Malaysian Journal of Computer Science, Vol. 17 No. 2, December 2004, pp. 40-54

DIGIT RECOGNITION USING NEURAL NETWORKS

Chin Luh Tan and Adznan Jantan
Faculty of Engineering
Universiti Putra Malaysia
43400 Serdang, Selangor Darul Ehsan
Malaysia
email: kailup4@hotmail.com
adznan@eng.upm.edu.my
Tel.: 603-80761953
Fax: 603-80761954

ABSTRACT

This paper investigates the use of feed-forward multi-layer perceptrons trained by back-propagation in speech recognition. Besides this, the paper also proposes an automatic technique for both training and recognition. The use of neural networks for speaker independent isolated word recognition on small vocabularies is studied and an automated system from the training stage to the recognition stage without the need of manual cropping for speech signals is developed to evaluate the performance of the automatic speech recognition (ASR) system. Linear predictive coding (LPC) has been applied to represent speech signal in frames in early stage. Features from the selected frames are used to train multilayer perceptrons (MLP) using back-propagation. The same routine is applied to the speech signal during the recognition stage and unknown test patterns are classified to the nearest patterns. In short, the selected frames represent the local features of the speech signal and all of them contribute to the global similarity for the whole speech signal. The analysis, design and development of the automation system are done in MATLAB, in which an isolated word speaker independent digits recogniser is developed.

Keywords: ***Digits recognition, Feed-forward back-propagation, Linear predictive coding, Neural networks, Speech recognition***

1.0 INTRODUCTION

In the field of speech recognition, a large number of algorithms and methods have been proposed for various purposes. The requirement of different applications drives the researchers to develop new algorithms or improve existing methods to serve the need in different situations. For example, speaker-dependent (SD) systems which accept the speech from specific speakers are usually applied in security systems. On the other hand, speaker independent (SI) recognisers are designed to recognise speech from different speakers such as speech to text engines in word processing programs, as a substitute to a keyboard.

Broadly speaking, speech recognition systems are usually built upon three common approaches, namely, the acoustic-phonetic approach, the pattern recognition approach and the artificial intelligence approach [1]. The acoustic-phonetic approach attempts to decide the speech signal in a sequential manner based on the knowledge of the acoustic features and the relations between the acoustic features with phonetic symbols. The pattern recognition approach, on the other hand, classifies the speech patterns without explicit feature determination and segmentation such as in the formal approach. The artificial intelligence approach forms a hybrid system between the acoustic-phonetic approach and the pattern-recognition approach.

The artificial intelligence approach becomes the field of interest after seeing the success of this approach in solving problems (especially classification problems) [2]. The application of artificial neural networks is proposed to meet the needs of an accurate speech recogniser. For example, the neural network approach to phoneme recognition [3, 4] is proposed in Japanese vowel recognition. Besides, the combination of neural networks and linear dynamic models is proven in achieving a high level of accuracy in automatic speech recognition systems [5]. Another problem in speech recognition is the increase of error in the presence of noise such as in a typical office environment. Some researchers propose the use of visual information such as the lip movement [6, 7]. In this case, image processing techniques and neural networks are applied to capture and analyse lip movement.

Digit recognition is one of the common applications in this field, for example, mandarin digit recognition systems have been actively developed by researchers in China [8, 9]. Different systems have been proposed to recognise digits of different languages.

In this paper, the application of neural networks in the pattern-recognition approach is discussed. We propose the use of a multilayer perceptron (MLP), which is trained using the back-propagation technique to be the engine of an automated digit recognition system. Firstly, the features of the training datasets are extracted automatically using the end-point detection function. The features are then used to train the neural network. The same function is used to extract the features of signals during the recognition stage. Several networks with different structures (different numbers of neurons) were trained with different numbers of samples and the performance in recognising the unknown input patterns were compared. The system was built using MATLAB [10] and an accuracy greater than 95% was achieved for the unknown patterns.

The following section discusses the stages in designing the automatic speech recognition system. Firstly, the speech signal properties are discussed, followed by the end point detection method in finding the region of interest from the raw speech data. After the start point and the end point of a speech signal have been detected, it is then analysed by various methods. The LPC method is used to represent the features of the speech signal which has been blocked into frames. Besides, by referring to the start point and end point of the signals, a finite number of frames is selected to become the input for the neural network. Finally, a comparison of the performance for various networks with different numbers of training datasets and different numbers of neurons was done.

2.0 RELATED RESEARCH

There are two basic approaches of using neural networks in speech classification, which are the static approach and the dynamic approach. In the static approach, the neural network accepts all input speech data at once, and makes a single decision. On the other hand, for the dynamic approach, the neural network processes a small window of the speech at one time, and this window slides over the input speech data while the network makes a series of local decisions, which have to be integrated into a global decision at a later time.

Both approaches are being applied in phoneme recognition as well as word level recognition. In this project, a neural network will be used to recognise digits at the word level. A few researches related to this method are discussed.

Peeling and Moore (1987) [11] applied Multilayer Perceptrons to digit recognition with excellent results. A static input buffer of 60 frames of spectral coefficients is applied in which the briefer words were padded with zeros and positioned randomly in the 60-frame buffer. By evaluating different network topologies, a single hidden layer with 50 units was found to perform efficiently. A performance of 99.8% was found in speaker-dependent experiments and 99.4% was found for multi-speaker experiments.

Kammerer and Kupper (1988) [12] found that single-layer perceptrons outperformed both multi-layer perceptrons and a dynamic time warping (DTW) template-based recogniser in many cases. A static input buffer of 16 frames was applied in which each word was linearly normalised, with sixteen 2-bit coefficients per frame. The system achieved the performance of 99.6% in speaker-dependent experiments and 97.3% for speaker-independent experiments.

Burr (1988) [13] applied Multilayer Perceptrons in a more difficult task, alphabet recognition. A static input buffer of 20 frames was applied, in which each spoken letter was linearly normalised, with 8 spectral coefficients per frame. Training on three sets of the 26 spoken letters and testing on a fourth set, the performance achieved was 85% in speaker dependent experiments, matching the accuracy of a dynamic time warping (DTW) template-based approach.

3.0 SPEECH SIGNAL

3.1 Speech Signal Representation

A speech signal is usually classified into three states. The first state is silence, where no speech is produced. The second state is unvoiced, in which the vocal cords are not vibrating and the resulting signal is random in nature. The

last state is voices, in which the vocal cords vibrate and produce a quasi-periodic signal. The silence state is usually the unwanted state and has to be removed in order to save the processing time of the speech recognition system as well as to improve the accuracy of the system.

In the time domain, the amplitude of the speech signal at each sampling time is plotted over time. This representation gives the picture on how a speech varies over time, and requires large storages.

Spectral representations illustrate the nature of speech signals in terms of their frequency contents. Fig. 1 shows the spectrogram of a speech signal which corresponds to performing a fast Fourier transform on every 256 samples (32ms) with the analysis advancing in intervals of 64 samples (8ms).

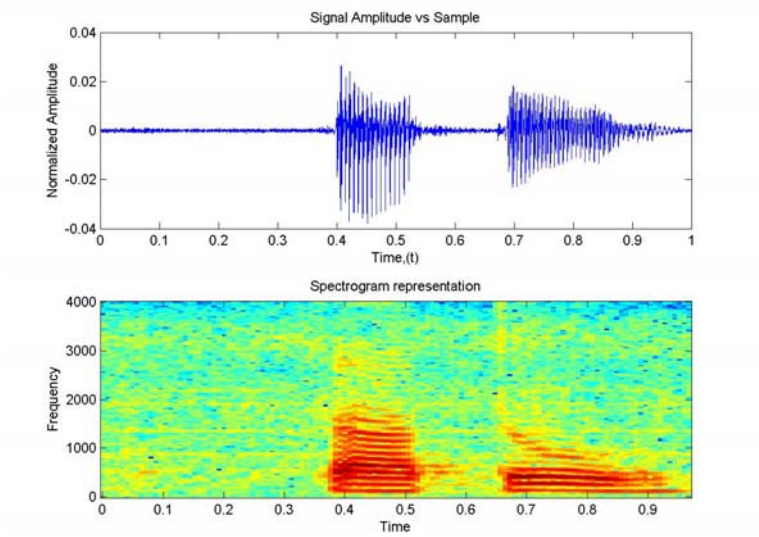

Fig. 1: Spectrogram analysis by using FFT

For the sake of analysis, the speech signal is usually broken into frames. This has been applied in the spectrogram shown above in which the frequency contents of all frames are arranged one next to the other to form a three dimensional representation (the colours represent the third dimension). The frequency information of a specific frame can also be obtained by taking the fast fourier transform of the specified frame. An analysis tool is built using MATLAB to perform this task. Fig. 2 shows the frequency contents of a frame with 256 samples.

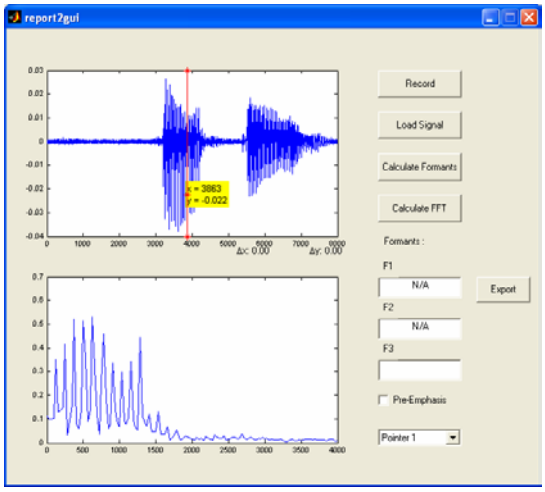

Fig. 2: Fast Fourier Transforms of a frame with 256 samples (from n=3863–127 to n=3863+128)

3.2 Endpoint Detection

The end point detection technique is applied to extract the region of interest from the speech signal. In other words, it removes the silent region in speech signals. The basic technique of end point detection is to find the energy level of a signal. Signal energy level is calculated in frames, where each frame consists of N samples. The frames usually overlap with the adjacent frames to produce a smooth energy line. Fig. 3 shows the energy plot of "One".

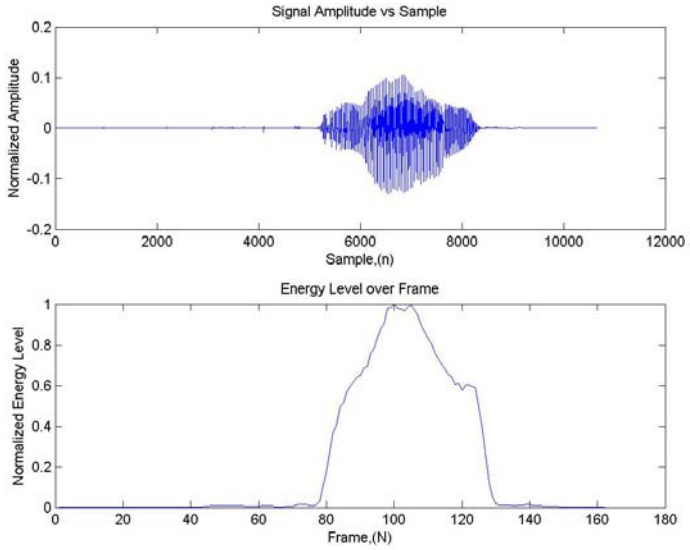

Fig. 3: (First Panel) Amplitude vs Time plot of "One" (Second panel) Energy level of the signal

Accurate end-point detection is important to reduce processing load and increase the accuracy of a speech recognition system. Basically there are two famous endpoint detection algorithms. The first algorithm uses signal features based on energy level and the second algorithm uses signal features based on the rate of zero crossings. The combination of both gives good results, but increases the complexity of the program and also the processing time. In this project, an end-point detection method that is based on the energy level is applied to reduce the pre-processing time [14].

Fig. 4 shows the signal of "one" sampled at 8000Hz for 10650 samples or 1.33 seconds. Before the speaking begins, the waveform started as silence for about 5000 samples. After the utterance, the signal remains in a silent state again for about 2000 samples. Throwing the unwanted silence region, the processing time can be improved to 3650/10650 * 100 = 34.3% by assuming all the frames in the region of interest have been processed. The energy level of the signal is inspected and a threshold value is determined from the energy plot. Fig. 5 shows the cropped signal, where the silence region has been eliminated, and the remaining regions of interest are used for further processing.

3.3 Speech Coding

Linear predictive coding (LPC) [15] is defined as a method for encoding a speech signal in which a particular value is predicted by a linear function of the past values of the signal. It is one of the methods of compression that models the process of speech production.

$$\tilde{s}(n) \approx a_1 s(n-1) + a_2 s(n-2) + a_3 s(n-3) + ... + a_p s(n-p) \tag{3.1}$$

The basic idea is that a given speech sample at time n, $s(n)$, can be approximated as a linear combination of the past p speech samples. The coefficients $a_1, a_2, \ldots, a_p$ are assumed constant over the speech analysis frame. The goal of this model is to predict the next sample of the signal by linearly combining the p most current samples while minimising the mean square error over the entire signal.

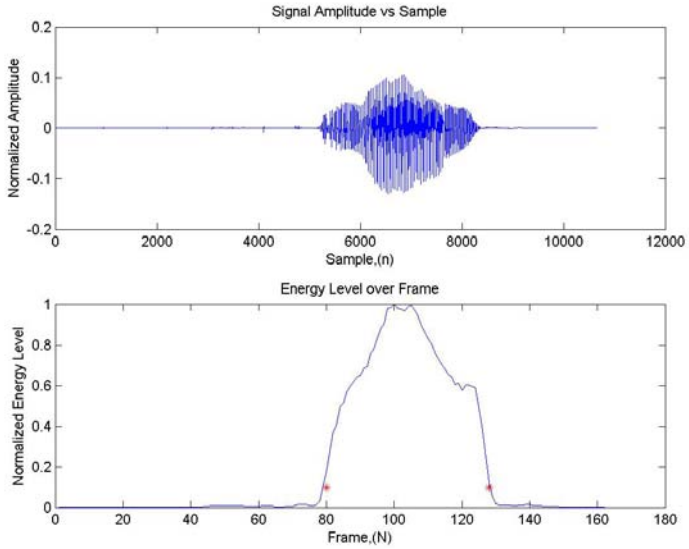

Fig. 4: (First panel) Original Signal, (Second panel) End-point detection by using the energy level of the speech signal

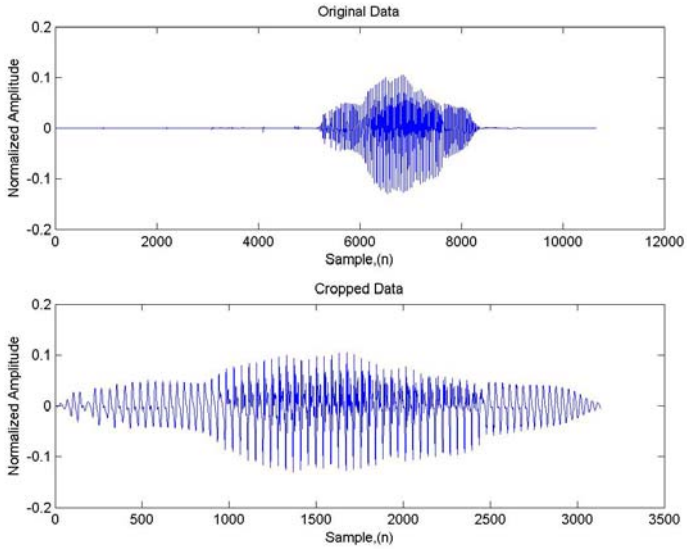

Fig. 5: (First panel) Detected End Point, (Second panel) Cropped Signal/Region of Interest

Fig. 6 shows the LPC estimation and its error for a frame of a speech signal with 256 samples.

This signal frame is a segment of "one", which was discussed in the previous section. By expressing Equation 3.1 in z-domain, including an excitation term GU(z), we get:

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + GU(z) \tag{3.2}$$

leading to the transfer function [16]:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{3.3}$$

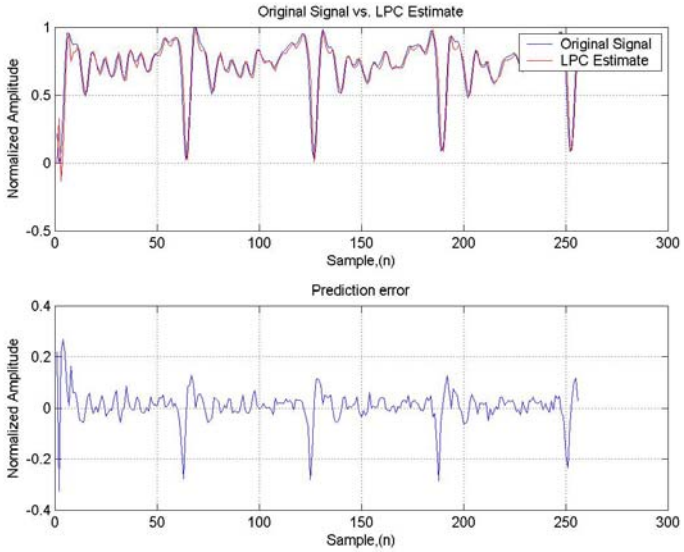

Fig. 6: LPC estimation for a speech signal frame with 256 samples

This will give the envelope spectra of the speech signal. The LPC spectrum can be obtained by plotting the H(z) as shown in the equation mentioned above. Fig. 7 shows the typical signal and the spectra for the LPC autocorrelation method for a segment of speech spoken by a male speaker. The analysis is performed using a $p = 8^{th}$ order LPC analysis over 256 samples at a sampling frequency of 8 KHz.

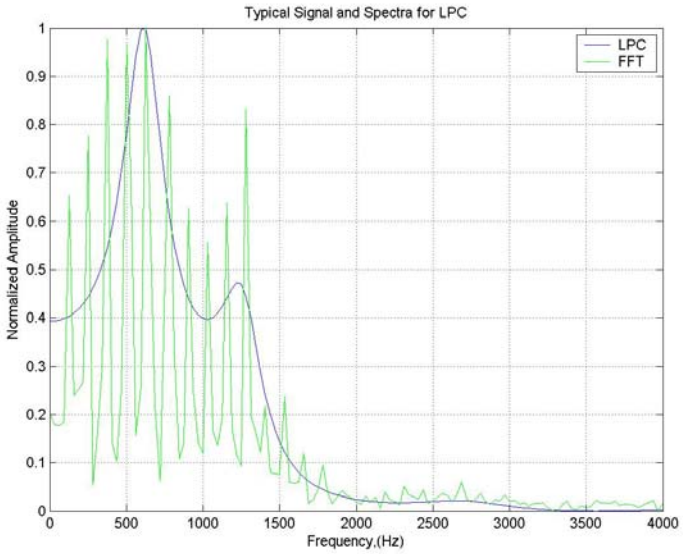

Fig. 7: Spectra for FFT and LPC autocorrelation method for a segment of speech by a male speaker

In other words, the transfer function of energy from the excitation source to the output can be described in terms of natural frequencies or resonances. Such resonances are called formants of the speech. From the LPC spectral, three resonances of significance can be noticed, and named as F1, F2 and F3 respectively. Mathematically, three formants can be obtained by taking the angle of roots of the denominator in Equation 3.3.

3.4 Frame Selection

Processing all frames in the region of interest as discussed in the previous section leads to few problems. Firstly, due to the various speaking rates, the number of frames is not equal between signals. Secondly, the processing time

for all frames is time consuming. In this paper, specific frames are selected to be presented to the neural network during the training process as well as during the recognition process. The frames are selected in linear distance with reference to the start point and the end point of the signal. Each frame consists of 256 samples of data.

Fig. 8 shows four frames that have been selected with reference to the start point and end point. The LPC coefficients of the selected frames are used as the inputs for the neural network which will be discussed in the next section.

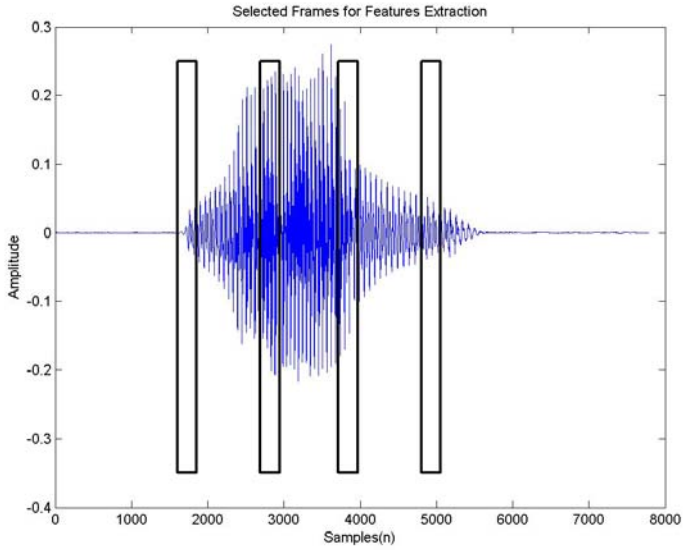

Fig. 8: Selected frames for features extraction

4.0 NEURAL NETWORK

4.1 The Multi-Layer Perceptron

Multi-layer perceptrons are one of many different types of existing neural networks. They comprise a number of neurons connected together to form a network. The "strengths" or "weights" of the links between the neurons is where the functionality of the network resides. Its basic structure is shown in Fig. 9.

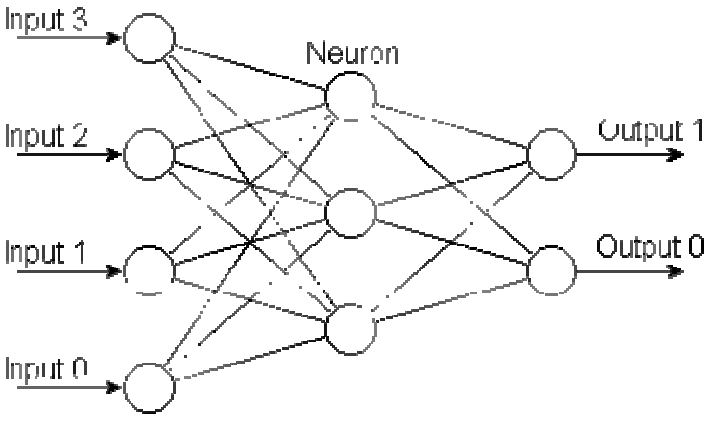

Fig. 9: Structure of a multi-layer perceptron

The idea behind neural networks stems from studies of the structure and function of the human brain. Neural networks are useful to model the behaviors of real-world phenomena. Being able to model the behaviors of certain phenomena, a neural network is able subsequently to classify the different aspects of those behaviors, recognise

what is going on at the moment, diagnose whether this is correct or faulty, predict what it will do next, and if necessary respond to what it will do next.

4.2 Feed-Forward Back-Propagation Network

Feed-forward networks [17] often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons. Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors.

Back-propagation was created by generalising the Widrow-Hoff learning rule to multiple-layer networks and nonlinear differentiable transfer functions. Input vectors and the corresponding target vectors are used to train a network until it can approximate a function, associate input vectors with specific output vectors. Networks with biases, a sigmoid layer, and a linear output layer are capable of approximating any function with a finite number of discontinuities.

4.3 Training

Standard back-propagation is a gradient descent algorithm, in which the network weights are moved along the negative of the gradient of the performance function. The term back-propagation refers to the manner in which the gradient is computed for nonlinear multilayer networks. There are a number of variations on the basic algorithm that are based on other standard optimisation techniques, such as conjugate gradient and Newton methods.

With standard steepest descent, the learning rate is held constant throughout training. The performance of the algorithm is very sensitive to the proper setting of the learning rate. If the learning rate is set too high, the algorithm may oscillate and become unstable. If the learning rate is too small, the algorithm will take too long to converge. It is not practical to determine the optimal setting for the learning rate before training, and, in fact, the optimal learning rate changes during the training process, as the algorithm moves across the performance surface.

The gradient descent algorithm for training the multi-layer perceptron was found slow especially when getting close to a minimum (since the gradient is disappearing). One of the reasons is that it uses a fixed-size step. In order to take into account the changing curvature of the error surface, many optimisation algorithms use steps that vary with each iteration.

In order to solve this problem, an adaptive learning rate [18] can be applied to attempt keeping the learning step size as large as possible while keeping learning stable. The learning rate is made responsive to the complexity of the local error surface. In this approach, new weights and biases are calculated using the current learning rate at each epoch. New outputs and errors are then calculated. As with momentum, if the new error exceeds the old error by more than a predefined ratio for example, 1.04, the new weights and biases are discarded. In addition, the learning rate is decreased. Otherwise, the new weights are kept. If the new error is less than the old error, the learning rate is increased. This procedure increases the learning rate.

In this paper, a feed-forward multi-layer perceptron with a single hidden layer and trained by gradient descent with momentum and an adaptive learning rate back-propagation method was applied to the digit classification problem.

5.0 NEURAL NETWORKS IN SPEECH RECOGNITION

5.1 Neural Network Structure

Fig. 10 illustrates the structure of the neural network in this project. The inputs of the network are the features extracted from the selected frames. The features can be the LPC coefficients or the first three formants of each frame. The training target is shown in the table where each of the digits will activate a different output neuron.

The whole database consists of:

a. 226 different speakers speaking at different rates
b. 10 digits (one to zero), for each speaker
c. 112 male speakers and 114 female speakers
d. Total number of utterances: 226x10 = 2260 utterances

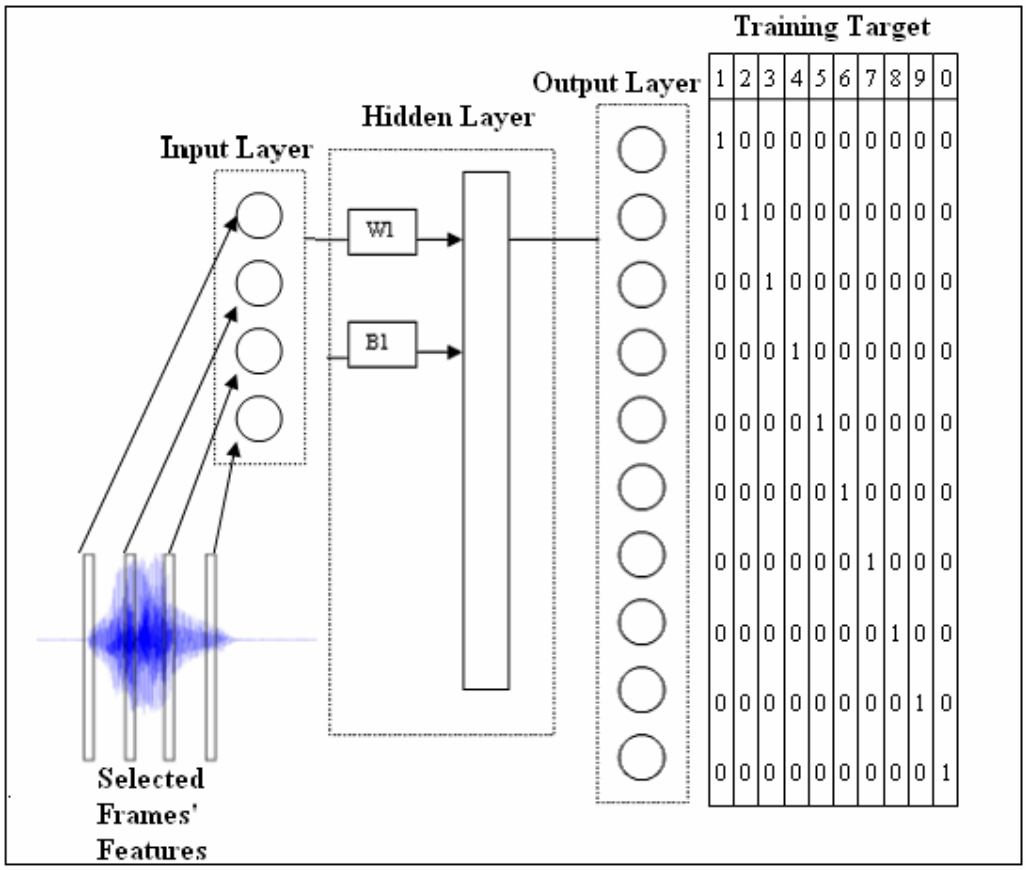

Fig. 10: Simplified Neural Network Architecture for digit recognition

The database is split into two groups, one for training the neural network, the other for testing the performance of the trained neural network. The first group, training database, comprises 56*10 = 560 male speakers' utterances and 57*10 = 570 female speakers' utterances. The second group, testing database, is comprised of same numbers of data from different speakers.

One of the common problems when using Multilayer Perceptrons is how to choose the number of neurons in the hidden layer. There are many suggestions on how to choose the number of hidden neurons in Multilayer Perceptrons. For example, the minimum number of neurons, h, can be:

$$h \geq \frac{p-1}{n+2} \tag{5.1}$$

where p is the number of training examples and n is the number of inputs of the network [19].

Equation 5.1 is used as a reference for choosing the number of neurons in the hidden layer. By referring to Equation 5.1, the number of hidden neurons must be around 34 if all training datasets are used since (113 speakers * 10 digits + 1) / (8 LPC * 4 + 2) ≈ 34.

In this paper, by referring to the numbers of hidden neurons proposed by Equation 5.1, neural networks with different numbers of hidden neurons (10, 30, 50 and 70 respectively) have been trained separately and the performance of each has been evaluated. Besides, the comparison also has been made among the networks trained with different numbers of datasets.

5.2 Automatic Speech Recognition System

Fig. 11 illustrates the stages in the automatic speech recognition system. Firstly, the end-point detection routine is applied to the raw data to find the region of interest for further processing (Fig. 11 b). This is followed by the selection of frames based on the starting and ending points (Fig. 11 c). The selected frames are then represented by LPC coefficients (Fig. 11 c). Finally, the features of all frames are fed into the neural network.

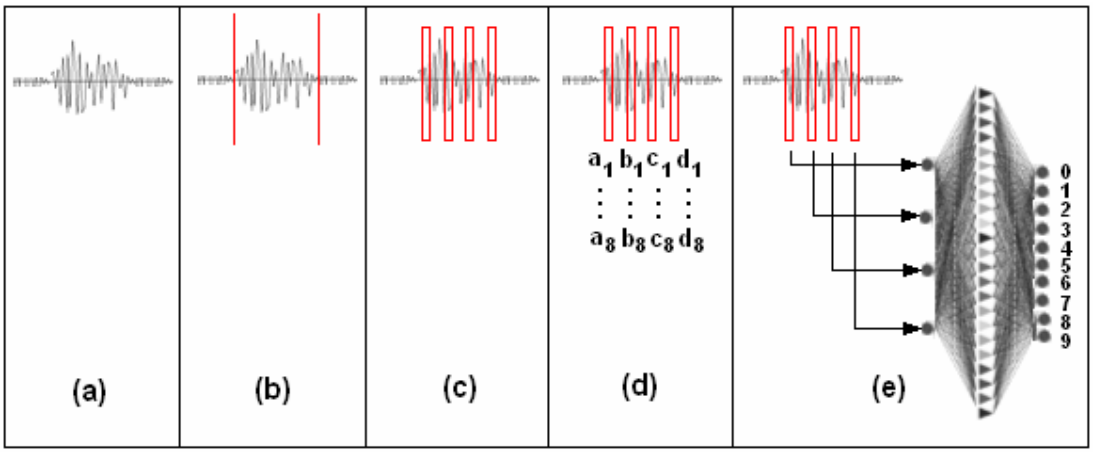

Fig. 11: Stages in the Automatic Speech Recognition System. (a) Raw data, (b) end-point detection, (c) frame selection, (d) feature representation (LPC), (e) feeding inputs to neural network

The same procedures are applied in the training and recognition stages. The only difference is that during the training stage, the targets are presented to the network so that the network is able to learn by examples. After the training, the network is used to recognise the untrained patterns and the results are discussed.

6.0 RESULTS AND ANALYSIS

Neural Networks with one hidden layer with sigmoid functions and the output layer with linear functions are used in this paper. There are 10 output neurons for all the networks while the numbers of hidden neurons vary from 10 to 70. The inputs of the network are the features of 4 selected frames with 256 samples per frame. Each frame is represented by either 8 LPC coefficients or the first 3 formants of the signal in the frame. All of the networks share the following common properties.

Table 1: Network Properties

Network Properties	**Information**
Training Method	Gradient descent with momentum and adaptive learning rate backpropagation.
Input layer	Features from 4 selected frames
Neurons Transfer Functions for hidden layer	Sigmoid transfer function
Neurons Transfer Functions for output layer	Linear transfer function
Epochs	20,000
Learning Rate	0.01
Momentum constant.	0.9
Maximum performance increase to reduce the learning rate	1.04
Ratio to increase learning rate	1.05
Ratio to decrease learning rate	0.7

Comparisons have been done in various ways:

a. Different numbers of hidden neurons
b. Different numbers of training datasets
c. Different features to represent the selected frames

Table 2 illustrates the results of comparing the performance of networks with different numbers of training datasets and different numbers of hidden layer neurons. Eight LPC coefficients (p=8) for each frame from the selected frames are the inputs to the neural network.

Table 2: Performance of neural networks with different numbers of hidden neurons and trained by different numbers of datasets. Network inputs are 8 LPC (p=8) coefficients * 4 frames = 32 (8 coefficients per frame)

Digit	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**0**	**Mean**
# Hidden Nodes	NN trained with 10 sets* data from male speakers and 10 sets* data from female speakers. LPC(p=8) 8 coefficients per frame.										
10	71.7	62.8	85.0	92.9	79.6	70.8	59.3	88.5	59.3	70.8	74.1
30	74.3	77.0	71.7	85.0	89.4	65.5	82.3	88.5	72.6	69.9	77.6
50	71.7	85.8	79.6	80.5	83.2	91.2	77.0	82.3	70.8	75.2	79.7
70	77.0	84.1	74.3	76.1	85.0	92.0	86.7	81.4	61.9	66.4	78.5
	NN trained with 20 sets* data from male speakers and 20 sets* data from female speakers. LPC(p=8) 8 coefficients per frame.										
10	84.1	84.1	88.5	97.3	92.0	81.4	89.4	90.3	60.2	92.9	86.0
30	90.3	88.5	88.5	92.0	92.0	92.9	92.9	90.3	92.9	88.5	90.9
50	85.8	83.2	91.2	95.6	93.8	88.5	97.3	92.0	84.1	89.4	90.1
70	88.5	92.9	92.0	96.5	93.8	93.8	92.0	86.7	91.2	94.7	92.2
	NN trained with 30 sets* data from male speakers and 30 sets* data from female speakers. LPC(p=8) 8 coefficients per frame.										
10	96.5	87.6	90.3	97.3	81.4	85.0	85.8	86.7	85.0	93.8	88.9
30	97.3	91.2	92.9	92.9	92.0	94.7	93.8	91.2	91.2	90.3	92.7
50	97.3	91.2	92.9	97.3	93.8	94.7	93.8	91.2	94.7	94.7	94.2
70	96.5	93.8	96.5	98.2	95.6	96.5	97.3	94.7	93.8	90.3	95.3
	NN trained with 56 sets* data from male speakers and 57 sets* data from female speakers. LPC(p=8) 8 coefficients per frame.										
10	96.5	88.5	93.8	97.3	92.9	80.5	92.9	93.8	92.0	93.8	92.2
30	99.1	91.2	97.3	99.1	96.5	94.7	93.8	92.9	91.2	94.7	95.0
50	100.0	94.7	100.0	97.3	96.5	94.7	98.2	94.7	92.9	94.7	96.4
70	99.1	94.7	98.2	99.1	96.5	93.8	95.6	95.6	95.6	92.9	96.1
* 1 set of data consists of 10 utterances from a speaker (zero to nine)											

Table 3 illustrates the results of comparing the performance of networks with different numbers of training datasets and different numbers of neurons in the hidden-layer. The first 3 formants of the frame signal are presented to the neural networks.

Table 3: Performance of neural networks with different numbers of hidden neurons and trained by different numbers of datasets. Network inputs are 3 formants * 4 frames = 12 (3 formants per frame)

Digit	1	2	3	4	5	6	7	8	9	0	Mean
# Hidden Nodes	NN trained with 10 sets* data from male speakers and 10 sets* data from female speakers. 3 formants per frame.										
10	83.2	85.0	82.3	88.5	85.0	78.8	67.3	67.3	62.8	89.4	78.9
30	75.2	85.8	85.0	92.0	92.9	77.0	64.6	72.6	59.3	89.4	79.4
50	77.9	88.5	84.1	80.5	84.1	82.3	73.5	69.9	61.9	88.5	79.1
70	69.9	85.8	81.4	85.0	90.3	78.8	77.9	77.0	61.9	90.3	79.8
	NN trained with 20 sets* data from male speakers and 20 sets* data from female speakers. 3 formants per frame.										
10	82.3	85.0	84.1	91.2	93.8	85.0	80.5	69.9	49.6	88.5	81.0
30	81.4	91.2	85.0	95.6	92.9	84.1	76.1	77.0	58.4	92.9	83.5
50	80.5	89.4	86.7	98.2	88.5	83.2	75.2	81.4	61.1	95.6	84.0
70	83.2	93.8	84.1	95.6	89.4	84.1	84.1	76.1	57.5	91.2	83.9
	NN trained with 30 sets* data from male speakers and 30 sets* data from female speakers. 3 formants per frame.										
10	81.4	92.0	85.0	95.6	92.0	77.9	69.0	69.0	41.6	91.2	79.5
30	85.8	90.3	79.6	94.7	92.9	85.8	80.5	74.3	61.9	89.4	83.5
50	85.0	89.4	91.2	95.6	93.8	84.1	83.2	77.0	61.9	92.9	85.4
70	85.0	95.6	90.3	94.7	94.7	87.6	83.2	83.2	73.5	95.6	88.3
	NN trained with 56 sets* data from male speakers and 57 sets* data from female speakers. 3 formants per frame.										
10	81.4	91.2	89.4	92.9	88.5	88.5	79.6	69.9	53.1	91.2	82.6
30	84.1	93.8	92.0	93.8	92.0	86.7	85.8	76.1	66.4	92.9	86.4
50	86.7	92.9	88.5	93.8	92.9	85.8	85.0	78.8	68.1	93.8	86.6
70	85.8	95.6	92.9	93.8	93.8	86.7	81.4	77.9	69.9	94.7	87.3
* 1 set of data consists of 10 utterances from a speaker (zero to nine)											

Fig. 12 and Fig. 13 summarise the performance of neural networks with different numbers of hidden neurons and trained by different numbers of datasets. The former figure uses the 8 LPC coefficients as the network's inputs and the latter takes 3 formants as the network's inputs.

From the figures, it is obvious that the LPC coefficients represent the speech signal better than the formants. This can be seen from the fact that the average performance in Fig. 12 is better than the average performance in Fig. 13. Besides, the performance of the system also improved with the increasing of the training data used to train the network. Both Fig. 12 and Fig. 13 show that the error reduces with the increasing of training data. Finally, the number of neurons in the hidden layer also affects the performance of the system. Equation 5.1 gives a reasonable reference for the number of hidden neurons.

7.0 CONCLUSIONS

In this paper, the approach of using neural networks for speaker independent isolated word recognition has been studied. Besides, an automatic speech recognition system has been designed using MATLAB programming. By the fully automated training and recognition process without the interference of manual cropping, an accuracy of more than 95% is achieved for unknown pattern (spoken by unknown speakers). This opens the door to the implementation in embedded systems, which requires small programs and simple algorithms for certain applications.

The results show that the performance of a network improves when more training datasets are used to train the network. Besides, the networks that use the LPC coefficients as inputs also perform better than the networks that use the first three formants as the networks' inputs.

For large vocabulary systems, this approach can also work together with other models to achieve higher accuracy. For example, it can be modified in order to recognise the phonemes in the speech signal and work with Hidden Marker Models (HMM) to recognise mandarin monosyllables [20].

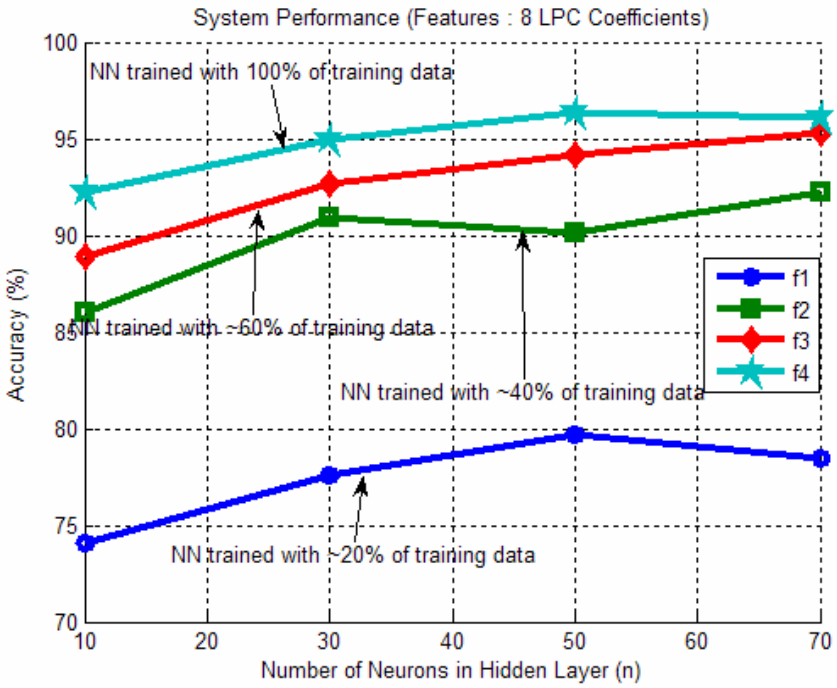

Fig. 12: Comparison of neural network performance with different numbers of hidden neurons and trained by different numbers of datasets. Networks' inputs are 8 LPC (p=8) coefficients * 4 frames = 32 (8 coefficients per frame)

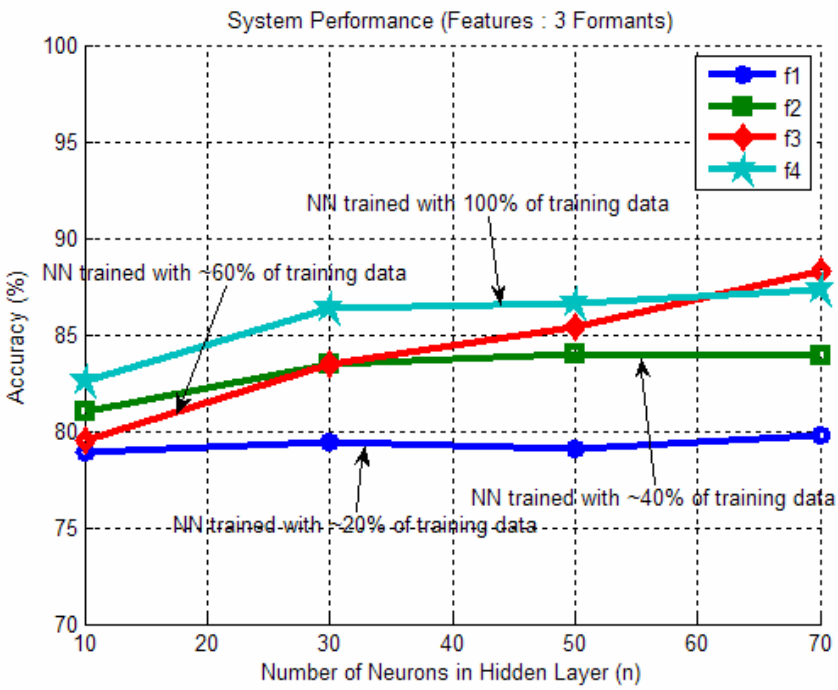

Fig. 13: Comparison of neural network performance with different numbers of hidden neurons and trained by different numbers of datasets. Networks' inputs are 3 formants * 4 frames = 12 (3 formants per frame)

REFERENCES

[1] L. R. Rabiner, B. H. Juang, *Fundamental of Speech Recognition*. Prentice Hall, New Jersey, 1993.

[2] P. G. J. Lisboa, *Neural Networks Current Application*. Chapman & Hall, 1992.

[3] B. A. St. George, E. C. Wooten, L, Sellami, "Speech Coding and Phoneme Classification Using MATLAB and NeuralWorks", in *Education Conference*, North-Holland University, 1997.

[4] M. Nakamura, K. Tsuda, J. Aoe, "A New Approach to Phoneme Recognition by Phoneme Filter Neural Networks". *Information Sciences Elsevier*, Vol. 90, 1996, pp. 109-119.

[5] J. Frankel, K. Richmond, S. King, P. Taylor, "An Automatic Speech Recognition System Using Neural Networks and Linear Dynamic Models to Recover and Model Articulatory Traces", in *Proceeding ICSLP*, University of Edinburgh, 2000.

[6] J. T. Jiang, A. Alwan, P. A. Keating, E. T. Auer L. E. Jr, Bernstein, "On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics". *EURASIP Journal on Applied Signal Processing*, Vol. 11, 2002, pp. 1174-1188.

[7] X. Z. Zhang, C.C. Broun, R. M. Mersereau, M. A. Clements, "Automatic Speechreading with Applications to Human-Computer Interfaces". *EURASIP Journal on Applied Signal Processing*, Vol. 11, 2002, pp. 1228-1247.

[8] H. S. Li, J. Liu, R. S. Liu, "High Performance Mandarin Digit Speech Recognition". *Journal of Tsinghua University (Science and Technology)*, 2000.

[9] H. S. Li, M. J. Yang, R. S. Liu, "Mandarin Digital Speech Recognition Adaptive Algorithm". *Journal of Circuits and Systems*, Vol. 4, No. 2, 1999.

[10] H. Demuth, M. Beale, *Neural Network Toolbox*. The Math Works, Inc., Natick, MA, 2000.

[11] S. Peeling, R. Moore, "Experiments in Isolated Digit Recognition Using the Multi-Layer Perceptron". *Technical Report 4073*, Royal Speech and Radar Establishment, Malvern, Worcester, Great Britain, 1987.

[12] B. Kammerer, W. Kupper, "Experiments for Isolated-Word Recognition with Single and Multi-Layer Perceptrons". *Abstracts of 1st Annual INNS Meeting*, Boston, 1988.

[13] D. Burr, "Experiments on Neural Net Recognition of Spoken and Written Text", in *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 36, 1988, pp. 1162-1168.

[14] L. R. Rabiner, M. R. Sambur, "An Algorithm for Determining the Endpoints for Isolated Utterances". *The Bell System Technical Journal*, Vol. 54, No. 2, 1975, pp. 297-315.

[15] Y. Shiraki, M. Honda, "LPC Speech Coding Based on Variable-Length Segment Quantization". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, No. 9, 1988.

[16] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*. Prentice Hall, USA, 2001.

[17] S. Haykin, *Neural Networks, A Comprehensive Foundation*. Prentice Hall, New Jersey, 1999.

[18] G. D. Magoulas, M. N. Vrahatis, G. S. Androulakis, "Improving the Convergence of the Backpropagation Algorithm Using Learning Rate Adaptation Methods". *Neural Computation*, Vol. 11, 1999, pp. 1769-1796.

[19] N. K. Kasabov, *Foundations of Neural Network, Fuzzy Systems, and Knowledge Engineering*. The MIT Press Cambridge, London, 1996.

[20] T. F. Li, "Speech Recognition of Mandarin Monosyllables". *The Journal of the Pattern Recognition Society,* 2003.

BIOGRAPHY

Chin Luh Tan received his Bachelor of Engineering in Electrical Engineering from Universiti Teknologi Malaysia. He has the experience of designing real-time control system and the implementation of direct digital controllers to real-time systems. Currently the author is undergoing his Master Program in Universiti Putra Malaysia and is working as a senior field application engineer in a software company. The author's research interests include speech recognition systems, image processing and automation systems, and embedded system design.

Adznan Jantan currently is a lecturer in Universiti Putra Malaysia (USM) under the Faculty of Engineering. Before that, he had been a lecturer in Universiti Sains Malaysia (USM), Multimedia University of Malaysia (MMU), Universiti Islam Malaysia (IIUM) and King Fahd University Petroleum Minwerals (KFUPM), Saudi Arabia. He obtained his Ph.D. from University College of Swansea, Wales, UK, in 1988. The author's research interests include speech recognition systems, data compression systems, human computer interaction systems, medical imaging, and smart school design systems.