THE OPTIMAL PERFORMANCE OF MULTI-LAYER NEURAL NETWORK FOR SPEAKER-INDEPENDENT ISOLATED SPOKEN MALAY PARLIAMENTARY SPEECH

Noraini Seman, Zainab Abu Bakar[1], Nordin Abu Bakar[1], Haslizatul Fairuz Mohamed[1], Nur Atiqah Sia Abdullah[1], Prasanna Ramakrisnan[1], Sharifah Mumtazah Syed Ahmad[2]

[1]*Department of Computer Science*
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA (UiTM), Shah Alam, Selangor, MALAYSIA
{aini, zainab, nordin, fairuz, atiqah, prasanna}@tmsk.uitm.edu.my

[2]*Department of System and Network Department*
College of Information Technology
Universiti Tenaga Nasional (UNITEN), Kajang, Selangor, MALAYSIA
smumtazah@uniten.edu.my

Abstract- This paper describes speech recognizer modeling techniques which are suited to high performance and robust isolated word recognition in speaker-independent manner. In this study, a speech recognition system is presented, specifically for an isolated spoken Malay word recognizer which uses spontaneous and formal speeches collected from Parliament of Malaysia. Currently, the vocabulary is limited to ten words that can be pronounced exactly as it is written and controlled the distribution of the vocalic segments. The speech segmentation task is achieved by adopting energy based parameter and zero crossing rate measure with modification to better locates the beginning and ending points of speech from the spoken words. The training and recognition processes are realized by using Multi-layer Perceptron (MLP) Neural Networks with two-layer feedforward network configurations that are trained with stochastic error back-propagation to adjust its weights and biases after presentation of every training data. The Mel-frequency Cepstral Coefficients (MFCCs) has been chosen as speech extraction approach from each segmented utterance as characteristic features for the word recognizer. The MLP performance to determine the optimal cepstral orders and hidden neurons numbers are analyzed. Recognition results showed that the performance of the two-layer network increased as the numbers of hidden neurons increased. Experimental result also showed that the cepstral orders of 12 to 14 were appropriate for the speech feature extraction for the data in this study.

Keywords- *Multi-layer Perceptron, Feedforward, Mel-frequency Cepstral Coefficients, Hidden Neuron, Target vector.*

1. INTRODUCTION

Speech is the most natural way of communication for humans. The aim of speech recognition is to create machines that are capable of receiving speech from humans (or some spoken commands) and taking action upon this spoken information [1]. Although it was once thought to be a straightforward

ISSN 2231-7473

problem, many decades of research has revealed the fact that speech recognition is difficult task to achieve. Several dimensions of difficulty due to the non-stationary nature of speech, the vocabulary size, speaker dependency issues, etc. [1]. However, there have been quite remarkable advances and many successful applications in speech recognition field, especially with the advances in computing technology beginning in the 1980s.

Various methods have been developed to classify and recognize the speech sounds. Multi-layer Perceptron [2-4], Hidden Markov Models [5-7], Recurrent Neural Network [8-9] and Dynamic Time Warping [10-12] are some common methods applied to recognize the speech signal. Neural network has been selected in this study and has many advantages compared to other methods of speech recognizers. It is a non linear computation method that can approximate non linear dynamic system. Secondly, it is robustness to noise. Finally, it has ability to learn [13-15]. Thus, the ability of the neural network in recognizing isolated spoken Malay utterances in a speaker-dependent manner is investigated in this paper. A lot of research has been carried out in adult speech recognition of Malay language [10, 16-17]. This study uses the Malay language, which is a branch of the Austronesian (Malayo-Polynesian) language family, spoken as a native language by more than 33,000,000 persons. The Malay language has been distributed over the Malay Peninsula, Sumatra, Borneo, and numerous smaller islands of the area and widely used in Malaysia and Indonesia as a second language [18].

Experiments are conducted to determine the optimal performance of the neural network in the parameters of cepstral order and hidden neuron number of neural network to recognize isolated spoken Malay utterances. The discussion topics of this study are decomposed into several sections, where Section 2, will explain the Malay speech materials. The details of the methods and implementation of the methods will be described in Section 3. Section 4 describes the results and discussions on the experimental of the feature extraction approaches. Lastly, in Section 5, the paper is ended with conclusions.

2. SPEECH COLLECTION

All experiments are conducted on the whole *hansard* document of Malaysian House of Parliament that consists of spontaneous and formally speeches. *Hansard* document is the daily record of the words spoken in the hearings of parliamentary committees. *Hansard* is not a verbatim (word for word) record of parliamentary business but is a useful record that enables interested people to check what members and senators have said and what witnesses have told parliamentary committees. The document comprises of live video and audio recorded speeches that consists of disturbance or interruption of speakers, and noisy environment from various speakers (Malay, Chinese and Indian). The selection of the data is due to its natural way and spontaneous speaking styles during each of Parliamentary session.

The speech signals that are recorded during the Parliamentary session are in the form of 44100 Hz with 16 bit per second. For the experiments, all the audio files were digitized at a sampling rate of 16 kHz, where the frame size is 256 kbps. Sampling rate of 16000 Hz is a high fidelity microphone, which has the capability of 16 kHz sampling rate of microphone speech [19-20]. This sampling frequency is adequate for complete accuracy and Nyquist rate. In this study, the most frequently words used during eight hours of one day Parliament session are determined. After some analysis, the quantitative information shows that only 50 words are most commonly used by the speakers with more than 25 repetitions. The selection of 50 words are the root words that formed by joining one or two syllables structures (CV/VC – consonant or vowel structure) that can be pronounced exactly as it is written and can control the distribution of the Malay language vocalic segments. However, the vocabulary used in this study consisted of ten words as given in Table 1, due to different selection according to their groups of syllable structure with maximum 25 repetitions and spoken by 20 speakers. Thus, the speech data set

consists of 2500 vocabulary of isolated Malay spoken words. All the signals data will be converted into a form that is suitable for further computer processing and analysis.

Table 1. Selected Malay Words As The Speech Target Sounds

No.	Word	Structure
1	ADA	V + CV
2	BOLEH	CV + CVC
3	DENGAN	CV + CCVC
4	IALAH	VV + CVC
5	KALAU	CV+CVC

No.	Word	Structure
6	ORANG	V+CVCC
7	SAH	CVC
8	SAYA	CV + CV
9	UNTUK	VC + CVC
10	YANG	CVCC

3. METHODS AND IMPLEMENTATION

The general idea towards this study is to generate a speech recognizer for isolated spoken Malay utterances to improve the recognition performance in an offline mode. The overall process of this model is briefly described as block diagram as shown in **Figure 1** below.

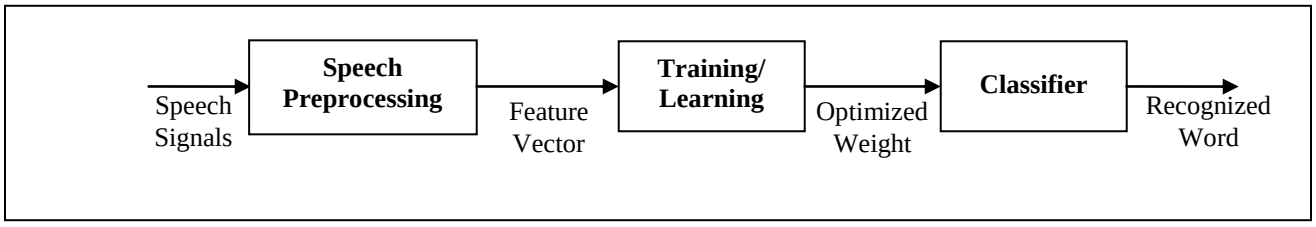

Figure 1. Block Diagram Of Isolated Spoken Malay Recognizer

3.1 Speech Processing

The pre-processing block designed in speech recognition aims towards reducing the complexity of the problem before the next stage start to work with the data. As mentioned above, the digitized speeches will be segmented manually into isolated spoken words according to their root words by using Cool Edit Pro (version 2.0) software. Furthermore, in order to extract the spectral characteristics of the vocabulary words, their short-time Fourier Transform (STFT) magnitudes or spectrograms are investigated since they best express the time-varying nature of the speech signals and combine both the time-domain and frequency-domain information into single, consistent and integrated framework [21]. The amplitude spectrum of speech signal is dominant at low frequencies ranges up to 4 kHz. **Figure 2** illustrates the spectrograms obtained for one trial of each word /SAYA/ and /YANG/ contained in the vocabulary spoken by the same adult male speaker.

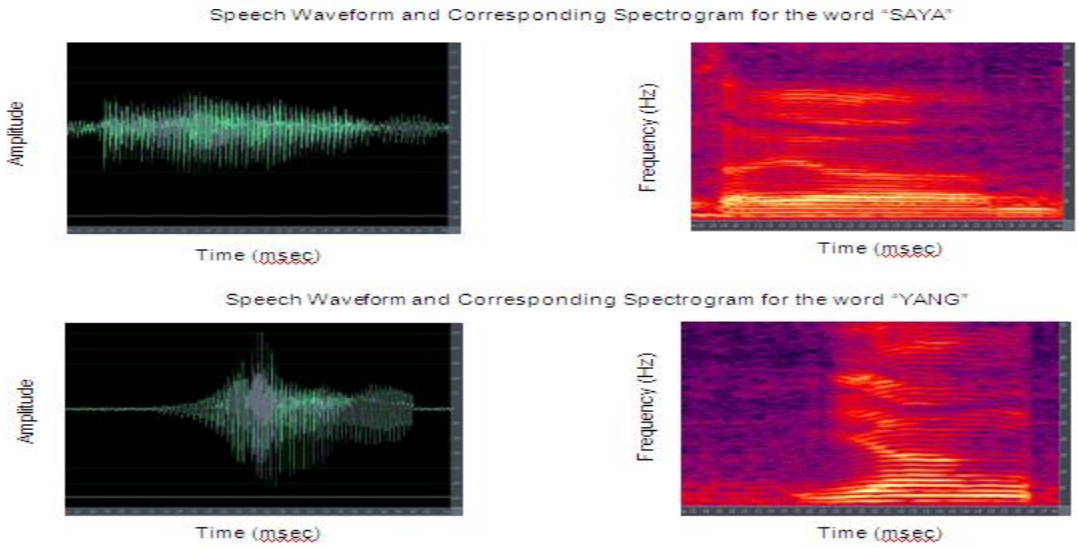

Figure 2. Speech Waveform (Top Plot) And Associated Spectrogram (Bottom Plot) Of The Word "SAYA" And "YANG"

A few important general observations, which is very important for the front-end processing of the data, can be made here:

a) There is a constant very low-frequency disturbance or "hum" present in all recorded speech spectrograms. This disturbance occupies the frequencies band between 0 and 100 Hz.

b) The voiced portions of the speech waveforms, where most of the signal energy is concentrated, have frequency information mainly ranging from 100 up to 2000 – 2500 Hz.

Usually, a one-coefficient simple digital filter, known as a pre-emphasis filter is used. A common form of the pre-emphasis filter is given in [21] as follows:

$$y(n) = s(n) - As(n-1)$$

(1)where $s(n)$ is the speech signal and *A* is typically chosen between 0.9 and 1.0, reflecting the degree of pre-emphasis.

However, recall that **Figure 2** showed that the Parliamentary spoken Malay utterances dampened high frequency components as most of the words energy content was contained in the frequency range between 100 Hz to 2500 Hz, thus, an alternative filtering technique, high-pass filtering, was preferred to mask the low frequency ambient noise. A 6th order infinite impulse response (IIR) elliptic high-pass filter was applied. The filter specifications were as follows:

a) The stopband: 0 - 60 Hz.

b) The passband: 100 – 4000 Hz (half of the sampling frequency).

c) The passband ripple: 0.5 dB.

3.1.1 Speech Endpoint Detection

The problem of locating the endpoints of an utterance in a speech signal is a major problem for speech recognizer. Efficiency of accurate endpoint detection has significant and direct effect on the performance of the entire recognition system [22]. In practice, the process of accurate endpoint detection is not stable and many recognition faults or misclassifications that can be traced back to poor endpoint detection [23]. A popular method for endpoint detection of speech signal was first published by Rabiner and Sambur [22]. The short-time energy (STE) and zero-crossing rate (ZCR) algorithms of speech signals have been extensively used to detect the endpoints of an utterance since then and promising of increased accuracy rate [24]. We adopted STE and ZCR algorithms and modified to better locate the beginning and the termination of speech from the Parliamentary data. This is a two-step search algorithm where the STE for a coarse search is first used. The, ZCR fine-tunes the coarse boundaries expanding forward and backward. The ZCR measure applied in the second search helps to detect low-energy phonemes at the beginning or end of the word, especially when dealing with weak fricative (/f/, /th/, /h/), plosive bursts (/p/, /t/, /k/) or final nasals (/m/, /n/, /ng/). The absolute short-time energy is being chosen as a parameter in short-time energy endpoint detection algorithm due to its simple implementation and efficiency [23]. Therefore, the short-time absolute energy and zero crossing rate can be computed as:

$$E_s = \sum_{n=m-N+1}^{m} |s(n)\, w(m-n)| \tag{2}$$

$$Z_s = \frac{1}{N} \sum_{n=m-N+1}^{m} \frac{|sgn[s(n)]| - sgn[s(n-1)]}{2} \tag{3}$$

where,

$$sgn\big(s(n)\big) = \begin{cases} 1, & s(n) \geq 0 \\ -1, & s(n) < 0 \end{cases}$$

The mean and standard deviation of the short-time energy and zero crossing measures are first computed during the first 50 ms of recording, assuming there is only background noise in that interval. An upper and lower threshold (T_u and T_l) for the short-time energy and another threshold for the zero crossing (T_{zc}) measures are based on these statistics and experimental findings as follows:

$$T_l = 16 \text{ x } MINSTE \tag{4}$$

$$T_u = 32 \text{ x } MINSTE \tag{5}$$

$$MINSTE = \min(IE, \text{mean}\,(STE) + \text{std}\,(STE)) \tag{6}$$

$$IE = 0.25 \tag{7}$$

$$T_{zc} = \min(IF, \text{mean}\,(ZCR) + \text{std}\,(ZCR)) \tag{8}$$

$$IF = 0.25 \text{ x } N \tag{9}$$

where, N is the frame length, *STE* and *ZCR* stand for short-time energy and zero crossing rate, respectively. **Figure 3** shows STE and ZCR measures for a typical utterance of the word /DENGAN/ from one male speaker.

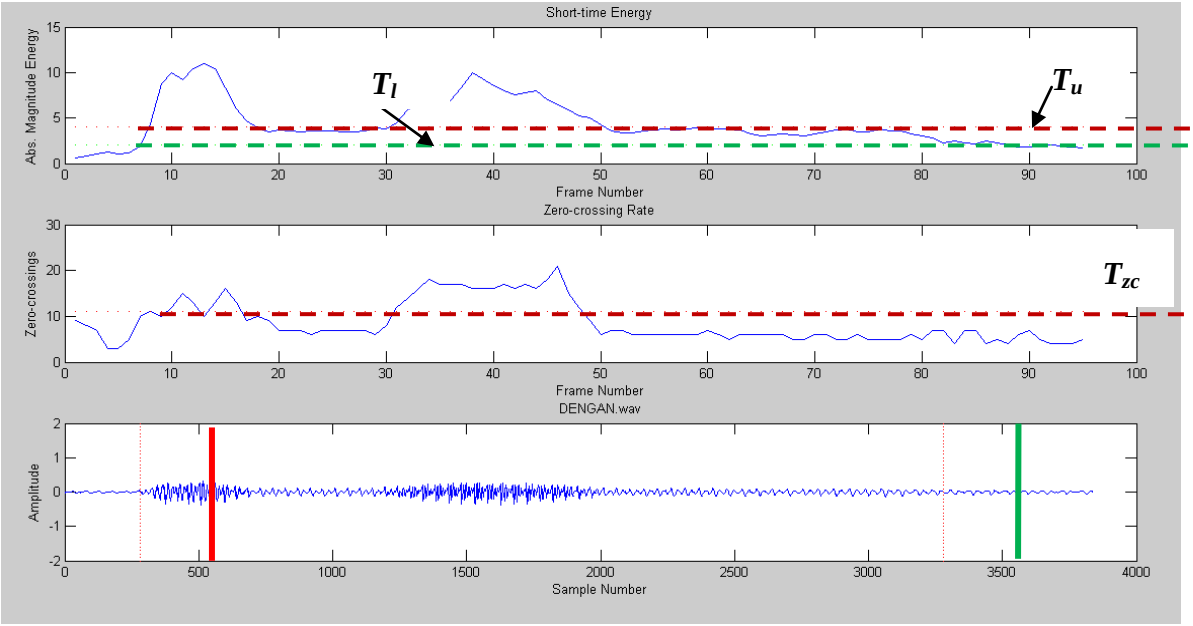

Figure 3. Upper (T_u) And Lower (T_l) Threshold Of Absolute Magnitude Energy Contour (Top Plot) And The Threshold Of Zero-Crossing Rate (T_{zc}) Contour (Middle Plot). Vertical Red And Green Lines Indicate The Beginning And End Points Of The /DENGAN/ (Bottom Plot)

3.1.2 Framing and Windowing

The speech signal is dynamic or time-variant in the nature. According to [5], the speech signal is assumed to be stationary when it is examined over a short period of time. In order to analyze the speech signal, it has to be divided into overlapping frames of *N* samples, with adjacent frames being separated by *M* samples.

$$s_{frame}(n) = s(n).w(n) \tag{10}$$

$$w(n) = \begin{cases} -1, & K.r < n \leq K.r + N, r = 0,1,2,\dots,M-1 \\ 0, & otherwise \end{cases} \tag{11}$$

where *M* is the number of frames, f_s is the sampling frequency, t_{frame} is the frame length measured in time, and *K* is the frame step.

$$N = f_s.t_{frame} \tag{12}$$

We use the f_s = 16 kHz sampling frequency in our system as show in **Table 2** below.

Table 2. Values Of Frame Length And Frame Step Interval Depending On The Sampling Frequency

Sampling frequency (f_s)	f_s = 16 kHz	f_s = 11 kHz	f_s = 8 kHz
Frame length (*N*)	400	256	200
Frame step (*K*)	160	110	80

Then, each frame is windowed to minimize the signal discontinuities at the beginning and ending of each frame or to taper the signal to zero at the beginning and ending of each frame. There are a number of different window functions to choose between to minimize the signal discontinuities. One of the most commonly used for windowing a speech signal is the Hamming window:

$$s_w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \tag{13}$$

The value of the analysis frame length *N* must be long enough so that tapering effects of the window do not seriously affect the result.

3.1.3 Feature Extraction

In broad sense, feature extraction aims for data reduction by converting the input signal into a compact set of parameters while preserving spectral or temporal characteristics of the speech signal information. In this study, Mel-Frequency Cepstral Coefficients (MFCC) has been chosen as speech features approach. The MFCC were first used for a speech recognition system with a dynamic-time warping algorithm (DTW) in a study by [25]. Their study revealed the fact that MFCCs outperform any other parametric representation such as Linear Predictive Coding (LPC) coefficients and Real Cepstrum (RC) coefficients. Since then, MFCC have become the most popular features due to the sensitivity of the low order cepstral coefficients to overall spectral slop and the sensitivity properties of the high-order cepstral coefficient [10]. The block diagram of feature extraction procedures using MFCC is shown in **Figure 4**.

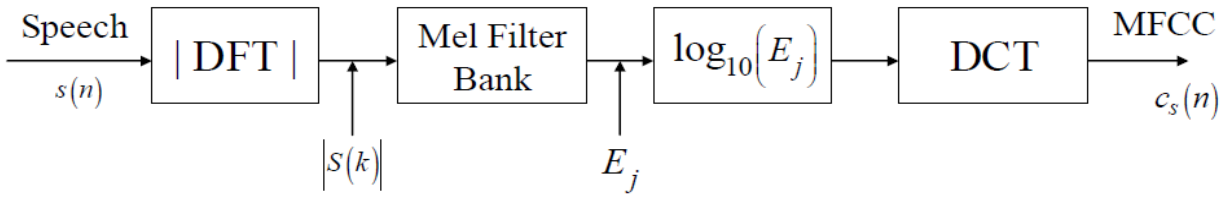

Figure 4. The MFCC Computation As A Block Diagram [26]

First, the spectral magnitude (or energy) of a speech signal or a frame of the speech signal *s(n)* is calculated as:

$$S_i = |S(k)|, \qquad \text{for } i=0,1,\ldots,N/2 \tag{14}$$

where *S(k)* is the *N*-point discrete Fourier transform (DFT) of the speech signal or a frame of the speech sign

$$S(k) = \sum_{n=0}^{N-1} s(n)e^{j2\pi kn/N}, \quad \text{for } k=0,1,\ldots,N\text{-}1 \tag{15}$$

The spectral energy, $|S(k)|^2$ can also be used in Equation (15) instead of the spectral magnitude. Next, the energy in each critical band is obtained by applying the conceptual triangular windows to the spectral magnitude in Equation (15):

$$E_j = \sum_{i=0}^{(N/2)-1} S_i \cdot h_j(i), \quad \text{for } j=1,\ldots,J \tag{16}$$

where J is the total number of triangular filters, h_j (i), used. Finally, MFCCs are calculated as:

$$c_s(n) = \sum_{j=1}^{J} \log_{10}(E_j) \cos\left[n(j+0.5)\frac{\pi}{J} \right] \tag{17}$$

where n is the number of MFCCs to be obtained, generally 8 to 14 of cepstral order [21]. The first coefficient $c_s(0)$ represents the average power in the speech signal. However, $c_s(0)$ is not often used in recognition applications since the average power varies considerably depending on the recording channel. The coefficients $c_s(n)$ give increasingly finer spectral details for each $n > 1$[21]. Experimental results will determine the optimal cepstral order coefficients, excluding the first coefficient that were sufficient to represent the spoken data and selected as feature vectors for classification stage of the recognizer. The cepstral orders were set to vary from 8 to 26 with a step of 2.

3.2 Multi-layer Perceptron (MLP) Neural Network

A two-layer Multi-layer Perceptron (MLP) with one hidden layer and one output layer is applied to recognize the 10 isolated spoken Malay utterances in a speaker-independent manner. The reason we did not evaluate architectures with more than one hidden layer due to two reasons:

a) Any function that can be computed by an MLP with multiple hidden layers can be computed by an MLP with just a single hidden layer, if it has enough hidden units [27]; and
b) Experience has shown that training time increases substantially for networks with multiple hidden layers [14].

The number of input layer will be calculated through the experiments by multiplying the cepstral order with the total frame number.

Total Frame Number = Signal Length/Shift – (Frame Length/Shift-1) (18)

Input Neuron Number = Cepstral order * Total Frame Number (19)

The number of hidden neurons (HNN) has a strong impact on the performance of an MLP. The more hidden neurons a network has, the more complex decision surfaces it can form, and hence the better classification accuracy it can attain. However, the HNN cannot be too many, otherwise, it cannot obtain convergence. If the number is too small, recognition error will be large. Therefore, several experiments were conducted in order to search for the optimal HNN in the hidden layer that varies from 20 to 300 with a step of 20.

The MLP is trained with stochastic error back propagation (BP) in a way to achieve the minimum training error at 0.01 or the maximum iteration of 1000 epochs. The training error is defined as mean square error as in Equation (20).

$$E_{mse} = \frac{1}{Q} \sum_{k=1}^{Q} (t_k - a_k)^2 \tag{20}$$

where Q is the number of training tokens, k is the number of output neurons, t_k is target value and a_k is the actual output. The error information terms, δ_k is calculated at the output layer.

$$\delta_k = (t_k - a_k) a_k (1 - a_k) \tag{21}$$

The δ_k then is used to calculate the weight correction term, Δw_k and bias correction term, Δb_k. The Δw_k will be used to update the connection weight, w_k and the Δb_k will be used to update the bias, b_k.

$$\Delta w_k = \eta \delta_k h \tag{22}$$

$$\Delta b_k = \eta \delta_k \tag{23}$$

where η is learning rate of the neural network and h is the hidden layer. The δ_k is served as delta input to the hidden layer. Each hidden neuron sums its delta input to give

$$\delta\_input_j = \sum_{k=1}^{Q} \delta_k w_k \tag{24}$$

The error information at the hidden layer, δ_j is calculated according to

$$\delta_j = \delta_{input_j}(h_j)(1-h_j) \tag{25}$$

The δ_j then is used to calculate weight correction term, Δw_j and bias correction term, Δbh_j. The Δw_j and the Δbh_j will be used to update w_j and bh_j respectively later.

$$\Delta w_j = \eta \delta_j x \tag{26}$$

$$\Delta bh_j = \eta \delta_j \tag{27}$$

Weights and biases are updated according to the following Equations (28-31).

$$w_k(t+1) = w_k(t) + \eta \delta_k h + \alpha \Delta w_k(t) \tag{28}$$

$$b_k(t+1) = b_k(t) + \eta \delta_k + \alpha \Delta b_k(t) \tag{29}$$

$$w_j(t+1) = w_j(t) + \eta \delta_h x + \alpha \Delta w_j(t) \tag{30}$$

$$bh_j(t+1) = bh_j(t) + \eta \delta_j + \alpha \Delta bh_j(t) \tag{31}$$

where α represents the momentum term. In the experiments learning rate and momentum are initialized at 0.1 and 0.9 respectively.

The hyperbolic tangent sigmoid function (tansig) was selected as the activation function for the hidden neurons, as it provides the necessary nonlinearities in the network to solve the classification problem. The log sigmoid function (logsig) was used for the neurons at the output layer in order to restrict the network outputs to the interval [0 , 1]. These can be calculated as:

$$f_{\tan}(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}, \qquad f_{\log}(z) = \frac{1}{1 + e^{-\alpha z}} \tag{32}$$

Network outputs are continuous between zero and one, as the logsig function is used in the output layer. In order to achieve 1-of-n coding, network outputs were converted to zeros and ones by passing them through a simple maximum detector which assigns one to the maximum output value and zero to the rest. In this study, 1-of- *n* coding was selected for the target representation, where $n = 10$ is the total number of classes. In the 1-of-10 representation, the output of one of the neurons at the output layer which corresponds to one of the ten classes is set to one, with the output of the rest set to zero. For instance, the word "ADA" is labeled as *Class 1*, and its associated target vector is defined as $[\,0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\,]^{T}$. The class number and target vector assignment for each word in the vocabulary are shown in **Table 3** below.

Table 3. Class Numbers And Target Vectors Associated With The Vocabulary Words

Vocabulary Words	Class Number	Target Vector
ADA	1	$[\,0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\,]^{T}$
BOLEH	2	$[\,0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\,]^{T}$
DENGAN	3	$[\,0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\,]^{T}$
IALAH	4	$[\,0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\,]^{T}$
KALAU	5	$[\,0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\,]^{T}$
ORANG	6	$[\,0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\,]^{T}$
SAH	7	$[\,0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\,]^{T}$
SAYA	8	$[\,0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\,]^{T}$
UNTUK	9	$[\,0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\,]^{T}$
YANG	10	$[\,1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\,]^{T}$

The conjugate gradient (CG) algorithm was selected as the backpropagation (BP) learning function. CG has been chosen to speed-up the convergence of the BP algorithm due to CG approach was computationally much faster and led to better classification results [28]. The architectures of two-layer neural networks implemented for the word recognition is shown in **Figure 5** illustrate the standard MLP configuration.

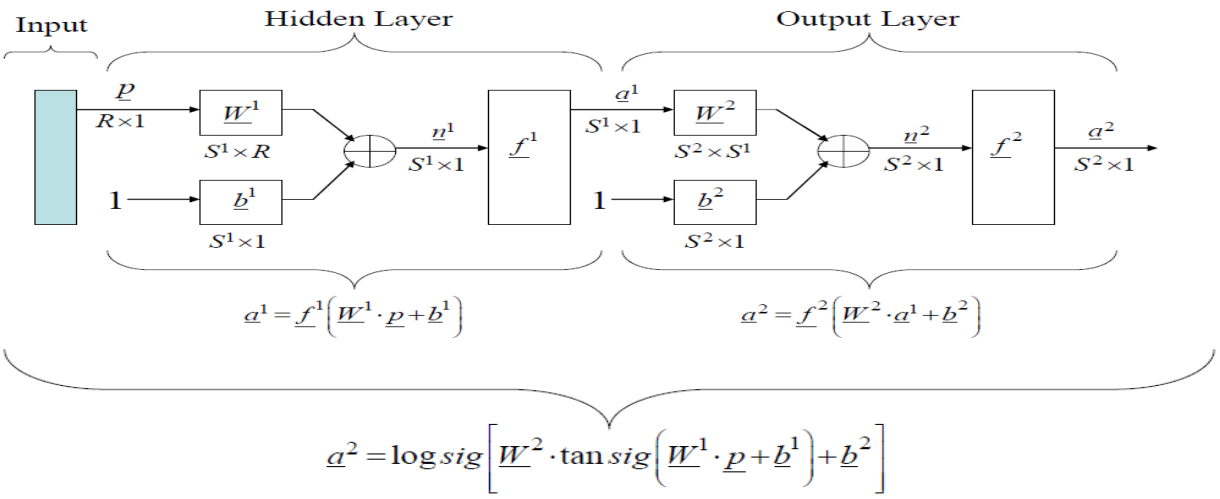

Figure 5. Two-Layer Feedforward Neural Network Architecture Configuration

The network configuration considered in the study was applied to the maximum number of 1000 epochs for network training to obtain statistically meaningful results. The training set was formed by randomly picking 15 repetitions of a word from ten speakers. As a result, the total size of the training set was 1500(10*10*15), since there are 10 speakers and ten words in the vocabulary. The remaining ten repetitions of a word for another 10 speakers were assigned to the testing set for iteration. As a result, the testing set comprises of 1000 vocabulary words.

4. RESULTS AND DISCUSSIONS

The word recognition results obtained with two-layer neural network structure considered is presented in this section. Performance measures of the MLP at different HNN and the optimal cepstral order were examined as shown in **Table 4**. In order to determine the optimal cepstral order, the average recognition rate was calculated over every cepstral order. Results show that the maximum average recognition rate was achieved at cepstral order of 60.80% while the minimum average recognition rate was achieved at cepstral order of 58.44%. Experimental result showed that the cepstral orders of 12 to 14 were appropriate for the speech feature extraction at the sampling rate of 16 kHz for the data investigated. **Table 4** also shows the maximum recognition rate achieved at different HNN. Generally, the accuracy was low at low HNN. The recognition rate increases with the increment of HNN. This is true, when MLP achieved its highest accuracy at HNN of 200 and suggesting that the MLP needed to be trained at higher HNN such as 300 to 400.

Table 4. Performance Measurement For MLP At Different HNN And Cepstral Order Coefficients

Cepstral Order (CO)	**Hidden Neuron Number (HNN)**										
	20	**40**	**60**	**80**	**100**	**120**	**140**	**160**	**180**	**200**	***Average CO (%)***
8	56.95	57.73	57.40	58.45	57.79	58.34	58.45	59.84	59.62	59.79	***58.44***
10	58.84	58.90	59.45	59.40	59.84	60.34	60.40	60.73	61.12	60.62	***59.96***
12	60.29	60.23	59.68	59.84	60.68	61.07	61.23	60.40	61.23	60.73	***60.54***

14	59.90	60.40	59.96	61.35	60.40	60.07	61.85	61.40	61.18	61.51	***60.80***
16	59.90	60.23	60.34	60.73	60.68	60.62	60.79	59.57	60.62	61.51	***60.50***
18	59.45	60.51	60.12	60.46	59.51	60.23	60.34	60.18	59.84	60.46	***60.11***
20	59.23	59.68	59.23	58.58	58.73	57.62	58.40	59.68	59.79	59.84	***59.08***
22	58.73	59.60	59.23	59.23	59.62	59.57	59.84	60.34	60.01	59.79	***59.60***
24	58.84	59.90	59.45	59.40	59.84	60.34	60.40	60.73	61.12	60.62	***60.06***
26	58.03	59.73	59.40	59.45	58.79	58.34	58.45	59.84	59.62	59.79	***59.14***
Average HNN (%)	***59.02***	***59.69***	***59.43***	***59.69***	***59.59***	***59.65***	***60.02***	***60.42***	***60.47***	***60.27***	

Finally, the recognition accuracy and confusion matrix of the optimal performance of the MLP is shown in **Table 5**. The Malay words /YANG/, /SAH/ and /SAYA/ were recognized with the highest accuracy of more than 80%, while Malay words /ORANG/ and /DENGAN/ were recognized with the lowest accuracy of 64% and 65% respectively. Recognition performances become very poor, due to the network is tested on types of data that were not trained on before. However, there might be multiple reasons for this performance degradation due to possible cause may be the gap between the voiced and unvoiced portion at the beginning and ending points of the some words that resulting in incorrect endpoint detection.

Table 5. Recognition Accuracy And Confusion Matrix Of MLP Performance

RECOGNITION ACCURACY (%)											
WORD	**ADA**	**BOLEH**	**DENGAN**	**IALAH**	**KALAU**	**ORANG**	**SAH**	**SAYA**	**UNTUK**	**YANG**	**Other**
ADA	**78**	2	4	1	3	1	2	4	3	1	1
BOLEH	5	**70**	2	3	1	2	2	2	4	9	0
DENGAN	3	8	**65**	2	4	3	3	5	3	3	1
IALAH	2	7	3	**68**	4	3	2	3	4	4	0
KALAU	3	4	4	3	**71**	2	6	2	2	2	1
ORANG	5	4	4	3	2	**64**	5	4	3	5	1
SAH	1	2	3	2	3	4	**82**	0	3	0	0
SAYA	1	0	3	4	3	4	3	**80**	0	1	1
UNTUK	1	1	1	5	5	3	2	3	**75**	3	1
YANG	1	3	2	2	2	1	0	0	3	**86**	0

5. CONCLUSION

In conclusion, the study shows that a two-layer feed forward neural network configuration can be used for isolated spoken Malay word recognition problem. The recognition performance of the multi-layer networks gradually increases as the number of hidden neurons in the hidden layers increases. However, the performance of the multi-layer networks degraded significantly under the testing data types which not be trained into the network before and should consider in expanding the vocabulary used for future experiments. Finally, we believe that there are still a wide variety of issues that can be addressed in order to build a robust and reliable speech recognizer in future.

6. ACKNOWLEDGMENT

This research is supported by the following research grant: E-Science Grant Scheme, Malaysian Ministry of Science, Technology and Innovation, E-Science Fund – 01-01-01-SF0266.

REFERENCES

[1] Hunt, Recurrent Neural Networks for Syllabification, *Speech Communication*, 13(1-2), 159-175, 1993.

[2] B.B. Mosbah, Speech Recognition for Disabilities People, in *Proc. of IEEE International Conference on Information and Communication Technologies*, 1, 864-869, 2006.

[3] Britannica, 2007, Encyclopedia Britannica Online http://www.britannica.com/eb/article-9050292.

[4] C.H. Lin, C.H. Wu, P.Y. Ting and H.M. Wang, Frameworks for Recognition of Mandarin Syllables with Tones Using Sub-syllabic Units, *Speech Communication*, 18, 175-190, 1996.

[5] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing*, 2nd Edition, Prentice Hall, 2008.

[6] F. Rosdi and R.N. Ainon, Isolated Malay Speech Recognition using Hidden Markov Models, in *Proc. of IEEE International Conference on Computer and Communication Engineering*, 721-725, 2008.

[7] G. Cybenko, Approximation by Superpositions of a Sigmoid Function, *Mathematics of Controls, Signals and Systems*, 2, 303-314, 1989.

[8] H. Qiang and Z. Youwei, On Prefiltering and Endpoint Detection of Speech Signal, in *Proceedings of The Fourth International Conference on Signal Processing*, 1, 749-752, 1998.

[9] H. Sakoe and S. Chiba, Dynamic Programming Algorithm Optimization for Spoken Word Recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1), 43-49, 1978.

[10] J. R. G. Deller, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, New York, 1993.

[11] J. Tebelskis, Speech Recognition Using Neural Networks. School of Computer Science, Carnegie Mellon University: PhD. Dissertation, 1995.

[12] L. Deng and D. O'Shaughnessy, *Speech Processing*, Marcel Dekker, New York, 2003.

[13] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition* , Prentice-Hall, New Jersey, 1993.

[14] L. R. Rabiner and M. R. Sambur, An Algorithm for Determining the Endpoints of Isolated Utterances," *The Bell System Technical Journal*, 54, 297-315, 1975.

[15] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, New Jersey, 1978.

[16] M. Ehab, S. Ahmad and A. Mousa, Speaker independent Quranic recognizer Based on Maximum Likelihood Linear Regression, in *Proceedings of World Academy of Science, Engineering and Technology*, 20, 100-106, 2007.

[17] M. T. Hagan, H. B. Demuth and M. H. Beale, *Neural Network Design*, Campus Publishing Service, University of Colorado, Boulder, Colorado, 1996.

[18] Q. Zhu and A. Alwan, Non-linear Feature Extraction for Robust Speech Recognition in Stationary and Non-Stationary Noise, *Computer Speech and Language*, 17(4), 381-402, 2003.

[19] R. Lippmann, Review of Neural Networks for Speech Recognition Neural Computation, 1(1),1-38, 1989.

[20] R. P. Lippmann, An Introduction to Computing with Neural Nets, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 4–22, 1987.

[21] S. B. Davis, and P. Mermelstein, Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, No. 4, pp. 357-366, August 1980.

[22] S.A.R. Al-Haddad, S.A. Samad, A. Hussain and K.A. Ishak, Isolated Malay Digit Recognition Using Pattern Recognition Fusion of Dynamic Time Warping and Hidden Markov Models, *American Journal of Applied Sciences*, 5(6), 714-720, 2008.

[23] T. F. Li, Speech Recognition of Mandarin Syllables, Pattern Recognition, 36, 2713-2721, 2003.

[24] T. Lee and P.C. Ching, Cantonese Syllable Recognition Using Neural Networks, *IEEE Transactions on Speech and Audio Processing*, 7(4), 466-472, 1999.

[25] T. Lee and P.C. Ching, Cantonese Syllable RecognitionUsing Neural Networks, *IEEE Transactions on Speech and Audio Processing*, 7(4), 466-472, 1999.

[26] T.S. Tan, A. Liboh, A.K. Ariff, C.M. Ting and Sh.H. Salleh, Application of Malay Speech Technology in Malay speech Therapy Assistance Tools, in *Proc. of IEEE Conference on Intelligent and Advanced Systems*, 330-334, 2007.

[27] V. Skorpil and J. Stastny, Back-Propagation and K-Means Algorithms Comparison, in *Proc. of 8 IEEE International Conference on Signal Processing*, 16-20, 2006.

[28] Z. Pablo, Speech Recognition Using Neural Networks. University of Arizona: Master Thesis, 1998.