Baghdad Science Journal

Volume 22 | Issue 1 Article 28

2025

Vowel Recognition for Rehabilitation Assessment of Speech Disorder Patients via Multi-source Frequency Spectrum Images

Nur Syahmina Ahmad Azhar
Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Melaka Malaysia, nursyahminaazhar@gmail.com

Nik Mohd Zarifie Hashim
Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Melaka Malaysia, nikzarifie@utem.edu.my

Masrullizam Mat Ibrahim
Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Melaka Malaysia, masrullizam@utem.edu.my

Mahmud Dwi Sulistiyo
School of Computing, Telkom University, West Java, Indonesia, mahmuddwis@telkomuniversity.ac.id

How to Cite this Article

Azhar, Nur Syahmina Ahmad; Hashim, Nik Mohd Zarifie; Ibrahim, Masrullizam Mat; and Sulistiyo, Mahmud Dwi (2025) "Vowel Recognition for Rehabilitation Assessment of Speech Disorder Patients via Multi-source Frequency Spectrum Images," *Baghdad Science Journal*: Vol. 22: Iss. 1, Article 28.
DOI: 10.21123/bsj.2024.9202
Available at: https://bsj.researchcommons.org/home/vol22/iss1/28

BAGHDAD SCIENCE JOURNAL 2025;22(1):319–334

RESEARCH ARTICLE

Vowel Recognition for Rehabilitation Assessment of Speech Disorder Patients via Multi-source Frequency Spectrum Images

Nur Syahmina Ahmad Azhar [1,*], **Nik Mohd Zarifie Hashim** [1], **Masrullizam Mat Ibrahim** [1], **Mahmud Dwi Sulistiyo** [2]

[1] Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Melaka Malaysia
[2] School of Computing, Telkom University, West Java, Indonesia

ABSTRACT

Communication impairments have a broad spectrum of medical causes, such as speech disorders, hearing loss, brain injury, stroke, and physical impairments. As a result, communication disorders can affect social development and interpersonal relationships. Speech impairments can benefit from early speech treatments; however, the majority of rehab facilities across the world still carry out this process manually. A wide range of studies has been conducted on speech processing for various human languages. Machine learning and deep learning have been applied to the medical and healthcare industry to enhance rehabilitation by utilizing the new technology. This study analyzed the classification accuracy of the designed network and other pre-trained models (VGG-Net, AlexNet, and Inception) and performed a complete comparative analysis to assess the classification accuracy of several pre-trained models. The sound is converted to the image as a new way to see them in the neural network via a newly proposed concept named image-profiled data. These image-profiled datasets that used a spectrogram and a Mel-frequency cepstral coefficient (MFCC) produced this study's best results and accuracy. This project aims to develop a new neural network that can successfully distinguish between the vowels from the voices of normal people, patients with speech disorders and the mix from the prior two groups using the six and twelve classes of Malay vowels. The designed network model, which used 6 batch sizes, 20 epochs, and ADAM as the optimizer, this study presented and achieved the maximum accuracy values of both classes for image-profiled audio data in analyses conducted.

Keywords: Convolutional neural network (CNN), Deep learning, Mel-frequency cepstral coefficient (MFCC), Rehabilitation, Spectrogram, Vowel recognition

Introduction

Communication is essential for every human in every possible way. If the disabled person can regain ways to communicate with their family, this may help to re-establish emotional bonds and support roles, and it may help to prevent frustration. A man who used to be self-sufficient may become enraged when he needs help and has difficulty requesting it for simple tasks. Some of the frustration might be relieved by introducing efficient communication strategies. Sound production is necessary, and the form of words will be produced to communicate. Sounds are produced by the vibration of the vocal cords in a human's mouth. There are numerous of study and research on the sound's production and vibration. Ladefoged and Johnstone's research [1] in 2015 is among the works that proposed the acoustic phonetics and formant frequencies of humans. The cavities above the larynx produce a sound, and these cavities resonate with specific frequencies, and some do not resonate. The frequencies that resonate are called formant frequencies.

Received 8 June 2023; revised 2 February 2024; accepted 4 February 2024.
Available online 1 January 2025

* Corresponding author.
E-mail addresses: nursyahminaazhar@gmail.com (N. S. A. Azhar), nikzarifie@utem.edu.my (N. M. Z. Hashim), masrullizam@utem.edu.my (M. M. Ibrahim), mahmuddwis@telkomuniversity.ac.id (M. D. Sulistiyo).

https://doi.org/10.21123/bsj.2024.9202

Formants or formant frequencies are the peaks or maximal points in the sound spectrum. Spectrum is the acoustic component that can identify a complex sound wave.[2] Thus, this study will use two types of visual representation of speech: Spectrogram and Mel-Frequency Cepstral Coefficient (MFCC). The aids of visual representation, or the image-profiled, are to analyse and interpret the dataset images to understand sounds with objective numbers. Vowels are letters that come in for speech sounds in which the air leaves the mouth unimpeded by the tongue, lips, or throat.[3] Speech sounds can be classified into two categories: those where the air is blocked by the lips, tongue, or throat before leaving the mouth or the air is not secured. Vowels signify unblocked sounds, while consonants represent blocked sounds. Crystal *et al.* state three formant frequencies, and every vowel, which is categorized into front, central, and back of the human's mouth, has its frequencies.[4]

Aphasia is a disorder resulting from damage to portions of the brain responsible for language for most people. These areas are on the left side of the brain.[5] Aphasia usually occurs suddenly, often following a stroke or a head injury, but it may also develop slowly due to a brain tumour or progressive neurological disease. It's, therefore, essential to note that aphasia is not a disease in itself but rather a constellation of symptoms concerning difficulties with expressing or comprehending language that could be due to multiple different underlying causes. Even though speech-language pathologists cover a wide range of treatments, it is required to have two separate areas for someone with a speech problem to diagnose.[6]

One is called apraxia of speech, and the other is called dysarthria. According to one definition, dysarthria is a group of speech disorders caused by problems in the muscles that control speech as one or more of the motor systems essential to speech production are impaired.[7] According to this definition, "dysarthria" only refers to speech disorders with a neurogenic origin. A motoric deficit, or a fundamental disturbance of movement, of the muscles in the speech production process results in dysarthria.[8] Apraxia of speech is a motor programming disorder characterized primarily by articulatory disturbances with associated compensatory prosodic disturbances.[9] Unlike dysarthria, apraxia of speech is not related to significant slowness, weakness, incoordination, or paralysis of the muscles of the speech production mechanism.

A spectrogram is a three-dimensional visual representation of speech. It shows the time horizontally while the frequency is vertical. Initially, it is two-dimensional, but the added dimensions come with the intensity of the information in the spectrogram.[10] The third dimension, represented by colors, depicts the signal strength (or loudness) for a given frequency at a given point. The color scheme may vary in each spectrogram, but the essential idea behind each remains constant.[11] MFCC is widely used in speech recognition as an aid visualization. It is popularly extracted from speech signals for use in recognition tasks. MFCC can represent the filler (vocal tract) in the source-filter speech model.[12] The frequency response of the vocal tract is relatively smooth, whereas the source of voiced speech can be modeled as an impulse train. MFCC uses a log, which transforms convolution between the excitation source and vocal tract (filter) into addition in the spectral domain.[13] The 'cepstral' in the MFCC is the adjective of the spectrum.

The conversion from the wav file audio to the spectrogram image profile is pre-processed by using the LibROSA library in Python. During this step, the signal was downsampled from 44100 Hz to 16000 Hz sampling rate. The length of the window used is 1024 samples, and after t, it was converted into a time-frequency by using short-time Fourier transform (STFT), and this method is calculated using the fast Fourier transform (FFT).[14] The proposed paper will focus on vowel recognition for normal persons and stroke patients with speech defects. Two groups of normal people and disorder patients were used to acquire the sound recording data. A few studies have been done so far to classify the pronunciation of Thai, Japanese, and Arabic vowels using an intelligent method. Therefore, this research aims to design and develop a system using a deep learning structure for Malay vowel speech recognition. The experiment in this study will be carried out over a network and will use five models of networks. This system was created to:

1. Address the issue of pronunciation problems in speech disorders, especially stroke patients.
2. Address the issue of the lack of specialists in teaching vowel pronunciation and rehabilitation.
3. Address the issue of the actual process, which is difficult and time-consuming and needs to provide results in real time.
4. Introduce an innovative solution to address the issue of practicing Malay vowel pronunciation for non-native speakers of the language or non-standard speakers.

This study uses a convolution neural network (CNN) is used to study vowel identification in the Malay language, especially for the speech disorder group. In place of the CNN customary sound file, the technique vowels are in the six and twelve groups.

The voices collected from different aspects in real-life circumstances, such as gender, age, accent, environment, and noise, made up the dataset utilised to train this model. The following are the main contributions to this work:

1. Initiated a comparative study for vowel recognition in two classes of vowels of six and twelve classes by using five network models network.
2. Compare the performance accuracy of two image-profiled spectrograms and MFCC between normal person and speech disorder patients for Malay vowel classification tasks.

Related works

Speech is the production of individual sounds that are put together to form words, while language is a combination or system used to convey ideas and understand others. Speech and language disorders may occur separately or together. It's important to note that these symptoms may be mistakenly identified as poor listening for attention, selective hearing or bad behavior. Still, people with speech difficulties often express their frustration by communicating through physical aggression or other disruptive behavior that may lead to poor relationships with peers, siblings or reluctance. The ability to learn sometimes worsens the patient's condition, which could indicate difficulties requiring professional evaluation and intervention. Early intervention has been shown to have a significant impact on speech development. It improves the patient's ability to communicate and interact with others and improves social and emotional development.

These days, numerous rehabilitation centers are available to help people overcome speech impairments. Developing the most effective ways to interact is the aim of rehabilitation. Speech rehabilitation is important because it helps people who struggle to speak to improve communication and remove communication barriers. Speaking more clearly, building up the speech muscles, and learning proper pronunciation are all objectives of speech therapy. There is a lot of research and methodology used in the field of speech therapy and rehabilitation as technology advances. Most therapy plans focus on particular features of developmental speech rehabilitation and employ treatment strategies to address them.

Research on Thai Vowel pronunciation recognition that uses deep learning methods has developed an automatic computer-assisted pronunciation training called CAPT. The automatic CAPT was created to address the need for qualified instruction professionals and the challenging vowel teaching methodology. The new dataset was collected, designed and examined to develop the classification of Thai vowels. By using spectrogram and MFCC image-profiled and CNN architecture with a deep learning model, they achieved the highest accuracy of spectrogram images with 98.61%, while the MFCC images were 94.44%. This study also conducted the MFCC with the baseline long short-term memory (LTSM) model with an accuracy of 94.44% and spectrogram LTSM with 90.00%.[15]

Vowel recognition is applied to all worldwide vowels. An approach recognition for Arabic phonemes also uses a similar deep learning approach. For this study, the crucial part of the experiments is the pronunciation of Arabic phonemes. Incorrect pronunciation of Arabic short vowels can completely alter a sentence's meaning. For this reason, both students and teachers of Classical Arabic (CA) must practice more while correcting their students' pronunciation of Arabic short vowels. This study has developed a model that can classify Arabic vowels using Deep Neural Networks (DNN). Similar to this study, they created and designed a new audio Arabic dataset, developed neural network architecture and achieved the best classification accuracy. Their proposed model has reached a testing accuracy of 95.77%.[16]

Javanese language vowels are unique to pronounce as they have vowels, semi-vowels and consonants. It can complicate new learners and beginners in pronouncing the language. This study aims to improve the model's accuracy by using weight initialization methods and weight functions. They have used three weight initialization and activation functions inside the CNN model. This study used MFCC and MFSC features to conduct a multinomial logistic regression model. They have proved that the optimal CNN model could be achieved by combining Xavier weight initialization and ReLU activation.[17] The classification accuracy gained from this study is 99.60%.

Materials and methods

This study's proposed method includes improving training and validation accuracy. A large collection of images, including images of normal people and speech disorder patients, is required to construct a new intelligent classification system for speech disorders of stroke patients. The recording, converting, and cropping processes produce data for normal people and stroke patients. A convolutional neural network will then be trained using the entire dataset.

Data collection

The steps for recording, gathering, and preparing the dataset are described in this section. To ensure the data's accuracy and quality, every action taken to prepare the dataset was highly crucial. For this study goal, there aren't any publicly accessible datasets for Malay vowels. Therefore, the dataset preparation was created for this study's purpose. The dataset was gathered from two groups: the normal person group and the stroke patients group with a speech impairment. All tests were run on this dataset. For the normal and healthy people, the speech dataset was collected from Malay speakers who speak the standard Malay dialect (10 males and 10 females between 18 and 27). The speech data was then gathered from 9 stroke patients from Perkeso rehab Ayer Keroh in Melaka, Malaysia, who also spoke the standard Malay dialect (6 male and 3 female).

The dataset contains 44,100 Hz of standard speech data recorded from a REMAX RP1 8GB Digital Audio voice recorder. Everyone records similarly with a 15 cm space between their mouths and the voice recorder. For the normal person group, 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. This research has divided the recording time into three periods for the normal person group to come up with the patients' voice pronunciation capabilities. Stroke patients may have difficulties pronouncing some vowel classes, so the recording for the normal person group is conducted in three periods to come up with the condition where patients with speech disorders pronounce the vowel for one to three seconds.

This study aims to evaluate how expanding the classification group of the dataset affects performance accuracy. There are six classes of vowels and twelve classes of vowels in this study's classification. Every vowel class in the first six classes combines male and female individuals; however, the next 12 classes separate male from female members. This study wants to observe whether adding more classification groups would improve performance or whether combining males and females would alter training performance. The distribution of the dataset collected in the studies is displayed in Table 1.

Table 1. Distribution of group vowel classes from collected datasets.

Quantity of classes	Class of vowels
6 class	/a/ (Male + Female)
	/e/ (Male + Female)
	/E/ (Male + Female)
	/i/ (Male + Female)
	/o/ (Male + Female)
	/u/ (Male + Female)
12 class	/a/ (Male)
	/e/ (Male)
	/E/ (Male)
	/i/ (Male)
	/o/ (Male)
	/u/ (Male)
	/a/ (Female)
	/e/ (Female)
	/E/ (Female)
	/i/ (Female)
	/o/ (Female)
	/u/ (Female)

Spectrogram

Instead of a real-time image, this study employed a spectrogram image profile since a spectrogram is a graphic depiction of a signal's strength over time at various frequencies that make up a waveform. The vertical axis of a spectrogram shows data depending on frequency. The lowest frequency is shown at the bottom, and the highest is displayed at the top.[18] Fig. 1 shows the spectrogram images for two groups, normal persons and speech disorder patients. Each vowel class in each group sample is represented in Fig. 1. Vowel classes are listed from left to right in the following order: /a/, /e/, /E/, /i/, /o/, /u/. After the conversion from the audio to images is complete, cropping will begin with each vowel's image, which has a size of 240 by 55 pixels and a bit depth of 32 bits.

Mel frequency cepstral coefficient (MFCC)

For use in tasks requiring recognition, MFCC are frequently derived features from voice signals. MFCCs are regarded as a representation of the filter (vocal tract) in the source-filter model of speech. During the conversions from audio to MFCC, the default acoustic features used are 20 Mel bands. The features of the speech sample are extracted after some basic processing. Pre-emphasis, frame blocking, and windowing are the three feature extraction stages used in MFCC.[19] The pre-emphasis speech signal x(n) must go via a high-pass filter. In the equation, the output signal is represented by the symbol y(n), and the value of "a" is typically between 0.9 and 1.0. Fig. 2 displays the cropped MFCC images of each class of normal person and speech disorder group. Each image has a size of 200 by 35 pixels and a bit depth of 32 bits. To ensure that essential information was retained during the conversion process, careful attention was given to the image quality and resolution of both spectrogram and MFCC images. The resolution was set at a level

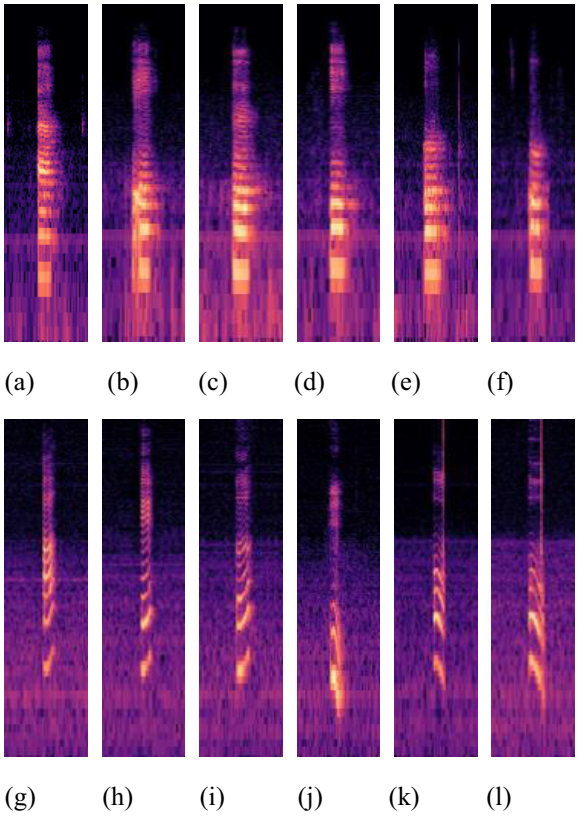

Fig. 1. Spectrogram images for vowels (a)-(f) for vowel /a/, /e/, /E/, /i/, /o/ and /u/ for normal person and vowels (g)-(l) for vowel /a/, /e/, /E/, /i/, /o/ and /u/ for disordered patient.

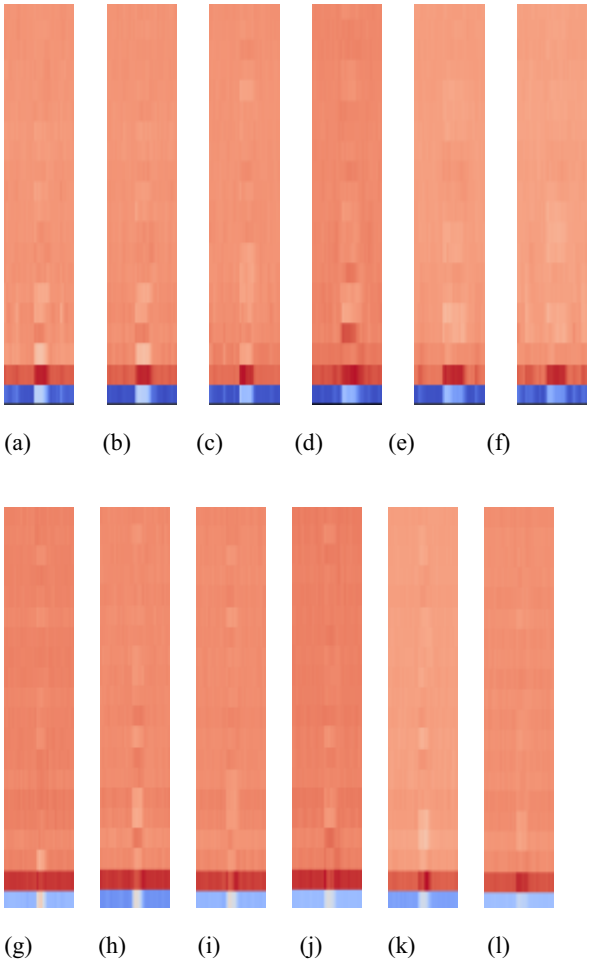

Fig. 2. MFCC images for vowels (a)-(f) for vowel /a/, /e/, /E/, /i/, /o/ and /u/ for normal person and vowels (g)-(l) for vowel /a/, /e/, /E/, /i/, /o/ and /u/ for speech disorder patient.

that allowed for clear visualization of important features while avoiding excessive noise or distortion.

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

The reasoning behind selecting specific image sizes for spectrogram (240 × 55) and MFCC (200 × 35) is based on the trade-off between resolution and computational efficiency. The selected sizes balance memory and processing power consumption while capturing sufficient frequency and time information, and are effective in various audio analysis applications.

Convolutional nueral networks

The Convolutional Neural Network (CNN) is a widely used model that comprises one or more convolutional layers, which can be pooled or fully linked, and is based on a variant of multilayer perceptrons. Further, these convolutional layers create feature maps that record a region of the image, which is ultimately broken into rectangles and sent out for nonlinear processing. The advantages of CNN are they offer very high accuracy in image recognition problems and are capable of automatically detecting important features without any human supervision compared to the ANN and RNN.[20] ANN has no specific rule for determining the structure of artificial neural networks, and appropriate network structure is achieved through experience and trial and error. Besides, the RNN model can't process very long sequences if using real as an activation function, and the training process of RNN is challenging. In conclusion, CNN is considered more potent than other models as it has high accuracy in image recognition problems and weight sharing.[21]

The structure of the designed network is described in this section. To address the vowel recognition problem, this study developed a designed network model. A CNN with an input image size of 240 × 55 and 200 × 35 was used to build the model, along with an ADAM classifier as the optimizer and SoftMax as the activation function. CNNs use SoftMax as the activation function and the ADAM classifier as the

optimizer. The final layer transforms the output into a probability distribution across classes, while the ADAM optimizer minimizes the loss function. These decisions improve CNN's performance and accuracy in image classification tasks. The CNN architecture includes activation functions, pooling layers, fully connected layers, dropout layers, and convolutional layers. Techniques for padding feature maps maintain size while boosting efficiency.

CNN architectures use convolutional layers to extract features, pooling layers to reduce spectral variability, and dropout layers to mitigate overfitting. A fully connected layer predicts image class using convolution output. CNN hyperparameters are established through a methodical process of testing and analysis. A literature review and past knowledge are used to identify a range of values for each hyperparameter, followed by a grid search or random search strategy to investigate various combinations.

The CNN model's performance was evaluated using metrics like accuracy or loss function, and the hyperparameters with the best results were chosen. Cross-validation was used to refine these hyperparameters, ensuring they were stable and not overfit to specific data subsets. Some analysis was conducted to examine how different hyperparameters affected the model's performance and enable further optimization if needed. The final hyperparameters were then selected based on this analysis.

The convolutional layers applied convolutional filters to the input before a nonlinear activation function, with Convolution Layer 1 being the first. Each layer had a kernel of size three, batch sizes of 32 and 64, and two distinct dropout values. The CNN model was created through a trial-and-error learning process. It was kept up to date by gradually raising testing accuracy, which increased testing accuracy to 90.0%.[22] This study presented a CNN model with six convolutional layers, three max-pooling layers, one flattened layer, and two fully connected layers as the optimum CNN model for Malay vowels. The model's specifications are as follows:

- The first convolutional layer consists of 32 filters (3 × 3), a relu activation function and considers the image is 238 by 53 pixels in size.
- The second convolutional layer consists of 32 filters (3 × 3), a relu activation function, max-pooling (2 × 2) for the images of 118 × 25, and dropout (0.25).
- The third convolutional layer consists of 32 filters (3 × 3), a relu activation function an image resolution of 116 × 23.
- The fourth convolutional layer consists of 32 filters (3 × 3), a relu activation function, max-pooling (2 × 2) size of 57 × 10, and dropout (0.25).
- The fifth convolutional layer consists of 64 filters (3 × 3), a relu activation function and recognizes that the image is 55 by 8 pixels.
- The sixth convolutional layer consists of 64 filters (3 × 3), a relu activation function, max-pooling (2 × 2) layer size is 26 by 3, and dropout (0.25).
- Finally, to create a dense layer, 1024 units of a dense layer are combined with 6 units of a dense SoftMax layer.

Table 2. 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.

The architecture of CNN was selected with simplicity and complexity in mind. Larger accuracy may be possible with more complicated models, but these models frequently have larger processing costs and are more susceptible to overfitting. Conversely, more capacity could be required for simpler models to fully capture complex patterns in the data. The complexity of the dataset and the intended level of abstraction served as the main factors in deciding the number of convolutional layers, pooling layers, etc. The CNN architecture was chosen after taking into account the dataset's size and the processing resources that were available. The selected design attempts to create a good trade-off between accuracy and processing speed by finding a balance between simplicity and complexity.

When it comes to learning rates, the training procedure usually includes the application of adaptive learning rate algorithms, like Adam or RMSprop. For this study, a learning rate of 0.0001 was employed. These methods dynamically adjust the learning rate during training to optimize convergence. As for loss functions, it depends on the specific task at hand. Common choices include mean squared error (MSE) for regression problems and categorical cross-entropy for classification tasks. Early stopping criteria are used to prevent over-fitting and involve monitoring a validation metric such as accuracy or loss. This study didn't use the early stopping class to prevent overfitting. Instead, the study relied on a small number of epochs to ensure that the model didn't train for too long and potentially over-fit the data. This approach helps strike a balance between training the model adequately and avoiding over-fitting.

Performance evaluation

A confusion matrix helps the user see the many outcomes of a categorization problem by displaying a table structure of the findings and predictions. Confusion matrices are useful because they allow for direct comparisons of variables like "True Positives," "False Positives," "True Negatives," and "False Negatives."[23] When the convolutional neural network (CNN) tests a particular data set, the confusion matrix is utilised to identify which classes the CNN properly predicts and which do not. Confusion matrices in the equations below describe crucial predictive characteristics, including recall, specificity, accuracy, and precision.

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

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

$$Recall = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}}{K} \quad (4)$$

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

In addition to accuracy, this study used precision, recall, and F1 scores as performance metrics to evaluate the model. Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. Recall calculates the proportion of correctly predicted positive instances out of all actual positive instances. The F1 score is the harmonic mean of precision and recall, providing a balanced measure between the two metrics. These additional metrics help us gain a comprehensive understanding of the model's performance beyond just accuracy.

Results and discussion

In this section, the performance of the proposed CNN will be compared to the other current network models with a confusion matrix to measure the performance accuracy in all analysis studies. The results were split into two analytical studies: (a) the comparison of pre-trained models by using spectrogram image-profiled; and (b) the comparison of pre-trained models by using MFCC image-profiled. A ratio of 80:10:10 was used to divide the vowel dataset into training, evaluation, and testing for the three analysis experiments in (a), and (b). This proposed study aims to validate the efficacy of pre-trained models (Designed model, VGG-16, VGG-19, AlexNet, and Inception) that employ images of 55 × 240 of spectrogram and 35 × 200 of MFCC dimensions and 32-bit depth. The Python approach was developed with the KERAS (Tensorflow) neural network computation framework.

The comparison of pre-trained models by using spectrogram image-profiled

The first study analysis includes 20 normal persons, ten male and ten female. A total of 10800 spectrogram images were collected to conduct this experiment. The second analysis consists of unequal speakers of speech disorder patients with a total of 1620 and 1080 total images for six and twelve classes, respectively. The last and third analyses combine normal person and speech disorder patient datasets called mixed analysis. A total of 12420 for six classes and 11880 and twelve classes for this analysis have been collected. The epoch and batch size were set to 20 and 6, respectively.

The dataset for the normal person group using spectrogram and MFCC image profile is balanced with the same amount of images that were collected. Similar to the post-stroke patient's group dataset, the two image profile contains a balanced and the same amount of data. However, the comparison of images between the groups is unbalanced due to the constraints that exist for the post-stroke patient to pronounce the vowel class during the recording session. So, due to this condition, the amount that can be collected in the post-stroke patient group is less compared to the normal group. However, an initiative has been made to progress the project where the dataset for the group of normal people is varied so that the system can learn more about a vowel class from the dataset of normal people and post-stroke patients.

Five different network models (Designed model, VGG16, VGG19, Inception, and AlexNet model) are used to apply for a comprehensive comparison of one to the other. For each of these networks, this study explained the experiment's findings. For a batch size of 6 and an epoch of 20, the designed network model in Fig. 3 and Fig. 4 achieves a classification accuracy of 92.96% and 93.70% for 6 class and 12 class of vowels, respectively. The 12 classes of designed model have higher validation accuracy than the 6 classes. Table 4 shows the result of the training and validation accuracy for both classes using five different networks.

Batch size six and epoch 20 had the highest accuracy in 2022, only 81%, in training and testing performed by Hashim *et al.*[24] and other network models for the same environment. Out of the six model

Table 3. Total of spectrogram image profile collected in every class of vowels and analysis.

Image profile	Analysis	Classes of vowel	Assessment	Total images in a group	Total images collected
Spectrogram	Normal	6 Class	Training (80%)	1440	8640
			Evaluation (10%)	180	1080
			Testing (10%)	180	1080
		Total Spectrogram Image Profile Collected			10800
		12 Class	Training (80%)	720	8640
			Evaluation (10%)	90	1080
			Testing (10%)	90	1080
		Total Spectrogram Image Profile Collected			10800
	Patient	6 Class	Training (80%)	216	1296
			Evaluation (10%)	27	162
			Testing (10%)	27	162
		Total Spectrogram Image Profile Collected			1620
		12 Class	Training (80%)	72	864
			Evaluation (10%)	9	108
			Testing (10%)	9	108
		Total Spectrogram Image Profile Collected			1080
	Mix	6 Class	Training (80%)	1656	9936
			Evaluation (10%)	207	1242
			Testing (10%)	207	1242
		Total Spectrogram Image Profile Collected			12420
		12 Class	Training (80%)	792	9504
			Evaluation (10%)	99	1188
			Testing (10%)	99	1188
		Total Spectrogram Image Profile Collected			11880

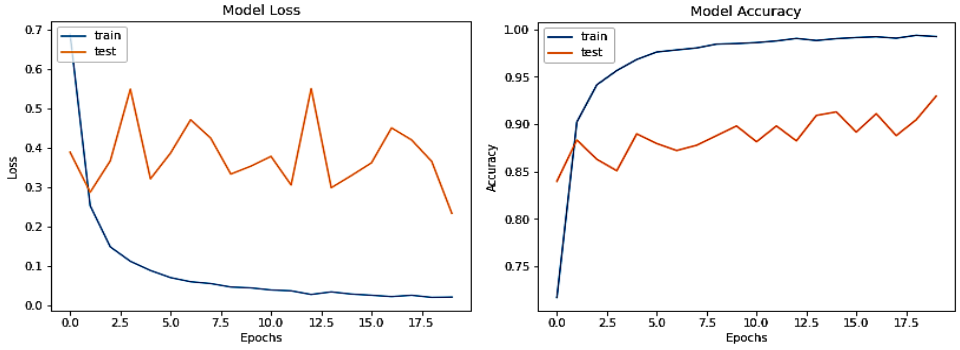

Fig. 3. Model loss and model accuracy of spectrogram image-profiled for 6 class of normal person.

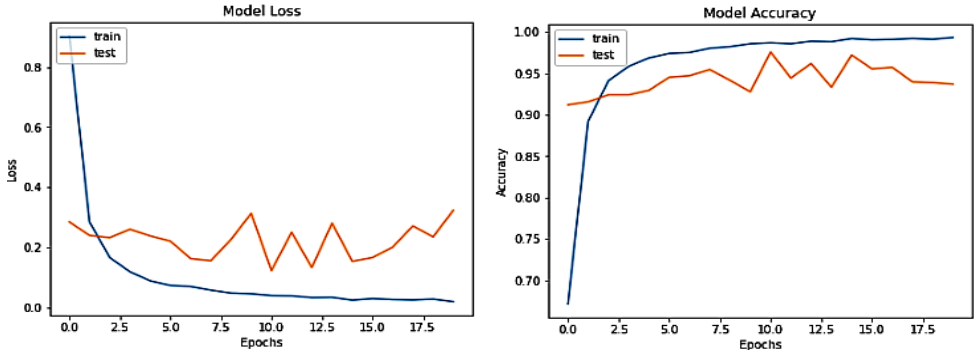

Fig. 4. Model loss and model accuracy of spectrogram image-profiled for 12 class of normal person.

Table 4. Results of multiple models accuracy by using spectrogram image profile.

Analysis	Class of vowels	Model	Epoch	Batch size	Training accuracy (%)	Validation accuracy (%)
Normal	6 Class	Designed	20	6	**99.95**	**92.96**
	12 Class				**99.78**	**93.70**
	6 Class	VGG16			87.88	68.52
	12 Class				74.55	74.54
	6 Class	VGG19			76.30	56.02
	12 Class				59.62	56.30
	6 Class	Inception			33.70	32.30
	12 Class				23.78	28.79
	6 Class	AlexNet			94.17	87.11
	12 Class				92.67	84.01
Patient	6 Class	Designed			**99.77**	**90.12**
	12 Class				**100.00**	**96.30**
	6 Class	VGG16			93.60	69.75
	12 Class				88.66	91.67
	6 Class	VGG19			80.71	64.81
	12 Class				68.63	65.74
	6 Class	Inception			41.37	33.83
	12 Class				45.97	45.37
	6 Class	AlexNet			90.33	76.76
	12 Class				79.90	60.13
Mix	6 Class	Designed			**99.97**	**88.33**
	12 Class				**99.96**	**95.37**
	6 Class	VGG16			84.33	65.22
	12 Class				75.95	76.43
	6 Class	VGG19			72.93	57.49
	12 Class				54.85	53.20
	6 Class	Inception			31.20	30.19
	12 Class				20.67	23.89
	6 Class	AlexNet			93.09	84.77
	12 Class				79.60	60.98

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%

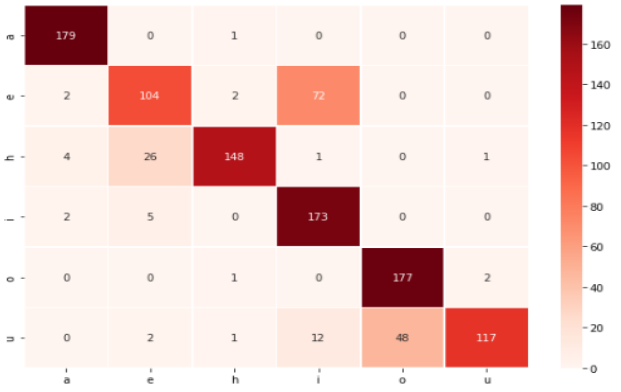

Fig. 5. Confusion matrix of spectrogram image-profiled for 6 class of normal person.

networks examined, this study could execute the maximum validation accuracy for epoch 20 and batch size 6 with the help of the newly proposed model in this proposed research, with 94.54%. This study has successfully improved upon Hashim's proposed network with a more accurate network for this study investigation.

Fig. 3 and Fig. 4 shows the line graphs of the model's accuracy and loss of the designed CNN models by using spectrogram image-profiled for 6 classes and 12 class of normal person. The visualization displays a line graph from 0 to 20 epochs that compares the accuracy and loss of the training and validation models. The designed model with 12 classes of vowels outperformed the designed model with 6 classes of vowels and achieved the best accuracy of 93.70% as shown in Fig. 4. Overall, the classification accuracy results for the designed network gained above 85% in normal, patients, and mix analysis. Compared to the designed network of other pre-trained models, it reached the highest accuracy in all performances conducted.

This study set the batch size and epoch size to be similar in the classification studies for the normal person, disorder patients, and mixed group (normal person + disorder patient). It is to investigate the

Table 6. Comparison of the testing for 6 class of vowels in the pre-trained designed model.

Classes of vowels	Analysis	Vowel	Precision (%)	Recall (%)	F1 (%)	Support
6 Class	Normal	Class /a/	96.00	99.00	98.00	180
		Class /e/	76.00	58.00	66.00	180
		Class /E/	97.00	82.00	89.00	180
		Class /i/	67.00	96.00	79.00	180
		Class /o/	79.00	98.00	87.00	180
		Class /u/	97.00	65.00	78.00	180
	Patient	Class /a/	100.00	100.00	100.00	27
		Class /e/	100.00	89.00	94.00	27
		Class /E/	93.00	100.00	96.00	27
		Class /i/	93.00	100.00	96.00	27
		Class /o/	78.00	93.00	85.00	27
		Class /u/	95.00	74.00	83.00	27
	Mix	Class /a/	90.00	99.00	94.00	207
		Class /e/	66.00	83.00	73.00	207
		Class /E/	98.00	57.00	72.00	207
		Class /i/	94.00	95.00	89.00	207
		Class /o/	69.00	98.00	91.00	207
		Class /u/	99.00	53.00	69.00	207

effectiveness of small epoch size, which is reliable to the actual application. In Fig. 5, the graph shows a stable model accuracy and model loss performance in all experiments conducted. Epoch has no ideal number, and this study has proved that the training phase for six classes and twelve classes is sufficient for epoch 20 without having an over-fitting graph in each experiment. The model will train faster per epoch with bigger batch size, but poor generalization will result from using more batches.[25] To mitigate overfitting in the model, this study implemented several techniques in addition to using dropout layers. Firstly, this study employed data augmentation, which involved generating additional training examples by applying random transformations such as arrangement, shifting, and shuffling to the existing data. This helped in increasing the diversity of the training set and reducing overfitting. Six vowels were tested in various ways using different epochs and batch sizes, according to a study by Hashim *et al.*[24] The purpose of Hashim's study is to employ a different number of epochs and batch sizes, and the graphic shows the training procedure when the model learned about the training and validation datasets.

From the performance result above, the classification performance for six vowel classes is less than 12 classes for the designed and VGG networks. Based on the observation, this study may infer that increasing the training groups of the dataset will improve the classification performance and accuracy. Besides that, the six classes of AlexNet and Inception network perform better than the 12 classes, as the larger the training dataset size, the higher the accuracy. In the analysis that consists dataset of speech disorder patients, the validation accuracy for six classes of vowels of the designed network was lower than the analysis for the normal person shown in Table 2. Due to their difficulties pronouncing vowels throughout the recording procedure, speech disorder patients' validation accuracy is lower than normal people. Some of them struggle with certain vowel groups. Vowels with similar sounds, like /e/ and /i/, might make it challenging to train data sets. However, the mixed analysis, which combines the dataset from patients with speech disorders and normal people, reveals a rise for 12 classes of vowels.

Table 6 displays the CNN model's precision, recall, and F1 scores for categorising each vowel in the six classes of vowels. Throughout experiments with all of the analyses and six classes of vowels utilizing spectrogram image profile, the dataset's lowest F1 score was obtained by the normal person analysis class /e/, with a score of 66%. Conversely, class /a/ has the highest F1 score, 98% and 100% for normal person analysis and speech disorder patients analysis, respectively. For the 12 classes of vowels, the confusion matrix can be evaluated using the F1 score findings. The lowest score gained by /e/ class of normal person analysis with 21% while the highest score on the dataset is 100% for /o/ class. Table 6 shows the results of precision, recall and F1 scores of the designed CNN model for six classes of vowels.

The designed model's confusion matrix is displayed in Fig. 5 and Fig. 6 for the error analysis. The confusion matrix to the /e/ vowel class shows the most miss-prediction. Vowel pairs with the sounds /e/, /h/, and /u/ are the most imprecise in the confusion matrix. These sounds are comparable, which medical theory explains because they have

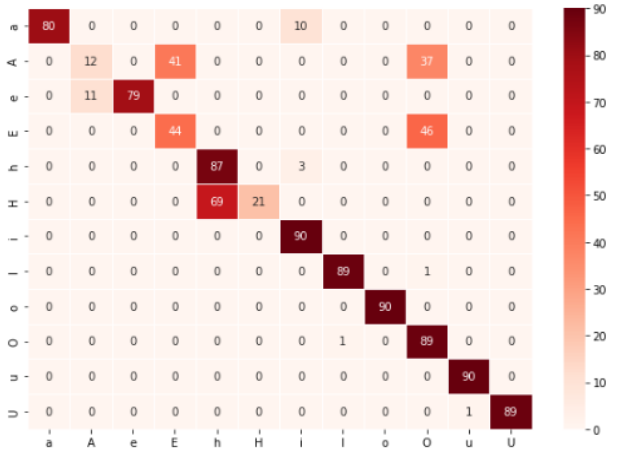

Fig. 6. Confusion matrix of spectrogram image-profiled for 12 group of normal person.

similar characteristics. In turn, the system's model recognition could be more precise. Dark red tones symbolize the accurate prediction, whereas light red tones symbolize the inaccurate prediction. [26]

The comparison of pre-trained models by using MFCC image-profiled

The MFCC image profiled has gathered 10800, 1620, and 12420 for normal, patient, and mixed analyses, respectively, in the six classes of vowels, which is an equivalent amount of data to the spectrogram. The twelve classes also carry out the same assessment in an 80:10:10 ratio of training, validation, and testing.

For the experiment conducted by using MFCC image-profiled, the designed model's accuracy gained the highest for all the analyses, coming in at 94.54%, 98.77%, and 98.00%, respectively, for normal, patient, and mixed analysis of six classes of vowels. The twelve classes of the designed model in all analyses also gained performance accuracy above 95%. Compared to the spectrogram accuracy, the designed model and other pre-trained model results show higher accuracy in all analyses. AlexNet and VGG16 models come in after the designed model with an average performance accuracy, while the accuracy of the Inception model, which was used in all the studies, is the lowest. [27] Fig. 7 and Fig. 8 shows the model loss and model accuracy for the designed model of 6 and 12 classes of vowels.

Most of the neural network models depend on hyper-parameter, which makes VGG16 a unique model since instead of having a large number of hyper-parameter, they focus on having a uniform structure. The 16 in VGG refers to the 16 layers that have weights. This network is extensive and has about 138 million (approx) parameters. VGG network usually takes a lot of time to train. It is because the network architecture is large, and due to this reason, smaller network architectures are more desirable to learn. Refers to the result gained in this study, VGG16 achieved an average result compared to the designed network model for all the experiments conducted. VGG19 is more complex than VGG16 and performs worse in accuracy and loss.

VGG19 network trains on more than a million images from the Imagenet database. It is an image database of 14 million images organized according to the Wordnet hierarchy. This network consists of 19 layers deep and can classify images into thousand object categories such as animals or objects. As a result, this network has learned rich feature representation for a wide range of images. This study can observe that AlexNet gained higher validation accuracy during the experiments. It is because AlexNet uses a stochastic gradient descent optimization function with a batch size of 128, momentum of 0.9, and weight decay of 0.0005. The learning rate of all the layers in AlexNet is 0.001.

The precision, recall, and F1 scores for the CNN model's classification of each vowel for MFCC image-profiled in each of the six classes of vowels are shown in Table 4. The dataset's lowest F1 score was attained by the normal person analysis class /a/ of normal person analysis, with a score of 98% while the lowest score is achieved by class /u/ for speech disorder patient analysis with 68%. Overall, class /a/ has the highest F1 score in all analysis. Fig. 9 and Fig. 10 shows the confusion matrix for six group of normal person analysis by using MFCC image-profile.

Conclusion

Numerous studies use machine and deep learning to detect pronunciation and recognise speech in various language datasets. Therefore, this study introduced CNN architecture of Malay vowel recognition. From the experiments conducted, the designed network achieved the highest classification accuracy in both classes and all analysis compared to the other pre-trained models. This study has constructed a new, more accurate model than the existing models. Additionally, with the aid of spectrogram and MFCC image-profiled, the vowels' recognition process has reached the purpose of clarifying between two gendered groups. Statistical results were presented with a mixture of analysis of normal persons and speech disorder patients. From here, this study can observe that the classification of 12 classes of vowels reached the highest accuracy of the six classes. Thus,

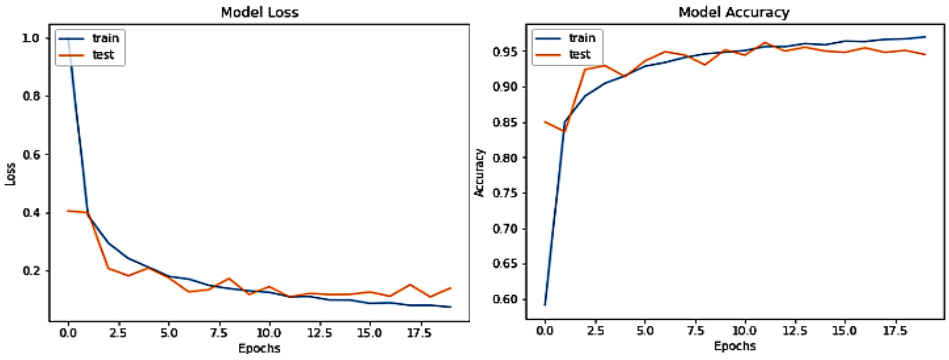

Fig. 7. Model loss and model accuracy of mfcc image-profiled for 6 class of normal person.

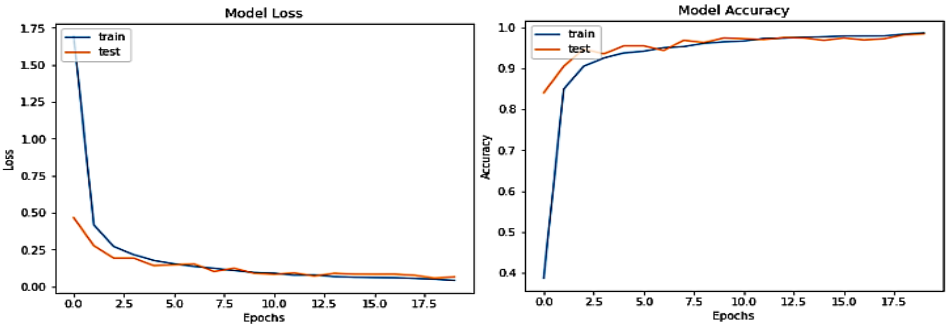

Fig. 8. Model loss and model accuracy of MFCC image-profiled for 12 class of normal person.

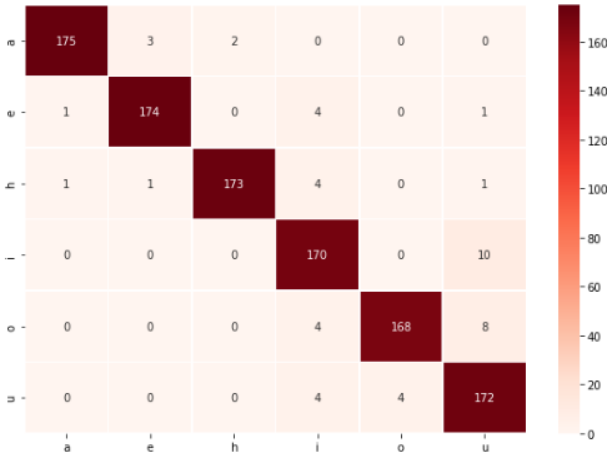

Fig. 9. Confusion matrix of of MFCC image-profiled for 6 group of normal person.

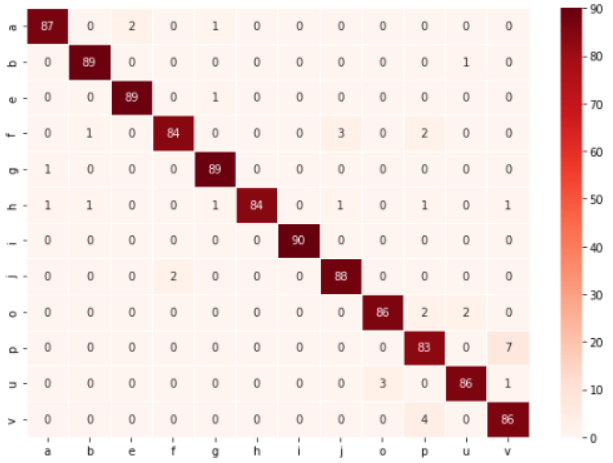

Fig. 10. Confusion matrix of MFCC image-profiled for 12 group of normal person.

it is shown that the system has successfully recognised vowels, especially for speech disorder patients.

Acknowledgment

We would like to express our gratitude to Universiti Teknikal Malaysia Melaka and the Perkeso Rehab Centre in Ayer Keroh, Melaka, for their assistance in providing funding and other resources to make this research attempt a success.

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any

Table 7. Total of MFCC image profile collected in every class of vowels and analysis.

Image profile	Analysis	Classes of vowel	Assessment	Total images in a group	Total images collected
MFCC	Normal	6 Class	Training (80%)	1440	8640
			Evaluation (10%)	180	1080
			Testing (10%)	180	1080
		Total MFCC Image Profile Collected			10800
		12 Class	Training (80%)	720	8640
			Evaluation (10%)	90	1080
			Testing (10%)	90	1080
		Total MFCC Image Profile Collected			10800
	Patient	6 Class	Training (80%)	216	1296
			Evaluation (10%)	27	162
			Testing (10%)	27	162
		Total MFCC Image Profile Collected			1620
		12 Class	Training (80%)	72	864
			Evaluation (10%)	9	108
			Testing (10%)	9	108
		Total MFCC Image Profile Collected			1080
	Mix	6 Class	Training (80%)	1656	9936
			Evaluation (10%)	207	1242
			Testing (10%)	207	1242
		Total MFCC Image Profile Collected			12420
		12 Class	Training (80%)	792	9504
			Evaluation (10%)	99	1188
			Testing (10%)	99	1188
		Total MFCC Image Profile Collected			11880

Table 8. Results of multiple models accuracy by using MFCC image profile.

Analysis	Class of vowels	Model	Epoch	Batch size	Training accuracy (%)	Validation accuracy (%)
Normal	6 Class	Designed	20	6	**97.87**	**94.54**
	12 Class				**99.59**	**98.52**
	6 Class	VGG16			59.21	62.50
	12 Class				62.73	67.59
	6 Class	VGG19			48.45	46.94
	12 Class				48.45	46.94
	6 Class	Inception			40.17	41.61
	12 Class				30.70	34.98
	6 Class	AlexNet			90.29	75.77
	12 Class				91.41	66.24
Patient	6 Class	Designed			**96.45**	**98.77**
	12 Class				**99.65**	**97.22**
	6 Class	VGG16			73.38	77.16
	12 Class				65.62	70.37
	6 Class	VGG19			52.47	52.47
	12 Class				39.12	30.56
	6 Class	Inception			56.31	59.35
	12 Class				58.20	63.06
	6 Class	AlexNet			68.35	42.28
	12 Class				69.46	20.78
Mix	6 Class	Designed			**98.00**	**95.57**
	12 Class				**99.07**	**96.55**
	6 Class	VGG16			59.14	69.65
	12 Class				64.89	70.56
	6 Class	VGG19			44.62	49.44
	12 Class				38.24	42.00
	6 Class	Inception			37.09	37.16
	12 Class				28.88	33.08
	6 Class	AlexNet			89.54	78.02
	12 Class				91.12	54.49

Table 9. Comparison of the testing for 6 class of vowels in the pre-trained designed model.

Classes of vowels	Analysis	Vowel	Precision (%)	Recall (%)	F1 (%)	Support
6 Class	Normal	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
	Patient	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
	Mix	Class /a/	96.00	96.00	95.00	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

Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at Universiti Teknikal Malaysia Melaka, Melaka Malaysia.

Authors' contribution statement

N.S.A.A. has contributed to the design, acquisition of data, and implementation of the analysis and drafting of the manuscript. N.M.Z.H. have contributed to the conception, acquisition of data, interpretation, and revision of the manuscript. M. M. I. and M. D. S. contributed to verify the analytical methods and all authors have discussed the result analysis and contributed to the manuscript.

References

1. Peter L, Keith J. A Course in Phonetics. 6th edition. Cengage Learning; 2010;1–336.
2. Julio Cesar CV, Anders E. Acoustic analysis of vowel formant frequencies in genetically related and non-genetically related speakers with implications for forensic speaker comparison. PLoS ONE. 2021;1–31. https://doi.org/10.1371/journal.pone.0246645.
3. Rebecca T, Victor B, Ruth T, Kira R. Children's phonology awareness: confusions between phonemes that differ only in voicing. J Exp Child Psychol. 1998;68(1):3–21. https://doi.org/10.1006/jecp.1997.2410.
4. Halil I. CERF-oriented probe into pronunciation: implications for language leaners and teachers. J Lang Linguist Stud. 2019;2(4):420–36. http://dx.doi.org/10.17263/jlls.586087.
5. Susan WJ, Dylan E. Postproke aphasia rehabilitation: why all talk and no action. Neurorehabil Neural Repair. 2019;33(4):235–44. http://dx.doi.org/10.1177/1545968319834901.
6. Perrotta G. Aphasia: definition, clinical contexts, neurobiological profiles and clinical treatments. Ann Alzheimers Dement Care. 2020;4(1):21–26. http://dx.doi.org/10.17352/aadc.000014.
7. Perrotta G. Dysarthria: definition, clinical contexts, neurobiological profiles and clinical treatments. Arch Community Med Public Health. 2020;6(2):142–45. http://dx.doi.org/10.17352/2455-5479.000094.
8. Aisha J, Fernando L, Omer R. Interaction between people with dysarthria and speech recognition systems: a review. Assistive Technology: Assist Technol. 2023;35(4):330–38. http://dx.doi.org/10.1080/10400435.2022.2061085.
9. Jung EP. Apraxia: review and update. J Clin Neurol. 2017;13(4):317–24. http://dx.doi.org/10.3988/jcn.2017.13.4.317.
10. Jeremy L, Alexander N, Yehoshua YZ. Classification of audio signals using spectrogram surfaces and extrinsic distortion measures. EURASIP J Adv Signal Process. 2022. https://doi.org/10.1186/s13634-022-00933-9.
11. Nurul AT, Siuly S, Hua W, Frank W, Kate W, Yanchun Z. A Spectrogram. Image based intelligent technique for automatic detection of autism spectrum disorder from EEG. PLoS ONE. 2021;16(6):e0253094. https://doi.org/10.1371/journal.pone.0253094.
12. Prabakaran D, Sriuppili S. Speech processing: MFCC based feature extraction techniques- an investigation. J Phys Conf Ser. 2021. http://dx.doi.org/10.1088/1742-6596/1717/1/012009.
13. Shikha G, Jafreezal J, Fatimah WA, Arpit B. Feature extraction using MFCC. Signal and Image Processing an International Journal. 2013;4(4):101–8. http://dx.doi.org/10.5121/sipij.2013.4408.
14. Shalbbya A, Safdar T, Syed SK, Naseem R. Mel frequency cepstral coefficient: a review. Proceedings of the 2nd International Conference of ICT for Digital, Smart, aand Sustainable

Development (ICIDSSD). 2021. http://dx.doi.org/10.4108/eai.27-2-2020.2303173.
15. Niyada R, Sunee P. An acoustic feature-based deep learning model for automatic thai vowel pronunciation recognition. Appl Sci. 2022. Vol and pages?? http://dx.doi.org/10.1109/iSAI-NLP48611.2019.9045520.
16. Amna A, Hamid M, Fatimah A, Hafiz FA, Abdulaziz A. An approach for pronunciation classification of classical arabic phonemes using deep learning. Appl Sci. 2022;12:1–19. https://doi.org/10.3390/app12010238.
17. Chandra KD, Afiahayati. Suitable CNN weight initialization and activation function for javanese vowels classification. Procedia Comput Sci. 2018;vol.?:124–32. https://doi.org/10.1016/j.procs.2018.10.512.
18. Md. N. A Spectrogram. Image based intelligent technique for automatic detection of autism spectrum disorder from EEG. PLoS ONE. 2021;16(6):1–20. https://doi.org/10.1371/journal.pone.0253094.
19. Shikha G, Jafreezal J, Wan FW, Arpit B. Feature Extraction using MFCC. Signal and Image Processing: An International Journal (SIPIJ). 2013;4(4):101–8. http://dx.doi.org/10.5121/sipij.2013.4408.
20. Luis CS, Sergio V, Omar L, Ana CC, Jhon S, Jan BR. Recognition of EEG signals from imagined vowels using deep learning methods. Sensors. 2021;21(9):6503. https://doi.org/10.3390/s21196503.
21. Nemanja M. Introduction to convolutional neural networks: with image classification using pytorch. Apress. 2020.
22. Shawn H, Sourish C, Daniel P, Jort FG. CNN architectures for large-scale audio classification. In IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings. 2017;p.131–35. https://doi.org/10.48550/arXiv.1609.09430.
23. Parashar D, Praveen D, Ahmad YJ, Vijay D. A near real-time automatic speaker recognition architecture for voice-based user interface. Mach Learn Knowl Extr. 2019;1(1):504–20. https://doi.org/10.3390/make1010031.
24. Zrifie H, Adilah HZ, Juzaila AL, Rostam AH, Farizal H, Maisarah K. Analysis on vowel /E/in malay language recognition via Convolution Neural Network (CNN). Theor Appl Inf Technol. 2022;5:1301–18.
25. Giulio P. Aphasia: Definition, clinical contexts, neurobiological profiles and clinical treatments. Psychologist sp.ed Strategic Psychotherapist. 2020;4(1):21–26. http://dx.doi.org/10.17352/aadc.000014.
26. Tarza HA, Fattah A, Berivan HA. COVID-19 diagnosis system using simpnet deep model. Baghdad Sci J. 2022;19(5):1078–89. https://doi.org/10.21123/bsj.2022.6074.
27. Osamah YF, Bashar SM, Ayad RA. Using VGG models with intermediate layer feature maps for static hand gesture recognition. Baghdad Sci J. 2023;20(5):1808–16. https://doi.org/10.21123/bsj.2023.7364.

التعرف على حروف العلة لتقييم تأهيل مرضى اضطراب الكلام عبر صور الطيف الترددي متعدد المصادر

نور شهمينا أحمد أزهر[1]، نيك محمد ظريفي هاشم[1]، مسرليزام بن مات إبراهي[1]، محمود دوي سوليستيو[2]

[1]*كلية التكنولوجيا والإلكترونيات وهندسة الحاسوب، جامعة ملقا التقنية الماليزية، ماليزيا.*
[2]*كلية الحاسبات، جامعة تيلكوم، جاوة الغربية، إندونيسيا.*

الخلاصة

هناك مجموعة واسعة من الأسباب الطبية لضعف الاتصال ، مثل اضطرابات الكلام ، وفقدان السمع ، وإصابات الدماغ ، والسكتة الدماغية ، والإعاقات الجسدية. نتيجة لذلك ، يمكن أن يؤثر اضطراب التواصل مدى الحياة على التنمية الاجتماعية والعلاقة الشخصية. يمكن أن تستفيد اضطرابات النطق من علاجات النطق المبكرة ؛ ومع ذلك ، لا تزال غالبية مرافق إعادة التأهيل في جميع أنحاء العالم تنفذ هذه العملية يدويًا. من وجهة نظر عالمية ، تم إجراء مجموعة واسعة من الدراسات حول معالجة الكلام لمختلف اللغات البشرية. نظرًا لأن رؤية الكمبيوتر قد أثرت على هذا المجال ، فقد تم تطبيق التعلم الآلي والتعلم العميق في الصناعة الطبية والرعاية الصحية لتعزيز إعادة التأهيل من خلال استخدام التكنولوجيا الجديدة. حللت هذه الدراسة دقة تصنيف الشبكة المصممة والنماذج الأخرى المدربة مسبقًا (VGG-Net و AlexNet و Inception) وأجرت تحليلًا مقارنًا كاملًا لتقييم دقة التصنيف للعديد من النماذج المدربة مسبقًا. في هذا العمل المقترح ، لإنجاز مهمة التصنيف هذه ، يتم تحويل الصوت لاحقًا إلى الصورة كطريقة جديدة لرؤيتها في الشبكة العصبية عبر مفهوم مقترح حديثًا يسمى بيانات ملف تعريف الصورة. أنتجت مجموعات البيانات التي تم تصنيفها عن طريق الصور والتي استخدمت مخططًا طيفيًا ومعامل تردد ميل التردد (MFCC) أفضل نتائج هذه الدراسة ودقتها. يهدف هذا المشروع إلى تطوير شبكة عصبية جديدة يمكنها التمييز بنجاح بين أحرف العلة من أصوات الأشخاص العاديين والمرضى الذين يعانون من اضطرابات الكلام والمزيج من المجموعتين السابقتين باستخدام الفئتين الستة والثاني عشر من حروف العلة الملايو. وفقًا للبيانات التجريبية التي تم إجراؤها ، ونموذج الشبكة المصمم ، والذي استخدم 6 أحجام دفعات ، و 20 حقبة ، و ADAM كمحسِّن ، قدم هذا المشروع وحقق قيم الدقة القصوى لكلا الفئتين لبيانات الصوت الخاصة بالصور في جميع التحليلات التي تم إجراؤها.

الكلمات المفتاحية: شبكة عصبية ملتوية (CNN)، التعلم العميق، معامل ميل التردد الرأسي (MFCC)، إعادة التأهيل، الطيف، التعرف على حروف العلة.