Spoken English Alphabet Recognition with Mel Frequency Cepstral Coefficients and Back Propagation Neural Networks

T.B. Adam
Comp. Science and Info. System
University Technology Malaysia
81300 Skudai, Johor Malaysia

Md Salam
Comp. Science and Info. System
University Technology Malaysia
81300 Skudai, Johor, Malaysia

ABSTRACT
Spoken alphabet recognition as one of the subsets of speechrecognition and pattern recognition has many applications. Unfortunately, spoken alphabet recognition might not be a simple task due to highly confusable set of letters as presented in the English alphabets. The highly acoustic similarities that contribute to the confusability may hinder the accuracy of speech recognition systems. One of the confusable set is called the E-set letters which consist of the letters B, C, D, E, G, P, T, V and Z. In this study, we present aninvestigation of isolated alphabet speech recognition system using the Mel Frequency Cepstral Coefficients (MFCC) and Back-propagation Neural Network (BPNN) for the E-set and for all the 26 English alphabets. Learning rates and momentum rates of the BPNN are adjusted and varied in order to achieve the best recognition rate for the E-set and all the 26 alphabets. By adjusting these parameters,we managed to achieve 62.28% and 70.49% recognition rate for E-set recognition under speaker-independent and speaker-dependent conditions respectively.

General Terms
Digital signal processing, speech processing, speech recognition, artificial intelligence, pattern recognition, human-computer interaction, neural networks, feature extraction, classification, spoken alphabet recognition, acoustic confusable letters.

Keywords
Mel-frequency cepstral coefficients, MFCC, Error back-propagation neural network, E-set

1. INTRODUCTION

Spoken alphabet recognition has been used in benchmarking many speech recognition systems especially to test accuracy of isolated speech recognition systems. Isolated speech recognition system may also use spoken digits to test its recognition accuracy. Spoken alphabet recognition may have several applications among them, automated directory assistance to retrieve information such as spelling names, telephone numbers addresses and ZIP codes [1, 2].Spoken alphabet recognition may be seen as a simple task for human beings but unfortunately, for machines this can be a challenging task due to high acoustic similarities among certain groups of letters[1-3]. High acoustic similarities may cause difficulty in classification while low acoustic similarities causes ease to discriminate among classes for speech recognition systems.An alphabet set which has been identified to be the most confusable for speech recognition is the so called E-set letters. The E-set letters consist of {B, C, D, E, G, P, T, V, Z} [1, 4]. However, there are still other sets of alphabet that may be quite confusable which are pointed out in [1, 3] and these confusability would increase in the presence of background noise.The E-set letters are so called because all the nine letters share the same /iy/ (as in 'E') vowel at the back end of its utterance[4, 5]. These set of letters are difficult to recognize because the distinguishing sound is short in time and low in energy [5].In order to be able to recognize these confusable letters, a need for an accurate classification scheme is a must. The two main steps that will produce such accurate results are either the feature extraction phase or the classification phase. In this paper, we will study the ability of Feed Forward Back Propagation (FFBP) with adaptive learning rate (FFBPALR) to classify the highly confusable E-set letters.Other experiments that have been conducted in the past seem to propose an improved method of an existing classifier or feature which is not the objective of this paper. The difference of this paper from others is that we try to use an already developed technique (FFBPALR) as the classification scheme and the mel-frequency cepstral coefficients (MFCCs) as the speech feature. We then experiment on some parameters of the FBPALR to gain increased recognition accuracy namely the learning rate and the momentum constant.

2. SOME EXISTING WORK

Previous work that have been using spoken alphabet recognition include that of [1, 3] while focused E-set experiments were conducted in [4, 6]. In [1] authors proposed a high performance alphabet recognition based on context-dependent phoneme hidden Markov models (HMM's) to address the problem by E-set letters and confusion caused by nasals (letters M and N). Here, phoneme HMM's were developed and tested against word based HMM's. As a final result, they achieved 95% recognition rate for speaker independent E-set recognition and overall alphabetrecognition of 97.3%.The EAR (English Alphabet Recognizer) system was described in [3, 7] that performs recognition of isolated alphabets.The EAR system first used a rule-based segmenter to segment the alphabets into four broad phonetic categories. Then features are extracted from these broad phonetic categories. The classification used back propagation neural

network (BPNN) with conjugate gradient optimization consisting of 617 input nodes, 52 hidden layer and 26 output nodes. The EAR system achieved high recognition rate of 96% for speaker-independent letters recognition.

Signal modeling for high performance and robust isolated word recognition were proposed in [2, 8]. The authors proposed a new technique for incorporating temporal and spectral feature within each word. The features computed by the proposed technique were then implemented using HMMs. Results showed an accuracy of 97.9% for speaker-independent isolated alphabet recognition. Experiments were also conducted for speech under additive Gaussian noise (15dB) and telephone speech simulation and achieve a recognition rate of 95.8% and 89.6% respectively.

In [6] compound wavelets were tested for speech recognition of the E-set letters. The speech signal is parameterized with compound wavelets and then a HMM's based recognizer was used for classification. Experiments were conducted by varying the compound level to note the increase in recognition rate. The best recognition rate obtained was 71.4% at compound level 4. However, tests were not conducted to observe the method under noisy environments.

Another proposed technique to outcome the confusability problem by the E-set letters was proposed in [4]. In their paper, the authors presented a technique to overcome the problem by means of time-extended features. The idea was to expand the duration of the consonants in order to gain high characteristic difference between confusable pairs in the E-set letters. To test the proposed technique, a continuous densitiy HMM (CDHMM's) were used as the classifier and the best results showed a recognition rate of 88.72%. Nevertheless, no tests were done for noisy speech.

In conclusion, we can say that some of the disadvantage of phoneme based recognizers as in [1] when compared to word based recognizer is complexity of the system and the word transcription must be known [2]. Also, from our review, many of these testing and experiments were done using HMMs and modified techniques. It is also hard to assume which proposed method may be superior to others as the experiments were done in different environments with different speech databases.

This work aims to observe the highest recognition possible by purely using mel-frequency cepstral coefficients and artificial neural network and their ability in recognizing confusable acoustic similarities presented in the E-set letters.

3. THE SPEECH RECOGNITION FRAMEWORK

Generally, almost all speech recognition (SR) system consist the following steps: Signal pre-processing, feature extraction and classification. These three steps are the most common in any SR system. Other auxiliary steps may be performed depending on the intended application of the system. In this section we briefly describe the three steps.

3.1. Signal Pre-processing

Pre-processing of a signal can be said as applying any required form of processing to the signal in time domain before the feature extraction phase.Normally, in the pre-processing stage the speech signal undergoes several common processes including analog to digital (A/D) conversion, enhancement, pre-emphasis filtering and usually for SR applications silence removal or end point detection (EPD).

The A/D process converts a sound pressure wave into its digital form. There are three steps in the A/D conversion process which is sampling, quantization and coding. The final product of this process is a digital version of the speech signal that can be processed by a computer.

In speech recognition and speech processing in general, speech enhancement is also conducted to remove or suppress unwanted noise from the speech signal. For SR application removing noise may increase the accuracy of the recognizer.In almost all SR application a pre-emphasis filtering step is conducted to the speech signal.

The pre-emphasis filter is used to emphasis the speech spectrum above 1 kHz which contains important aspects of the speech signal and equalizes the speech propagation trough air [9, 10].

3.2. Feature Extraction

One of the most important steps in a SR system is extracting certain important information from the speech signal. Feature extraction could be seen as extracting certain mathematically parameterized information from the original source signal.

There are many feature extraction techniques that may be used. Example includes fast fourier transform (FFT) coefficients, perceptual linear prediction (PLP), linear predictive cepstral coefficients (LPCC) and mel-frequency cepstral coefficients (MFCCs). In this investigation, we have opted to use MFCCs as the features.

3.2.1. Mel-Frequency Cepstral Coefficients

The mel-frequency cepstral coefficients (MFCCs) introduced by Davis and Mermelstein is perhaps the most popular and common feature for SR systems [11]. This may be attributed because MFCCs models the human auditory perception with regard to frequencies which in return can represent sound better [12]. Figure 1 shows the block diagram of the MFCCs.

To obtain the MFCCs of a speech signal, the signal is first subjected to pre-emphasis filtering with the following finite impulse response (FIR) filter given by [10] as;

$$H_{pre}(z) = \sum_{k=0}^{N} a_{pre}(k) z^{-k} \quad (1)$$

Its corresponding Z-transform;

$$H_{pre}(z) = 1 + a_{pre} z^{-1} \quad (2)$$

The value of the coefficient a_{pre} usually takes the value between -1.0 to -0.4. However, in speech recognition systems values that are almost near to -1.0 are usually used[10]. The speech is processed on a frame-by-frame basis in what is called framing. Normally, a frame size of 20ms to 30ms is used and windowing of these frames are done to compensate discontinuities within the speech signal as a result of segmentation and overlapped frames. A hamming window is used by equation (3);

$$w(n) = 0.54 + 0.46\cos(\frac{2\pi t}{T}) \quad (3)$$

Windowing means multiplying the window function $w(n)$ with the framed speech signals$s(n)$ to obtain the windowed speech signal $s_{0w}(n)$;

$$s_{0w}(n) = s(n)w(n) \quad (4)$$

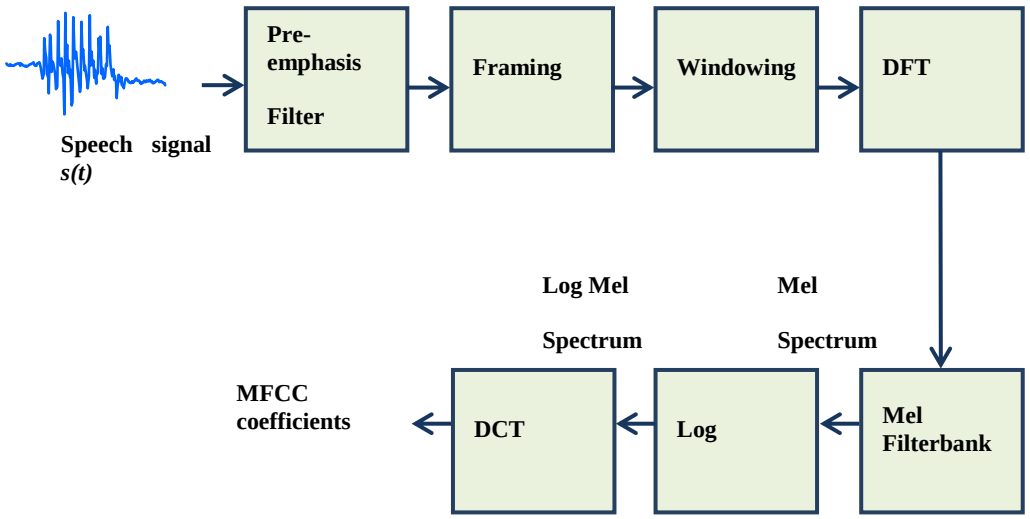

Figure 1: MFCC block diagram

The discrete Fourier transform (DFT) of the windowed speech signal is then computed by the following equation;

$$\hat{S}_{0w}(k) = \sum_{n=0}^{N-1} s_{0w}(n) e^{\frac{-j2\pi kn}{N}} \quad (5)$$

The mel-filterbankis a triangular bandpass filter which is equally spaced around the Mel-Scale. A Mel is a unit of perceived pitch or frequency of a tone.The mapping between real frequency (Hz) and Mel frequency is given by the following equation as;

$$f_{mel} = 2595 \cdot \log\left(1 + \frac{f}{700}\right) \quad (6)$$

The power spectrum from the DFT step is then *binned* by correlating it with each triangular filter in order to reflect the frequency resolution of the human ear.Binning here means multiplying the power spectrum coefficients with the triangular filter gain or coefficients and summing the resultant values to obtain the mel-spectral coefficients[13] as in equation (7) ;

$$G(k) = \sum_{n=0}^{N/2} \mathfrak{y}_{kn} \cdot \left|\hat{S}_{0w}(k)\right|^2 \quad (7)$$

Where $\mathfrak{y}_{kn}$ is the triangular filter coefficients, $k = 0, 1, 2, \ldots k-1, n = 0, 1, 2, \ldots \frac{N}{2}$ and $G(k)$ is the mel-spectral coefficients.

After that, The log of the mel-spectral coefficients $G(k)$, is taken. This step is to smooth unwanted ripples in the spectrum and done by the following equation;

$$m_k = \log G(k) \quad (8)$$

Finally, DCT is applied to the log mel-cepstrumm_k as in equation (9) to obtain the Mel-Frequency Cepstral Coefficients (MFCC) c_i of the ith frame;

$$c_i = \sqrt{\frac{2}{N}} \sum_{k=1}^{N} m_k \cos\left(\frac{\pi i}{N}(k - 0.5)\right) \quad (9)$$

3.3. Classification

The final part in an ASR system is the classification stage.This stage involves classifying the input speech (test signal) to determine whether the input speech uttered matches the desired targeted speech. Some of the categories of classification schemes are statistical and artificial intelligence approaches. In this study, we chose neural networks (NN) as one of the artificial intelligence approach.

3.3.1. Recognition using Artificial Neural Networks

Neural networks (NN) are parallel distributed information processing structure with processing elements connected through unidirectional signal channels called connections[14].ANNs consist of simple interconnected processing elements that are called neurons that perform weighted summation of inputs.

The NN model used for this experiment is the back-propagation neural network with adaptive learning rate (BPNNALR). The scheme for training the network is back-propagation with mean squared error as in equation (10);

$$E = \frac{1}{2} \sum_{l=1}^{L} [y_d(p) - y(p)]^2 \quad (10)$$

Where $y_d(p)$, $y(p)$is the desired and actual output at neuron in the network. Updating the weights in the network is done through minimizing the error E. The weight updating is done by equation (11);

$$W_{jk}(p+1) = W_{jk}(p) + \Delta W_{jk}(p) \qquad (11)$$

Where W_{jk} is the weight connection of neuron k at the output layer to neuron j in the hidden layer. The term $\Delta W_{jk}(p)$ is the weight correction and is defined as;

$$\Delta W_{jk}(p) = \eta y_j(p)\delta_k(p) + \alpha \Delta W_{jk}(p-1) \qquad (12)$$

In which η is the learning rate, $\delta_k(p)$ is the error gradient at neuron k of iteration p and α is a constant called momentum rate. Figure 2 shows the neural network structure. In the hidden layer and output layer nodes an activation function is used. Typical functions used are the sigmoidal, hyperbolic tangent and the linear function.

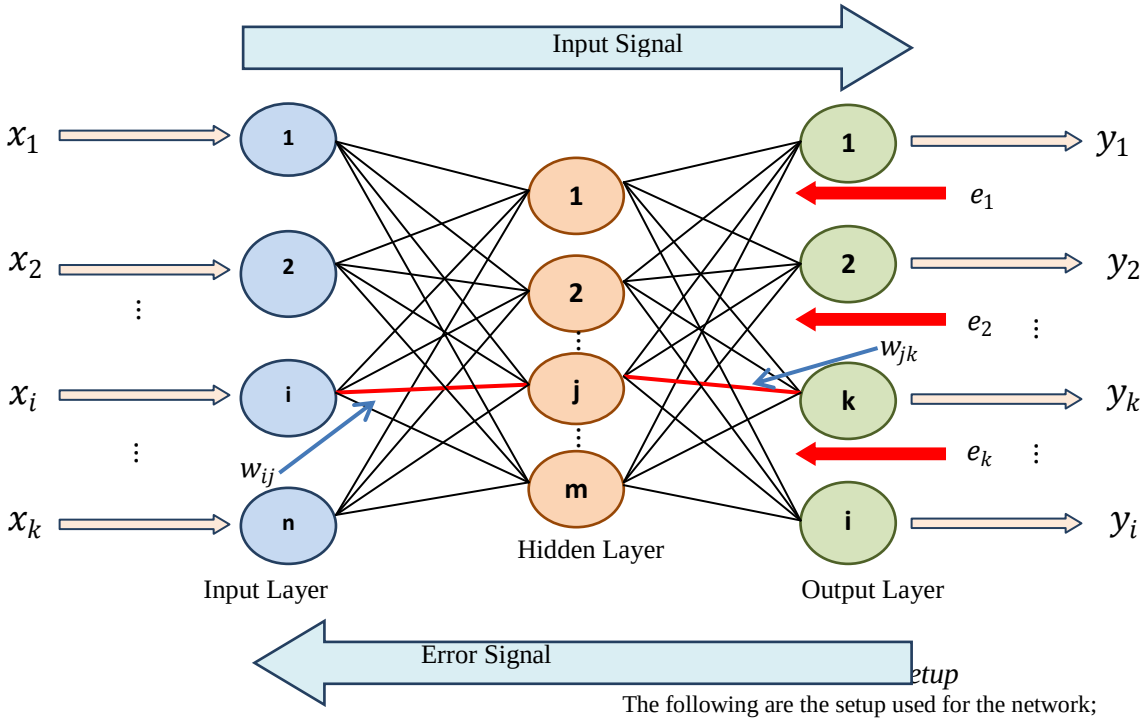

Figure 2: NN topology

4. METHODOLOGY

In this section we discuss the steps taken to conduct the experiments. Experimental setups and steps will be present in in this section.

4.1. Experimental Setup

The Experiment conducted uses Matlab 2009b with Neural Network toolbox and Speech and Audio Processing (SAP) toolbox from[15]. In this section, speech data, MFCC setup and Neural Network setup used for the experiment are presented.

4.1.1. Speech Data

In this study the speech data(wave files '.wav') are taken from the TI46 database isolated alphabet called TI ALPHA. The TI ALPHA consists of eight male and female speakers. The files were further divided into training and testing sets. For training, there are 16 patterns for each alphabet A to Z while for testing there are 10 patterns for each alphabet.

4.1.2. MFCC Setup

The MFCC features were computed with the following parameters;

- Pre-emphasis coefficients (a_{pre}) = -0.95
- Frame size = 256 samples (16ms)
- Frame overlap = 85 samples (5.3ms)
- Number of Triangular bandpass filters = 20
- Number of MFCCs = 12

etup

The following are the setup used for the network;

- Input layers = 400 nodes
- Hidden layers = 150 nodes
- Output layers = 26/9 nodes
- Hidden layer transfer function = Hyperbolic tangent
- Output layer transfer function = Linear

The input nodes are fixed to 400 by means of zero padding [16] the MFCC features while the output nodes are set to 26 in order to recognize all the 26 alphabets A to Z. For the E-set recognition, output nodes are set to nine in order to recognize the nine letters of the E-set.

The hidden layers are chosen by using the formula $h = \sqrt{n \times m}$[17]where n and m are the. number of input nodes and output nodes respectively. By using the formula, we found the hidden layer should be

around 102. However, this is only used for our initial guess, we increased the hidden layer to 150 based on this value. Increasing the number of hidden layer may help the network learn more complex problems.

Table 1: Parameter settings

Set	Learning rate (η)	Momentum rate (α)	Increment multiplier (η^+)	Decrement multiplier (η^-)
Set 1	0.25	0.5	1.05	0.7
Set 2	0.5	0.75	1.05	0.7
Set 3	1.0	0.9	1.05	0.7
Set 4	0.1	0.9	1.05	0.7

Table 1 shows the setting for learning rate (η) and momentum (α) rate variation for this experiment.

4.2. Procedure

In the training phase, all of the wave files from speaker F1, F2, F3, F4, F5, M1, M2 and M3 (F = female, M = male) of the TI ALPHA database are gathered in one folder. Each of the speakers utters the 26 alphabets 10 times. Thus, for eight of the speakers we have $10 \times 26 \times 8 = 2080$ files/pattern of speech for the vocabulary size. MFCCs will be extracted from these files for training. For the E-set, vocabulary size is 720.

Before extracting the MFCCs the speech signals are subjected to end point detection (EPD) to remove the silence before and after the voiced region as in figure 3.

The speech signals were also subjected to normalization by [18];

$$S_{Pi} = \frac{S_i - \mu}{\sigma} \qquad (13)$$

Where S_i is the i^{th} element of the signal S, μ and σ are the mean and standard deviation of vector S. After these pre-processing were done the MFCCs were computed and normalized between -1 and 1.

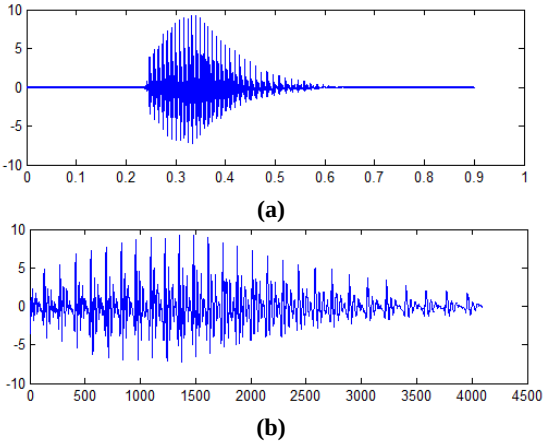

Figure 3: A speech signal (a) Original signal (b) After EPD

5. RESULTS

In this study, our objective was to evaluate the effectiveness of the BPNNALR in classifying the letters by adjusting the learning rate and momentum rate. In order to evaluate the effectiveness, we observed the recognition rate as the main indicator for this purpose.

5.1. Results for All Alphabets

The experiment was conducted for both speaker dependent and speaker independent. For speaker-dependent speakers for training and testing are the same however, for speaker-independent testing speakers are different form training.Figure 4 and figure 5 show the average recognition rate (RR) for all 26 alphabets A to Z.

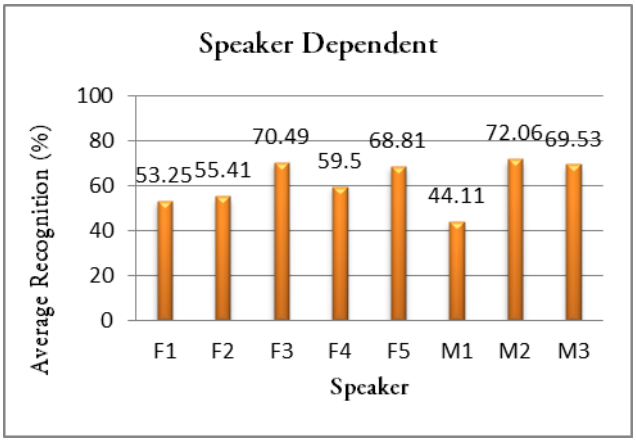

Figure 4: Recognition rates for speaker dependent

In order to observer which pair of learning rate and momentum rate that achieved the best recognition rate we plotted the average recognition rate achieved by the four settings. Figure 6 and figure 7 shows the results obtained.

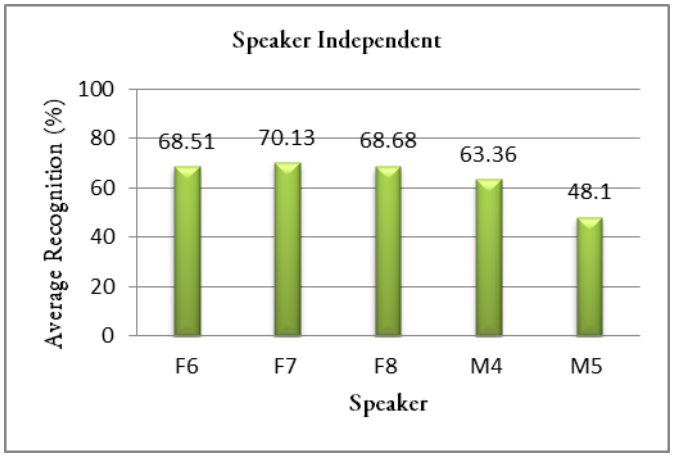

Figure 5: Recognition rates for speaker Independent

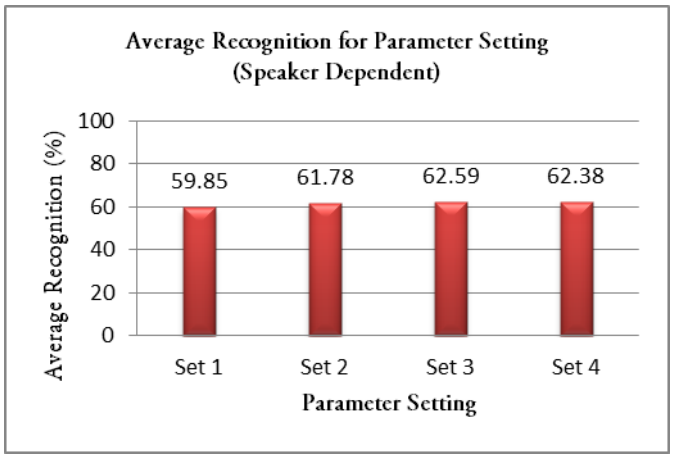

Figure 6: Speaker dependent average recognition rates achieved for each parameter

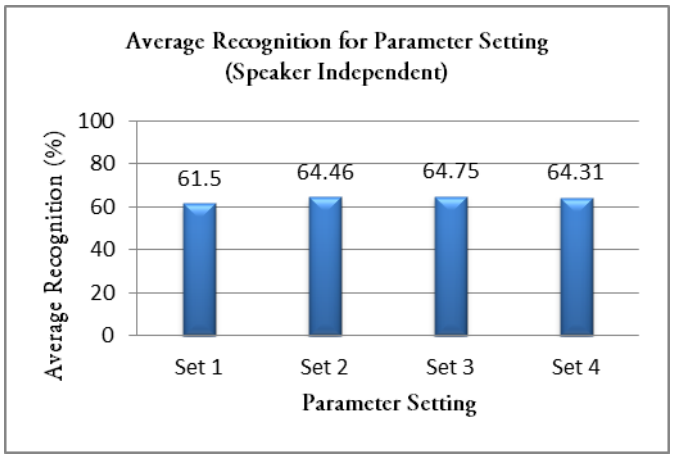

Figure 7: Speaker Independent average recognition rates achieved for each parameter

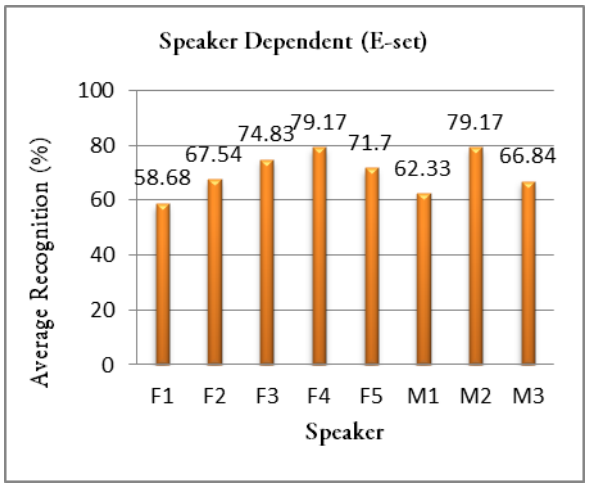

Figure 8: Speaker dependent average recognition rates for E-set

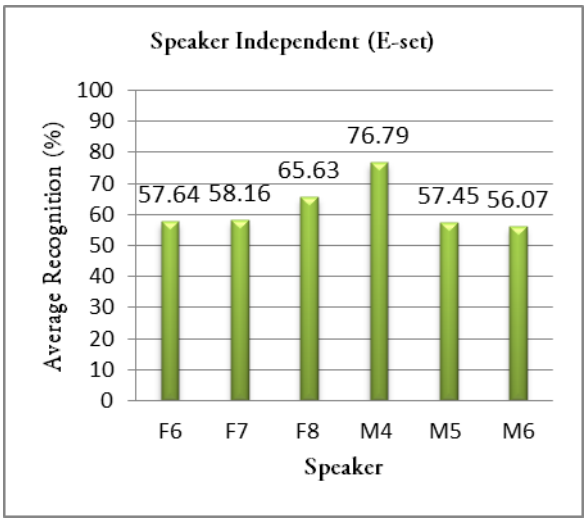

Figure 9: Speaker independent average recognition rates for E-set

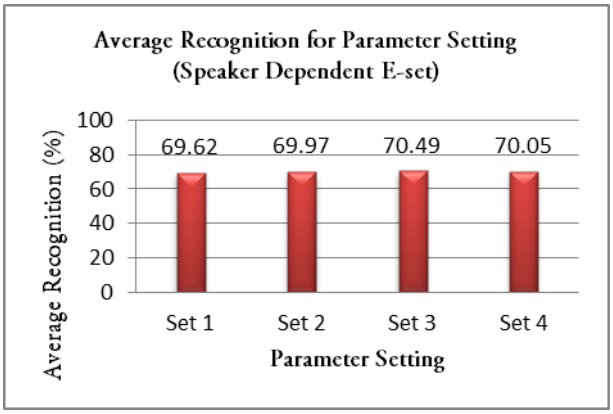

Figure 10: Speaker dependent average recognition rates achieved for each parameter for the E-set

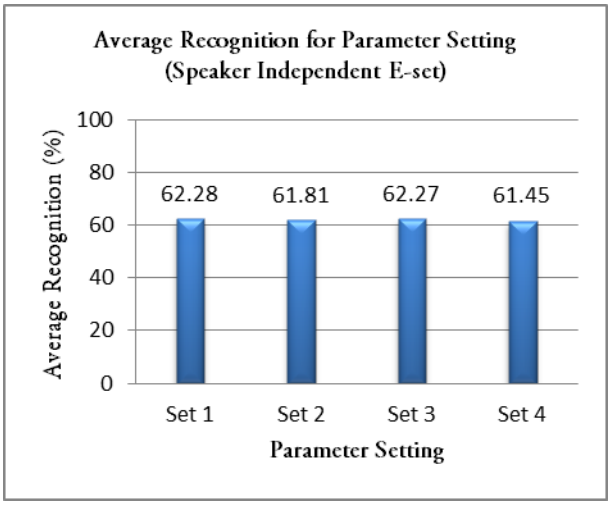

Figure 11: Speaker Independent average recognition rates achieved for each parameter for the E-set

5.2. Results for E-set Alphabets

Tests were also done for the E-set alphabets in the same manner. We tested for speaker-dependent and speaker-independent. Figure 8 and figure 9 shows the results obtained.

6. DISCUSSION

For recognition of all the 26 alphabets, it is shown that the BPNNALR did not achieve significant overall recognition for either speaker dependent or speaker independent. This may be attributed because of the vocabulary size. In this case, 2080 patterns all together. This could be seen for both speaker dependent and speaker independent tests (figure 4 and 5). Recognition rates never exceeded 73% for both tests.Another reason that may be attributed to the low recognition rate may be related to the E-set utterances present in the 26 Latin alphabets A to Z.

By varying the parameters of the learning rate and momentum rate (Table1) we managed to increase the average recognition rate from 59.85% to 62.59% (Figure 6) for the speaker dependent tests. While for speaker-independent, an increase from 61.5% to 64.75% was managed to achieve.

For the E-set results, the highest accuracy achieved for speaker-dependent was 79.17% (Figure 8) which is quite reasonable for a confusable vocabulary of letters. Meanwhile, the highest recognition dropped to 76. 79% (Figure 9) for the speaker-independent test.

Although the E-set is acoustically confusable, the reduce in vocabulary size from 2080 to only 720 might be the reason for the acceptable accuracy.

Variation in learning rate and momentum rate showed no significant increase in recognition rates as shown in figure 9 and figure 10.

As for the parameter setting (Table 1) in can be observed that the learning rate and momentum rate of setting 3 results in the best pair for the BPNNALR for this experiment.

7. CONCLUSION

In this experiment we conclude that the BPNNALR can be useful in speech recognition system with small vocabularies.

The highly confusable E-set presented in the 26 alphabets effects the accuracy of the system. In order to increase the accuracy and the classification power of the BPNNALR we varied two parameters and found that by varying these parameters only a small increase can be achieved however, the best pair of learning rate and momentum rate was found to be {1.0, 0.9}.

Finally, we recommend that further studies be conducted with hybrid classification schemes to investigate the English alphabet recognition and the E-set alphabets.

8. REFERENCES

[1] P. C. Loizou and A. S. Spanias, "High-Performance Alphabet Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 430-445, 1996.

[2] M. Karnjanadecha and S. A. Zahorian, "Signal Modeling for Isolated Word Recognition," presented at the Proceedings of the Acoustics, Speech, and Signal Processing (ICASSP), 1999.

[3] R. Cole, M. Fanty, Y. Muthusamy, and M. Gopalakrishnan, "Speaker-Independent Recognition of Spoken English Letters," in *International Joint Conference on Neural Networks (IJCNN)*, 1990, pp. 45-51

[4] M. D. Ibrahim, A. M. Ahmad, D. F. Smaon, and M. S. H. Salam, "Improved E-set Recognition Performance using Time-Expanded Features," presented at the Second National Conference on Computer Graphics and Multimedia (CoGRAMM), Selangor, Malaysia, 2004.

[5] K. J. Lang, A. H. Waibel, and G. E. Hinton, "A Time-Delay Neural Network Architecture for Isolated Word Recognition," *Neural Networks*, vol. 3, pp. 23-43, 1990.

[6] R. F. Favero, "Compound Wavelets: Wavelets for Speech Recognition," in *International Symposium on Time-Frequency and Time-Scale Analysis*, 1994, pp. 600-603.

[7] M. Fanty and R. Cole, "Spoken Letter Recognition," presented at the Proceedings of the conference on Advances in neural information processing systems Denver, Colorado, United States, 1990.

[8] M. Karnjanadecha and S. A. Zahorian, "Signal Modeling for High-Performance Robust Isolated Word Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 647-654, 2001.

[9] M. E. Ayadi, M. S.Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, pp. 572-587, 2011.

[10] J. W. Picone, "Signal Modeling Techniques in Speech Recognition," *Proceedings of the IEEE*, vol. 81, pp. 1215-1247, 1993.

[11] D. O'Shaughnessy, "Invited Paper: Automatic Speech Recognition: History, Methods and Challenges," *Pattern Recognition*, vol. 41, pp. 2965-2979, 2008.

[12] Z. Razak, N. J. Ibrahim, M. Y. I. Idris, E. M. Tamil, Z. M. Yusoff, and N. N. A. Rahman, "Quranic Verse Recitation Recognition Module for Support in j-QAF Learning: A Review," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 8, pp. 207-216, August 2008.

[13] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*: Microsoft Corporation, 2000.

[14] A. S. Pandya and R. B.Macy, *Pattern Recognition with Neural Networks in C++*. Florida: CRC Press, 1996.

[15] J.-S. R. Jang. *Speech and Audio Toolbox*. Available: http://mirlab.org/jang/matlab/toolbox/sap/

[16] M. S. H. Salam, D. Mohamad, and S. H. S. Salleh, "Temporal Speech Normalization Methods Comparison in Speech Recognition Using Neural Network," presented at the International Conference of Soft Computing and Pattern Recognition (SoCPaR), Melacca, Malaysia, 2009.

[17] M. S. H. Salam, D. Mohamad, and S. Salleh, "Malay Isolated Speech Recognition Using Neural Network: A Work in Finding Number of Hidden Nodes and Learning Parameters," *The International Arab Journal of Information Technology*, vol. 8, pp. 364-371, October 2011.

[18] K. Daqrouq, "Wavelet Entropy and Neural Network for Text-Independent Speaker Identification," *Engineering Applications of Artificial Intelligence*, vol. 24, pp. 796-802, 2011.