doi.org/10.24191/jeesr.v19i1.004

A Systematic Approach for Malay Language Dialect Identification by Using CNN

Mohd Azman Hanif Sulaiman, Nurhakimah Abd Aziz, Azlee Zabidi, Zuraidah Jantan, Ihsan Mohd Yassin, Megat Syahirul Amin Megat Ali, and Farzad Eskandari

Abstract*— As Malaysia moves forward towards the Industrial Revolution (IR 4. 0), computer systems have become part of everyday life, leading to increased man-machine interactions. Verbal communication is a convenient means to interact with computers. Speech recognition systems need to be robust to cater for various languages and dialects in order to interact better with humans. Dialects within a spoken language present a challenge for computers require a speech recognition system to translate these verbal commands to computer understanding of the underlying meaning from spoken words. In this paper, works on Malay language dialect identification are presented using Convolution Neural Network (CNN) trained on Mel Frequency Cepstral Coefficient (MFCC) features. Data was collected from 12 native speakers. Each speaker was instructed to utter 10 carefully selected words to emphasize the dialect nuances of the eastern, northern and central (standard) Malay dialect. The MFCC features were then extracted from the recorded audio samples and converted to graphical form. The images were then used to train a custom CNN neural network to differentiate between the various spoken words and their dialects. Results demonstrate that CNN was able to effectively differentiate between the spoken words with excellent accuracy (between 85% and 100%).

Index Terms—**Convolution Neural Network, Mel Frequency Cepstrum Coefficient, speech recognition, dialect recognition**

I. Introduction

AUSTRONESIAN languages are spoken by about 386 million people. Among them, The Malay language is a major language in Austronesian family people commonly spoken by 290 million people in South East Asia (SEA) countries, such as Malaysia, Indonesia, Javanese, Singapore, Brunei, and parts of Thailand [1], [2], [3]. The language is known by many names, such as Bahasa Malaysia (Malaysia), Bahasa Melayu (Singapore & Brunei) and Bahasa Indonesia (Indonesia) [4].

To illustrate its significance, the Malay language is more widely spoken than Japanese, German and Thai language [5]. Recently, the Malay language is one of the minority languages that has risen to international prominence due to the recent political and economic developments in Southeast Asia. A review by [5] recognized the Malay language's importance to China due to the country's need to interface with Malay-speaking countries as part of its Belt & Road Initiative.

The Malay language has a rich history, and like other languages, it has evolved over time. Experts have classified the Malay Language evolution into five important categories, namely Old Malay, the Transitional Period, the Malacca Period (Classical Malay), Late Modern Malay and modern Malay [6]. Additionally, the language is categorized into several dialects as a result of diverse geographical regions that use it as a communication language [7]. The different dialects are also attributed to the different pronunciations and vocabulary unique to the area. In linguistics, pronunciation belongs to articulatory phonetics, which involves the organ and area of articulation, while auditory phonetics refers to interpretation of a sound. In this process, the transmission of acoustic phonetic happens between the sound signals to the listener speaker during communication [8]. Both articulatory and auditory phonetics is important for effective communication.

Dialect recognition is particularly important in designing a robust speech recognition system tailored to specific languages. The insight of a dialect recognition system would help speech recognition systems to understand and interact with the speaker better. In the last decade, much research has been done to discover solutions for speech and dialect identification using various techniques [9]. As the raw audio data of three class subset of emotion (angry, sad and neutral) was taken from the German Corpus Berlin Database and contained 271 labeled recording with length 783 seconds as the total. The architechture for this Speech Emotion Recognition (SER) was used Deep Neural Network (DNN) with convolutional, pooling and fully connected layers[10]. However, this conventional machine-learning techniques, has limitation in their performance to process natural data from the raw form. So, new machine learning technique like Deep learning has an ability to allow multiple processing with multiple layers by learning representations of data with multiple levels of abstraction [11]. This, in part, is driven by a better need for interaction between human and machine systems.

This manuscript is submitted on 18th February 2021 and accepted on 28th April 2021. M.A.H.Sulaiman, N.A.Aziz, A.I.M.Yassin and M.S.A.M.Ali are with the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor (azman2858@uitm.edu.my). .Zaibidi was with Faculty of Computing, University Malaysia Pahang (UMP) (email : azlee@ump.edu.my).

Zuraidah Jantan was with Academy Language of Study (ALS), University Technology MARA (UiTM). (email : zuraidah0420@uitm.edu.my). Farzad Eskandari is Proffesor at Alahmeh Tabatabai University with his research area on statistic and data science. (email : askandari@atu.ac.ir).

This paper presents work on Malay dialect recognition using Convolution Neural Network (CNN). The methods presented here utilizes CNN to recognize Mel Frequency Cepstrum Coefficient (MFCC) features extracted from sound files of native dialect speakers on selected vocabulary. The MFCC extraction and CNN training process is described.

The rest of this paper is organized as follows: Section II presents about dialect in Malaysia, section III present about relevant methods for dialect classification. Section IV about features representation of MFCC, Section V presents faeatures extraction of MFCC, Section V about deep learning, Section VII the methodology section, describing the details of data collection, CNN structure and training parameters. This is followed by Section VIII, which discusses the results of the proposed method. Then, conclusions are presented in Section IX and finally X is about the future work.

II. THE MALAY LANGUAGE AND DIALECTS

The standard Malay Language has various official names as the national language of several countries in SEA. In Malaysia it is called Malaysian language (Bahasa Malaysia) or Malay Language (Bahasa Melayu) whereas in Singapore and Brunei, the language is known as the Malay Language (Bahasa Melayu). In Indonesia, the language is known as the Indonesia language (Bahasa Indonesia). Despite the various names, there exists a Standard Malay Language independent of speakers from these countries, which standardizes the language in its unique system to accommodate dialects related to the standard language [12]. In Malaysia and Indonesia, the Malay language is spoken by 290 million people across the Malaysian coast, eastern Sumatran coast in Indonesia, Sabah and Sarawak, and across the Strait of Malacca [3].

From the 36 phonemes in Malay language, 27 of them are consonants, three are diphthongs and six are vowels. The word structures present in the language are V, CV, CVC, CCV, CCVC, CCCV and CCCVC, where V is vowel and C is consonant [13].

Dialects are variations of a language uniquely spoken by the people in a certain place or country [14]. Malaysia, with an estimated population of 32. 4 million people [15], uses Malay as the official language. However, diversity of Malay races and regions has resulted in different dialects across the region. Each state in Malaysia has its own local dialects , but they can be categorized into eight primary dialects:- Johor, Melaka, Negeri Sembilan, Perak, Kedah, Kelantan, Terengganu, and Pahang [16].

III. RELATED RESEARCH

This section lists some relevant research. In [17], a deep learning-based system for screaming sound detection was presented. MFCC features were extracted from audio recordings, the presented to the DL classifier to detect the screaming sound. The dataset consists of 130 scream sounds and 110 normal sounds as the dataset, 92 scream and 78 normal sounds are used for training other than that, 38 scream and 32 normal sounds, are used for testing of the proposed system. The results indicate that can significantly improve the precision of audio recognition systems to the 100% [18], [19].

Another technique that has been used in previous research was Gaussian Mixture Model (GMM) for dialect recognition of Javanese and Sundanese, after features extraction function has been perform, the selected segment that contain speech data by using Voice Activity Detection (VAD). VAD will select certain threshold to predict the dialect class with maximum value using log-likelihood ratio were collected and subsequently, the data is then trained by I-Vector as the classifier, after the extraction process completed, it will train and test the dataset then weighted by using Logistic Regression (LR) to ensure Posterior Probability value as input signal for confidence measurement for I-Vector model. From the maximum value of the certain threshold, it were compare with the posterior value given by LR on each test data against the dialect class. prior to be classified as a Posterior Probability input signal by the weighting of LR [20].

GMM operates in two phases with the first phase is converting the speech expression into features vector for each dialect and the second phase is involving the recognition process, where the speech expression is compared with each GMM. In this process, the unknown speech expression were identified and computed where the model that represent the [21]. On the other hand, one of the common modeling methods that has been used in dialect identification is Support Vector Machine (SVM). SVM works in two phases, the first is to transform the data into for maximum variability, the defining a hyperplane to maximally separate between two or more data classes [22]. Other approaches such as CNN will reduce over fitting or poor generalization by multi stage training. A multilayer perceptron structure in CNN provides an advantage that able to recognize the 2D or 3D pattern signals [23].

In [17] the combination of MFCC and DLNN was used to detect screaming sounds. First, a dataset of screaming sounds was collected. MFCC features, extracted from the screaming sounds were used to train several classifiers - DLNN, and SVM with GMM kernel to detect screaming sound. The result shows that the DLNN achieved better results compared to the SVM with GMM kernel with 100% and 91. 14% accuracy respectively. New parts of semi-supervised DLNN for lung sound analysis focusing on wheeze and crackle may be a crucial component of pulmonary disease diagnostics for essential care and common persistent observing for telemedicine was presented in [24]. The data were collected from 284 pulmonary patients from four separate clinics and 890 four-second recordings were randomly selected from 11, 627 sound files. The signals were then passed through low-pass filters with Hamming window. The dataset was then used to train the semi-supervised deep-learning to analyze the lung sound. The result shows by using semi-supervised DLNN was able to automatically identify lung sounds from a large number of patients.

In [18], ensemble and DLNN classifier were used for detection of abnormal heart sounds. Phonocardiogram (PCG) data was tested on two combinations of classifiers which is AdaBoost and CNN in order to detect abnormal heart sounds. First, a PCG recording of solid subjects and pathological patients was collected. The two cases were trained and evaluated on a blind test dataset using AdaBoost alone, CNN alone and combination of AdaBoost and CNN. The combined classifier obtained an increase of classification sensitivity by 18% compared to AdaBoost alone and 9% compare to CNN alone but decreased in specificity by 6% compare to AdaBoost and 4% compared to CNN alone.

A recent research by [19] also has been conducted to compare the environmental sound detection between DLNN methods and non-neural models (GMM and I-Vectors). The dataset was taken from IEEE Challenge on Detection and Classification of Acoustic Scenes and events to compare the detection with the three mentioned methods [19]. The DLNN achieved 84. 2% average accuracy meanwhile GMM and I-Vector both score at 72. 5% and 81. 7% average accuracy. However, merging the method improve the performance by 88. 1% average accuracy.

In [25], improved speaker recognition system for stressed speech using DLNN was presented. The researchers used a GMM-Universal Background Model (GMM-UBM) and DLNN based classifiers. Data was collected from the Buckeye corpus of conversational speech dataset consisting of 40 speakers. Features were extracted using MFCC. The training system (DNN and GMM-UBM) was applied to both breathing sound and model speech. The results show that both GMM-UBM and DLNN performed well when trained and tested with best equal error rate (EER) of 0. 76 and 0. 67, respectively with GMM-UBM show higher reduce in their perfomance. On degradation side GMM-UBM based system was higher than DLNN-based system.

Similarly, in [10], DLNN was used for speech emotion recognition. The research begins with collecting data from German Corpus dataset, containing approximately 800 sentences with different emotion of (angry, neutral, sad) and genders. With network structure consisted of only fully connected layers. The first layer with size 480, the second layer with size 240 and the last layer was an output Softmax layer with 3 output neurons. All pooling layers were used with pool size 2. ? Training was performed using the Gradient Descent algorithm. The results indicate that the DLNN archived 96. 97% accuracy on speech emotion recognition task and average prediction confidence of 69. 55%.

Some research has highlighted the speech recognition that focusing on word recognition, where a word spoken by the user were then need to be recognized by the speech recognition system for Automatic Speech Recognition (ASR)

Applications [26]. ASR systems works to convert human speech or audio signals into text, and this system has been adopted in such embedded system as interactive voice response (IVR), verbal document retrieval or transcription tools [27].

Several studies were focusing in dialect recognition. Research by [20] used GMM to identify Indonesian Java and Sunda dialects using MFCC features classified with GMM and I-Vector classifiers. The results indicate that the classification error (CE) for I-Vector was 25%, outperforming the GMM with 50% classification error. Another work in Batak, Javanese and Minang language recognition [28], the researchers discovered accuracies rate with less than 20% for Bataknese, 50% to 80% accuracy for Javanese and 70% to 90% accuracy for Minang dialect. So this method is succeed to identify the Javanese and Minang with good precision.

In [29], a Malay ASR was research utilizing the Malay Grapheme to Phoneme (G2P) tool. A (G2P) tool was used as the. Malay ASR system achieves word error rate(WER) of 16. 5%, which is as it were 1. 9% higher compared to the utilization of a pronunciation dictionary that are manually verified. This paper focus is on Malay pronunciation and phonology.

In [30], the dialect of the speaker was verified by doing some preliminary analysis on the distinguishability of the Slovak dialects by a basic automatic classifier. The sound file of the Slovak Dialects has been chosen a suitable set of recordings from the Slovak Parliament, and made a database, for dialect-specific acoustic models training and testing. A standard GMM based recognizer was utilized for the classification tests. The tests were performed with three distinctive degrees of coordinate between the scope of (sub-) dialects within the training and testing data.

Therefore [31], performance comparison of DLNN frameworks in image classification problems using Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) was presented. In this research DLNN frameworks that used are Tensor Flow, Theano and Torch. The database MNIST and CIFAR-10 was collected and used. MNIST database is 2 subsets of handwritten database combined together meanwhile, CIFAR-10 is set of data containing 60000 tiny color images. The training was applied to all three frameworks to decide which frameworks are better. For the result, Torch framework show the lowest time required for iteration in both CPU and GPU settings. Meanwhile, Tensor flow score higher than Theano in all cases except testing time of CIFAR-10.

IV. FEATURE REPRESENTATION USING MFCC

MFCC is a popular feature extraction method to represent audio [32]. The method has been proven to provide a precise and accurate representation of human-audible sounds [33], in which it is capable to filter banks model characteristics of the human auditory system [34]. Previous researchers have been utilized MFCC as a feature representation method for audio recognition and classification [35], [36]. The application of MFCC is gaining more attention among researchers due to its reliable features which able to give result over 80% accuracy to recognize the voice. For example, MFCC has been used together with Support Vector Machine (SVM) for Pashto dialect, as well as combination of MFCC and Gaussian Mixture Model (GMM) for Bangladeshi dialect [28]. Reference [37] highlighted that MFCC coefficients produced better result in terms of accuracy for the classification of anger and happiness with accuracies of 99. 44% and 99. 71%, respectively. Reference [32] has demonstrated MFCC's speed and accuracy compared to DWT for road type classification based on sound.

In [38], proficiency of MFCC and distinctive number of channel and Hamming window or rectangular window has been utilized for speech recognition. Initially, the element was extracted and contrasted between the highlights with number of channels which are 22, 32 and 42 channels and following by windowing strategy which are Hamming and Rectangular window. At that point, vector quantization procedure was done to recognize similitude and divergence for the two speaker signals. The outcomes demonstrate that MFCC with 32 channels and Hamming window score most elevated with 87. 5% efficiency.

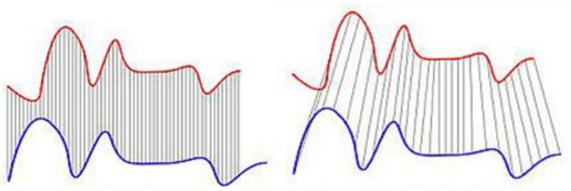
V. MFCC FEATURE EXTRACTION

MFCC Features Extraction steaps are windowing, quick Fourier change, Mel frequency sifting, take logarithm and discrete cosine transform process [36]. As MFCC is replicating human sound related frameworks, it's very subject to its extraction parameters. A too-high or too low MFCC parameter setting may cause overshoot or under setting (both during data preprocessing and training) of the classifier, the classification accuracy were reduced.

A. Dynamic Time Warping

In speech recognition tasks, although the same person doing voice recording repeating the same word, there's no guarantee that the length for each words is constantly the same. In order to ensure the data is in the same length, Dynamic Time Warping (DTW) is the perfect technique to accommodate differences in timing between sample words and templates. In order to calculate the optimal warping path between the two data from data sound, Dynamic Time Warping (DTW) algorithm can be used to warping the values and distance of the output. Warping path can be determine as the difference of two patterns that been produced during comparison of two patterns. If the warping path is small, it's mean that the two patterns can be exactly the same [39].

To reduce distortion that occur in the process of alignment two time series sequences that produce stretched and compressed sections, DTW algorithm was used in the sequence of feature vectors [40]. DTW can really help in order to solve the problem regarding time series recognition especially in various pattern length and noise in a feedforward architecture [41]. Fig. 1 below will show illustration of the DTW method [39]:



(a) Original alignment (b) Alignment with DTW

Fig. 1. DTW alignment developed

For a better result and accuracy, DTW is the simples and reliable algorithm. In [42] show that the basic principle is to identify feature values through a random process with a good alignment. Reference speech signal working as input at the pre-processing unit, will leave a fixed signal after filter out unnecessary information and known as reference template. Then, another speech signal to be tested same as the reference signal and obtain the feature vector sequence where we call test template.

DTW algorithm has received its reputation when it's has a very good ability as the time series similarity technique that can extent and reduce the effects of noise or shifting in order to find the same shape by the different phases [43]. DTW algorithm works by performing comparison of the parameters from the unidentified pronounced words with the reference template words parameters. When more reference template has been use for the same word, it will leads for the highest accuracy of recognition rate. However, by increasing the number of reference template also can leads to the computation time and increasing memory [26].

VI. DEEP LEARNING

Deep Learning (DL) one of the family from machine learning and also known as deep structured learning or hierarchical learning, their learning process can be supervised, semi supervised or unsupervised [44], [11]. The DLNN ability is to recognize more on visual pattern than other model, hence of their ability mimic on human, it robustly learning and find the data from sample [45]. There's an advantages from CNN over the other type of neural network is it can extract features from the input data very strongly and significant [46]. The "Deep" itself which in "Deep Learning" is subject to the layer's number where the data is transformed process. The DLNN systems have a significant credit assignment path (CAP) depth that also can be describe as sequence of revolutions from input to the output. In theoretically, CAP is connection between input and output [44].

By being one of the most prominent method used by the researcher or developer, DLNN has shown a successful performance than others method. The DL concept resembles from the human brain activity. With this method, process of non-structured information such as sound, image, signal and videos can be cracking using DLNN. Plus, the convolutional and recurrent models have been solving problem in computer vision application very efficiently [31]. DLNN is used in machine learning in classification and pattern analysis which is supervised and unsupervised deep learning.

A. Convolution Neural Network

CNN has been used in this research area for a long time ago, especially in image processing and speech recognition [47]. The ability of CNN that very special is convolution networks are combining both feature extraction and the classification part. The CNN network in Figure 2 show it was compress with five different layers in order to convolve with multiple layers with a certain weight, and then the pooling section layer will reduce the size but still maintain the data or information. The last two layer of the CNN will compose all the feature extraction and combine in the fully connected layer before output it in the output layer. Fig. 2 below will give some brief on what is CNN [48].

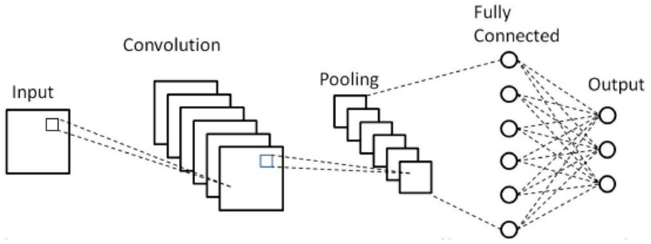


Fig. 2. CNN block diagram

CNN filters can absorb input character, specific patterns or sound as their learning process that related for a specific task. In the training process, every steps is fed into the network to generate forecast value and if there is an error the networks will back propagated to ensure the weights at every layer are restructured [49]. CNN as the one of the sub from Multilayer Perceptron (MLP) was a famous technique that been use to do classification in DL. It can extract the features automatically rather than MLP that need to do it manually [50].

VII. METHODOLOGY

The methodology of this study is presented in this section. The main steps involved include data collection, design and implementation of networks based on deep learning, training, testing and analysis of network performance. The flowchart for the methodology is shown in Fig. 3.

Due to the novel application of using Malay language datasets in this study, the data were collected from volunteer subjects that are fluent with the language and dialects studied. For CNN, MFCC feature extraction was used to extract the features. Next, these features were represented in graphical form (top view of the cepstrum features), and used to train the CNN. Implementation was done using MATLAB r2019. Training was performed by using Graphics Processing Unit (GPU) NVIDIA GeForce GTX 1080 Ti, Processor Intel® Core™ i5-6400 CPU 2.7 GHz with Random Acces Memory (RAM) of 20 GB. The performance several networks were evaluated and compared in order to find the best network for dialect recognition.

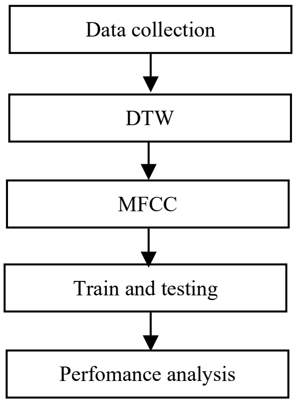


Fig. 3. Experiment methodology

A. Data Collection

Before the network start extracting the data and train the network, first of all we need to gather or find the data that useful towards this research area before transforming it into dataset and is important to get or gather the right data. A few steps are used in this data collection section such as:

i. Create sample words that follow the dialect

Several words were carefully selected to represent significant differences between their spoken representation. 20 words was selected for the differrent dialects, namely standard, eastern and northern. All the sample words were vetted by one of the co-authors a Malay Language lecturer from the Academy of Language Studies, UiTM.

Table I describes the International Phonetic Association (IPA) representation of selected words used in this experiment. From the recording process, all the words given to the subject also successfully follow the International Phonetic Association (IPA) transcription sound chart. This can help researcher in order to clarify the words given is really follow the International standard in language research area.

ii. Speech recording

To ensure optimal recording conditions, recording was performed in a controlled environment. Recording was performed inside a silent room without any disturbance. Each subject (six male and six female native dialect speakers) was required to utter the the each of same word ten times. Each recording was saved into the raw *wav* format, resulting in 2,400 samples divided into standard, eastern and northern dialects. Additionally, the age of the subjects ranged from 20 to 35 years old in order to obtain clear and fluent pronounciation of their native dialect. To establish their native dialect, each subject was selected based on their state of origin and/or the period of stay in the state.

The dataset were pre-processed to remove recording artefacts. Due to the spontaneous nature of the speech, the recordings suffered from filled pauses and elongation that need to be determined and removed to preserve the dataset quality. The sampling frequency of the collected words was set to 16 kHz in the mono channel with 16-bit bit resolution. Framing was done to the collected speech by blocking the speech signal into different frames.

The voice recorder used was the SONY ICD-UX560F Digital Voice Recorder with WAV/MP3/LPCM format file were used in speech recording session, which equipped with a high-sensitivity Stereo-Microphone and noise removal function available with high pass filter.

TABLE I
WORDS FOR RECORDING

Standard	IPA Standard Transcription	Eastern	IPA Eastern Transcription	Northern	IPA Northern Transcription
Atas	[atas]	atah	[atah]	atah	[atah]
Awal	[ʔawal]	awa	[awa]	awai	[awaj]
Belajar	[bəlaǰar]	belaja	[bəlaja]	belaǰaq	[bəlaǰaʕ]
Beras	[bəras]	berah	[bəɤah]	berah	[bəɤah]
Besar	[bəsar]	Besa	[bəsa]	Besaq	[bəsaʕ]
Biar	[biar]	Bia	[biʲa]	biaq	[biaʕ]
Buaya	[buaja]	boyo	[bƆjƆ]	boya	[bƆja]
Bungkus	[bungkus]	bukuh	[bukuh]	bungkuih	[buŋkujh]
Hitam	[hitam]	Hite	[itƐ]	itam	[itam]
Kakak	[kakaʔ]	Kakok	[kakƆʔ]	kakaq	[kakaʔ]
Keluar	[kəluar]	Kelua	[kəlua]	Keluaq	[kəluaʕ]
Kerbau	[kərbau]	Kuba	[kuba]	kebaw	[kəbaw]
Lapar	[lapar]	lapa	[lapa]	lapaq	[lapaʕ]
Lepas	[ləpas]	lepah	[lepah]	lepah	[ləpah]
Pahit	[pahit]	pahik	[pahiʔ]	payt	[pajt]
Saya	[saya]	Sayo	[sayƆ]	Saya	[saya]
Tawar	[tawar]	tawa	[tawa]	tawaq	[tawaʕ]
Tebal	[təbal]	teba	[təba]	tebaiyh	[təbaj]
Tikus	[tikus]	tikuh	[tikuh]	tikuyh	[tikujh]
Ular	[ular]	Ula	[ula]	ulaq	[ulaʕ]

B. Dynamic Time Warping

As the words were uttered multiple times, there is a strong possibility that the words were of different length. DTW was used to align the words correctly and process them such that they have equal length.

C. Features Extraction using MFCC

MFCC is a feature representation method that replicates human audio characteristics. Five important steps for MFCC extraction are windowing, Fast Fourier Transform (FFT), Mel frequency filtering, and discrete cosine transform.

MFCC is very sensitive to its extraction parameters, namely the number of filter banks and the number of coefficients. A too-high or too low parameter setting may cause suboptimal representation of the features, resulting in reduced classification accuracy. Fig. 4 shows the flow chart regarding the MFCC process.

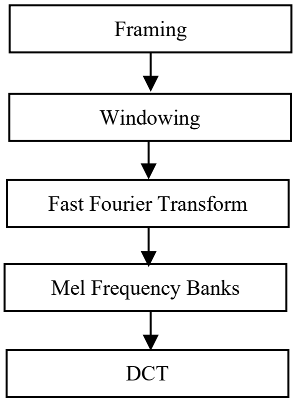


Fig. 4: MFCC Flow Chart Process

i. Framing & Windowing

The first step of MFCC is framing, which decomposes the raw audio signals into smaller pieces. The typical frame size is between 20 to 40 ms, while the standard is 25 ms [51].

Windowing minimizes the effect of signal discontinuity at the beginning and end of each frame [28]. The windowing process pre-multiplies the signal with Hamming window that gradually diminish the amplitude of the to zero at the beginning and end of each frame. Hamming window was chosen since it smoothens the attenuation of the signal more gradually.

As the result from the MFCC process, the signal segments was multiply with the Hamming window where the width of 25ms and succeeding the frames overlapped by 50% and FFT was applied on every frame.

ii. Fast Fourier Transform & Mel Frequency Banks Extraction

FFT represents each of the time-domain frames into frequency domain. This process is necessary for the next Mel frequency extraction. Mel frequency banks show that human ear can receive certain frequency bands. The formula for the mel frequency bands is shown in Eq. (1):

$$mel(f) = 2595\left[ln\left(1 + \frac{f}{700}\right)\right] \qquad (1)$$

iii. Discrete Cosine Transform (DCT)

To convert the frequency back to the normal form, the following equation of DCT need to be performed. $C(n)$ represents the MFCC, m is the number of coefficient, N the numbers of triangular band pass filter, M the total of Mel scale cepstral coefficients and $Y(m)$ the result from spectrum multiplication with conjugate.

For this research, the number of filter banks were set between 20 to 40 triangular filters was used with only 10-20 coefficients were computed from the filter banks. The approximate formula is shown in Eq. (2):

$$c(n) \quad = \sum_{m=1}^{M} Y(m) cos\left[\frac{mn(m-\frac{1}{2})}{N}\right] \qquad (2)$$

D. Classification

The MFCC coefficients were plotted and the top-view was was used as features for the CNN classifier. Each plot contains a distinct signature depicting the patterns that can be distinguished by the classifier. An advantage of using the graphical representation is that the features are consistently sized, which makes fitting them to the network easier.

The structure of the CNN classifier is shown in Table II. The CNN networks were tested using various combinations of MFCC parameters.

TABLE II
CNN PARAMETERS

Layer	Filter size D x H x W	Others Parameters
Convolution 1	16 x 7 x 7	Stride = 1, padding = 0
Batch normalization	-	-
ReLU	-	-
Max pooling	16 x 3 x 3	Stride = 1, padding = 2
Convolution 2	32 x 5 x 5	Stride = 1, padding = 0
Batch normalization	-	-
ReLU	-	-
Max pooling	32 x 3 x 3	Stride = 1, padding = 1
Convolution 3	64 x 3 x 3	Stride = 1, padding = 0
Batch normalization	-	-
ReLU	-	-
Max pooling	64 x 3 x 3	Stride = 1, padding = 1
Convolution 4	128 x 3 x 3	Stride = 1, padding = 0
Batch normalization	-	-
ReLU	-	-
Fully connected+Softmax	60	-

The CNN network consists of 16 layers. The convolution layers are feature extractors, while the ReLU (Rectified Linear Units) are the activation functions of the layer. Pooling layers help reduce the dimensionality of the features passing through each convolution layers. The convolution layers integrate more meaningful features deeper into the network. Finally, the fully connected and softmax layers at the back learn the features and output the classification results.

TABLE III
CNN TRAINING PARAMETERS

Layer	Others Parameters
Training Algorithm	Adaptive Moment Estimation (ADAM)
Execution Environment	GPU
Max Epoch	1,000
Mini Batch Size	512
Learn rate Schedule	piecewise
L2 Regularization	0.05

The training algorithm chosen was the Adaptive Moment Estimation (ADAM) algorithm algorithm [52]. The ADAM algorithm is a hybrid between the Stochastic Gradient Descent with Momentum (SGDM) and Root Mean Square Propagation (RMSprop) algorithm, thus inheriting both the advantages of the algorithms and improving the traversal of the solution space.

Training utilized the GPU to perform calculations. The advantages of using parallel processing for training CNN are well-known in literature due to the GPU's ability to accelerate training by utilizing its multi-core processor architecture.

The mini-batch size is GPU-memory dependant. A higher mini-batch size would increase training speed at the cost of GPU memory. This is because a higher mini-batch would load more training data at each epoch. The mini batch size was set to 512 based on a trial and error basis. Initially, a low value is set, and the GPU memory usage was measured. The mini-batch value was the increased incrementally, resulting in higher GPU memory usage. The optimal value for the mini-batch was determined when the GPU memory usage is approximately 80%. A higher value was not attempted to protect training stability, as there will be some fluctuations in GPU memory usage. If a too high value is set, the training may abruptly stop due to data unable to be copied into the GPU for training.

The maximum number of epochs adjusts the number of times the training is performed. This value is considered sufficient since the CNN consistently demonstrated above 90% accuracy using this setting in acceptable time.

L2-regularization was used to avoid overfitting, a condition when the CNN memorizes the training data while performing poorly on previously unseen data. L2-regularization adjusts the learning rate of the training algorithm so small movements are made across the error surface towards the minimum solution near the final epochs.

E. Performance comparison

To evaluate the performance of the DLNNs, the evaluation was done using confusion matrix:

1. The first confusion matrix describes the classification results of different DLNNs. The format of the classification matrix is shown in Figure 5 [53]. For this research a 20x20 confusion matrix was used to show the accuracy of each word and dialect.
2. The second clustering matrix was defined into 3x3 classes to indicate the accuracy of dialect prediction.

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Fig. 5. Confusion matrix table

Positive sample will represent the selected dialect we categorise and the negative sample will represent non selected dialect we use.

- True Positive (TP) class will predict the right assumption on true positive class is the selected dialect.
- False Positive (FP) class will predict the wrong assumption on false positive class is the selected dialect.
- True Negative (TN) class will predict the right assumption on true negative class is the non-selected dialect.
- False Negative (FN) class will predict the wrong assumption on false negative class is the non-selected dialect.

This measurement calculation strategy is broadly utilized in DNN to classify the lesson. The estimation utilize is based on extent of precise classes archive with the full test sample by the equation:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Precision is mean out of all the positive classes we have predicted correctly, how many are actually positive.

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

Specificity was known as the proportion of the Truly Negative Rate (TNR) and it also known as truly negative by the equation given below:

$$Specificity = \frac{TN}{TN+FP} \tag{5}$$

Sensitivity, known as the proportion of the truly positive that the positive class is the selected dialect or also known as True Positive Rate (TPR) and the equation given below:

$$Sensitivity = \frac{TP}{TP+FN} \tag{6}$$

VIII. RESULT & DISCUSSION

The results section is divided into several subsections namely DTW, MFCC, and classification results. They are described in detail below.

A. *Result for DTW process*

In order to accommodate the differences in length of the recordings, DTW was used to adjust the audio into equal length recordings. An example of the DTW process is shown in Fig. 6. It shows that two different recorded voice for the same word has length 1200ms for data A and 10190ms for data B. Then after the DTW process the length for both data A and data B change to 12374ms. The similar lengths would allow the MFCC to be consistent across the dataset.

B. *Result for MFCC process*

The MFCC results can be visualized as a three-dimensional graph. The graph was rotated so that the top view is visible. The top view depicts distinctive patterns that can be classified using the CNN. The top view was saved as a Portable Network Graphics (PNG) and used to train the CNN. The process is shown in Fig. 7.

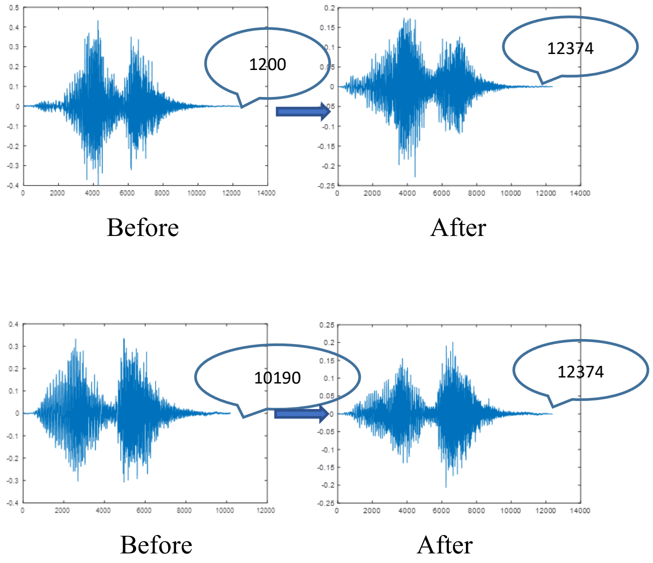


Fig. 6. Data length before and after DTW process

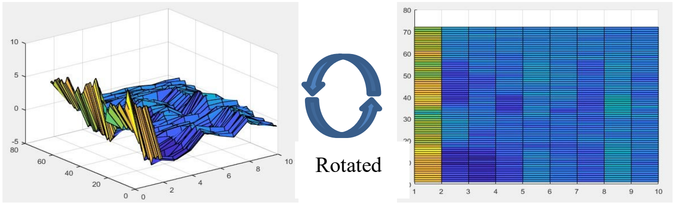


Fig. 7. Process of rotated Mel Cepstrum

C. *Classification result*

The confusion matrices for all the 20 classes are shown in Fig. 8 and Fig. 9. The correct classifications are highlighted. The classification values were generally above 80%, indicating the the CNN was able to discriminate between the classes well. A summary of the class accuracy is shown in Table IV. As can be seen, the CNN consistently performs well on the classes, both for the training and testing sets. A good performance on the testing set indicates that the CNN was able to generalize well, even for previously unseen cases.

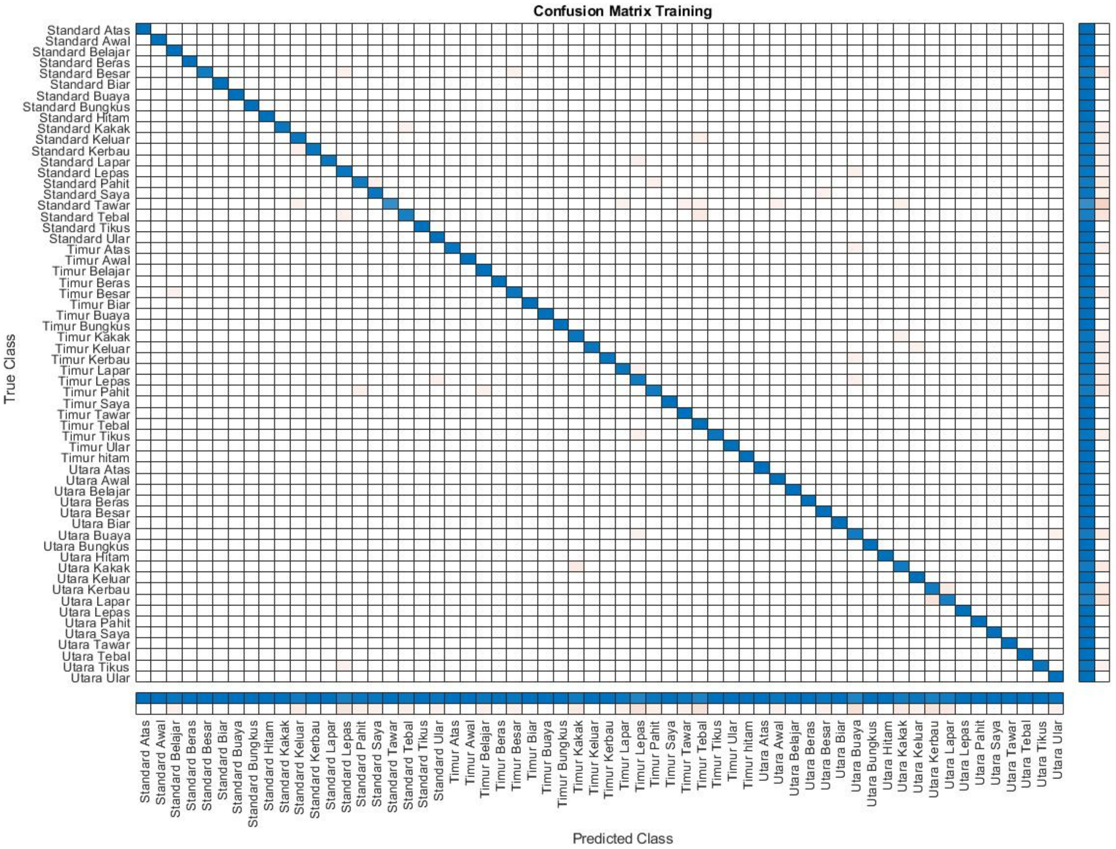


Fig. 8. Confusion Matrix (Training Set). Blue tones indicate intensity of correct classifcations, while red tones indicate intensity of incorrect classifications

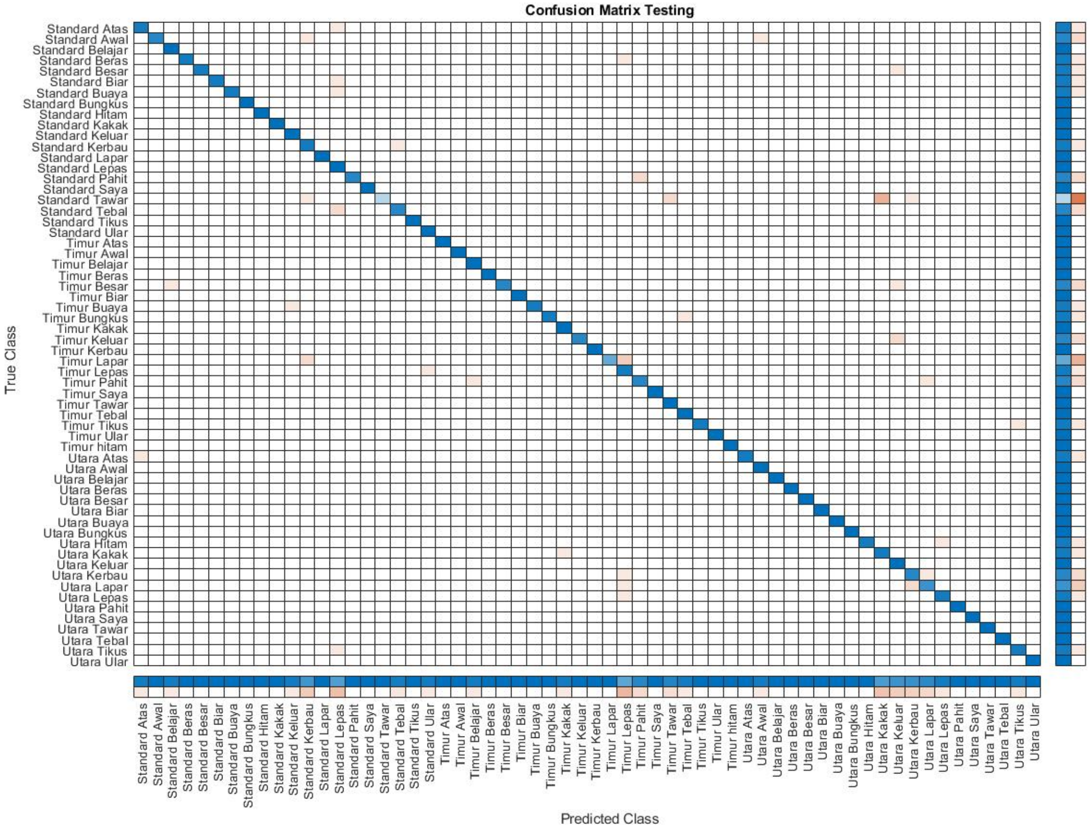


Fig. 9. Confusion Matrix (Testing Set). Blue tones indicate intensity of correct classifcations, while red tones indicate intensity of incorrect classifications

TABLE IV
ACCURACY RESULT FROM THE CONFUSION MATRIX TABLE

Dialect	Standard		Eastern		Northern		Range
Word	Train	Test	Train	Test	Train	Test	
Atas	100	91.67	96.43	100	100	91.67	91.67 to 100
Awal	100	83.33	100	100	100	100	83.33 to 100
Belajar	100	100	100	100	100	100	100
Beras	100	100	100	100	100	100	100
Besar	92.86	91.67	96.43	83.33	100	100	83.33 to 100
Biar	100	91.67	100	100	100	100	91.67 to 100
Buaya	100	91.67	100	91.67	92.86	100	91.67 to 100
Bungkus	100	100	100	100	100	100	100
Hitam	100	100	96.43	100	100	91.67	91.67 to 100
Kakak	96.43	100	96.43	83.33	92.86	91.67	83.33 to 100
Keluar	96.43	100	96.43	100	100	100	96.43 to 100
Kerbau	96.43	91.67	96.43	58.33	92.86	83.33	83.33 to 100
Lapar	96.43	100	92.86	91.67	89.29	75	75 to 100
Lepas	96.43	100	92.86	83.33	100	91.67	83.33 to 100
Pahit	92.86	83.33	100	100	100	100	83.33 to 100
Saya	96.43	91.67	100	100	100	100	91.67 to 100
Tawar	78.57	83.33	100	100	100	100	78.57 to 100
Tebal	89.28	100	96.43	91.67	100	100	89.28 to 100
Tikus	100	100	100	100	96.43	91.67	91.67 to 100
Ular	100	100	100	100	100	100	100
Range	78.57 to 100		92.86 to 100		89.29 to 100		

Fig. 8 and Fig. 9 depicts the confusion matrix of the training and testing sets. The y-axis and x-axis depicts the number of actual and predicted cases. The blue and red tones indicate the number of correct and mis-classification. The high intensity diagonal blue boxes, and low intensity of other boxes indicate high correct classification rates for all classes. The percentage of correct classifications for each word and dialect is shown in Table IV. A majority of correct classifications are above 90%, further confirming the generalization effectiveness of the proposed algorithm for previously seen and unseen cases.

D. *Perfomance analysis*

The precision and accuracy of the classification are shown in Fig. 10 and 12 and the specificity and sensitivity of the classifications are shown in Fig. 11 and 13. The values for the dialects were higher than 90%, indicating high precision and recall for all dialects.

The high accuracy across all dialects suggest that the dialects and words were classified with minimal errors. This indicates that the graphical representation of MFCC (top view representation of the MFCC features) is representative of the dialect features. The results show that MFCC (traditionally represented using continuous-valued cepstral features) can be adapted to graphical form successfully for effective classification by CNN over a wide variety of Malay words and dialects.

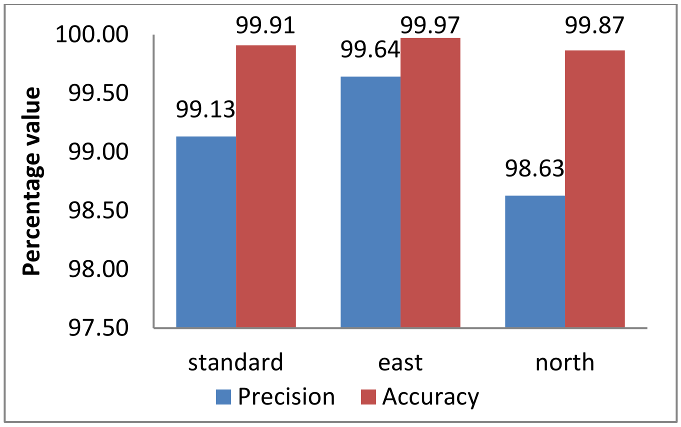


Fig. 10. Precision and accuracy of CNN on dialect training data

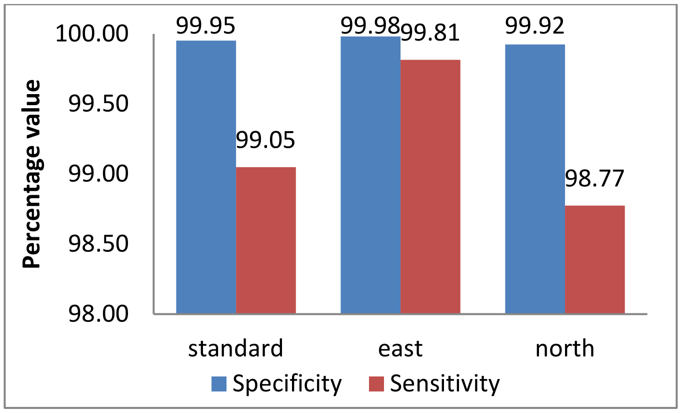


Fig. 11. Specificity and sensitivity of CNN on dialect training data

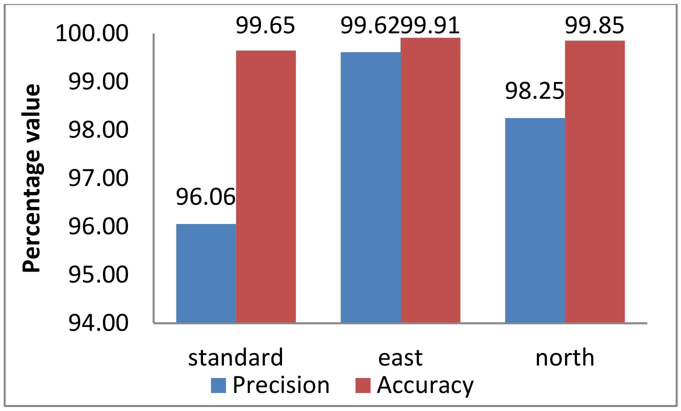


Fig. 12. Precision and accuracy of CNN on dialect testing data

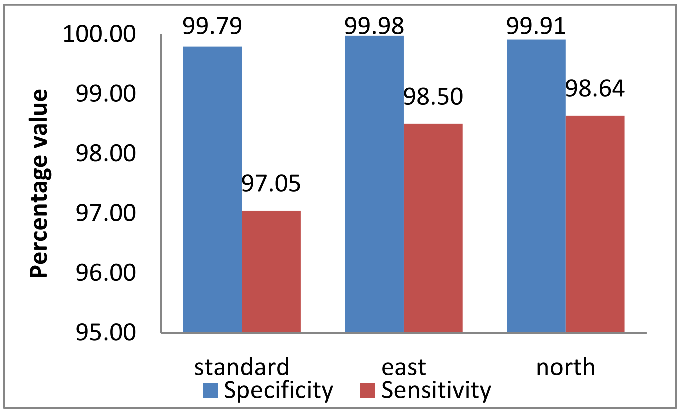


Fig. 13. Specificity and sensitivity of CNN on dialect testing data

IX. CONCLUSION

By conducted systematic literature review shed a few light around the beginning point to investigate in term of SA for Malay dialect. The looked into component offers the clear thought with regard to the different interests, proposed classification method and diverse substance has been utilized. As we see, both result for the training and testing data set show that accuracy and specificity score the highest classification than precision and sensitivity. This result show that CNN network can classify the dialect for the 20 words as resulting 20 classes in the confusion matrix table for 3 dialects.

This also showed that the dialect it self have their own niche pronounciation as we can conclude the standard dialect is the most highest classification because of the dialect is very clear without any special intonation like the eastern dialect and northen dialect. The eastern dialect score more higher than northern where archieved the lowest classification because of their dialect has their own intonation and pronounciation.

X. REFERENCES

[1] T. E. Strahan, “Laurie Bauer, The Linguistics Student’s Handbook. Edinburgh: Edinburgh University Press, 2007. Pp. ix + 387.,” *Nord. J. Linguist.*, vol. 32, no. 1, pp. 165–174, 2009.

[2] O. A. Monographs and R. Blust, *Asia-Pacific Linguistics The Austronesian languages Revised Edition*. 2009.

[3] M. L. Weiss, Ed., *Routledge Handbook of Contemporary Malaysia*. Routledge, 2014.

[4] A. Clynes and D. Deterding, “Standard Malay (Brunei),” *J. Int. Phon. Assoc.*, vol. 41, no. 2, pp. 259–268, 2011.

[5] J. J. Zhai, S. S. Wu, and Y. B. Li, “Research on Synthesis of Speech Parameter and Emotional Speech for Malay Language Using LSTM RNN,” *Proc. - Int. Conf. Mach. Learn. Cybern.*, vol. 2, pp. 598–603, 2018.

[6] D. T. Wurm, Stephen; Mühlhäusler, Peter; Tryon, *Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas*, vol. 51, no. 2. Berlin, New York: DE GRUYTER MOUTON.

[7] T. P. Tan, S. S. Goh, and Y. M. Khaw, “A malay dialect translation and synthesis system: Proposal and preliminary system,” *Proc. - 2012 Int. Conf. Asian Lang. Process. IALP 2012*, pp. 109–112, 2012.

[8] S. M. Zin *et al.*, “Electropalatography Contact Pattern in the Production of Consonant /m/ among Malay Speakers,” *Proc. 2018 7th Int. Conf. Comput. Commun. Eng. ICCCE 2018*, pp. 79–82, 2018.

[9] Q. Zhang and J. H. L. Hansen, “Language/Dialect recognition based on unsupervised deep learning,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 5, pp. 873–882, 2018.

[10] P. Harar, R. Burget, and M. K. Dutta, “Speech emotion recognition with deep learning,” *2017 4th Int. Conf. Signal Process. Integr. Networks, SPIN 2017*, pp. 137–140, 2017.

[11] N. Rusk, “Deep learning,” *Nat. Methods*, vol. 13, no. 1, p. 35, 2015.

[12] Asmah Haji Omar, *Bahasa standard dan standardisasi bahasa Melayu*. .

[13] S. S. Juan, L. Besacier, and T. P. Tan, “Analysis of malay speech recognition for different speaker origins,” *Proc. - 2012 Int. Conf. Asian Lang. Process. IALP 2012*, pp. 229–232, 2012.

[14] M. Pei and F. Gaynor, *A dictionary of Linguistic*. Philosophical Library, 1958.

[15] D. of S. Malaysia, “Department of Statistics Malaysia Press Release,” *Dep. Stat. Malaysia*, no. June, pp. 5–9, 2018.

[16] AH Omar, *The Encyclopedia of Malaysia*. Archipelago Press, 1998.

[17] M. Z. Zaheer, J. Y. Kim, H. G. Kim, and S. Y. Na, “A preliminary study on deep-learning based screaming sound detection,” *2015 5th Int. Conf. IT Converg. Secur. ICITCS 2015 - Proc.*, 2015.

[18] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy, “Ensemble of Feature:based and Deep learning:based Classifiers for Detection of Abnormal Heart Sounds,” *2016 Comput. Cardiol. Conf.*, vol. 43, pp. 621–624, 2017.

[19] “A Comparison Of Deep Learning Methods For Environmental Sound Detection Juncheng Li *, Wei Dai *, Florian Metze *, Shuhui Qu , and Samarjit Das,” *IEEE Int. Conf. Acoust. Speech, Signal Process. 2017*, pp. 126–130, 2017.

[20] R. Rahmawati and D. P. Lestari, “Java and Sunda dialect recognition from Indonesian speech using GMM and I-Vector,” *Proceeding 2017 11th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2017*, vol. 2018-Janua, pp. 1–5, 2018.

[21] P. P. Das, S. M. Allayear, R. Amin, and Z. Rahman, “Bangladeshi dialect recognition using Mel Frequency Cepstral Coefficient, Delta, Delta-delta and Gaussian Mixture Model,” *Proc. 8th Int. Conf. Adv. Comput. Intell. ICACI 2016*, pp. 359–364, 2016.

[22] J. Ibrahim and D. P. Lestari, “Classification and clustering to identify spoken dialects in Indonesian,” *Proc. 2017 Int. Conf. Data Softw. Eng. ICoDSE 2017*, vol. 2018-Janua, pp. 1–6, 2018.

[23] D. Guiming, W. Xia, W. Guangyan, Z. Yan, and L. Dan, “Speech recognition based on convolutional neural networks,” *2016 IEEE Int. Conf. Signal Image Process. ICSIP 2016*, pp. 708–711, 2017.

[24] D. Chamberlain, R. Kodgule, D. Ganelin, V. Miglani, and R. R. Fletcher, “Application of semi-supervised deep learning to lung sound analysis,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2016-Octob, pp. 804–807, 2016.

[25] S. H. Dumpala and S. K. Kopparapu, “Improved speaker recognition system for stressed speech using deep neural networks,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, pp. 1257–1264, 2017.

[26] C. B. Kare and V. S. Navale, “Speech recognition by Dynamic Time Warping,” pp. 12–16, 2015.

[27] T. Zoughi, M. M. Homayounpour, and M. Deypir, “Adaptive windows multiple deep residual networks for speech recognition,” *Expert Syst. Appl.*, vol. 139, p. 112840, 2020.

[28] A. Mawadda Warohma, P. Kurniasari, S. Dwijayanti, Irmawan, and B. Yudho Suprapto, “Identification of Regional Dialects Using Mel Frequency Cepstral Coefficients (MFCCs) and Neural Network,” *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 522–527, 2018.

[29] T.-P. Tan and B. Ranaivo-Malançon, “Malay Grapheme to Phoneme Tool for Automatic Speech Recognition Tien-Ping,” *Third Int. Work. Malay*, pp. 1–6, 2009.

[30] S. Darjaa, R. Sabo, M. Trnka, M. Rusko, and G. Múcskova, “Automatic recognition of slovak regional dialects,” *DISA 2018 - IEEE World Symp. Digit. Intell. Syst. Mach. Proc.*, pp. 305–308, 2018.

[31] R. D. Fonnegra, B. Blair, and G. M. Diaz, “Performance comparison of deep learning frameworks in image classification problems using convolutional and recurrent networks,” *2017 IEEE Colomb. Conf. Commun. Comput. COLCOM 2017 - Proc.*, 2017.

[32] M. Atibi, I. Atouf, M. Boussaa, and A. Bennis, “Comparison between the MFCC and DWT applied to the roadway classification,” *Proc. - CSIT 2016 2016 7th Int. Conf. Comput. Sci. Inf. Technol.*, 2016.

[33] S. Gaikwad, B. Gawali, P. Yannawar, and S. Mehrotra, “Feature extraction using fusion MFCC for continuous marathi speech recognition,” *Proc. - 2011 Annu. IEEE India Conf. Eng. Sustain. Solut. INDICON-2011*, 2011.

[34] Wu Jun, “A speaker recognition system based on MFCC and SCHMM,” pp. 88–92, 2013.

[35] A. Sukhwal and M. Kumar, “Comparative study between different classifiers based speaker recognition system using MFCC for noisy

environment," *Proc. 2015 Int. Conf. Green Comput. Internet Things, ICGCIoT 2015*, pp. 955–960, 2016.

[36] A. Zabidi, W. Mansor, L. Y. Khuan, I. M. Yassin, and R. Sahak, "Three-dimensional particle swam optimisation of Mel Frequency Cepstrum Coefficient computation and multilayer perceptron neural network for classifying asphyxiated infant cry," *ICCAIE 2011 - 2011 IEEE Conf. Comput. Appl. Ind. Electron.*, no. Iccaie, pp. 290–293, 2011.

[37] M. Murugappan, N. Q. I. Baharuddin, and S. Jerritta, "DWT and MFCC based human emotional speech classification using LDA," *2012 Int. Conf. Biomed. Eng. ICoBE 2012*, no. February, pp. 203–206, 2012.

[38] K. Bhattarai, P. W. C. Prasad, A. Alsadoon, L. Pham, and A. Elchouemi, "Experiments on the MFCC application in speaker recognition using Matlab," *7th Int. Conf. Inf. Sci. Technol. ICIST 2017 - Proc.*, pp. 32–37, 2017.

[39] Y. Permanasari, E. H. Harahap, and E. P. Ali, "Speech recognition using Dynamic Time Warping (DTW)," *J. Phys. Conf. Ser.*, vol. 1366, no. 1, 2019.

[40] A. Brahme and U. Bhadade, "Marathi digit recognition using lip geometric shape features and dynamic time warping," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2017-Decem, pp. 974–979, 2017.

[41] B. Kenji, V. Frinken, and S. Uchida, "Knowledge-Based Systems DTW-NN : A novel neural network for time series recognition using dynamic alignment between inputs and weights ☆," *Knowledge-Based Syst.*, no. xxxx, p. 104971, 2019.

[42] H. U. Zhi-qiang, Z. Jia-qi, W. Xin, L. I. U. Zi-wei, and L. I. U. Yong, "Improved algorithm of DTW in speech recognition Improved algorithm of DTW in speech recognition," 2019.

[43] P. Senin, "Dynamic Time Warping Algorithm Review," *Science (80- .).*, vol. 2007, no. December, pp. 1–23, 2008.

[44] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[45] J. Wang and D. Pei, "Kernel-based deep learning for intelligent data analysis," *1st Int. Conf. Electron. Instrum. Inf. Syst. EIIS 2017*, vol. 2018-Janua, pp. 1–5, 2018.

[46] A. Abdalmisreb, A. F. Abidin, and N. M. Tahir, "Maxout based deep neural networks for Arabic phonemes recognition," *Proc. - 2015 IEEE 11th Int. Colloq. Signal Process. Its Appl. CSPA 2015*, pp. 192–197, 2015.

[47] J. Yang and J. Li, "Application of deep convolution neural network," *2016 13th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. ICCWAMTIP 2017*, vol. 2018-Febru, pp. 229–232, 2017.

[48] L. Hertel, E. Barth, T. Kaster, and T. Martinetz, "Deep convolutional neural networks as generic feature extractors," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2015-Septe, 2015.

[49] M. Najafian, S. Khurana, S. Shan, A. Ali, and J. Glass, "Exploiting Convolutional Neural Networks for Phonotactic Based Dialect Identification," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 5174–5178, 2018.

[50] "An Algorithm For Measuring Pterygium ′ S Progress In Already Diagnosed Eyes { rgm , emnf }@ cin . ufpe . br," pp. 733–736, 2012.

[51] M. M. Oo, "Comparative Study of MFCC Feature with Different Machine Learning Techniques in Acoustic Scene Classification," *Int. J. Res. Eng.*, vol. 5, no. 7, pp. 439–444, 2018.

[52] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[53] M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis.," *BMC Genomics*, vol. 13 Suppl 4, no. Suppl 4, p. S2, 2012.

Mohd Azman Hanif Sulaiman received his Diploma in Electrical Electronic Engineering from Politeknik Sultan Salahuddin Abdul Aziz Shah (PSSAAS) and BSc in Electrical Engineering (Power) from Universiti Teknologi Malaysia (UTM) in 2015. His working as assistant lecturer at Faculty of Electrical Engineering, UiTM. His research interest are in Optimization and Artificial Intelligence and he is currently working toward the M. Sc. degree in Electrical Engineering with the Faculty of Electrical Engineering, UiTM as a part time student.

Nurhakimah Binti Abd Aziz received his BSc. And MSc. In Electrical Engineering from Universiti Teknologi MARA (UiTM), Malaysia in 2012 and 2016, respectively. She is a lecturer at the Pre University Department, INTI International College Subang. Her research interests include Signal and Image Processing, Artificial Intelligence (AI) and Machine Learning. Nurhakimah is currently pursuing PhD in Electrical Engineering as a part time student in Faculty of Electrical Engineering, UiTM, Malaysia.

Azlee Zaibidi received his Bachelor of Electrical Engineering (Hons), Master of Electrical Engineering and Doctor of Philosophy in Electrical Engineering from Universiti Teknologi MARA (UiTM). Currently he is a senior lecturer at Universiti Malaysia Pahang (UMP) and hold Professional Technology in Electrical and Amp (Electronics) and also President of Society for Advancement of Science & Technology.

Zuraidah Jantan received his Bachelor of Literature (Linguistic) and M. Sc in social science and humanities (Linguistic) from Univrsiti Kebangsaan Malaysia (UKM) . Currently she is Coordinator for the Department of Malay Studies and Resource Person (RP) for the Introduction of Phonetic and Phonology (Bahasa Melayu) at Universiti Teknologi MARA (UiTM). She also associate member the entity of excellence, LinGKOM.

Ihsan Mohd Yassin received his BSc. in Electrical Engineering (Information Systems) from Universiti Tun Hussein Onn in 2004, MSc and PhD in Electrical Engineering from Universiti Teknologi Mara, Malaysia in 2008 and 2014, respectively. His research interests are in Artificial Intelligence, System Identification, Blockchain Technology and Optimization. Ihsan is an active Senior Member of IEEE, holding various position in the IEEE IA/IE and IEEE CSS Malaysia Section. He is also registered with the Engineering Council, IET as a Chartered Engineer. He is also a registered Professional Engineer with the Board of Engineers Malaysia.

Megat Syahirul Amin Megat Ali received his B. Eng. (Biomedical) from University of Malaya, Malaysia, M. Sc. in Biomedical Engineering from University of Surrey, United Kingdom, and Ph. D. in Electrical Engineering from Universiti Teknologi MARA, Malaysia. He is a senior lecturer at the Faculty of Electrical Engineering, Universiti Teknologi MARA. His research interests include biomedical signal processing and artificial intelligence. Dr. Megat is currently a research fellow at the Microwave Research Institute, Universiti Teknologi MARA.

Farzad Eskandari is a Vice Chancellor of Allameh Tabatabai University Tehran, Iran and also a Professor at Department of Mathematical Science and Computer. His research interest are in Bayesian Statistical Inference, Non-parametric Statistical Inference, Machine Learning and Bayesian Network, Graphical Model Analysis, Data Science and Data Mining Modelling.