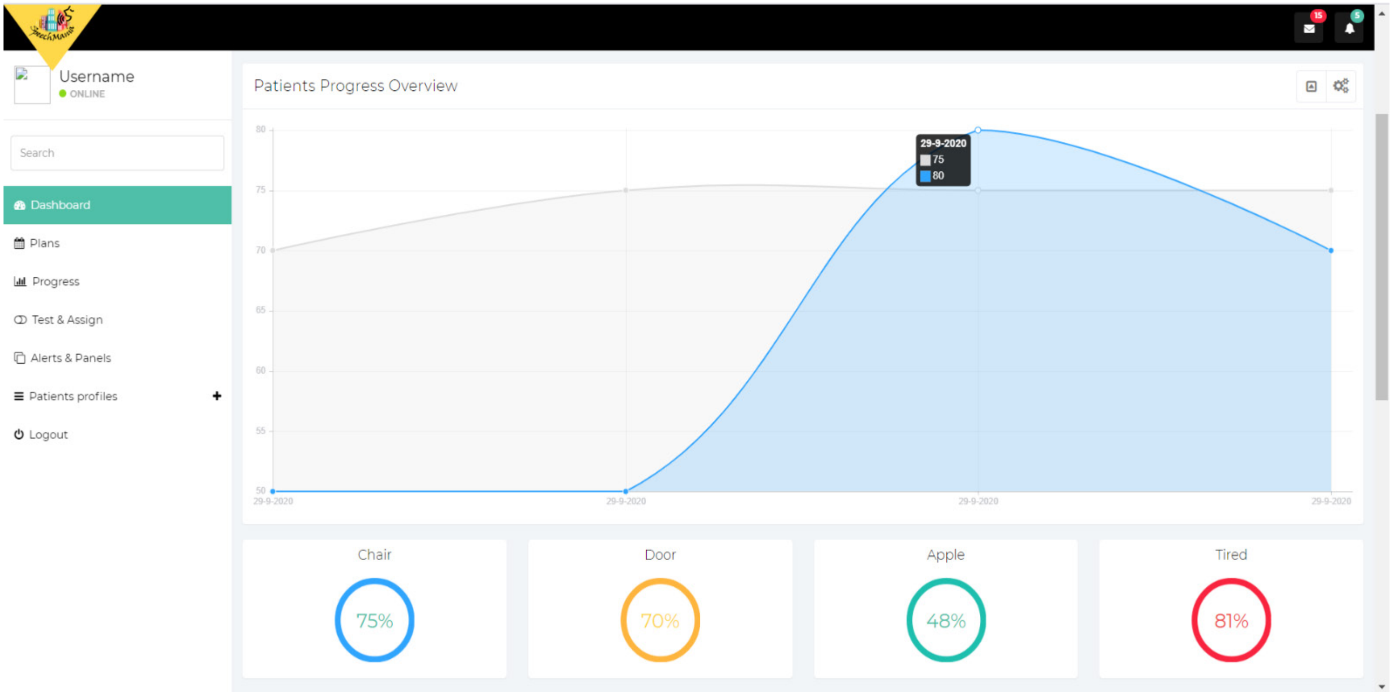


Figure 8: The MFCCs representation of a speech audio signal for “chair”.



Firstly, we organized the dataset and divided it into train and validation sets with a validation ratio of 0.25. Next, we implemented CNN model considering a loss function to be local minimum so, we used standard Adam optimizer minimizing categorical cross-entropy and monitor accuracy in the course of training.

With regard to the training parameters, we used a number of epochs = 15 and batch size = 64. Moreover, early stopping has been added in order to stop the training once validation loss function starts to increase.

3.3. Feedback System

During the game, the voice of the patients is continuously taken to a cloud server using Artificial Intelligence of Things (AIOT) providing real-time feedback to their speech improvement. The feedback is divided into two parts, the first one is an on-screen feedback to guide the patients and encourage improvement in speech and comprehension.

The second one is to be shared via a website with the actual therapist for remote follow-up and to see the progress report of his patients using the AIOT system and assign needed tasks and effective rehabilitation plan.

3.3.1. Website Architecture

The website was created using HTML, CSS, JavaScript and Bootstrap. It consists of three main parts; the home page, therapist authentication page and the most important part is the therapist's dashboard shown in Figure 9. It enables therapists to follow up their patients' progress through the rehabilitation timeline and assign needed tasks.

The feedback is visualized by a curve for each task along the rehabilitation timeline and accumulation of the improvement percentages for learning words, considering the good experience and emphasizing the functionality.

3.3.2. Cloud Server

After evaluating different cloud service providers, we chose Google Firebase mainly because of the ease of use. It offers the needed storage and real-time database for our system so we can store and retrieve data in a game made with Unity3D and share resources between the AI model, the rehabilitation environment, and the website.

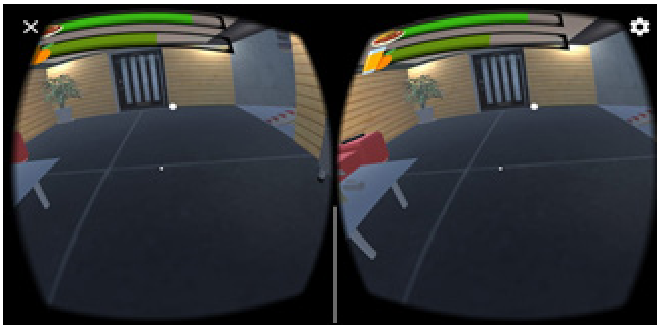


Figure 10: Final VR application on android operating system.

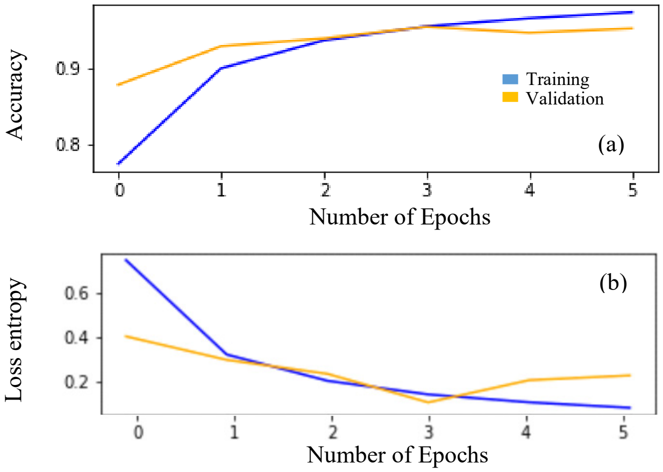


Figure 11: Results of the model on the spectrograms dataset after applying augmentation. (a) Classification Accuracy and (b) Cross Loss Entropy.

4. Results

4.1. VR environment build

Finally, the whole integrated system is built as an interactive computer-aided system on regular PCs and as VR mobile phone application consolidating immersivity concept. Figure 10 shows a final look of the system built on an android device.

Table 1: Comparison of the model accuracy using the two techniques used in feature extraction applied on both datasets.

Feature extraction method	**Dataset**	
	Without augmentation	Augmented data
Log Spectrograms	92.17 %	95.2 %
MFCC	90.7 %	92.6 %

4.2. Speech recognition model results

We tested the CNN model using the two approaches of feature extraction. First, we started with the spectrograms' dataset calculating the accuracy over validation set to obtain the validation accuracy. The final accuracy and loss on the validation set of spectrograms were obtained at the 3rd epoch where the accuracy was 92.17 % and the validation loss was 0.33. Then, we tested the model on the dataset after extracting MFCCs from it and also got the accuracy and loss on the validation set. The final accuracy and loss on the validation set were obtained at the 5th epoch where the accuracy was 90.7 % and 0.39 validation loss.

After applying the augmentation techniques to the original data, we extracted spectrogram and MFCC features and again tested the model on them. The final accuracy and loss on the validation set of spectrograms are shown in Figure 11 where an accuracy of 95.2 % and a validation loss of 0.22 were obtained at the 5rd epoch. Extracting the MFCC features results in an accuracy of 92.6 % and the validation loss was 0.39.

Results of recognition accuracy of both techniques applied on the dataset before and after augmentation are summarized and compared in Table 1.

Table 2: Precision, recall and f1-score calculated for the testing set of the 10 words (classes) on each one of the four models used.

	Log-Spectrogram						MFCC					
	Before Augmentation			**After Augmentation**			**Before Augmentation**			**After Augmentation**		
Classes	**Precision**	**Recall**	**f1-score**	**Precision**	**Recall**	**f1-score**	**Precision**	**Recall**	**f1-score**	**Precision**	**Recall**	**f1-score**
تفاح /Apples	0.500	0.40	0.44	1.00	1.00	1.00	0.40	0.40	0.40	0.57	0.80	0.67
كرسي / Chair	0.00	0.00	0.00	0.43	0.75	0.55	0.17	0.25	0.20	0.33	0.25	0.29
باب / Door	1.00	0.50	0.67	0.57	1.00	0.73	0.67	0.50	0.57	0.50	0.50	0.57
أكل / Eat	0.00	0.00	0.00	1.00	0.75	0.86	0.00	0.00	0.00	0.50	0.25	0.33
عصير /Juice	0.08	0.25	0.13	1.00	0.75	0.86	0.50	0.25	0.33	1.00	0.25	0.40
جنيه / Pound	0.50	0.25	0.33	1.00	0.50	0.67	0.33	0.50	0.40	0.60	0.75	0.67
تاكسي /Taxi	0.00	0.00	0.00	0.33	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
تعبان /Tired	0.36	0.80	0.500	1.00	0.60	0.75	0.38	0.60	0.46	0.60	0.60	0.60
شجرة / Tree	0.50	0.25	0.33	1.00	0.50	0.67	1.00	0.25	0.40	0.50	0.25	0.33
مايه / Water	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.60	0.75	0.67	0.80	0.73

The testing process is done with 15 speakers with the same ratio of males and females in the training set. We obtained about 2 or 3 records of mimicked sounds of multiple cases of patients from each speaker; thus, a total of 40 data points is used to evaluate the performance of the model. Then, we applied all the preprocessing steps on each audio in the test set. Table 2 shows the precision, recall, and f1-score in the classification report of the test data when applying the four trained models.

5. Conclusion

This paper addresses the problems of language and speech dysfunction after stroke or traumatic brain injury. We proposed a computer aided Arabic-based rehabilitation system for people with aphasia as a regular PC interactive game and a VR-based mobile application. Language comprehension and speech production are assessed and measured using a CNN algorithm to provide feedback for both patients and the SLPs via a website for remote follow-up. In the purpose of speech assessment, we build four speech recognition models. Experimenting two feature extraction methods and apply them on the collected audio dataset and one more time after augmenting these data.

The test results showed that the best model was obtained when applying data augmentation techniques on the spectrogram dataset with a training accuracy of 95.2 %.

In future work, more testing and enhancing of the model will be applied, adding more categories of words and more tasks for training and a clinical study will be done to validate our system performance and patient's acceptance of this new rehabilitation method.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to thank Dr. Nahla Rifaie, Professor and head of phoniatrics unit at faculty of medicine, Ain Shams University for contributing with all needed information about the medical condition and the idea of accurate lip syncing, in addition to arranging periodic consultations for validation. We also grateful for each participant who devoted his/her time to help us in collecting the dataset and for Cairo University, the Academy of Scientific Research & Technology, and ITIDA as well for providing us the needed fund to proceed with the project.

References

[1] W. Johnson, O. Onuma, M. Owolabi, S. Sachdey, "Stroke: a global response is needed," Bulletin of the World Health Organization, **94**(9), 634-634A, 2016, doi: 10.2471/BLT.16.181636.

[2] M. Lindsay, B. Norrving, R. L. Sacco, M. Brainin, W. Hacke, S. Martins, J. Pandian, V. Feigin, "World stroke organization (WSO): global stroke fact sheet 2019" International Journal of Stroke, **14**(8), 806–817, 2019. doi: 10.1177/1747493019881353.

[3] United Kingdom, State of the nation: stroke statistics, JN 1718.250, 2018.

[4] National Institute on Deafness and Other Communication Disorders, NIDCD fact sheet | voice, speech and language: Aphasia, NIH Pub. No. 97-4257, 2015.

[5] National Aphasia Association, aphasia definitions fact sheet.

[6] K. Hilari, S. Byng, "Health-related quality of life in people with severe aphasia," International Journal of Language and Communications Disorders, **44**(2), 193–205, 2009, doi: 10.1080/13682820802008820.

[7] O. L. Backus, L. D. Henrry, J. L. Clancy, H. M. Dunn, Aphasia in Adults: The Rehabilitation of Persons with Loss or Disturbance of the Faculty of Speech Resulting from Brain Injury, University of Michigan Official Publication, 1945.

[8] I. K. Hamed, M. Elmahdy, S. Abdennadher, "Collection and Analysis of Code-switch Egyptian Arabic-English Speech Corpus," in the 11th International Conference on Language Resources and Evaluation, 2018.

[9] L. R. Cherney, S. V. Vuuren, "Telerehabilitation, virtual therapists, and acquired neurologic speech and language disorders," Seminars in Speech and Language, **33**(3), 243-57, 2012, doi: 10.1055/s-0032-1320044.

[10] S. V. Vuuren, L. R. Cherney, "A virtual therapist for speech and language therapy," In: Bickmore T., Marsella S., Sidner C. (eds) Intelligent Virtual Agents. Lecture Notes in Computer Science, **8637**, Springer, Cham, 2014, doi: 10.1007/978-3-319-09767-1_55.

[11] J. R. Galliers, S. Wilson, J. Marshall, R. Talbot, N. Devane, T. Booth, C. Woolf, H. Greenwood, "Experiencing eva park, a multi-user virtual world for people with aphasia," ACM Transactions on Accessible Computing, **10**(4), 15, 2017, doi: 10.1145/3134227.

[12] H. S. Lee, Y. J. Park, S. W. Park, "The effects of virtual reality training on function in chronic stroke patients: A systematic review and meta-analysis," Biomed Research International, 2019, doi: 10.1155/2019/7595639.

[13] J. Thomas, C. France, S. Leitkam, M. Applegate, P. Pidcoe, S. Walkowski, "Effects of real-world versus virtual environments on joint excursions in full-body reaching tasks," IEEE Journal of Translational Engineering in Health and Medicine, **4**, 1-8, 2016, doi: 10.1109/JTEHM.2016.2623787.

[14] S. Parsons, S. Cobb., "State-of-the-art of Virtual Reality technologies for children on the autism spectrum," European Journal of Special Needs Education , **26**(3), 355-366, 2011, doi: 10.1080/08856257.2011.593831.

[15] N. S. Rosenfield, K. Lamkin, J. Re, K. Day, L. Boyd, E. Linstead, "A virtual reality system for practicing conversation skills for children with autism," Multimodal Technologies and Interaction, **3**(2), 28, 2019, doi: 10.3390/mti3020028.

[16] B. Brundage, A. B. Hancock, "Real enough: using virtual public speaking environments to evoke feelings and behaviors targeted in stuttering assessment and treatment," American Journal of Speech-Language Pathology, **24**(2), 139-149, 2015, doi: 10.1044/2014_AJSLP-14-0087.

[17] Y. Zhang, P. Chen, X. Li, G. Wan, C. Xie, X. Yu, "Clinical research on therapeutic effect of virtual reality technology on broca aphasia patients," in IEEE International Conference on Information Technology, 2, 2017, doi: 10.1109/INCIT.2017.8257880.

[18] M. Horváth, C. Dániel, J. Stark, C. Sik Lanyi, "Virtual reality house for rehabilitation of aphasic clients," Transactions on Edutainment III, **5940**, 231–239, 2009, doi: 10.1007/978-3-642-11245-4_20.

[19] N. Sebkhi, D. Desai, M. Islam, J. Lu, K. Wilson, M. Ghovanloo, "Multimodal speech capture system for speech rehabilitation and earning," IEEE Transactions on Biomedical Engineering, **64**(11), 2639-2649, 2017, doi: 10.1109/TBME.2017.2654361.

[20] K. Grechuta, B. Rubio Bellaster, R. Esp´ın Munn´e, T. Usabiaga Bernal, B. Molina Herv´as, R. San Segundo, P. F.M.J. Verschure, "The effects of silent visuomotor cueing on word retrieval in broca's aphasies: a pilot study," in IEEE International Conference on Rehabilitation Robotics, 2017, doi: 10.1109/ICORR.2017.8009245.

[21] J. Gibson, M. Van Segbroeck, S. Narayanan, "Comparing Time-Frequency Representations for Directional Derivative Features," in Interspeech, 612–615, 2014.

[22] M. Alsulaiman, G. Muhammad, Z. Ali, "Comparison of Voice Features for Arabic Speech Recognition," in IEEE 6th International Conference on Digital Information Management, **6**, 2011, doi: 10.1109/ICDIM.2011.6093369.

[23] S. Prasomphan, "Improvement of speech emotion recognition with neural network classifier by using speech spectrogram," in International Conference on Systems, Signals and Image Processing ,2015, doi: 10.1109/IWSSIP.2015.7314180.

[24] N. Jmour, S. Zayen, A. Abdelkrim, "Convolutional neural networks for image classification" in IEEE International Conference on Advanced Systems and Electric Technologies, 2018, doi: 10.1109/ASET.2018.8379889.