Jurnal Teknologi Maklumat & Multimedia **7**(2009): 17 - 29

Endpoint Detection Enhancement for Speaker Dependent Recognition

UMMU SALMAH MOHAMAD, SITI MARIYAM SHAMSUDDIN
& RAMLAN MAHMUD

ABSTRACT

The automatic speech recognition (ASR) field has become one of the leading speech technology areas today. Various methods have been introduced to develop an efficient ASR system. The Neural Network (NN) approach is one of the more popular methods that is widely used in this field. Another Multilayer perceptron (MLP) model which is popularly used in the ASR field is the NN model. However, the current problems faced by MLP and most NN models in the ASR field is the long duration of training. Furthermore, the robustness of the isolated digit recognition is not trivial because it has been widely used in many applications. This study focuses on improving the training time and robustness of the MLP neural network for the Malay isolated digit recognition system by proposing variance endpoint detection to accelerate the convergence time of the NN and to produce the highest recognition accuracy. The proposed endpoint method have shown very promising results over experiments carried out. The overall performance for the Malay data set is 99.83% with a convergence time of 82 seconds.

Keyword: Automatic Speech Recognition, Multilayer Perceptron, Endpoint Detection, Artificial Neural Network

ABSTRAK

Bidang pengecaman pertuturan automatik telah menjadi salah satu bahagian teknologi pertuturan yang utama masa kini. Pelbagai kaedah telah diperkenalkan untuk membangun sistem pengecaman pertuturan automatik yang efisien. Rangkaian neural merupakan salah satu pendekatan terkenal yang sering digunakan dengan meluas di dalam bidang ini. Perseptron multi aras merupakan model rangkaian neural yang popular dalam bidang pengecaman pertuturan. Walau bagaimanapun, salah satu masalah yang dihadapi oleh perseptron multi aras dan model rangkaian neural lain dalam bidang pengecaman pertuturan ialah masa latihan yang terlalu lama. Di samping itu, ketepatan pengecaman pertuturan digit terpencil juga tidak boleh diabaikan kerana ianya digunakan dengan meluas dalam banyak aplikasi. Kajian ini memfokus terhadap pembaikan masa latihan dan ketegapan pengecaman bagi perceptron multi aras dengan mencadangkan kaedah pengesanan titik hujung varians bagi melajukan masa penumpuan dan menghasilkan ketepatan pengecaman yang tertinggi. Cadangan kaedah pengesanan titik hujung telah menunjukkan keputusan yang memberangsangkan bagi keseluruhan eksperimen yang dijalankan. Keseluruhan prestasi untuk data set Melayu adalah 99.83% dengan masa penumpuan 82 saat.

Kata kunci: Pengecaman Pertuturan Automatik, Perseptron MultiAras, Pengesanan Titik Hujung, Rangkaian Neural.

INTRODUCTION

An automatic speech recognition (ASR) system has been a goal in speech research for more than 6 decades (Salam *et al*., 2009). The ASR research began in the 1900's and has attracted much interest in the market recently because of the advancements in its methods, algorithm, and related technology (Srinivasan and Brown, 2002; Padmanabhan and Picheny, 2002). The ultimate goal in ASR is to establish a natural spoken language with an independent spoken style in all environments. However, achieving such a goal is not free from obstacles. Such obstacles can be handled easily by impressive human speech production. A similar approach with biological nerve cells in the human brain is a promising way to overcome those problems (Vilda *et.al*, 2009). One of the more effective methods associated with the human brain is a NN. A neural network consists of interconnected nerve cells (Kevans and Rodman, 1997).

The Neural network is capable of classifying noisy data, various patterned data, variable data streams, multiple data and overlapping, interacting and incomplete cues. Neural networks have made great progress in isolated word recognition and other ASR fields (Beufays et al., 2001) but the main obstacles that are faced by the NN model is the long training duration which increases with the data set. Besides that, the robustness of isolated digit recognition is not trivial because it has been widely used for many applications (Karnjanadecha and Zahorian, 2001; Beritelli et al., 2002). These applications include recognizing telephone numbers, spelled names and addresses, zip codes, and as a spelling mode for use with difficult words and out-of-vocabulary items in a continuous speech recognizer, mobile equipment and voice command-based consumer electronic products. Therefore in this study, an enhancement is done to improve the training time and robustness of the Multilayer Perceptron (MLP) NN for the Malay isolated digit recognition. The enhancement emphasises on the endpoint detection phase.

ENDPOINT DETECTION

Endpoint detection plays an important role in speech application and has been studied for several decades. It is an essential task in speech recognition systems to separate the speech segments from non-speech segments (Zhang et al. 1997; Revathi et al. 2007). Non-speech segments consist of speech, silence and other background noises. The method of detection of the speech signal embedded in various types of non-silence and background noise is also known as speech detection or speech activity detection (Li et al. 2002). The process of inaccurate detection at the beginning and ending boundaries of test and reference pattern will be the cause of errors in speech recognition (Shin et al. 2000; Ying et al. 1993). The main function of the endpoint detection is to discard the extraneous data in order to increase the recognition rate and to accelerate the computation time (Hahn and Park, 1992). Generally, it is difficult to detect the beginning and end of an utterance, especially when the following occurs:

- Weak fricatives (/f/, /th/, /h/) at the beginning or end.
- Weak plosive bursts (/p/, /t/, /k/) at the beginning or end.
- Nasals at the end.
- Voiced fricatives that become devoiced at the end of words.
- Trailing off of vowel sounds at the end of an utterance.

There are various methods to perform the task of endpoint detection. A short list includes energy threshold, pitch detection, spectrum analysis, zero crossing rate, periodicity measure, hybrid detection and fusion. Two widely used methods are energy and energy and zero crossing (conventional method), and these methods are the earliest methods, introduced by Rabiner and Sambur (1975) to detect the speech boundary.

In this study, the endpoint detection is emphasized in high signal to noise ratio (SNR). In other words, the endpoint detection needs to remove the silence speech only because the entire recording is done in silence condition. Few researchers have proposed several related works on enhancing algorithm for high SNR condition.

Hahn and Park (1992) introduced an improved speech detection algorithm for isolated Korean utterances. They utilized segmentation parameters and threshold value based on three basic frame based features. These include logarithmic energy, zero crossing rate and the modified zero crossing. These approaches were compared with energy and energy and zero crossing method. The results of the proposed method were very encouraging.

Zhang et al. (1997) introduced a fast endpoint detection algorithm for Chinese isolated word recognition (100 words). They presented hybrid method, which was developed by combining several conventional methods. First, the energy and zero crossing were utilized to acquire the references endpoints and the principle of variable frame rate was adopted. Finally cepstrum was used to accurately define the boundaries of isolated words. The performance of this method was moderate.

Ying et al. (1993) examined and successfully improved the Teager energy, which is developed by Teager (Ying et al. 1993). In this study, the improved Teager energy detects the speech boundary accurately.

He and Yu (2002) proposed two novel methods to perform the endpoint detection task. The first task was based on time domain features (energy and zero crossing). They introduced a novel approach for continuous speech recognition. In their study, the speech boundary was detected correctly at 96%. He and Yu (2002) also proposed another endpoint detection method, which was based on continuous multi band spectral features. This approach was introduced to overcome the problems in the previous works and it was found that the average detection correctness of boundary was 97%.

Hussain et al. (2000) considered using NN approach to detect the silence activity. They utilized Adaptive Linear Neuron (ADALINE) and MLP to perform the endpoint detection. The NN approach was compared with energy and zero crossing method and with the improved energy and zero crossing method. It was reported that the MLP detected the endpoint accurately compared to the other methods. However it dropped in rejection rates compared energy and zero crossing and improved energy and zero crossing method.

Most of the related research above made comparison of their proposed method with energy and energy and zero crossing methods**.** It can be concluded that energy and energy and zero crossing method have a few limitations when applied to practical problems. The major problem with these methods is that the threshold seldom restricts each other (He and Yu, 2002). For instance if we set the energy and zero crossing threshold larger, the utterance or some portion of the utterance with low energy or zero crossing will not be detected. On the contrary, too much information on unvoiced will be captured by zero crossing method if the threshold is smaller. Hence, it leads to poor performance. This is because the unvoiced consists of noise at constriction part.

In this study, a new endpoint detection method based on statistic approach that is variance method is proposed. Throughout this study the MLP obtains very encouraging result by using the variance method compared to energy and energy and zero crossing methods (conventional methods). The process of ASR in this study consists of the preprocessing, training, and recognition phase. The preprocessing phase is intended to process the data set before being fed to the NN.

Data Preprocessing

The preprocessing module prepared the speech signal in a digitized form before being processed by the NN module. This phase comprised of five stages: recording, endpoint detection, time axis normalization, feature extraction, and normalization. These stages are explained accordingly. The recording session is carried out to develop the Malay language corpus for the purpose of this study. Endpoint detection is used to separate the speech segments from non-speech segments (Zhang et al. 1997). Time-axis normalization is implemented to normalize the speech length to get a fixed number of neurons in order to cope with the NN structure. The

Linear Predictive Code (LPC) method is used to represent the speech model. The normalization stage is the final stage in preprocessing before the speech data is fed into the NN module. In this study, we used 2 sets of data; Malay dataset and TI46 dataset.

MALAY DATA SET

The Malay data set is a speaker-dependent corpus, comprising of isolated speech words from a single male speaker. The speech samples acquired at a 10kHz sampling rate were digitized into 8-bit resolutions. The speaker uttered numbers "sifar" to "sembilan" (0 to 9) at least 100 times and the total number of datasets taken was 1000 samples. The first 400 speech samples were used as the training set and the remaining 600 speech samples were used for data testing.

TI46 DATA SET

The TI46 data set is a benchmark corpus for English text-dependent speaker recognition (Chen *et al*., 2002). It contains isolated speech words from 16 speakers, eight males and eight females. They were divided into 2 sets: TI20 and TI_Alpha. The TI20 was used in this study that comprised of digits from "0" to "9". This dataset had been captured at a 12,500Hz sampling rate and digitized to 16-bit resolution. Each speaker uttered number "0" to "9" 10 times. The first 5 repetitions were used as the training set while the remaining 5 as the testing set. The recording was done in a high signal-to-noise ratio condition (SNR). The TI46 data set was split into male and female data sets. Thus, the total number of samples in the training set was 400 and the number of samples in the test data set was 400 for both males and females respectively. The male and female datasets were trained separately in this study

ENDPOINT DETECTION

For the purpose of this study, three methods were used. These included the Rabiner and Sambur approach (energy and energy and zero crossing methods) and the variance method (proposed method). The measurements of the speech signal using those methods are time-domain measurements and calculated on a frame-to-frame basis. The time domain measurements directly involve the waveform.

ENERGY METHOD

Energy measures the loudness of the sounds and provides a basis to differentiate voiced speech segments from the unvoiced speech segments. Commonly, the energy for voiced data such as **a**, **e** and **o** is much higher than silence. The energy of unvoiced for instance, **s**, **and f** is lower than voice sounds but higher than silence. The energy rate of the speech is defined in this study as the number of average magnitude of energy per 30ms interval and overlapping at 10ms. This type of interval is sufficient to capture important speech features (Rabiner and Juang, 1975).
where

$$x_t = \sum_{n=0}^{N} \left| S\left(ml + n\right) \right| \tag{1}$$

$$M_l = \frac{x_l}{N - 1} \tag{2}$$

S indicates the speech samples.

M speech samples overlapping at 10ms.

M average magnitude of energy.

N 0,1,…N-1 (N speech samples)
L 0,1,…L-1 (Intervals)
N speech samples at 30ms per interval

Before the endpoint detection process, the mean of the average magnitude was computed to give a statistical characterization of the background noise. It was assumed that the first 100ms of interval contained no speech. This information was further used to compute the peak energy (*IMX*) for the entire interval in each speech sample and the silence energy (*IMN*). Subsequently, the *IMX* and *IMN* were used to set two thresholds: upper threshold (*ITL*) and lower threshold (*ITU*), by using:

$$I1 = .003 \times (IMX - IMN) + IMN, \quad (3)$$

$$I2 = 4 \times IMN, \quad (4)$$

$$ITL = MIN\,(I1, I2), \quad (5)$$

$$ITU = 5 \times ITL. \quad (6)$$

Equation (3) shows *I1* to be a level, which is 3% of the peak energy, whereas Equation (4) shows *I2* to be a level set at four times the silence energy. The lower threshold, *ITL* Equation (5), is the minimum of this conservative energy threshold, and the upper threshold, *ITU* Equation (6), is five times the lower threshold (Rabiner and Sambur, 1975).

The average magnitude profile was searched to find the interval, in which it exceeded the conservative threshold, *ITL*. It is assumed that the beginning and end points lie outside this interval. Then working backwards from the point at which E_n first exceeded the threshold *ITU*, the point (labeled *N1* in Figure 1) where E_n first fell below a lower threshold *ITL,* is selected as the beginning point.

A similar procedure was followed to find the tentative endpoint *N2*. This double threshold (ITL and ITU) was performed to avoid dips in the average magnitude function. It is to ensure that it does not falsely signal the endpoint. For *energy* and *zero crossing* algorithms, at this stage, it is reasonably safe to assume that the beginning and ending points are not within the interval *N1* to *N2*. The final result was gained after implementing the zero crossing algorithms (Rabiner and Sambur, 1975).

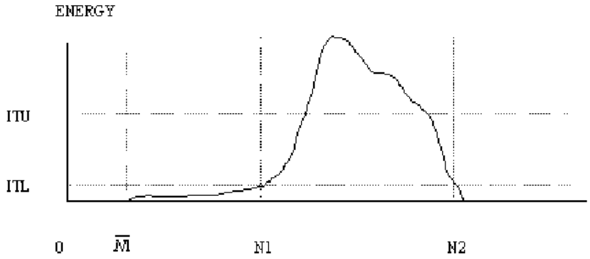

FIGURE 1. Typical Example of Energy for a Word Beginning with a Strong Fricative

ZERO CROSSING METHOD

This method was used to count the frequent of the signal that crosses the zero axes. In other words, a zero crossing occurs if successive samples have different algebraic signs. Zero crossing is very useful method for detecting the occurrence of unvoiced speech. The unvoiced speech is produced due to excitation of the vocal tract by the noise-like source at a point of construction in the interior of the vocal tract and shows a high zero crossing count. Due to this reason, the endpoint detection algorithm was refined by zero crossing after the energy measurements located the actual begining or end point. Throughout this study, the zero (level) crossing rate of the speech, Z_l, was defined as the number of zero (level) crossing per 30ms interval and overlapping at 10ms.

$$Z_l = \frac{1}{2}\sum_{n=0}^{N}\left|\operatorname{sgn}(S\left[ml+n\right]) - \operatorname{sgn}(S\left[ml+n\right]\right| \tag{7}$$

where

$$\operatorname{sgn}\left(S\left[ml+n\right]\right) = \begin{bmatrix} 1 & S\left[ml+n\right] & \geq 0 \\ -1 & S\left[ml+n\right] & < 0 \end{bmatrix}$$

Z zero crossing rate
m speech samples overlapping at 10ms.
n 0,1,...N-1 (N speech samples)
l 0,1,...L-1 (Intervals)
N speech samples at 30ms per interval

The zero crossing count of silence was expected to be lower than unvoiced speech, but relatively comparable to that of voiced speech. The standard deviation zero crossing was analyzed to obtain statistical characterization of the background noise as in energy method. This information was used to compute the zero crossing threshold (*IZCT)* by assuming that the first 100msec contained no speech or silence speech. A zero crossing threshold (*IZCT)* for unvoiced speech was chosen as the minimum of a fixed threshold. The *IZCT* was obtained by the sum of the mean zero crossing rate during silence ($\overline{IZC}$), plus twice the standard deviation of the zero crossing rate during silence Equation (8). Rabiner and Sambur (1975) give clear explanation about zero crossing method.

$$IZCT = MIN\left(IF, \overline{IZC} + 2\sigma\,\overline{IZC}\right) \tag{8}$$

Using the zero crossing algorithms, the searching starts backwards from *N1* (forward from *N2*) comparing the zero crossing rates to a threshold *IZCT* (determined from the statistics of the zero crossing rates for the background noise) as shown in Figure 2. This is limited to 25 frames preceding *N1* (following *N2*). If the zero crossing rates exceeded the threshold by 3 or more times, the beginning point *N1* will move back to the first point at which the zero crossing threshold exceeded ($\overline{N1}$). Otherwise, *N1* is defined as the beginning. A similar procedure is followed at the end of utterance to remove the silence samples.

VARIANCE METHOD

The variance indicates the variability of a list of samples. It is an average distance from the mean of provided observations. In this study, the mean value is represented by average magnitude, which is calculated from frame to frame. If the values are grouped near to the average magnitude (mean), the variance will be low. The average magnitude of the signal was computed over the window of 30 ms with an overlap ratio of 10ms. Equations (9) computed the average magnitude of the k_{th} frame.

$$M_k = \frac{1}{N}\sum_{n=0}^{N} S\left(ml+n\right), \tag{9}$$

The variance in each frame is computed in two ways. The first one uses the original equation as shown in Equation 10.

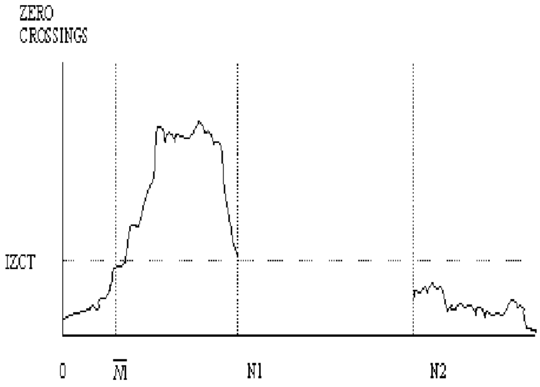

FIGURE 2. Typical Example of Zero Crossing Rate for a Word Beginning with a Strong Fricative

In the second way, the formula is modified by replacing the power with modulus syambol, as in Equation 11. The best recognition is yielded by the modified formula. Hence, this formula is used in the variance method to calculate the frequency of the samples in each interval for both Malay and TI46 data sets.

$$\sigma_k = \frac{\sum_{n=0}^{N}\left(S(ml+n)-M_k\right)^2}{N}, \tag{10}$$

$$\sigma_k = \frac{\sum_{n=0}^{N}\left|S(ml+n)-M_k\right|}{N}, \tag{11}$$

where

m speech samples overlapping at 10ms.

M_k is the average magnitude for k_{th} frame

S is a speech samples

σ is a variance for k_{th} frame

N is the total samples in k_{th} frame or speech samples at 30ms per interval

n 0,1,…N-1 (N speech samples)

l 0,1,…L-1 (intervals)

A thorough search was performed on the speech signal from the forward and backward directions on frame-by-frame basis to locate the beginning and ending points. The starting and ending points are established when the value of the variance for that frames exceeded the variance threshold (e.g. variance for frame 6 is 50 and the threshold is 7.0). In this study, the threshold was fixed after 30 experiments. The range of experimented threshold was 1 to 10. Normally this range contains a big amount of silence samples. After the experiments, the values 1.0 and 7.0 were tentatively selected as the best threshold because both values produced highest recognition rate for Malay and TI46 data set respectively. The different type of threshold was obtained due to digitization resolution. The 8 bits resolution consists of a small variation of sample numbers, which is from 0 to 256. Therefore, the silence detection is not accurate because the silence only represents limited value. Hence, 1.0 threshold was sufficient to produce good recognition rate. Meanwhile, for 16 bits resolution, the variation sample numbers are –32768 to 32767. Thus, the silence detection is more accurate than 8 bits, because the range of samples is wide. As such, the value of 7.0 was used as optimum threshold to obtain good recognition rate.

The algorithm of proposed variance method is given below.

Step 1: Read the speech data (in samples form)
Step 2: Implement frame blocking and windowing process
Step 3: Detect the start point from the beginning/start of the speech data
Step 4: For each frame of speech data, do 4-6.
Step 5: Compute average magnitude (refer Equation 9).
Step 6: Compute variance (refer Equation 11).
Step 7: Compare the variance value for each frame with certain threshold.
If the variance is below the threshold, the frame is assigned as a silence portion and discards it.
Step 8: Perform the endpoint from the backward/ending of the speech data
Step 9: For each frame of speech data, do 9-11.
Step 10: Compute average magnitude (refer Equation 9).
Step 11: Compute variance (refer Equation 9).
Step 12: Compare the variance value for each frame with certain threshold.
If the variance is below the threshold, the frame is assigned as a silence portion and discards it.
Step 13: Store the new value of start and ending point for the speech data in a file.

Normalization was performed to scale speech data (LPC parameter) into a certain range before being fed to the NN. Usually, the range is between [–1,1] and [0,1] to suit the output from the neurons.

Findings and Discussion

The endpoint detection experiments were carried out to compare the proposed methods with conventional methods in term of recognition rate and convergence times using MLP. The proposed method comprises of variance method. Meanwhile, the conventional methods consist of energy and energy and zero crossing method. Both Malay and TI46 data set are used as input. From the produced results, the best endpoint detection method for Malay and TI46 data set, which produced highest recognition rate, was selected.

MALAY DATA SET EXPERIMENTS

The experimental results on endpoint detection for Malay data set are shown in Table 1. From Figure 3, it shows that the variance method gave the fastest convergence times compared to other methods. It took 68s for BP to converge towards the minimum error rates. Meanwhile energy and energy and zero crossing took 91s and 92s respectively for BP to reach the minimum error rates. Convergence times are defined as the time that has been taken by MLP to achieve the minimum error rate in learning phase.

Figure 4 describes the recognition rate using various endpoint detection based on Malay data set. It shows that the variance, energy and energy and zero crossing methods produced 100% recognition rate for learning data. These results changed when they were tested on test data**.** Here, the variance method achieved the highest recognition rate with 99.83%. The variance method is superior to energy and energy and zero crossing methods by 4.17% and 5.33% respectively.

The experimental results showed that the proposed method performed better in terms of convergence time and recognition rate. Hence, the variance method was used as endpoint detection method for Malay data set.

TABLE 1. The Convergence Times and Recognition Rate For Endpoint Detection Methods *(Malay Data Set)*

Endpoint Detection Method	**Time (s)**	**Recognition Rate (%)**
Energy	91	95.66
Variance	68	99.16
Energy and Zero Crossing	92	94.50

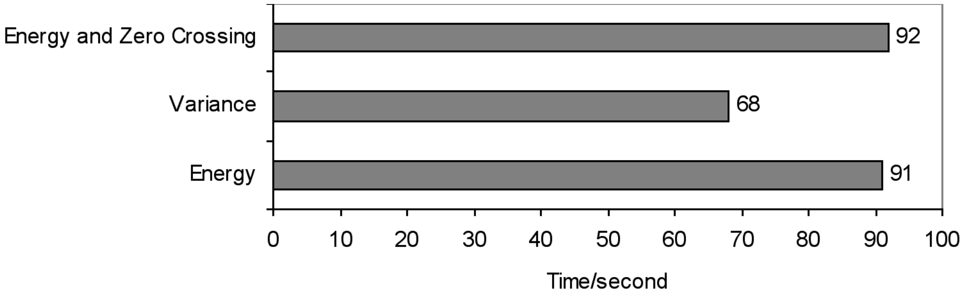

FIGURE 3. The Convergence Times for Various Endpoint Detection Methods *(Malay Data Set)*

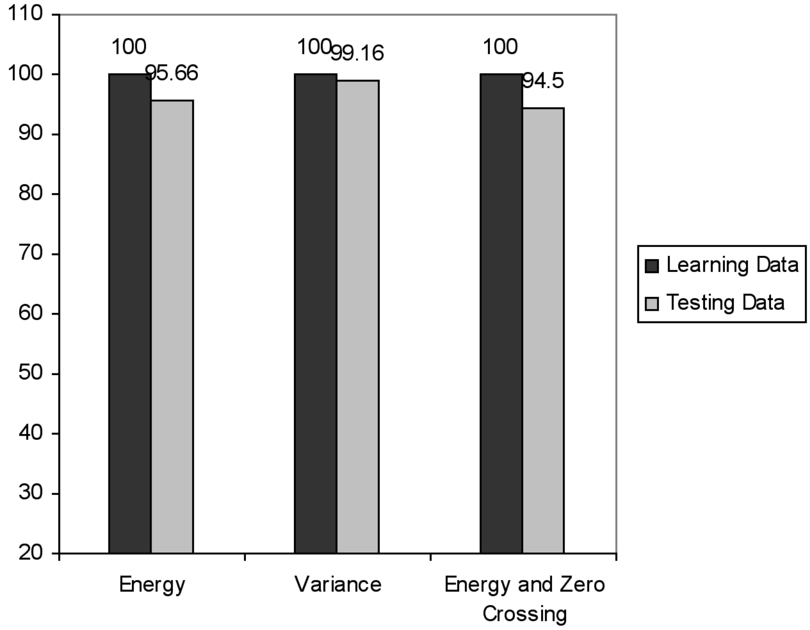

FIGURE 4. The Recognition Rate for Various Endpoint Detection Method *(Malay Data Set)*

TI46 DATA SET EXPERIMENTS

In this section, the experiments for TI46 data set is presented as shown in Table 2 and 3. Figure 5 presents the convergence time for various endpoint detection methods. It shows that the variance method produced the fastest times with 105s compared to other methods for standard BP to converge towards the solution. The energy and zero crossing and energy method took 111s and 129s for BP to complete its learning process.

Figure 6 shows the recognition rate of male data set using various endpoint detection methods. It seemed that, the endpoint detection methods produced 100% for learning data. For test data, the energy method produced highest recognition rate with 94.75% and 93.75% respectively. However the energy method superior to proposed method (Variance) is by 1% only. The energy method produced the highest recognition rate but slower convergence times compared to variance method. Variance method obtained the fastest convergence times but lower in recognition rate compared to the energy method. The energy and zero crossing method showed the lowest recognition rate among other methods with 87.75%.

TABLE 2. The Convergence Times and Recognition Rate For Endpoint Detection Methods *(Male Data Set)*

Endpoint Detection Method	Time(s)	Recognition Rate (%)
Energy	111	94.75
Variance	105	93.75
Energy and Zero Crossing	129	87.75

TABLE 3. The Convergence Times and Recognition Rate for Endpoint Detection Methods *(Female data set)*

Endpoint Detection Method	Time (s)	Recognition Rate (%)
Energy	120	92.25
Variance	133	96.50
Energy and Zero Crossing	146	85.50

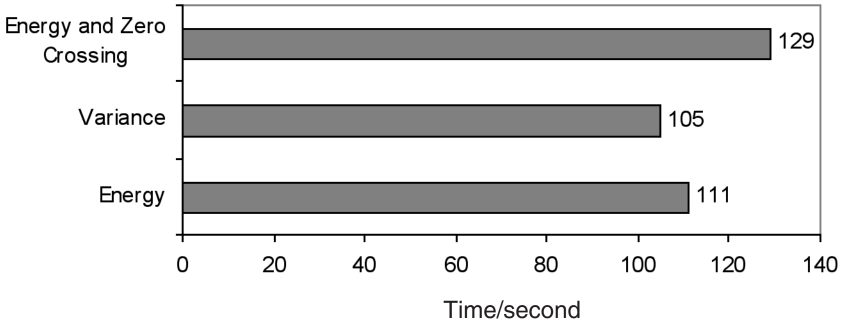

FIGURE 5. The Convergence Times for Various Endpoint Detection Method *(Male Data Set)*

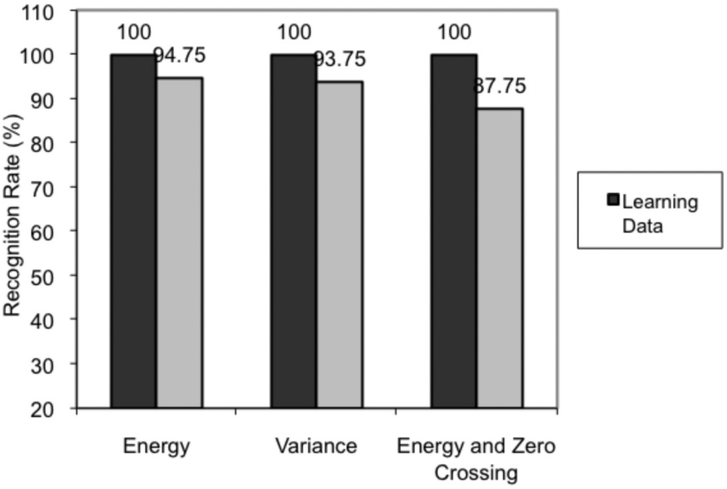

FIGURE 6. Recognition Rate for Various Endpoint Detection Method
(MaleData Set)

The experiment for female data set was carried out to determine the best endpoint detection method. Figure 7 shows that the energy method produced the fastest convergence times compared to other methods. It took 120s for BP to converge to the minimum error rates. The variance method took 133s with slow convergence time compared to energy method. The energy and zero crossing method were relatively slow compared to the other methods with 146s for standard BP to achieve towards the minimum error rates.

Figure 8 shows the recognition rate for female data sets using various endpoint detection methods. The recognition for learning data was 100% in this domain. However, for test data, the variance method gave the highest result with the recognition rate of 96.5% compared to energy, 92.25%. Meanwhile energy and zero crossing method gave 85.5% recognition rate.

From the observation, the proposed method (variance) successfully captured the meaningful information for Malay and female data set to gain highest recognition rate. In addition, this method produced fastest convergence time for Malay and male data set. Even though the proposed method did not produce the highest recognition rate for male data set, the results were inferior to energy method by only 1%.

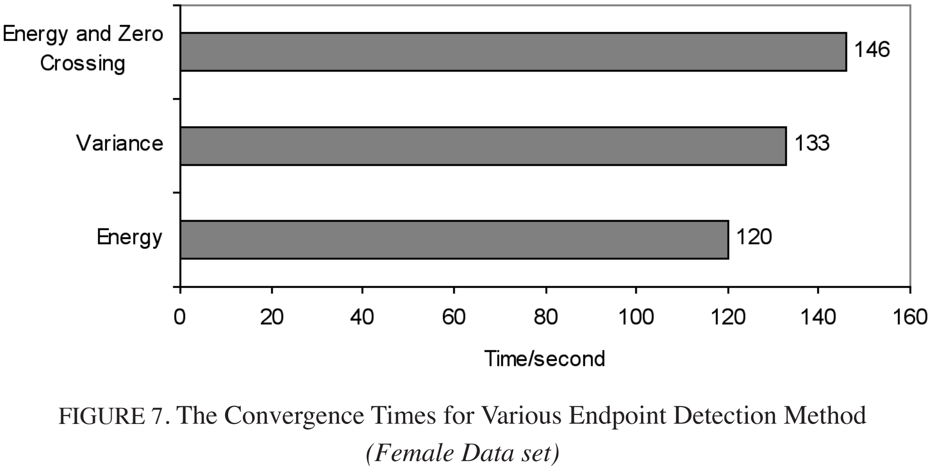

FIGURE 7. The Convergence Times for Various Endpoint Detection Method
(Female Data set)

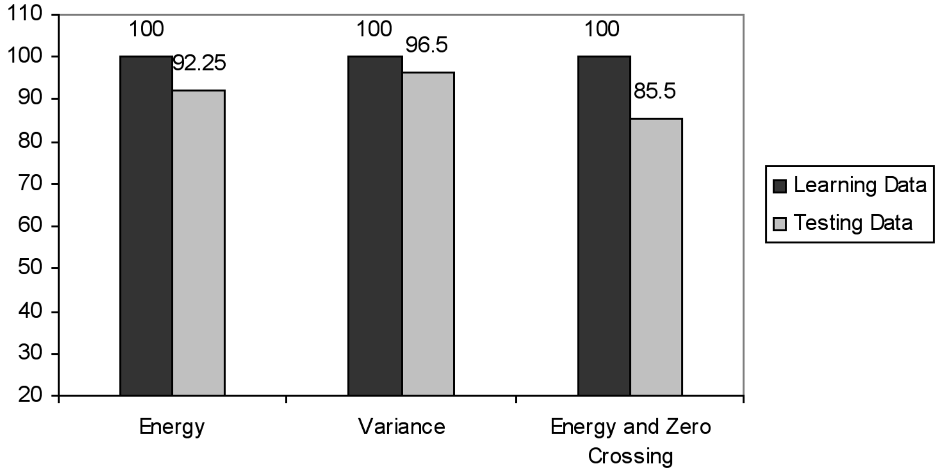

FIGURE 8. The Recognition Rate for Various Endpoint Detections Method *(Female Data Set)*

The variance method successfully removed the silence samples and produced the highest recognition rate for Malay and female data set. For male experiment, it slightly dropped in recognition rate compared to energy, but the difference was only 1%. In addition, the variance method yielded fastest convergence time compared to other methods in male data set experiment. The variance method performed consistently better in all experiments because it was capable of eliminating the silence accurately for all data set by successfully differentiating between the silence and speech. This condition contrasted with energy and energy and zero crossing methods. The energy method successfully detected the silence accurately only in male data set, with slightly higher in recognition rate compared to variance method as described before. Meanwhile the energy and zero crossing did not perform better for all data sets. From these observations, three obstacles had led both methods to produce low recognition rate. The *energy* and *energy* and *zero crossing* detected the speech as a silence in some of the waveform. Thus the valuable information is probably discarded. Another reason is that, the silence was not precisely detected by both methods, and hence silence occurred in some of speech template as a noise, processed by the MLP. Lastly, the *energy* and *zero crossing* method captured too much redundant information of unvoiced sounds. While the unvoiced sounds contained noise at a constriction especially for fricatives sounds, leading to poor recognition.

CONCLUSION

Difficulty in finding an appropriate threshold value in conventional methods, led to finding an alternative method, namely variance method. This method was capable of eliminating the silence and produce satisfactory results for Malay and female TI46 data set. The Malay data set has shown very encouraging results and proved the ability of the NN in the ASR field. Additionally, the standard data set, TI46, also yielded good accuracy in this study. Overall performance shows that the Malay and TI46 data sets achieved 99.83% and 96.75% accuracy respectively. Consequently, the proposed endpoint detection had shown a potential performance to remove the silence and become an alternative method for endpoint detection.

REFERENCES

Aksoy, S. & Haralick, R.M. 2000. Probabilistic vs. Geometric Similarity Measures for Image Retrieval. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 2 : 357 -362 .

Baron, R.; Girau, B. 1998. Parameterized Normalization: Application to Wavelet Networks. *Proc. International Joint Conference on Neural Networks Proceedings*. Anchorage, AK, USA. 4-9 May. 1433 –1437.

Beaufays, F., Bourlard, H., Franco, H. & Morgan, N. 2000. Neural Networks In Automatic Speech Recognition. *In The Handbook of Brain Theory and Neural Networks*. The MIT Press.

Beritelli,F., Casale,S.& Serrano,S. 2002. A Robust Speaker Dependent Algorithm for Isolated Word Recognition. *Proc. 14th International Conference on Digital Signal Processing*.Vol. 2, pp. 993 –996.

Chen, K., Wu, T.Y. & Zhang, H.J. 2002. On The Use of Nearest Feature Line for Speaker Identification. *Pattern Recognition Letters*. 23 : 1735-1746.

Aini Hussain. 1997. Development of Phoneme Based Malay Speech Recognition Systems Using Modular Artificial Neural Networks, Ph.D.Diss., National University of Malaysia, Bangi, Malaysia.

Karnjanadecha, M.; Zahorian, S.A.; 2001. Signal Modeling for High-Performance Robust Isolated Word Recognition. *IEEE Transactions on Speech and Audio Processing*. 9(6): 647 –654.

Kevans, L. & Rodman, R.D. 1997. *Voice Recognition*. London :Artech House.

Padmanabhan, M. & Picheny, M. 2002. Large Vocabulary Speech Recognition Algorithms. *Computer*. 35(4) : 42–50.

Pedersen, M.W. 1997. Training Recurrent Networks. *Proceedings of the 1997 IEEE Workshop,* Amelia Island, FL, USA. 355 –364.

Vilda, P.G., Vicente, J.M.F., Biarge,V.R. & Baillo, R.F. 2009. Time-Frequency Representations in Speech Perception. Neurocomputing 72(2009) : pp820-830.

Rafiq, M. Y., Bugmann, G. & Easterbrook, D. J. 2001. Neural Network Design for Engineering Applications. *Computers & Structures*. 79 (17) :1541-1552.

Revathi, A., Chinnadurai, R. & Venkataramani, Y. 2007. *End-point detection of speech under low SNR*. Electronic Engineering Times India. pp 1-3.

Salam, M.S., Mohamad D. & Salleh, S.H. 2009. Improved Statistical Segmentation Using Connectionist Approach. *Journal of Computer Science*. 5(4) : pp 275-282.

Salleh, S.H.S., Ismail, I. & Akmal, S. 1994. Investigation of Digit Recognition Using The Backpropgation. *Jurnal Elektrika*. 2-4 (4) : pp. 32-40.

Srinivasan, S. & Brown, E. 2002. Is Speech Becoming Mainstream? *Computer*. 35 (4): pp. 58 –66.

Vieira, K., Wilamowski, B. & Kubichek, R. 1997. Speaker Verification for Security Systems Using Artificial Neural Networks". *International Conference on Industrial Electronics, Control and Instrumentation (IECON97)*. New Orleans, LA, USA.1102-1107.

Zhang, G., Patuwo, E. & Hu, Y.H. 1998. Forecasting with Artificial Neural Networks: The State of The Art. *International Journal of Forecasting*. 14(1) : pp.35-62.

Ummu Salmah Mohamad , Siti Mariyam Shamsuddin & Ramlan Mahmud
Faculty of Information System, Kolej Islam Darul Ridzuan, Kuala Kangsar, Perak .
ummusalmah93@gmail.com

Siti Mariyam Shamsuddin
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia, 81300 Skudai, Johor, Malaysia.
mariyam@utm.my

Ramlan Mahmud
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia, Serdang Selangor.
ramlan@fsktm.upm.my