World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:13, No:11, 2019

# Neural Network Based Speech to Text in Malay Language

H. F. A. Abdul Ghani, R. R. Porle

Open Science Index, Cognitive and Language Sciences Vol:13, No:11, 2019 publications.waset.org/10010905/pdf

***Abstract*—**Speech to text in Malay language is a system that converts Malay speech into text. The Malay language recognition system is still limited, thus, this paper aims to investigate the performance of ten Malay words obtained from the online Malay news. The methodology consists of three stages, which are preprocessing, feature extraction, and speech classification. In preprocessing stage, the speech samples are filtered using pre emphasis. After that, feature extraction method is applied to the samples using Mel Frequency Cepstrum Coefficient (MFCC). Lastly, speech classification is performed using Feedforward Neural Network (FFNN). The accuracy of the classification is further investigated based on the hidden layer size. From experimentation, the classifier with 40 hidden neurons shows the highest classification rate which is 94%.

***Keywords*—**Feed-Forward Neural Network, FFNN, Malay speech recognition, Mel Frequency Cepstrum Coefficient, MFCC, speech-to-text.

## I. Introduction

SPEECH recognition is the process of extracting essential information from the input speech signal to make precise decisions about the corresponding text. Speech signal carries a lot of information; this encourages many researchers to develop the system that can process the speech automatically such as speech recognition, speech enhancement, and speech synthesis. Moreover, speech recognition can be categorized as speaker dependent and speaker independent. With the help of speech recognition mechanism, computer can follow the human voice commands and recognize human languages [1].

Speech is one of the dominant technologies of communication as speech has become a medium in many applications such as text-to-speech (TTS) that converted text to spoken language, and speech recognition (SR) that have the capability of extracting and interpreting the user speeches [2]. It can be defined as the area of the spoken audio in the audio signal when a user is speaking [3].

There are several applications that applied the capability of speech. The mobile phone is one of the speech applications where a user can mention the name of the recipient in their phone directory making a phone call. Besides, car navigation system can be used to request for directions without setting the destination. SR is useful for users with disabilities such as the users that have problems in spelling and recognizing words [2]. Furthermore, SR can aid the user in interacting with computer without touching a keyboard or a screen.

Speech-to-text system has become more popular nowadays. With the use of SR system, this technology has been applied in many different environments such as in human, robot or animated Avatar and in computer gaming. Regardless of the rapid improvement in this technology, SR for Malay language has been implemented that can be simply combined with other components for controlling task [4].

Fadhilah and Raja [5] indicated that the advance of the SR application offers helpful contributions to this research field in recent years. They also stated that speech would be a better interface than other computing devices used for example keyboard or mouse. Many researches have been conducted for the SR in various languages; nevertheless, not much research has been done for Malay SR [5]. Malay language is the division of the Austronesian (Malayo-Polynesian) language family that is spoken by more than 33 million people across Malaysia and Indonesia [6].

Yogita and Sushima [1] also explained that the SR system in Malay language is still limited and most of the SR technologies are designed for English language. People may be discouraged of using computer technology due to the language barrier. Therefore, the use of native language in the processing of computer will encourage more users to use this technology.

## II. Methodology

### A. Data Input

Ten Malay words obtained from online news are used in the experimentation and Table I shows their English translation. Each word has ten samples, so the total number of samples in this paper is 100. Each of the samples has a duration of 1 second and the sampling rate is set to 44.1 KHz.

TABLE I
Malay Words Used for Datasets

| Malay Words | English Translation | Malay Words | English Translation |
|---|---|---|---|
| *Akan* | will | *Itu* | that |
| *Bagi* | for | *Pada* | when |
| *Dalam* | in | *Tidak* | not |
| *Dan* | and | *Untuk* | for |
| *Ini* | this | *Yang* | whereby |

### B. Preprocessing

In this section filtering is applied to remove the noise in each of the speech signals. Pre-emphasis filter is applied to the Malay words in order to flatten the signal of the Malay words [1]. It is also used to balance the loss of the high frequency

R. R. Porle is with the Universiti Malaysia Sabah, 88400 Kota Kinabalu, Malaysia (corresponding author, phone: +6088-320000 ext 3083; fax: (+6088320348); e-mail: rlyn39@ums.edu.my).

H. F. A. Abdul Ghani was an undergraduate student of Universiti Malaysia Sabah.

component of the Malay words signal [7]. This filter uses (1):

$$y(n) = x(n) - 0.95\,x(n) \quad (1)$$

where y(n)= Output Pre emphasis filter; x(n)= Input speech.

Fig. 1 shows the waveform of the original signals and pre emphasis signal of words "akan1". The amplitude of the word "akan1" has decreased from the original signal after applying pre emphasis filters.

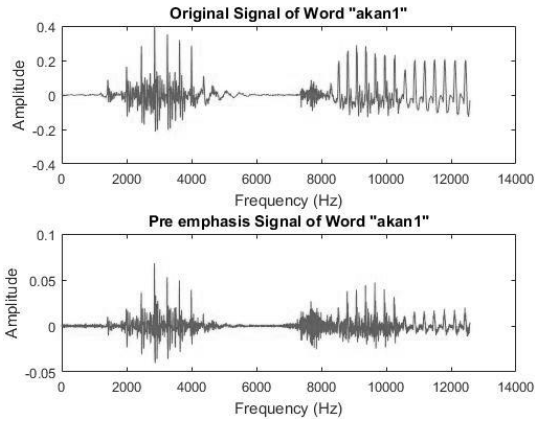



Fig. 1 Waveform of the Original Signal of Word "akan1" and Output Signal of Pre Emphasis Filter

### C. Feature Extraction

Feature extraction is one of the important steps in SR system as the main components of the speech can be identified throughout this process. Besides, by applying feature extraction process, the input speech can be compressed into features [1]. In this research, MFCC was used to extract the feature from the preprocessed data. The flow process of the feature extraction will undergo several steps as shown in Fig. 2.

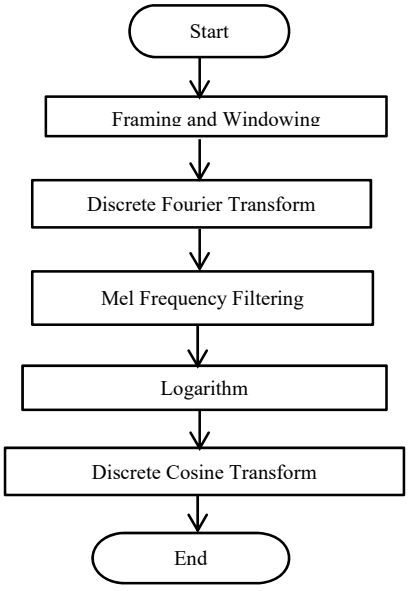



Fig. 2 Flowchart of feature extraction using MFCC

Framing and windowing are the first step that will be applied in the feature extraction process using MFCC. In order to evade any discontinuities in speech segment, the windowing process is applied to the speech signals. Hamming window is the most popular window function in windowing in SR. The Hamming window is applied by using (2). The word "akan1" was divided into 125 frames

$$. \; W_n = 0.54 + 0.46\cos\left(2\pi\frac{n}{N}\right), 0 \le n \le N \quad (2)$$

where $N$ = the number of sample in each frame; $W_n$= nth coefficient of hamming window, where n= 1, 2 ...N.

The Discrete Fourier Transform (DFT) is the next step in MFCCs feature extraction after framing and windowing. This step is useful for the estimation of the spectral coefficients of the speech. Spectral coefficients of speech are complex numbers consist of magnitude and stage. To implement DFT, Fast Fourier Transform (FFT) is used to obtain the magnitude frequency response of each frame.

As the ability of the human ear is limited towards the understanding of frequency contents of speech signals in linear scales (f), Mel scale is used on measuring the pitch where the Mel frequency scale is below 1000 Hz. Equation (3) shows the formula that can be used to transform linear frequency (f) into Mel frequency.

$$Mel(f) = 2595 * log10\left(1 + \frac{f}{700}\right) \quad (3)$$

The next step is to apply logarithm after transforming the linear frequency (f) into Mel frequency. This step is used to approximate the non-linearity in the loudness of the speech.

### D. Classification using FFNN Classifier

For data classification using the Neural Network classifier, the data were classified into ten classes. The target value for each word was set as shown in Table II. Logic numbers were used to set the target value, for instance, if the target value of class 1 was set as "1", the other classes will be set as "0" and so on.

TABLE II
TARGET VALUE OF MALAY WORDS

| Malay Words | Target Value |
|---|---|
| *Akan* | 1000000000 |
| *Bagi* | 0100000000 |
| *Dalam* | 0010000000 |
| *Dan* | 0001000000 |
| *Ini* | 0000100000 |
| *Itu* | 0000010000 |
| *Pada* | 0000001000 |
| *Tidak* | 0000000100 |
| *Untuk* | 0000000010 |
| *Yang* | 0000000001 |

100 samples with 360 feature vectors from each sample were taken from MFCC feature extraction to be used as the input data for the Neural Network classifier. The input data

Open Science Index, Cognitive and Language Sciences Vol:13, No:11, 2019 publications.waset.org/10010905/pdf

Open Science Index, Cognitive and Language Sciences Vol:13, No:11, 2019 publications.waset.org/10010905/pdf

were rearranged in one column for each sample so that the input data can be matched with the target value. Besides, the samples for each word or class were arranged randomly so that the Neural Network can learn more effectively during the training process. 100 samples of the input data were divided into three, for training, validation, and testing data. 80% of the input data will be used for training data, and 10% of the input data will be used for validation data. The other 10% of the input data will be used as testing data.

## III. Results and Discussion

### A. Training Data Using Neural Network Classifier

The sizes of hidden layer or the hidden neuron were tested from 10 to 50 to analyse which hidden layer size can give the best performance. The hidden layer with size equal to 40 was chosen because it shows the highest accuracy compared to others. Table III shows the accuracy of the classification based on the hidden layer size and Fig. 3 shows the structure of the Neural Network, where it contains 360 input and produces 10 outputs.

TABLE III
The Accuracy of the Classification Based on Hidden Layer Size

| Hidden Layer Size | Accuracy (%) |
|---|---|
| 10 | 92 |
| 20 | 90 |
| 30 | 91 |
| 40 | 94 |
| 50 | 90 |

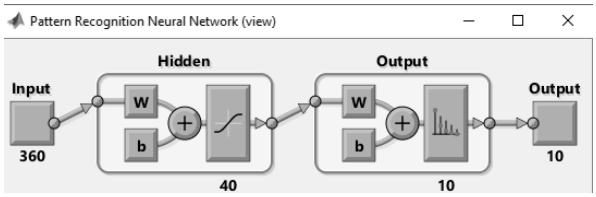



Fig. 3 Structure of Neural Network

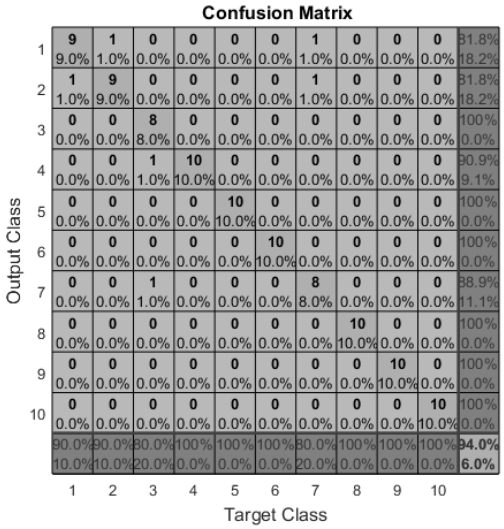



Fig. 4 Confusion Matrix of the Neural Network with Hidden Layer Size Equal to 40

The highest accuracy for the overall data classification using Neural Network is 94% with hidden layer size equal to 40. Fig. 4 shows the confusion matrix of the Neural Network with hidden layer size equal to 40.

Based on the confusion matrix in Fig. 4, the diagonal green cells show the number of samples that correctly predicted for each class. From the confusion matrix, five classes were correctly predicted for all samples which are worded “dan”, “ini”, “itu”, “tidak”, “untuk”, and “yang”. The word “akan” and “bagi” got one sample that was wrongly classified while word “dalam” and “pada” got two samples that were wrongly classified. Overall, 94% of the samples for ten classes were correctly classified while another 6% were wrongly classified.

### B. Testing Data using Neural Network Classifier

After the data samples have been trained using the Neural Network classifier, 11 samples consisting of one sample for each class and additional sample that is not included in the class were used to test the trained data. The sample will undergo testing stage where the loaded sample will be classified based on the class it owned. The class will display parameter “1” if the sample was recognized and if the sample was not recognized, the class will display parameter “0”. The message box will display the samples class and if the sample was not recognized, the message box will display “Unrecognized sound”. All the samples were correctly recognized by its class except for class 3. This is due to poor classification rate for word “dalam” in the trained stage. It also shows that the other sound sample loaded was wrongly recognized with class 6 which is word “itu”. This may cause by the same pitch between the other sound and word “itu”.

## IV. Conclusion

In this paper, the main objective to examine the performance of the conversion speech to text has been achieved. Table III shows the testing results of the datasets using FFNN classification. In the testing stage, nine out of 11 samples were correctly recognized. These may be due to the poor results of trained network data or the loaded samples have the same pitch as the class that was wrongly recognized. For the future work, another type of feature extraction and classification can be used instead of the methods that have been used in this paper to study the performance of the method for speech to text conversion system. Besides, instead of using isolated word as the datasets, one whole Malay sentence can be used as the dataset to widen the scope of speech to text system.

## References

[1] Yogita, H.G. and Sushama, D.S. “Speech to Text Conversion for Multilingual Languages,” *International Conference on Communication and Signal Processing,* 236-240. 2016.

[2] Hanifa, R.M., Isa, K. and Mohamad, S. “Malay Speech Recognition for Different Ethnic Speaker: An Exploratory Study,” *IEEE Symposium on Computer Application & Industrial Electronics (ISCAIE),* 91-96, 2017.

[3] Izzad, M., Nursurianti J. and Zainab A.B. “Speech/Non-Speech Detection in Malay Language Spontaneous Speech,*” IEEE International Conference,* 219-224. 2013.

[4] Rami, A.A. and Rini, A. “Speech to Text Translation for Malay

Language," *IOP Conference Series: Materials Science and Engineering,* 1-9. 2017.

[5] Fadhilah, R. and Raja, N. A. "Isolated Malay Speech Recognition Using Hidden Markov Models," *Proceedings of the International Conference on Computer and Communication Engineering,* 721-725. 2008.

[6] Noraini, S., Zainab, A.B. and Nordin, A.B. "An Evaluation of Endpoint Detection Measures for Malay Speech Recognition of an Isolated Words," *IEEE International Conference,* 1628-1635. 2010.

[7] Alireza, Z., Hua, N.T. and Seyed, M.M. "Gender Classification in Children Based on Speech Characteristics: Using Fundamental and Formant Frequencies of Malay Vowels," *Journal of Voice* 27, 201-209. 2013.

Open Science Index, Cognitive and Language Sciences Vol:13, No:11, 2019 publications.waset.org/10010905/pdf