Journal of Theoretical and Applied Information Technology
15th July 2023. Vol.101. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

VOWEL RECOGNITION FOR SPEECH DISORDER PATIENT VIA ANALYSIS ON MEL-FREQUENCY CEPSTRAL COEFFICIENT (MFCC) IMAGES

NUR SYAHMINA AHMAD AZHAR[1], NIK MOHD ZARIFIE HASHIM[2*], MASRULLIZAM BIN MAT IBRAHIM[3], MAHMUD DWI SULISTIYO[4]

[1,2,3] Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[4]School of Computing, Telkom University, West Java, Indonesia
*Corresponding Author Email: nikzarifie@utem.edu.my

ABSTRACT

An individual with a speech disorder, autism, brain injury, autistic spectrum disorders, and stroke usually has trouble producing or forming the spoken sounds necessary for effective interactions. As a result, patients' rehabilitation and treatment typically take a long time and involve ongoing medication, physical activity, and rehabilitation training. However, this rehabilitation process is still done manually in most rehab centers worldwide. Since the impact of computer vision on this profession, machine learning and deep learning have been introduced to the medical industry to improve rehabilitation using the new technology. Convolutional Neural Network (CNN) models have been proven in countless studies to be precise at classifying performance in various fields, including visual field, computer vision, audio, and text defects. This study analyzed the classification accuracy of different pre-trained models (Designed network, VGG-Net, AlexNet, and Inception). We created a thorough comparative analysis to compare the accuracy of several CNN models. The image-profiled sound uses the Mel-frequency Cepstral Coefficient (MFCC) to produce the best results and accuracy. This study aims to create a neural network that can discriminate between the vowels in the voices of normal persons and speech disorder patients. According to experimental results, the designed model had the highest accuracy of 94.54% by using 6 batch sizes, 20 epochs, and ADAM as the optimizer. Furthermore, we discovered that combining various hyper-parameters and fine-tuning the pre-trained models deliver impacts the performance of deep learning models for this classification task.

Keywords: *Convolutional Neural Network (CNN), Deep Learning, Mel-Frequency Cepstral Coefficient (MFCC), Speech Disorder Patients, Vowel Recognition*

1. INTRODUCTION

A brain or biological neural network is considered the most well organized system that processes information from different senses such as sight, hearing, touch, taste, and smell efficiently and intelligently. One of the key mechanisms for information processing in a human brain is that complicated high-level information is processed collaboration. The collaboration consists of connections called synapses and large number of structurally simple elements called neurons [1].

The capability of the brain charms is communication between humans. Communication is the process of exchanging and sharing information and ideas [2]. Encoding, transmitting, and decoding messages effectively are all aspects of communication. Consequently, speech and language only represent a small part of communication. Other communication related aspects could exceed both of those parts. These aspects can be divided into paralinguistic, non-linguistic, and metalinguistic. Paralinguistic mechanisms can modify the form and meaning of a sentence. Intonation, stress, rate of delivery, and delay or hesitation are examples of these mechanisms that indicate attitude or mood. Some aspects of non-linguistic behavior can influence communication. Gestures, body posture, facial expression, eye contact, head and body movement, and physical distance, are non-linguistic clues.

Based on intuition about the appropriateness of statements, metalinguistic cues indicate the state of communication. In other words, metalinguistic abilities allow humans to discuss, examine, consider,

Journal of Theoretical and Applied Information Technology
15th July 2023. Vol.101. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

dissociate from, and evaluate language [3]. Speech is a linguistic medium for expressing one's thoughts or conveying meaning. Speech demands precise neuro-

Table 1: Acoustic Vowels' Characteristics.

Tongue Height	Tongue Position		
	Front	Centre	Back
High	/i/		/u/
Mid-high	/e/	/E/	/o/
Mid-low			
Low	/a/		

muscular coordination since it is the outcome of specific motor behaviors. Voice quality, intonation, pace, and sound combinations contribute to speech. These elements are each used to change the speech message. The smallest unit of speech is called phoneme [4-5]. The phoneme, a group of sounds with similar enough in perceptual qualities to be differentiated from other phonemes. A phoneme can be either vowels or consonants.

The definition says vowels are produced with a largely unrestricted airflow in the vocal tract. While at the same time, consonants require a closed or narrowly constrained path that results in friction and air turbulence. Phonemes are now classified using two approaches [4-5].

In this system, vowels use the uppermost part of the tongue, the tongue's position from front to back, and the rounding of the lips. Height of the tongue can be classified as high, mid, or low depending on where the highest part of the tongue is located. This highest position's location might be categorized as front, central, or back. Table 1 shows the vowel's characteristic of human's tongue [6-8]. A vowel, for instance, can be described as high front or low back. An additional description for classifying vowels is lip rounding. As the lips round, a small portion of lips protrudes, creating an "o" shape. Some back vowels have a rounded sound [7]. Figure 1 illustrates the human's vowel by tongue position graphically.

In medical terms, there are two types of Aphasia. An aphasic condition can be fluent and non-fluent. A non-fluent Aphasia lacks articulatory precision prosody in speech and, they like hesitant and slow with many pauses while communicating [9]. Mostly the causes of Aphasia is from stroke and brain damage. A stroke occurs when the blood supply to the brain is interrupted or reduced. It will prevent the brain tissue from getting enough oxygen and nutrients. The symptoms of someone having a stroke are headache, trouble walking and problems seeing with in one or both eyes. In addition, the signs of a stroke are like having some trouble in speaking and understanding what others are trying to say, and it can make someone paralysis and numbness in the face, arms, and legs [9-10].

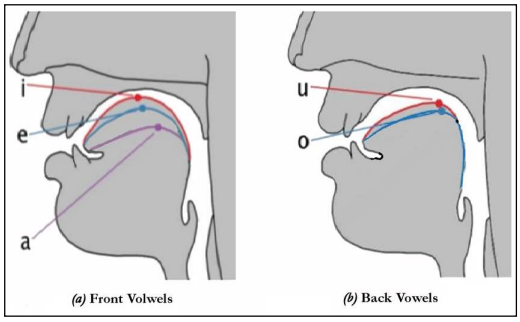


Figure 1: Vowel Production by Tongue Position

The two main motor speech disorders are dysarthria and speech apraxia. Regarding apraxia of speech, the patient has difficulty speaking while having unaffected muscle strength. Apraxia of speech is defined mainly by articulation errors, such as stress, intonation, or rhythm [11]. These errors are considered a main symptom of the disorder as many researchers believe it is a form of compensation [12]. Dysarthria refers to a group of linked speech disorders that are caused by disturbances in the muscular control of speech. According to this definition, the term "dysarthria" only refers to speech disorders with a neurogenic origin and excludes diseases related to somatic anatomical defects or other disorders [13]. A motoric impairment, or a basic disturbance of movement, the muscles used in the speech production process is the outcome of dysarthria [14].

This proposed paper will focus on vowel recognition for normal and disordered people. The data of sound recording have been collected within two groups, normal person, and disorder patient. We conducted a new study on the performance of normal person and speech disordered patients, and six kinds of vowels were also used in the evaluation. This study uses a convolution neural network to provide stroke patients with Malay language vowel recognition (CNN). The technique used Mel-Frequency Cepstral Coefficients (MFCC) image-profiled in place of CNN's conventional sound file to distinguish all six vowels in the Malay language. Five models of networks will be employed in this paper's experiment, which will be conducted via a network.

Firstly, the recording part of vowel pronunciation by the two is conducted, and it will be stored in the

data collection. Then, the conversion from the audio to the image profile and the cropping process to the fixed dimension is done. Lastly, the process of training, evaluating, and testing the recorded data by comparing the accuracy result of disorder people to normal people is carried out by using the proposed network model. The paper delivers two contributions:

1) we initiated a comparative study for vowel recognition via newly proposed and collected Mel-Frequency Cepstral Coefficients (MFCC) dataset images, and
2) the newly designed network outperformed other existing comparative vowel recognition methods via MFCC image-profile dataset.

Dealing with speech therapy assessment, particularly employing the vowel's reliability, accuracy, noise robustness, processing time, and automatic speech recognition, is still the main concern. For providing systematic training and testing for voice recognition via rehabilitation activities, this paper assesses various alternative design options using deep learning. Furthermore, improving the learning network model is crucial to produce a better result in vowel recognition, which would help us gain a reliable result in an actual application.

2. RELATED WORKS

Many treatment facilities and rehabilitation centers offer rehabilitation services to address the communication deficit. Speech rehabilitation is a service that focuses on the communication issues frequently experienced by people with disorders who have lost their capacity to communicate properly. As technology evolves, a lot of research and techniques involved in the speech therapy and rehabilitation field. Most therapy plans are based on the specific characteristics of developmental speech rehabilitation and use treatment methods to address them.

Some techniques that target the oral motor/speech systems are traditional therapy approaches, sensory awareness, reduced speech rate, motor drills, phonetic placements, sequencing and systematic drills, and the PROMPT system of therapy [15]. The purpose of the speech therapy exercises is for the patient to recover. The speech therapist frequently evaluates the patients' cognitive and communicative abilities during face-to-face therapy session. Since Malaysia has a shortage of speech therapists, alternative therapy provided through systems and applications is seen as a more effective treatment. There are speech system treatment programs, but none are accessible to post-stroke patients.

Before deep learning, a handful of rehabilitation techniques concentrated on speech impairments. In order to deal with all levels of the identified underlying impairment, therapy of this condition requires a comprehensive treatment approach. The development of delayed speech may be impacted by the delay in motor skill acquisition, especially in young children, according to studies that show a parallel development path for speech acquisition and fundamental motor skills. According to Van Riper's Speech Correction in his book, speech and motor skills are connected. To help those who lack communication to develop accurate pronunciation, the articulation drill and motor learning technique focused on motor practice of the tongue movement and synchronization of other articulators, such as the lips and jaw [16].

In addition to motor learning and articulation training, phonological/lexical interventions are employed with words and phrases. These interventions seek to boost word-level productivity [16]. This strategy varies from the Articulation Drill and Motor Learning strategies, which concentrates on where and how each spoken sound is produced utilizing the body. There is a reference about developmental of speech is increasingly acknowledged to be composed of several features. Unfortunately, delayed improvement in treatment plans is one of the primary indicators of the developmental of speech. The patients' improvements are indeed expected to continue over time [17].

Recently, there have been methods of speech rehabilitation using machine learning. Deep learning is currently one of the most popular machine-learning (ML) based techniques, and CNN is the major deep learning (DL) architecture used for image processing. Deep learning applications in rehabilitation speech treatment scenarios are another current technique for speech therapy in rehabilitation. Using a Malay Speech Intelligibility Test (MSIT) method, Yusof et al. investigated the speech comprehensibility of deaf children in Malaysia. Syllabification algorithms based on the Malay syllable structure were researched by researchers from Universiti Malaysia Sarawak [18]. Instead of the normal Malay Language, it was utilized to generate the Iban and Bidayuh syllable list and speech corpus.

A recent proposal by M. Y. S. Azmi called for gender vowel recognition and a modest improvement of 1.71%. The findings of the tests conducted on the pronunciation application suggested that it could aid users in assessing and improving their word pronunciation in Malay. However, the results indicated that the vowels /i/, /e/, /o/, and /u/ are frequently mispronounced due to pronunciation patterns.

A convolutional neural network has been trained to recognize a few phrases inside the odd and unusual speech using a speaker-dependent approach. They focused on identifying single words for native Italian with dysarthria speakers. They collected audio data from people with speech impairments as they performed articulation exercises using an already-existing mobile app to aid in speech therapy [19]. By implementing computer-based speech therapy systems in the Clinic of Audiology and Speech Sciences, Kay Elemtrics VisiPitch, and IBM Speech Viewer, Universiti Kebangsaan Malaysia (UKM) pioneered the practice in Malaysia [20]. Although not used for training or articulation treatment, these systems are used for voice therapy.

Other programs like OLTK (Optical Logo-Therapy Kit) and VATA (Vowel Articulation Training Aid) [21] exist in addition to the UKM voice therapy initiative. These methods have flaws and aren't strong enough to handle vowel detection in real time. A Malay Speech Therapy Assistance Tool (MSTAT) was created by Tan et al. in 2007 [22] in order to assist therapists in the diagnosis of language disorders in children and the training of children who stutter. It was created using a small word list from the Malay language. A Malay Language dialect translation and synthesis system created by Tan et al. later in 2012, however, it is still in its infancy.

Currently, the Recurrent Neural Network (RNN) advancement motivates [23,24] the proposed research to explore the Malay Language vowel recognition sub-field of speech recognition. The majority of voice recognition software analyses speech to break it down into parts that can be understood using Natural Language Processing (NLP) [25, 26] and deep learning neural networks. Here, the sound wave is represented as an image by connecting this RNN with vowel speech data. Despite the numerous studies on Malay language phoneme recognition, there is still more to be done for those who suffer from disorders, particularly post-speech disorder patients.

3. PROPOSED WORKS

This study's suggested method covers improving training and validation accuracy. A huge dataset of images, including both images of healthy and normal people and speech disorder patients, is to construct a new intelligent categorization system for speech disorder patients. The recording, converting, and cropping processes produce data for normal people and speech disorder patients. A convolutional neural network will then be trained using the entire dataset.

Table 2: Distribution of Group of Person from Collected Datasets.

Group of Person	Quantity	No. of record (images)
Normal person	20	10800
Speech disorder patient	9	1620

3.1 Dataset

All the experiments were performed on the dataset collected within two groups of persons, a normal person and a speech disorder patient. The dataset contains images from 20 normal person and images from 9 speech disorder patients. Both groups of persons have 12420 MFCC images collected to conduct three different experiments. These images were acquired in the bit depth of 32, and the dimension used is fixed at 35 x 200 pixels.

3.1.1 Audio recording

Recording should begin with the vowel sounds /a/, /e/, /E/, /i/, /o/, and /u/. The vowels /a/, /e/, /E/, /i/, /o/, and /u/ on healthy people and speech disorder patients were recorded using a REMAX RP1 8GB Digital Audio voice recorder. Everyone records similarly with a 15 cm space between their mouths and the voice recorder. For normal persons, every vowel must be pronounced in three distinct segments, short, middle, and long. The lengths of the recordings for the short-period, middle-period, and long-period signals are 1, 2, and 3 seconds, respectively. The group of speech disorder patients they are free to pronounce the vowel for between one to three seconds. This condition is caused by the fact that speech disorder patients have difficulty and struggle to pronounce vowel sounds. To come up

Journal of Theoretical and Applied Information Technology
15th July 2023. Vol.101. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

with these patients' voice pronunciation capabilities, we set the recording period for the normal person in three-period segments.

3.1.2 Mel-Frequency Cepstral Coefficients

The type of image-profiled used in these experiments is Mel-Frequency Cepstral Coefficients

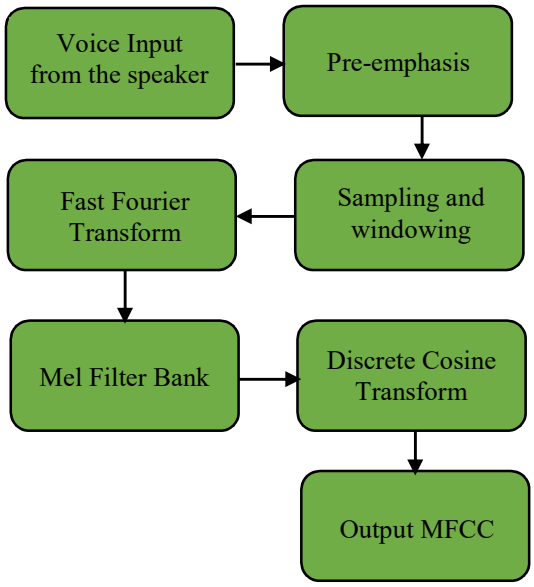


Figure 2: The Flow of Generating MFCC Images

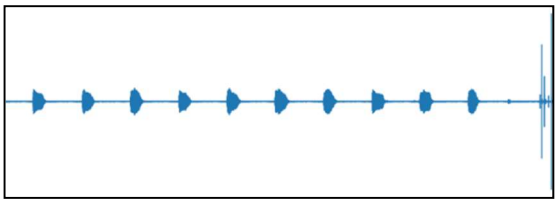


Figure 3: A Sample of Amplitude vs. Time Audio in Wav.file

(MFCC). MFCC, which has been employed in biomedical applications as well as voice-based processing applications like speaker identification, voice recognition, and gender identification utilizing the voice. The block diagram in Figure 2 describes how the MFCC is computed step by step. The method used to identify a voice from a speaker is known as input speech. Figure 3 shows an input speech of real-time audio which recorded in wav.file from a speaker to pronounce 10 repeated sounds of vowel.

After some preliminary processing, the characteristics of the speech sample are extracted. There are three feature extraction steps in MFCC which are pre-emphasis, frame blocking and windowing [27]. Pre-emphasis is a speech signal *x(n)* must be sent through high-pass filter. The equation shows the output signal labelled as *y(n)* and the value of 'a' is usually between the value 0.9 and 1.0.

$$y(n) = x(n) - a * x(n-1) \quad (1)$$

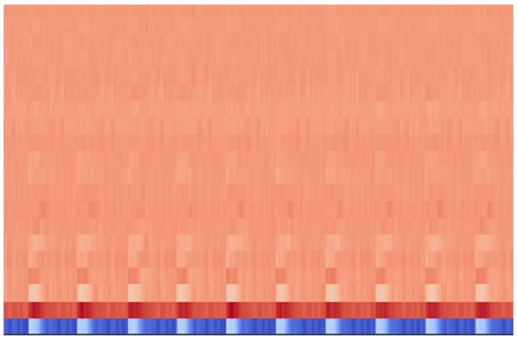


Figure 4: A Sample of MFCC Image Constructed from Sound of Normal Person

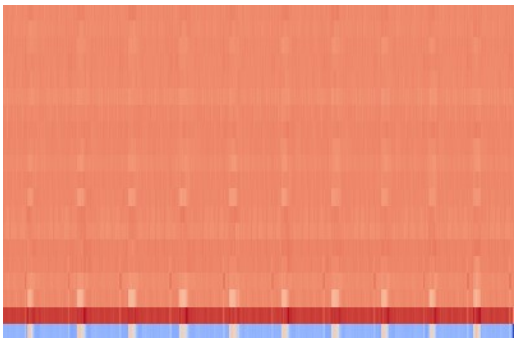


Figure 5: A Sample of MFCC Image Constructed from Sound of Speech Disorder Patients

The process of frame blocking is the speech signal is divided up into frames, which are 20–30 ms long, each while windowing is a speech signal is divided into temporal fragments by using a technique that is frequently employed in signal processing. Figure 4 and 5 shows the MFCC images of 10 repeated sounds of vowel for normal person and speech disorder patients [27-28].

3.1.3 Conversion from wav.file audio to Image-profile

Once the conversion from real-time audio to MFCC image-profile is done, the cropping process will be conducted. As mentioned, each vowel's

image have a size of 200 x 35 pixels and a bit depth of 32 bits. The conversion process is done by Python coding. Figure 5 shows the cropped MFCC images of normal person and disordered patients in every class of vowels.

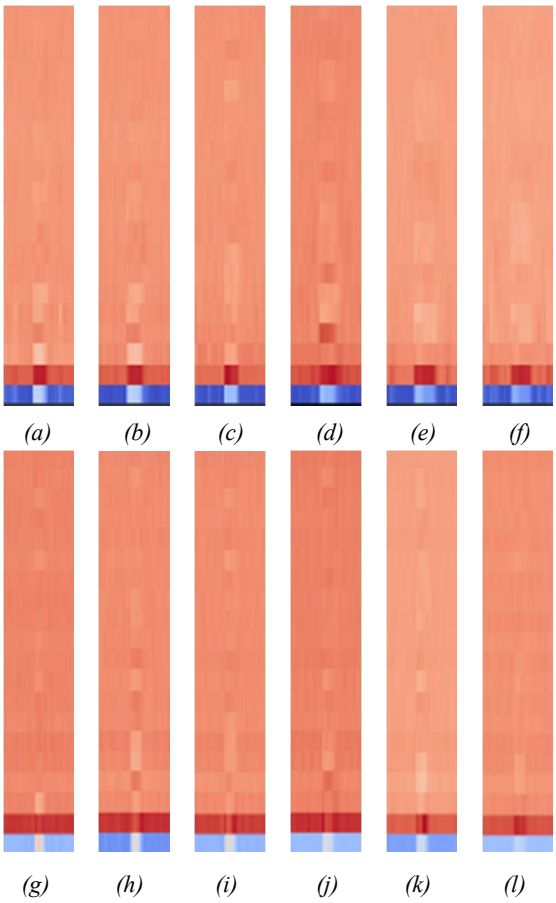


Figure 6: MFCC images for vowels (a)-(f) for vowel /a/, /e/, /E/, /i/, /o/ and /u/ for normal person and vowels (g)-(l) for vowel /a/, /e/, /E/, /i/, /o/ and /u/ for speech disorder patient

3.2 The Designed Network

This section includes a description of the structure for the designed network. We designed a basic network model to address the vowel recognition issue. Using the images from the six vowels collection, this paper used fix input image with a resolution of 200 x 35 pixels. The designed network, which is based on the CNN design, consists of activation functions, a pooling layer, a fully connected layer, a dropout layer, and a convolutional layer. Convolution layer 1 is the first one. The second convolution layer, Conv1-1, uses 32 filters and considers the image is 238 by 53 pixels in size. The convolution layer used in this experiment includes 32 filters and a maximum pooling size for the images of 118 x 25.

Conv2 is the second convolution layer and has 32 filters. The Conv2-1 is the next, with a maximum pooling of 32 filters and an image resolution of 116 x 23. The Conv2-2 has a maximum pooling size of 57 x 10 and employs 32 filters concurrently. The final convolutional layer has 64 filters and recognizes that the image is 55 by 8 pixels. The images' maxpooling layer size is 26 by 3. Finally, to create a dense layer, 1024 units of a dense layer are combined with 6 units of a dense SoftMax layer. The model was created using a CNN with an input image dimension of 200 x 35, an ADAM classifier as the optimizer, and SoftMax as the activation function.

Table 3: Confusion Matrix Used to Compute Error Metric

Predicted/True	Segment	Non-segment
Segment	TP	FP
Non-segment	FN	TN

TP: True positive, FP: False positive, TN: True negative, FN: False negative

3.3 Characterization of Errors

A confusion matrix aids in visualizing the various outcomes of a classification task by presenting a table layout of the results and predictions. They provide direct comparisons of variables like 'True Positives', 'False Positives', 'True Negatives', and 'False Negatives', confusion matrices are beneficial [33]. A confusion matrix, which identifies which classes the convolutional neural network (CNN) correctly predicts and which classes it predicts poorly, is used to determine when the model is testing a certain data.

Important predictive indicators like recall, specificity, accuracy, and precision are represented using confusion matrices shown in the equations below.

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} \frac{TN_k + TP_k}{TN_k + TP_k + FN_k + FP_k}}{K} \quad (2)$$

$$Precision = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FP_k}}{K} \quad (3)$$

(4)

$$Recall = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}}{K}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision^{-1} + Recall^{-1}} \quad (5)$$

4. EXPERIMENTAL RESULTS

This section will present the results of the overall performance of the proposed CNN with other existing network models with performance accuracy using confusion matrix. The results were divided into three analysis studies: (a) the comparison of pre-trained models for the normal person, (b) the comparison of pre-trained models for the speech disorder patient, and (c) the comparison of the pre-trained model for the both of normal person and speech disorder patient. The vowel dataset for the three analysis studies in (a), (b) and (c) were split into training, evaluation, and testing with a ratio of 80:10:10, respectively.

4.1 Pre-trained Models for the Normal Person

The goal of the first study analysis is to verify the effectiveness of pre-trained models (Designed model, VGG-16, VGG-19, AlexNet and Inception) that use images with dimensions of 55 x 200 and 32-bit depth. The KERAS (Tensorflow) neural network computing framework was used to write the method in Python.

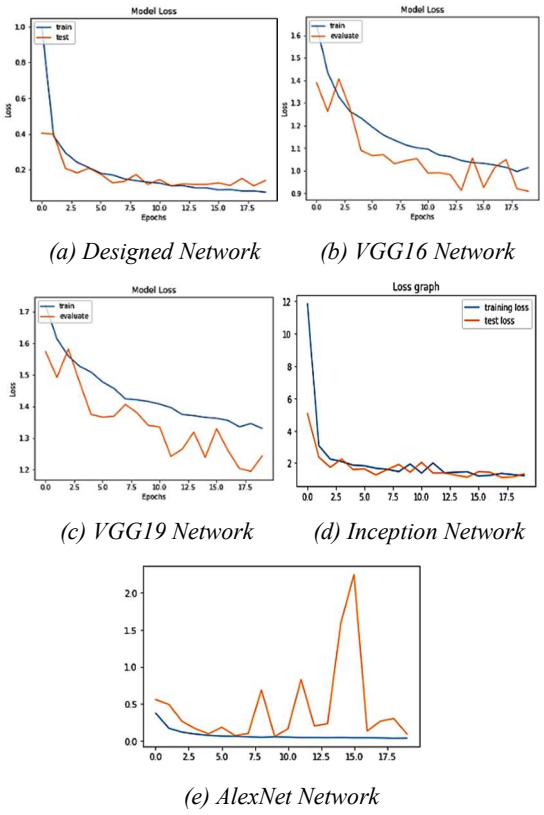
20 healthy subjects were included in the first study analysis and a total of 10800 MFCC images were gathered. We set the batch size and epoch size to be similar in the classification studies for the normal person, speech disorder patients, and mixed group (normal person + speech disorder patient), in order to investigate the effectiveness during the small epoch size, which is reliable to the actual application.

These models are implemented to compare each other completely using five different network models, designed model, VGG16, VGG19, Inception, and AlexNet model. We described the results of the experiment for each of these network. When applying the designed network model in Figure 7 (a) for batch size = 6 and epoch 20, the classification accuracy is 94.54%.

The validation accuracy of each model is compared in Table 4. According to the analysis, the designed network has the highest accuracy compared to the other networks. Figure 7 shows the graph for every pre-trained model loss performance while Figure 8 shows the accuracy performance graphs for every pre-trained model.

Table 4: Results of Multiple Models in the First Study Analysis.

Epoch	Batch Size	Model	Accuracy Percentage	
			Training (%)	Validation (%)
20	6	Designed	**97.87**	**94.54**
		VGG16	59.21	62.50
		VGG19	48.45	46.94
		Inception	40.17	41.61
		AlexNet	90.29	75.77



(a) Designed Network *(b) VGG16 Network*

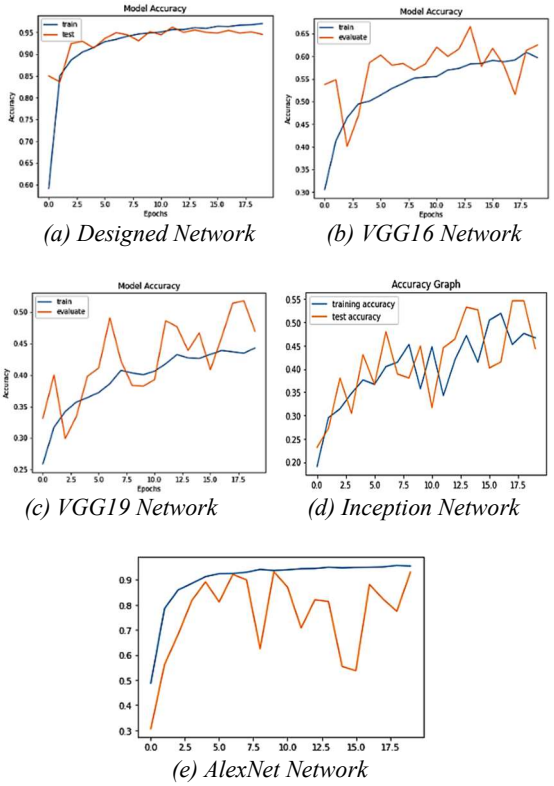
(c) VGG19 Network *(d) Inception Network*

(e) AlexNet Network

Figure 7: Model Loss Performance in the First Study Analysis

Journal of Theoretical and Applied Information Technology
15th July 2023. Vol.101. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(a) Designed Network (b) VGG16 Network

(c) VGG19 Network (d) Inception Network

(e) AlexNet Network

Figure 8: Model Accuracy Performance in the First Study Analysis

For the same setting in training and testing conducted by Hashim et al. [34] and other network models, batch size six and epoch 20, had the highest accuracy in 2022, only 81%. However, with the help of the newly designed model in this proposed paper, out of the six model networks tested, we performed the highest validation accuracy for epoch 20 and batch size 6, with 94.54%. We have successfully created a better accuracy network for this study investigation than Hashim's proposed network [34]. The dataset gathered for healthy and normal people is larger in this study compared to Hashim's dataset [34].

Table 5: The Comparison of Result Accuracy of the Designed Network Models.

Designed Network Model	Classification Accuracy (%)
Hashim et al. (2022)	81.00%
Proposed Network Model	**94.54%**

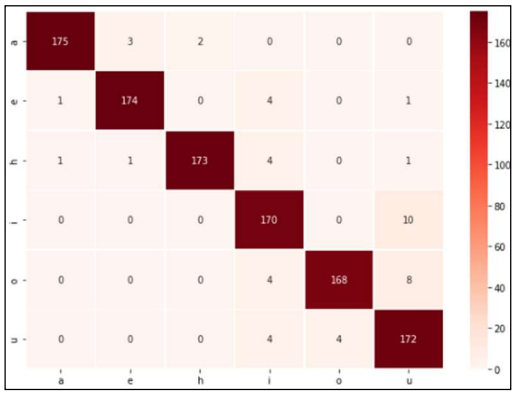


Figure 9: Designed Model Confusion Matrix in the First Study Analysis

Table 6 displays the accuracy of each vowel's class according to designed models. The confusion matrices for the six groups of vowels from the categorization are shown in Figure 9 of this paper, which identified six different types of vowel classes. The /h/ class shown in the confusion matrix is referred as /E/ class.

Table 6: Comparison of the Testing for Every Classes of Vowels in the Pre-trained Designed Model during First Study Analysis.

Vowel	Precision (%)	Recall (%)	F1 (%)	Support
Class /a/	99.00	97.00	98.00	180
Class /e/	98.00	97.00	97.00	180
Class /E/	99.00	96.00	97.00	180
Class /i/	91.00	94.00	93.00	180
Class /o/	98.00	93.00	95.00	180
Class /u/	90.00	96.00	92.00	180

4.2 Pre-trained Models in Speech Disorder Patient

The second study analysis is carried out to evaluate how well the dataset performs among speech disorder patients. In the second study, a dataset with nine speech disorder patients was included. We gathered a total of 1620 MFCC images from the second investigation. From the first study analysis experiments, the designed model has the highest accuracy in comparison to the others.

Journal of Theoretical and Applied Information Technology
15th July 2023. Vol.101. No 13

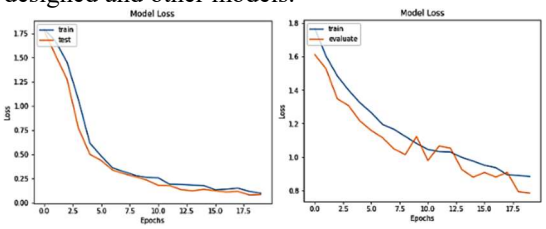
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The carried out training procedure uses the same 20 epoch and 6 batch size as the initial study. 1620 MFCC pictures from nine speech disorder patients are included in the second investigation. The 10% validation process used 162 out of the 1296 MFCC images from the six classes of vowels, and others used in the 80% training process. The remaining 162 MFCC images made up the remaining 10% of the testing procedure.

Table 7: Results of Multiple Models in the Second Study Analysis.

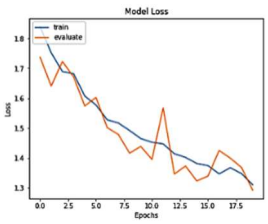
Epoch	Batch Size	Model	Accuracy Percentage	
			Training (%)	Validation (%)
20	6	Designed	**96.45**	**98.77**
		VGG16	73.38	77.16
		VGG19	52.47	52.47
		Inception	56.31	59.35
		AlexNet	68.35	42.28

When this result is compared to the result on the first study analysis, which is depicted in Table 4, the validation accuracy for the designed network was higher in the second study analysis compared to the validation accuracy for the designed model in the first study analysis. During conducting the experiment for the second study which contains a dataset of nine speech disorder patients, we can observed that the validation accuracy of speech disorder patients group is higher than that of normal persons group. Figure 10 and 11 compares the training and validation accuracy graphs for the designed and other models.

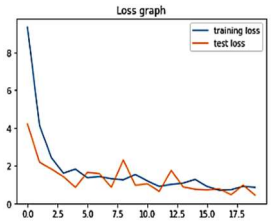


(a) Designed Network

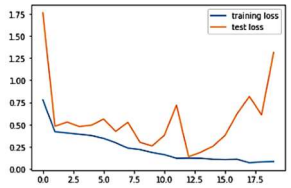
(b) VGG16 Network



(c) VGG19 Network

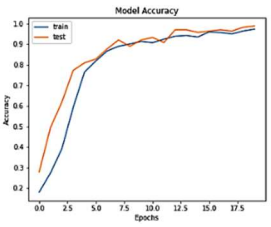


(d) Inception Network

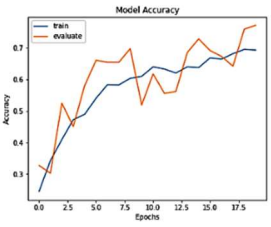


(e) AlexNet Network

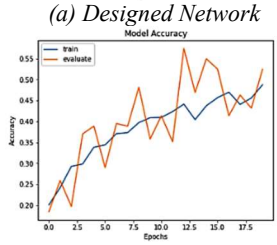
Figure 10: Model Loss Performance in the Second Study Analysis



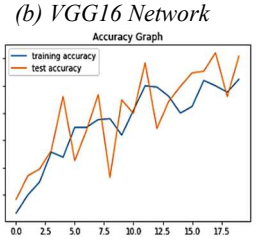
(a) Designed Network



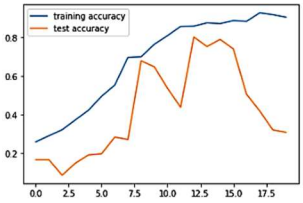
(b) VGG16 Network



(c) VGG19 Network



(d) Inception Network



(e) AlexNet Network

Figure 11: Model Accuracy Performance in the Second Study Analysis

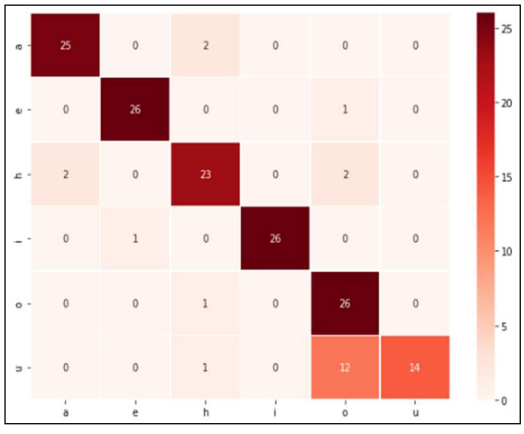


Figure 12: Designed Model Confusion Matrix in the Second Study Analysis

The dataset for the second and third studies analysis consists speech disorder patients and the combination of both group respectively. The designed model and other network models reliability, consistency, and performance have been observed during the training process. In comparison to previous networks, the planned training was more accurate and required less layers through each layer. Figure 12 shows the confusion matrix of every classes of vowels for stroke patient.

Table 8: Comparison of the Testing for Every Classes of Vowels in the Pre-trained Designed Model during Second Study Analysis.

Vowel	Precision (%)	Recall (%)	F1 (%)	Support
Class /a/	93.00	93.00	93.00	27
Class /e/	96.00	96.00	96.00	27
Class /E/	85.00	85.00	85.00	27
Class /i/	100.00	96.00	98.00	27
Class /o/	63.00	96.00	76.00	27
Class /u/	100.00	52.00	68.00	27

4.3 Pre-trained Models in Normal Person and Speech Disorder Patient

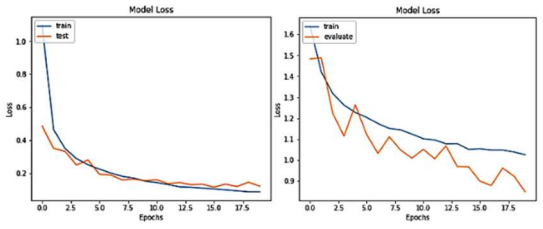
The third study analysis combines the datasets from a normal persons group and speech disorder patients group. The approach required 80% training using 9936 MFCC images and 10% evaluation with 1242 images, totaling 12420 MFCC images. The third study analysis is carried out to evaluate the performance of the combination dataset in a normal person and speech disorder patient. From the first study and second study analysis experiments that the designed model has the highest accuracy in comparison to the others. The carried out training procedure uses the same 20-epoch and 6-batch size as the initial study.

Table 9: Results of Multiple Models in the Third Study Analysis.

Epoch	Batch Size	Model	Accuracy Percentage	
			Training (%)	Validation (%)
20	6	Designed	**98.00**	**95.57**
		VGG16	59.14	69.65
		VGG19	44.62	49.44
		Inception	37.09	37.16
		AlexNet	89.54	78.02

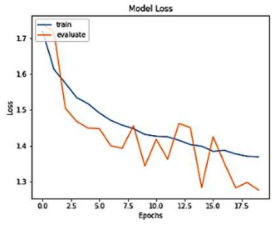
The designed network model is contrasted with VGG16 model, VGG19 model, Inception model, and AlexNet model to show the classification performance. In the third study analysis, a total of 12420 MFCC images were employed. For the training and evaluation phases, 9936 photos from 20 normal persons group and 1242 images from 9 speech disorder patients group, respectively, were used.

According to the result accuracy table shown in Table 9 for the third study analysis, the designed model has the highest accuracy (95.57%), followed by the AlexNet model (78.02%). With a 37.16% accuracy rate, the Inception model performs the least accurately of all the investigations. Deep neural networks used in Inception must be extremely powerful. For a neural network to be considered substantial, there must be multiple additional network layers and units inside those layers. Figure 13 and 14 shows the model loss and model accuracy graphs for the designed and other models.

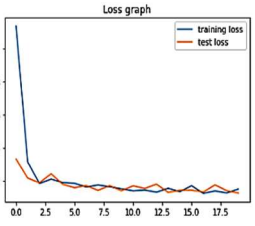


(a) Designed Network

(b) VGG16 Network



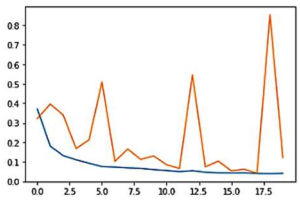
(c) VGG19 Network



(d) Inception Network

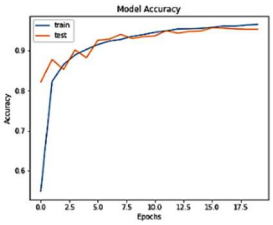
Journal of Theoretical and Applied Information Technology
15th July 2023. Vol.101. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

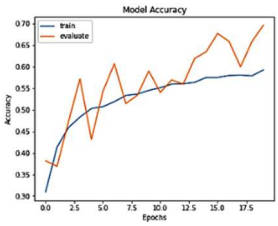


(e) AlexNet Network

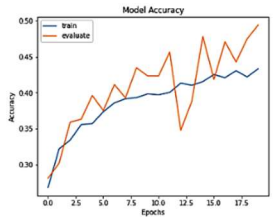
Figure 13: Model Loss Performance in the Third Study Analysis



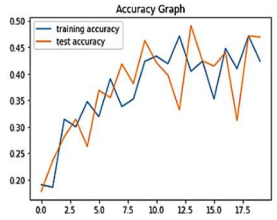
(a) Designed Network



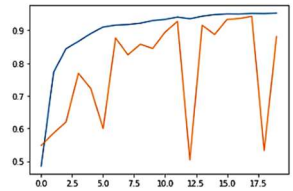
(b) VGG16 Network



(c) VGG19 Network



(d) Inception Network



(e) AlexNet Network

Figure 14: Model Accuracy Performance in the Third Study Analysis

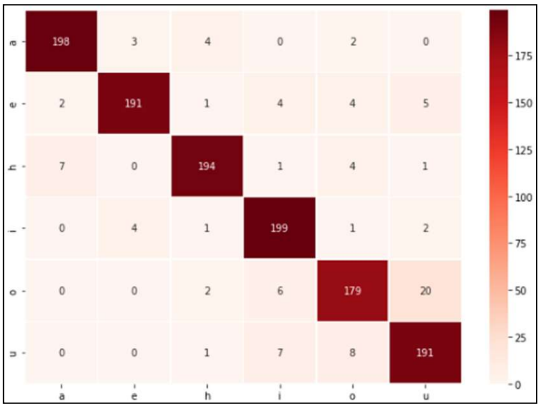


Figure 15: Designed Model Confusion Matrix in the Third Study Analysis

Table 10: Comparison of the testing for every class of vowels in the pre-trained designed model during third study.

Vowel	Precision (%)	Recall (%)	F1 (%)	Support
Class /a/	96.00	96.00	96.00	207
Class /e/	96.00	92.00	94.00	207
Class /E/	96.00	94.00	95.00	207
Class /i/	92.00	96.00	94.00	207
Class /o/	90.00	86.00	88.00	207
Class /u/	87.00	92.00	90.00	207

From the observation conducted in every study analysis, we can come with a comparative analysis for strengths and weaknesses in the experiment. From the result gained in every study analysis, we can observed that the designed model reached the highest accuracy compared to others. We would like to discuss the findings critically in form of Plus Minus Interesting Facts (PMI). The PMIs is a brainstorming and a critical thinking tool to examine the result gained for an experiment. We considered CNN method as the plus point to this study. A CNN with multiple convolution layers was used in the designed network. In the first convolutional layer, we trained 32 kernels. For images sized 35 x 200, we used a kernel size of 3 x 3. Max-pooling with kernels of 2 x 2 was performed on the convolved images.

The output images of the max-pooling layer served as input to the next layer, where 32 kernels of size 3 x 3 were used. In addition, max-pooling with kernels of size 2 x 2 was performed on the convolved images. Last convolutional layer consists of 64 kernel of size 3 x 3 and max-pooling kernels of size 2 x 2. The output of the last layer of max-pooling served as input to a fully-connected layer with 1024 nodes. The output layer of the CNN computed the soft-max activation function. The outputs of this function can be understood as the posterior probability for the classes because it transforms its inputs to positive values that sum to 1 [29].

As the minus point attached to this study, the others model network shown a low accuracy compared to the designed network. After being built, the AlexNet model consistently displayed great accuracy in all experiments after the designed model.

It is because Alexnet is a convolutional neural network that consist of 8 layers deep. VGG-16 comes after Alexnet in proving their reliability performing on the dataset. Because there is only one convolutional layer and one max-pooling layer in the layers for VGG-16, the model performed well because the extracted features and weights from the source are consistent with the target in every experiments. Comparatively speaking, the accuracy of the VGG-19 and Inception model is the lowest. The biggest drawback was the size of the network in terms of the number of parameters that needed to be trained.

Lastly, the experiments of various epochs and batch sizes of six vowels are displayed in a study by Hashim et al. [34]. Compared to them, we attached a fixed epoch and batch size as the model will train more quickly each epoch with a bigger batch size, but poor generalization will result from using more batches. The dataset is sufficient for epoch 20 during the training phase without an over-fitting graph in each research, and the number of epochs has no ideal number. This context point gave the interesting points of this study to encourage a better result and accuracy in every study analysis.

5. CONCLUSION

All the results from each of the conducted experiment, which contained images from normal people, speech disorder patients, and a mix of the normal and patients, have been reported. The newly designed model outperformed other pre-trained models in experiments achieving a performance of 94.54% based on 6 batch sizes, 20 epochs, with ADAM as the optimizer in the first study analysis. It also outperformed other comparative networks, which the designed network gained accuracy of 98.77% and 95.57% in the second study and the third study respectively although with a small dataset in the second study analysis to recognize six classes of vowels.

Overall, this proposed paper provides a comprehensive analysis using Mel-Frequency Cepstral Coefficients (MFCC) and the usage of batch size, period sizes, a variety of classes, and differences for a thorough understanding of the distinction of the vowel, especially for disorder patients. At the end of the experiment, we managed to initiate a comparative study for vowel recognition via newly proposed and collected Mel-Frequency Cepstral Coefficients (MFCC) dataset images. In addition, we have successfully constructed a designed network which can outperform other existing comparative vowel recognition methods via MFCC image-profile dataset. We expected to develop and test the more complex network model in a wide-ranging experimental setting for future tasks.

ACKNOWLEDGEMENTS

We also like to thank the Perkeso Rehab Center in Ayer Keroh, Melaka, and Universiti Teknikal Malaysia Melaka for their support in providing financing and other resources to make this endeavor for the research a success.

REFERENCES:

[1] R. D. Hayes, A. Begum, David T., "Functional Status and All-Cause Mortality in Serious Mental Illness," PLoS ONE, vol. 7, no. 9, 2012.

[2] Delafield, Colwyn, Jonathan, Trevarthen, "Theories of The Development of Human Communication," Handbook of Communication Science, 2013.

[3] Ephratt, Michal, "Linguistic, Paralinguistic and Extralinguistic Speech and Silence," Fuel and Energy Abstracts, vol. 43, pp. 2286-2307, 2011.

[4] Rebecca Treiman, Victor Broderick, Ruth Tinco, Kira Rodrigue, "Children's Phonology Awareness: Confusions between Phonemes that Differ Only in Voicing," Journal of Experimental Child Psychology, vol. 68, no. 1, pp. 3-21.

[5] Ibrahim Halil, "CERF-oriented Probe into Pronunciation: Implications for Language Leaners and Teachers," Journal of Language and Linguistic Studies, vol. 2, no. 4, pp. 420-436, 2019.

[6] N. Narasimhan, "Vowel Space Area in Speech of Children with Hearing Impairment," International Journal of Health Sciences & Research, vol. 9, no. 8, pp. 97-102, 2019.

[7] Taqi, Hanan Algharabally, Rahima, "The Realization of English Vowels by Kuwaiti Speakers," International Journal of English Linguistics, vol. 8, no. 3.

[8] N. Amir, O. Tzenker, O. Amir, J. Rosenhouse "Quantifying Vowel Characteristics in Hebrew and Arabic," Afeka Conference for Speech Processing, 2012.

[9] S. Wortman-Jutt, D. Edwards, "Poststroke Aphasia Rehabilitation: Why All Talk and No Action?," Neurorehabilitation and Neural Repair, vol. 33, no. 4, pp. 235-244, 2019.

[10] Jane Marshall, "Classification of Aphasia: Are there Benefits for Practice?," Aphasiology, pp. 408-412, vol. 24, no. 3, 2010.

[11] C. Code, "Contemporary Issues in Apraxia of Speech," Aphasiology, vol. 35, no. 4, pp. 391-396, 2021.

[12] Chris Code, "Contemporary Issues in Apraxia of Speech," Aphasiology, vol. 35, no. 4, pp. 391-396, 2021.

[13] Waber, D. P., Boiselle, E. C., Yakut, A. D., Peek, C. P., Strand, K. E., "Developmental Dyspraxia in Children With Learning Disorders: Four-Year Experience in a Referred Sample, " Journal of Child Neurology, vol. 36, no.4,pp. 210-221, 2021.

[14] Waber, Boiselle, Yakut, "Developmental Dyspraxia in Children With Learning Disorders: Four-Year Experience in a Referred Sample", Journal of Child Neurology, vol. 36, no. 3, pp. 210-221.

[15] S. Ayers, Carrie D. Liewellyn, Cambridge Handbook of Psychology, Health and Medicine: Third edition, Cambridge: Cambridge Handbooks in Psychology, 2019.

[16] C. Van Riper, Speech Correction: An Introduction to Speech Pathology and Audiology / Charles Van Riper, Robert L. Erickson. — 9th ed. p. cm. Needham Heights, MA: A Simon Schuster Company, 1995.

[17] S. P. Rosenbaum S, Speech and Language Disorders in Children: Implications for the Social Security Administration's Supplemental Security Income Program, Washington (DC): National Academies Press (US), 2016.

[18] T. P. Tan, S. S. Goh and Y. M. Khaw, "A Malay Dialect Translation and Synthesis System: Proposal and Preliminary System," International Conference on IEEE Asian Language Processing (IALP), pp. 109-112, 2012.

[19] Marini, Marco, Viganò, Mauro, Corbo, Massimo, Zettin, Marina and Simoncini, "IDEA: An Italian Dysarthric Speech Database, "Inproceedings, pp. 1086-1093, 2021.

[20] H. Ting, J. Yunus, S. Vandort, and L. Wong, "Computer-based Malay Articulation Training for Malay Plosives at Isolated, Syllable and Word Level", Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003.

[21] A. Hatzis, "Optical Logo-Therapy (OLT): Computer-based Audio-visual Feedback using Interactive Visual Displays for Speech Training, 1999.

[22] Z. M. Yusof, R. Hussain, and M. Ahmed, "Malay Speech Intelligibility Test (MSIT) for Deaf Malaysian Children," International Journal of Integrated Engineering, vol. 5, no. 3, 2014.

[23] Karita, S., Chen, Hayashi, T., Hori, T., Inaguma, H., Jiang, "A comparative Study on Transformer vs RNN in Speech Applications", IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 449-456, 2019.

[24] Hu, H., Zhao, R., Li, J., "Exploring Pre-Training with Alignments for RNN Transducer Based End-to-End Speech Recognition", ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7079-7083, 2020.

[25] Dhakal, P., Damacharla, P., Javaid, A. Y., & Devabhaktuni, V., "A near real-time automatic speaker recognition architecture for voice-based user interface", Machine Learning and Knowledge Extraction, vol. 1, no. 1, pp. 504-520, 2019.

[26] Terzopoulos, G., & Satratzemi, M., "Voice Assistants and Artificial Intelligence in Education. In Proceedings of the 9th Balkan", Conference on Informatics, pp. 1-6.

[27] Shikha Gupta, Jafreezal Jaafar, Arpit Bansal, "Feature Extraction using MFCC," An International Journal (SIPIJ), vol. 33, no. 4, pp. 101-244, 108.

[28] Shalbbya Ali, Dr. Safdar Tanweer, "Mel Frequency Cepstral Coefficient: A Review," EAI, 2021.

[29] N. Milosevic, "Introduction to Convolutional Neural Networks," Introduction to Convolutional Neural Networks, pp. 1-31, 2020.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[30] S. Hershey, Shawn Chaudhuri, "CNN Architectures for Large-scale Audio Classification," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 131–135, 2017.

[31] Akbar,Sumaiya B., Sughanthi., "Combining the Advantages of AlexNet Convolutional Deep Neural Network Optimized with Anopheles Search Algorithm based Feature Extraction and Random Forest Classifier for COVID-19 Classification," Concurrency and Computation: Practice and Experience, vol. 34, no. 15, 2022.

[32] Sinha, Amresh, "The Ideology of Inception," Film and Philosophy, vol. 21, pp. 91-112, 2017.

[33] Masyitah Abu, Nik Adilah, Amiza Amir, "A Comprehensive Performance Analysis of Transfer Learning Optimization in Visual Field Defect Classification," Multidisciplinary Digital Publishing Institute (MDPI), 2022.

[34] Hashim N. M. Z., Zahri N.A.H., Latif M.J.A., Hamzah R.A., Hashim N.F., Kamal M., Sulistiyo M.D., Kamaruddin A.I., "Analysis on Vowel /E/ in Malay Language Recognition Via Convolution Neural Network (CNN), " Journal of Theoretical and Applied Information Technology, vol. 5, no. 1301-1318, p. 100, 2022.