A Speech Recognition System for Malaysian English Pronunciation Using Neural Network

Paulraj M P[1],Sazali Bin Yaacob[1], Ahamad Nazri[2], Sathees Kumar[1]
[1]School of Mechatronic Engineering, Universiti Malaysia Perlis, Perlis, Malaysia.
[2]Centre for Communication Skills & Entrepreneurship, University Malaysia Perlis, Perlis, Malaysia.
Email: satheesjuly4@gmail.com

Abstract- **The English language as spoken by Malaysians varies from place to place and differs from one ethnic community and its sub-group to another. In this paper, an automatic vowel classification system based on linear predictive coding (LPC) and neural network is presented to understand the English pronunciation as spoken by Malaysians. A database consisting of 11 words recorded from 10 speakers is created and used in this work. The input signal is pre-emphasised and frames features are extracted using LPC; a simple feedforward neural network trained by conventional backpropagation procedure in four different modes of activation functions is also proposed. To stabilize the cumulative error versus epoch training and to minimize the training time, a systole activation function is also proposed. The results obtained from the neural network trained by systole activation function are compared with the sigmoidal activation functions.**

Keywords - **LPC Coefficients, Back Propagation Neural Network, Systole Activation Function.**

I. INTRODUCTION

Speech recognition started as early as the 1950s, [5] One of the major problems in speech recognition is to find suitable front end features. Various front end features used by different researchers are linear prediction coefficients (LPC) reflection coefficients (RC), the linear prediction cepstrum coefficients (LPCC), mel-frequency cepstrum coefficients (MFCC), and linear frequency cepstrum coefficients (LFCC) [6]. Since many years, the two most common and successful approaches for speaker recognition are based on modelling the speech by Gaussian Mixture Models and Hidden Markov Models [7]. These methods are attractive for their phonetic discrimination capacity [8].

In this research work the LPC coefficients [9] extracted from the speaker phonemes act as discriminative features. Linear predictive coding (LPC) is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at low bit rate. It provides extremely accurate estimates of speech parameters, and is relatively efficient for computation. In this research work the LPC coefficients are used for extracting the features and are used for classification.

Artificial Neural Network (ANN) provides an alternative form of computing that attempts to mimic the functionality of the human brain. One of the most used learning method in ANN is back propagation [10-11]. The back propagation method (BP) is a learning procedure for training multilayer, feed forward neural networks. BP is being used in a wide variety of applications such as information processing, pattern recognition, signal processing and control applications [12]. BP procedure can be considered as a non-linear regression technique which trains a neural network to acquire an input output association using limited number of samples chosen from a population of input output patterns. However, BP is very slow in convergence. To overcome this problem several modifications have been suggested in this work.

In this paper, the hidden and output neurons of the neural network are activated by various combinations of sigmoidal activation functions such as, binary-binary, bipolar-bipolar, binary-bipolar, bipolar-binary sigmoidal activation functions and the results are compared. Then, to stabilize the cumulative error versus epoch training and to minimize the training time the hidden and output neurons of the neural network are activated using the proposed systole activation function. The results of the neural network model activated by systole activation function are compared with neural network model activated by the conventional sigmoidal activation function.

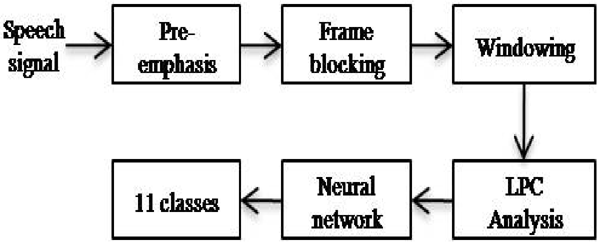

Fig. 1. Shows the block diagram of a simple speech phonemes classification system.

I. EXPERIMENTAL DATA SET

The English language as spoken by Malaysians varies from place to place and differs from one ethnic community and its sub-groups to another and it is quite different from Standard English as it is being influenced by their mother tongues. A significant level of variation in their pronunciation can be observed from the different ethnic communities. In this research work the data is collected from ten individuals from different ethnic communities and of different races and classified into 11 classes depending upon the positions of the tongue, tongue tension and front, central and back positions of lips as shown in Table 1.

TABLE 1
VOWEL CLASSIFICATIONS

Tongue positions	Tongue tension	Front position	Central position	Back position
High	Tense Relaxed	*'beet' /iy/* *'bit' /ih/*	- -	*'boot' /uv/* *'book' /uh/*
Mid	Tense Relaxed	*'bait' /ey/* *'bet' /eh/*	- *'but' /ah/*	*'boat' /ow/* *'bought' /ao/*
Low	Not applicable	*'bat' /ae/*	*'pot' /aa/*	

Table-1 shows the list of phonemes that are recorded and used for vowel classification. These 11 word phonemes are recorded from 10 peoples of different ethnic communities at a sampling frequency of 16Khz., This sampling frequency was chosen to minimize the effects of *aliasing* in the analog-to-digital conversion [15]. LPC features are then extracted from the pre-emphasised signals and then the features are partitioned into training and testing set. The features are then applied to the feed forward neural network with different activation functions and simple neural networks are developed and their performances are compared.

II. FEATURE EXTRACTION USING LPC

One of the more powerful analysis techniques is the method of linear prediction. [4] Linear predictive analysis of speech has become the predominant technique for estimating the basic parameters of speech. Linear predictive analysis provides both an accurate estimate of the speech parameters and also an efficient computational model of speech. The basic idea behind linear predictive analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples [2]. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and linear predicted values a unique set of parameters or predictor coefficients can be determined [1]. The recorded speech samples are then read and the following process are performed

A. Pre Emphasis

The digitized (sampled) speech signal s(n) is located through a first order pre-emphasis filter with the following transfer function

$$H(z) = 1 - az^{-1} \text{ where } a = 0.9375.$$

The pre-emphasised speech signal is blocked into number of frames with window size 1024 and with a frame rate of 50% is chosen from the experimental observations. The process of frame blocking is followed by windowing in order to reduce the energy at the edges and decrease the discontinuities at the edges of each frame.

B. LPC Coefficients

Each frame of the windowed signal is then auto correlated. The highest autocorrelation value *p* is the order of the LPC analysis. The next processing step is the LPC analysis, which converts each frame of *p*+1 autocorrelations [1] into an LPC parameter set in which the set consists of LPC coefficients. The formal method of converting from autocorrelation coefficients to a LPC parameter set is known as DURBIN's method [14]. By applying the above described procedures for each frame a set of LPC coefficients is computed. For each speech signal 12 LPC coefficients are calculated [16] and used as a feature set to model in the neural network.

III. FEATURE CLASSIFICATION USING THREE LAYER FEED FORWARD NEURAL NETWORK

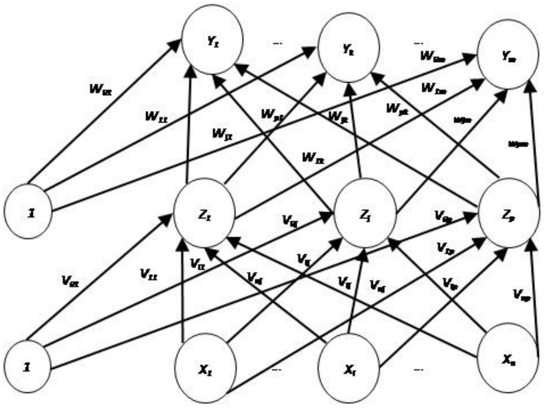

Fig. 2. Three Layer Feed Forward Neural Network

The extracted LPC coefficients are partitioned into two training sets which includes 60% and 70% of the total samples (110 samples) and the total samples are used as a testing set. A three layer feed forward neural network (FNN) having 30 neurons in the first hidden layer and 25 neurons in the second hidden layer and 11 neuron in the output layer is considered and a simple schematic network representation is shown in Figure 1. The hidden and the output neurons are activated using sigmoidal activation function

$$f(x) = 1/[1+\exp^{-x}]^{-1}$$

where *x* is the net input to the neuron.

The initial weights for the above network are randomized between -0.5 and 0.5. Depending upon the type of activation function used in the hidden and output layers, binary normalized or bipolar normalized input and output training pairs are used. The different types activation functions along with the type of normalization method used are depicted in Table-2. The input and output data are normalised and represented both in bipolar and binary forms using the following equation. [13]

$$x_n = \left(1.8\frac{x - x_{min}}{x_{max} - x_{min}}\right) - 0.9$$

where x_n is the normalised data, x_{max} *and* x_{min} are the maximum and minimum value of the data.

$$x_n = \left(0.8\frac{y - y_{min}}{y_{max} - y_{min}}\right) + 0.1$$

where y_n is the normalised data, y_{max} *and* y_{min} are the maximum and minimum value of the data respectively.

TABLE 2
NORMALISATION METHOD

Activation Function		Normalisation Method	
Hidden layer	output layer	Input Data	Output Data
Binary	Binary	Binary	Binary
Bipolar	Bipolar	Bipolar	Bipolar
Bipolar	Binary	Bipolar	Binary
Binary	Bipolar	Binary	Bipolar

When the hidden and output neurons are activated by a systole activation function, the network is trained using bipolar normalised input and output data set.

While training the neural network, a Mean Squared Error (MSE) tolerance of 0.1 is used [10].

$$MSE = \sum_{p=1}^{p} \sum_{k=1}^{k} (t_{k,p} - o_{k,p})^2 \qquad (3)$$

where *p* is the total number of patterns in data set, *k* is the output units, $t_{k,p}$ is the target value at the *k*th output neuron for the *p*th sample. The learning rate and momentum factor are chosen as 0.1 and 0.7 respectively. The values for learning rate, momentum factor and number of neurons in the hidden layers are chosen by experimental observations in order to get better classification accuracy. The network is trained with twenty five such trials under each activation functions and the epoch, network training parameters and the mean classification rate are tabulated and shown in Table 2.

TABLE 2
TRAINING RESULTS OF BINARY AND BIPOLAR SIGMOIDAL ACTIVATION FUNCTIONS.

<table>
<tr><td colspan="6">Number of Input Neurons- 12
Number of Hidden Neurons in first layer- 30
Number of Hidden Neurons in first layer- 25
Number of Output Neurons-11
Epochs Set-600
Performance Goal-0.01
Learning Rate 0.1</td></tr>
<tr><td>Percentage of Samples</td><td colspan="2">Activation Function</td><td>Mean Number of Epochs</td><td>Mean Time (sec)</td><td>Classification Rate (%)</td></tr>
<tr><td rowspan="4">60%</td><td>binary</td><td>binary</td><td>38</td><td>110</td><td>95.75</td></tr>
<tr><td>binary</td><td>bipolar</td><td>40</td><td>113</td><td>93.58</td></tr>
<tr><td>bipolar</td><td>binary</td><td>23</td><td>65</td><td>76.61</td></tr>
<tr><td>bipolar</td><td>bipolar</td><td>41</td><td>119</td><td>87.36</td></tr>
<tr><td rowspan="4">70%</td><td>binary</td><td>binary</td><td>31</td><td>101</td><td>96.29</td></tr>
<tr><td>binary</td><td>bipolar</td><td>50</td><td>163</td><td>94.9</td></tr>
<tr><td>bipolar</td><td>binary</td><td>25</td><td>83</td><td>79.64</td></tr>
<tr><td>bipolar</td><td>bipolar</td><td>43</td><td>140</td><td>88.5</td></tr>
</table>

The resulting mean square error (mse) versus epoch graph for bipolar sigmoidal and binary sigmoidal activation functions are shown in Figures 2 and 3.

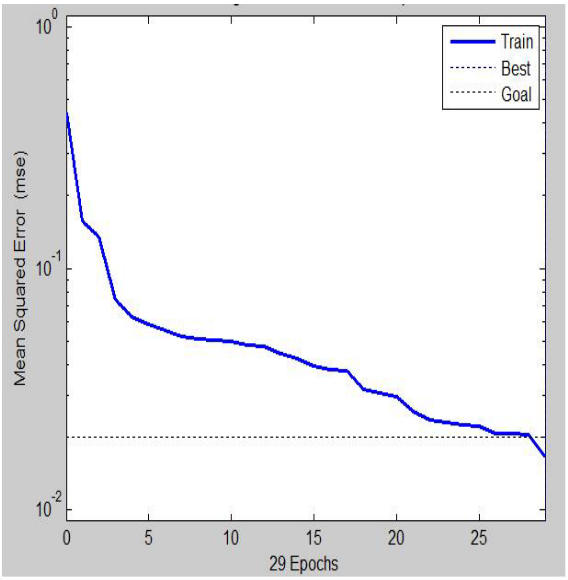

Fig.2. Mean square error *versus* epoch graph for bipolar sigmoidal activation functions.

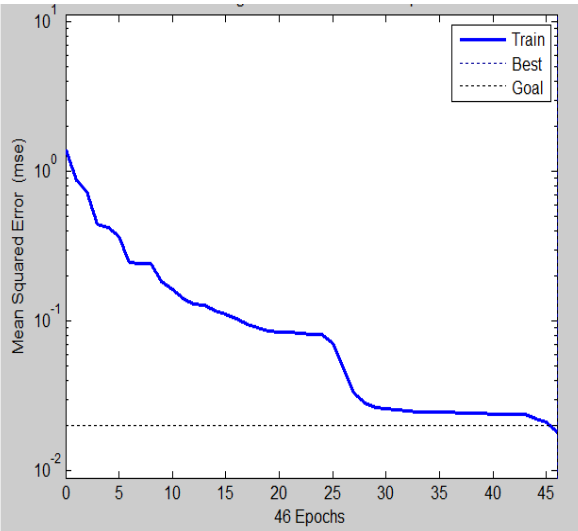

Fig.3. Mean square error (mse) versus epoch graph for binary sigmoidal activation function is shown in

IV. Systole Activation Function

In this method, instead of using binary and bipolar sigmoidal activation functions, the hidden layer neurons and the output neurons are activated by means of a systole activation function. The systole activation function can be represented as

$$f(x) = k_1 x e^{(-x^2 k_2)}$$

where x is the net input to a neuron k_1 and k_2 are the gain and the slope parameters of the systole activation function respectively. The network is trained with $k_1 = 1.7$ and $k_2 = 0.5$ respectively.

For each experimental study the network is trained for 25 different initial weights under systole activation functions and the results are studied and tabulated. The resulting mean cumulative error versus epoch graph is shown in Figure .3.

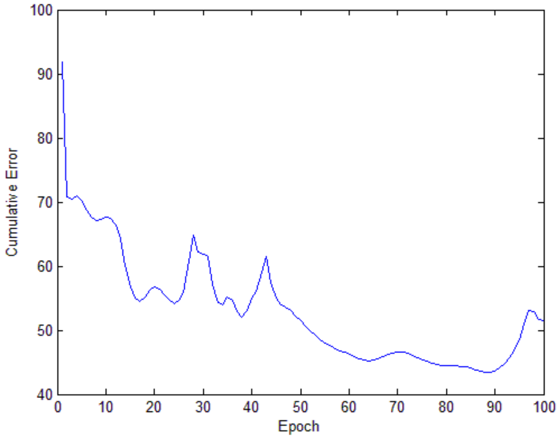

Fig.3. Cumulative error *versus* epoch graph for systole activation function

The network training parameters and the mean classification rate for systole activation function are shown in Table 3.

TABLE 3
Systole Sigmoidal Activation Function Training Result

<table>
<tr><td colspan="2">Number of input neurons- 12
Number of hidden neurons in first layer- 30
Number of hidden neurons in second layer- 30
Number of output neurons-11
Tolerance- 0.01</td><td colspan="2">Momentum factor- 0.1
Learning Rate-0.1
Epochs set-100
performance Goal-0.01
SYSTOLE Activation Function</td></tr>
<tr><td>Percentage Of samples (%)</td><td>Mean Number of Epocs</td><td>Time (sec)</td><td>Classification Rate (%)</td></tr>
<tr><td>60%</td><td>100</td><td>7.0553</td><td>89.80147</td></tr>
<tr><td>70%</td><td>100</td><td>8.1717</td><td>91.37932</td></tr>
</table>

V. Result and Discussion

In the experimental study, the 11 classes of voice samples classification are obtained from 10 individuals. LPC features are extracted from the recorded sound waves. These coefficients are then used as sample input pattern to the neural network.

The initial weights are randomized between -0.5 and +0.5 and the input and output data are normalized depending upon the activation function used in hidden and output layer. The FFNN is trained with the BP algorithm. The network is trained with 60%and 70% of total samples (110) against total testing samples. The average result for the network trained with sigmoidal activation function is shown in table 2 and the average result for the network trained with systole activation function is shown in table 3.

From table 2 and 3, it is observed that the network trained with systole activation function has an edge over the network trained with binary and bipolar sigmoidal activation function. The minimum and maximum training time for sigmoidal activation function is 23 sec and 50 sec and the classification rate is from 76.61 to 96.29. The minimum and maximum training time for systole activation function is 7 sec and 8 sec and the range of classification accuracy is from 89.80 to 91.37. It is observed that the systole activation function improves the training time and also the classification rate and also the misclassification.

VI. References

[1] N. S. Jayant and P. Noll, "*Digital Coding of Waveforms*. Englewood Cliffs", NJ: Prentice-Hall, 1984.

[2] A. Gersho, "Advances in speech and audio compression," *Proc. IEEE*, vol. 82, pp. 900–918, June 1994.

[3] L.R. Rabiner and B.H. Juang, "*Fundamentals of Speech Recognition*", Prentice-Hall, Englewood Cliffs, N.J., 1993.

[4] S.E. Levison, D. B. Roe , "A Perspective On Speech Recognition", IEEE Comm.mag. 1990, pp. 28-34.

[5] K. H. Davis, R. Biddulph, and *S.* Balashek, “Automatic recognition of spoken digits,” J. *Acoust. Soc. Am.*, vol. 24, no. 6, pp. 637-642, 1952.

[6] Steven B. Davis and Paul Mermelstein, “Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 28, no. 4, pp. 357-366, Aug. 1980.

[7] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T.F. Quatieri, "Speaker Verification using Text-Constrained Gaussian Mixture Models," Proc. of IEEE ICASSP, May 2002, vol. 1, p. 677-680.

[8] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Mixture Models", Digital Signal Processing, vol. 10, pp 181 -202, 2000.

[9] B. Jacob, "Automatic speech recognition", Doctorat, Paul Sabatier university, Toulouse, September 2003.

[10] L. V. Fausett, Englewood Cliffs, “*Fundamentals of Neural Networks: Architectures, Algorithms, and Applications”*, NJ: Prentice Hall, 1994.

[11] Limin, Fu. 1994. “Neural Network in computer Intelligence”, McGraw-Hill, New York, USA.

[12] Sivanandam, S. N., and paulraj M. 1999. “An Approach for stabilisation of class of neural network using slope parameter and error feed back”, International conference on cognitive systems, New delhi, india, pp: 250-261.

[13] Robert, A. Jacobs. 1998. “Increase Rates of Convergence through Learning Rate Adaption”, Neural networks, Vol 1, pp: 295-307

[14] B. S. Atal and L. S. Hanauer, ``Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, pp. 637-655, 1971.

[15] An automatic speaker recognition system
http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition/speaker_recognition.html

[16] Ibrahim I. Ibrahim, El-adawy M. I.ï.” Determination Of the Lpc Coefficients Using Neural Networks”, IEEE Trans. On neuralnetworks,