Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers

Bishnu Prasad Das[1], Ranjan Parekh[2]

[1] *School of Education Technology, Jadavpur University, India*
[2] *School of Education Technology, Jadavpur University, India*

ABSTRACT

This paper proposes an approach to recognize English words corresponding to digits Zero to Nine spoken in an isolated way by different male and female speakers. A set of features consisting of a combination of Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Zero Crossing Rate (ZCR), and Short Time Energy (STE) of the audio signal, is used to generate a 63-element feature vector, which is subsequently used for discrimination. Classification is done using artificial neural networks (ANN) with feed-forward back-propagation architectures. An accuracy of 85% is obtained by the combination of features, when the proposed approach is tested using a dataset of 280 speech samples, which is more than those obtained by using the features singly.

***Keywords* – isolated word recognition, linear predictive coding, mel frequency cepstral coefficients, zero crossing rate, short time energy, artificial neural networks.**

I. INTRODUCTION

Speech recognition is a popular and active area of research, used to translate words spoken by humans so as to make them computer recognizable. It usually involves extraction of patterns from digitized speech samples and representing them using an appropriate data model. These patterns are subsequently compared to each other using mathematical operations to determine their contents. In this paper we focus only on recognition of words corresponding to English numerals zero to nine. Some typical applications of such numeral recognition are voice-recognized passwords, voice repertory dialers, automated call-type recognition, call distribution by voice commands, directory listing retrieval, credit card sales validation, speech to text processing, automated data entry etc. The main challenges of speech recognition involve modeling the variation of the same word as spoken by different speakers depending on speaking styles, accents, regional and social dialects, gender, voice patterns etc. In addition background noises and changing of signal properties over time, also pose major problems in speech recognition. This paper proposes an approach for identifying spoken words corresponding to English digits zero to nine using a combination of features. The paper is organized as follows: section II provides reviews of earlier work in this area, section III describes the proposed approach, section IV tabulates details of experimentations done and results obtained, section V provides an analysis of the current work vis-à-vis earlier works, section VI provides overall conclusions and outlines future scopes.

II. PREVIOUS WORKS

Over the years a number of different methodologies have been proposed for isolated word and continuous speech recognition. These can usually be grouped in two classes : speaker-dependent and speaker-independent. Speaker dependent methods usually involve training a system to recognize each of the vocabulary words uttered single or multiple times by a specific set of speakers [1, 2] while for speaker independent systems such training methods are generally not applicable and words are recognized by analyzing their inherent acoustical properties [3, 4]. Hidden Markov Models (HMM) have been proven to be highly reliable classifiers for speech recognition applications and have been extensively used with varying amounts of success [5, 6, 7]. Artificial Neural Networks (ANN) have also been demonstrated to be an acceptable classifier for speech recognition [8, 9, 10]. Support Vector Machines (SVM) classifiers have been used to classify speech patterns using linear and non-linear discrimination models [11]. Various features have been used singly or in combination with others to model the speech signals, ranging from dynamic time warping (DTW) [12], Linear Predictive Coding (LPC) [9, 13], Mel Frequency Cepstral Coefficients (MFCC) [12, 14, 15]. Often a combination of several features as mentioned above, have shown improvement in recognition accuracies as compared to single features [16], as well as using other associated features like formant frequency and Zero Crossing Rate (ZCR) [10], Discrete Wavelet Transform (DWT) [17], especially in noisy environments [7, 15]. A review of speech recognition techniques can be found in [18].

III. PROPOSED APPROACH

This paper proposes an approach to recognize automatically digits 0 to 9 from audio signals generated by different individuals in a controlled environment. It uses a combination of features based on Short Time Energy (STE), Zero Crossing Rate (ZCR), Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficient (MFCC). A neural network (multi-layer perceptron : MLP) is used to discriminate the speech data models into respective classes.

3.1 Preprocessing

An audio speech signal is represented as a collection of sample values. Each speech signal represents a spoken sample of a digit between 0 to 9, is typically of duration 0.3 seconds and is recorded using a pre-defined sampling rate Fs. The digitized audio is symbolically represented as a n-dimensional vector :

$$x = [x_1, x_2, \ldots, x_n] \quad (1)$$

The pre-processing stage involves temporal domain filtering using a uniform one-dimensional filter H with an m-element coefficient vector $b = [b_1, b_2, \dots, b_m]$. The filtered output y is given by :

$$y = H(b, 1, x) \qquad (2)$$

The effect of the temporal filtering is to produce an output represented as a linear combination of the input and the filter coefficients i.e. the i-th output element is given by :

$$y_i = b_1 x_i + b_2 x_{i-1} + \cdots + b_m x_{i-m+1} \qquad (3)$$

3.2 Short Time Energy (STE)

The energy content of a set of samples is approximated by the sum of the square of the samples. To calculate STE the filtered signal is sampled using a rectangular window function of width ω samples, where $\omega << n$. Within each window, energy e is computed as follows :

$$e = \sum_{i=1}^{\omega} x_i^2 \qquad (4)$$

The energy for each window is collected to generate the STE feature vector having $W = \frac{n}{\omega}$ elements

$$E = \bigcup_{j=1}^{W} e_j \qquad (5)$$

3.3 Zero Crossing Rate (ZCR)

ZCR of an audio signal is a measure of the number of times the signal crosses the zero amplitude line by transition from a positive to negative or vice versa. The audio signal is divided into temporal segments by the rectangular window function as described above and zero crossing rate for each segment is computed as below, where $sgn(x_i)$ indicates the sign of the i −th sample x_i and can have three possible values: +1, 0, -1 depending on whether the sample is positive, zero or negative.

$$z = \sum_{i=1}^{\omega} \frac{|sgn(x_i) - sgn(x_{i-1})|}{2} \qquad (6)$$

The value for each window is collected to generate the ZCR feature vector having $W = \frac{n}{\omega}$ elements

$$Z = \bigcup_{j=1}^{W} z_j \qquad (7)$$

3.4 Linear Predictive Coding (LPC)

Linear prediction is a mathematical operation which provides an estimation of the current sample of a discrete signal as a linear combination of several previous samples. The prediction error i.e. the difference between the predicted and actual value is called the residual. If the current sample x_i of the audio signal be predicted by the past p samples and x'_i be the predicted value then we have :

$$x'_i = -a_2 x_{i-1} - a_3 x_{i-2} - \cdots - a_{p+1} x_{i-p} \qquad (8)$$

Here $\{1, a_2, \dots, a_p, a_{p+1}\}$ are the $(p+1)$ filter coefficients. In this case the signal is passed through an LPC filter which generates a $(p+1)$ element feature vector L_A and a scalar L_G which represents the variance of the predicted signal.

3.5 Mel Frequency Cepstral Coefficients (MFCC)

The signal is divided into overlapping frames to compute MFCC coefficients. Let each frame consist of N samples and let adjacent frames be separated by M samples where $M < N$. Each frame is multiplied by a Hamming window where the Hamming window equation is given by :

$$\omega(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad (9)$$

In the third step, the signal is converted from time domain to frequency domain by subjecting it to Fourier Transform. The Discrete Fourier Transform (DFT) of a signal is defined by the following :

$$X_k = \sum_{i=0}^{N-1} x_i . e^{-j\frac{2\pi k i}{N}} \qquad (10)$$

In the next step the frequency domain signal is converted to Mel frequency scale, which is more appropriate for human hearing and perceptions. This is done by a set of triangular filters that are used to compute a weighted sum of spectral components so that the output of the process approximates a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. The following equation is used to calculate the Mel for a given frequency :

$$M = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (11)$$

In the next step the log Mel scale spectrum is converted to time domain using Discrete Cosine Transform (DCT). DCT is defined by the following, where $\propto$ is a constant dependent on N :

$$X_k = \propto . \sum_{i=0}^{N-1} x_i . \cos\left\{\frac{(2i+1)\pi k}{2N}\right\} \qquad (12)$$

The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficients is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vectors. Out of all the coefficients usually the first q MFCC coefficients are retained, leading to a q element MFCC vector M_C

The final feature vector for modeling each speech signal consists of the collection of the $(p+1)$ element LPC vector L_A, the 1 element LPC scalar L_G , the q element MFCC vector M_C, the W element ZCR vector Z, the W element STE vector E. Hence feature vector length is : $(p+1) + 1 + q + W + W$.

$$F = \{L_A, L_G, M_C, Z, E\} \qquad (13)$$

3.6 Classification Scheme

A word class i consists of set of j utterances by k speakers. For each utterance a combined feature vector is computed as per equation (13). A word class is characterized by the collection of its feature values obtained during a training phase. A test utterance with its computed feature vector is said to belong to a specific class if the probability of it being a member of that class is maximum. Class probability is determined by Artificial Neural Network (ANN) classifiers using feed-forward and back-propagation architectures.

IV. EXPERIMENTATIONS AND RESULTS

4.1 Dataset

The dataset consists of 280 speech samples recorded by 28 speakers each uttering the name of 10 digits, from 0 to 9, in English. Out of 28 speakers 14 are male and 14 female. The speech samples are recorded directly over microphone in a controlled environment. All the audio signals are stored in the WAV format with sample rate of 22050 Hz, bit rate of 16 bits and in mono (single channel) format.

4.2 Training Phase

The training set consists of 200 speech samples spoken by 20 speakers, 10 male and 10 female, each uttering the name of the 10 digits. Each speech file is subjected to a temporal domain filtering with a uniform one-dimensional filter H with a 2-element coefficient vector [1, -0.95]. Fig. 1 depicts pictorial representation of one of the speech files before and after temporal filtering. The filtered signal is shown in red, while the original signal in blue.

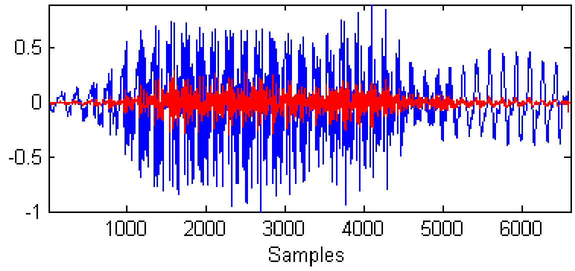

Fig. 1: Original speech signal (blue) and after temporal filtering (red)

LPC coefficients are extracted from the speech file using an LPC order of 16. This generates a 17-element vector L_A and a scalar L_G leading to an 18-element LPC vector. MFCC coefficients are then generated from the speech signal using an MFCC order of 15, which generates a 15-element MFCC vector M_C. Finally using a rectangular window of size $W = 15$, ZCR and STE vectors, each of 15 elements, are computed and added to the feature vector which becomes (18+15+15+15) = 63 elements in size.

Fig. 2 depicts feature plots of the 10 digits for an average of 20 speakers of the training set. The 63 elements of the feature vector are shown along the X-axis while their corresponding values are shown along the Y-axis.

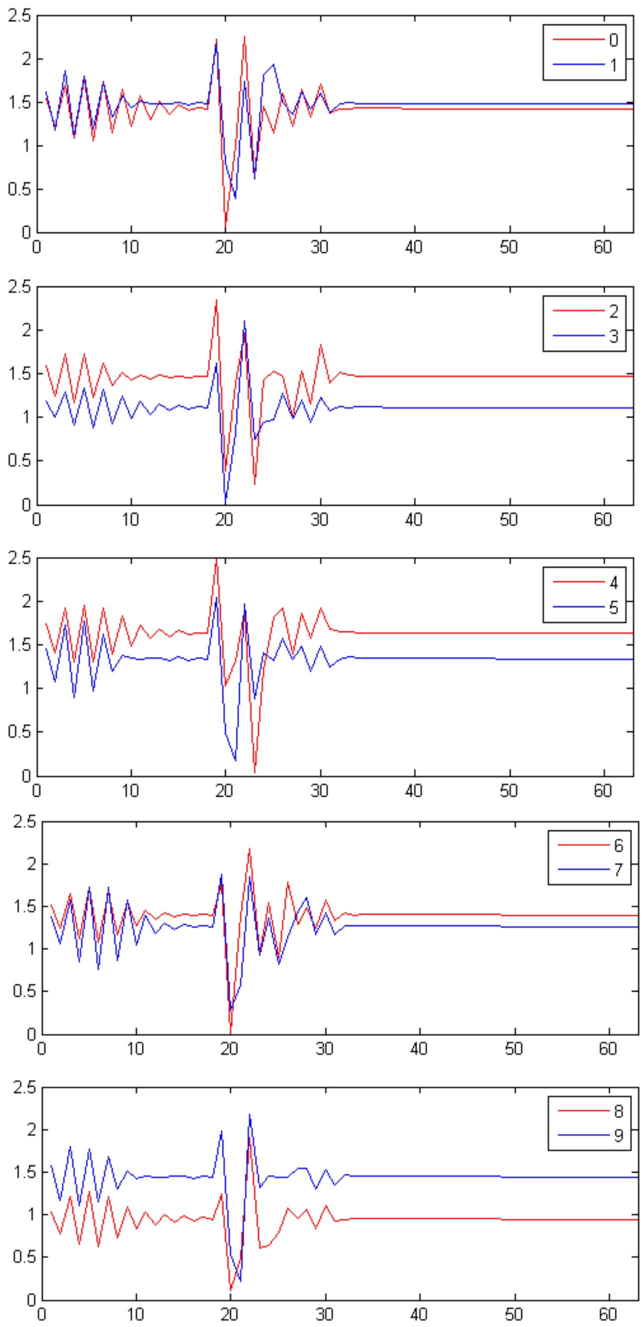

Fig. 2: Plots of the 63-element feature vector for digits 0 to 9 averaged over 20 speakers of the training set

4.3 Testing Phase

The testing set consists of 80 speech samples spoken by 8 speakers, 4 male and 4 female, each uttering the name of the 10 digits. Each speech file is subjected to the same steps of temporal filtering, followed by the extraction of the 63-element feature vector as described in the training phase. Fig. 3 depicts feature plots of the 10 digits for an average of 8 speakers of the testing set.

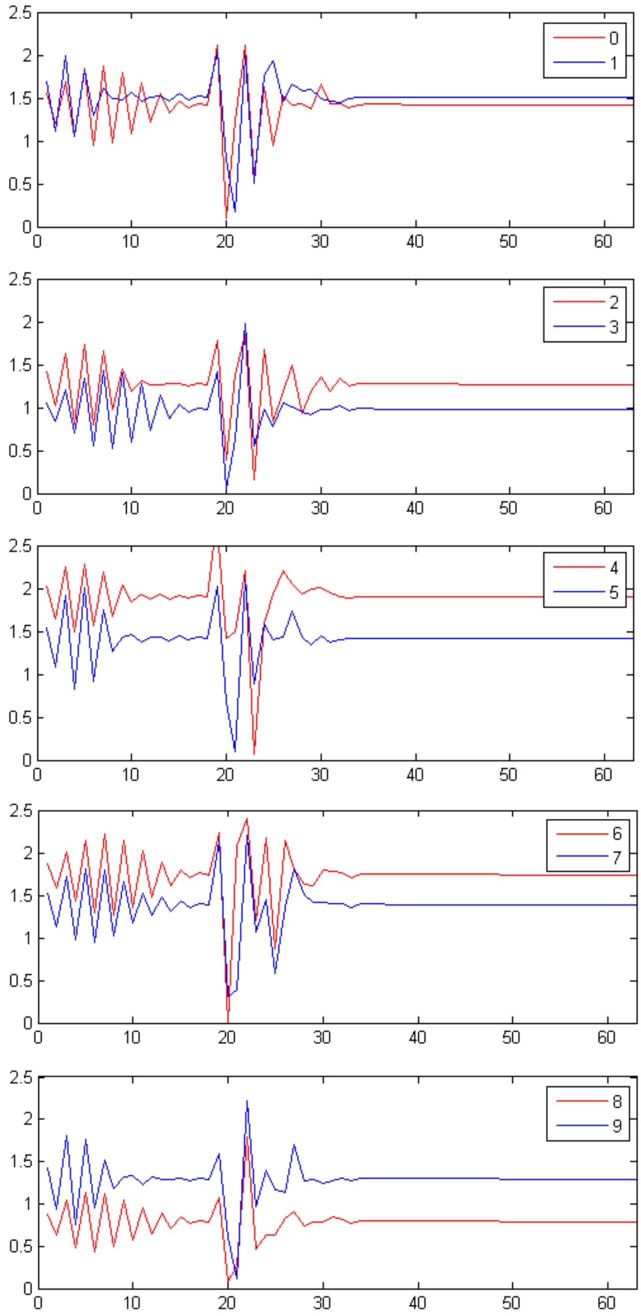

Fig. 3: Plots of the 63-element feature vector for digits 0 to 9 averaged over 8 speakers of the testing set

4.4 Classification

Classification of the speech signals is done by using a neural network (MLP : multi-layer perceptron). The MLP architecture used is 63-299-10 i.e. 63 input nodes (for the 63-element feature vector), 299 nodes in the hidden layer, and 10 output nodes (for discriminating between 10 words), log-sigmoid activation functions for both the neural layers, learning rate of 2.0 and Mean Square Error (MSE) threshold of 0.005 for convergence. The convergence plot and the MLP output are shown in Fig. 4. The accuracy obtained is 85% and requires 1097 epochs for convergence.

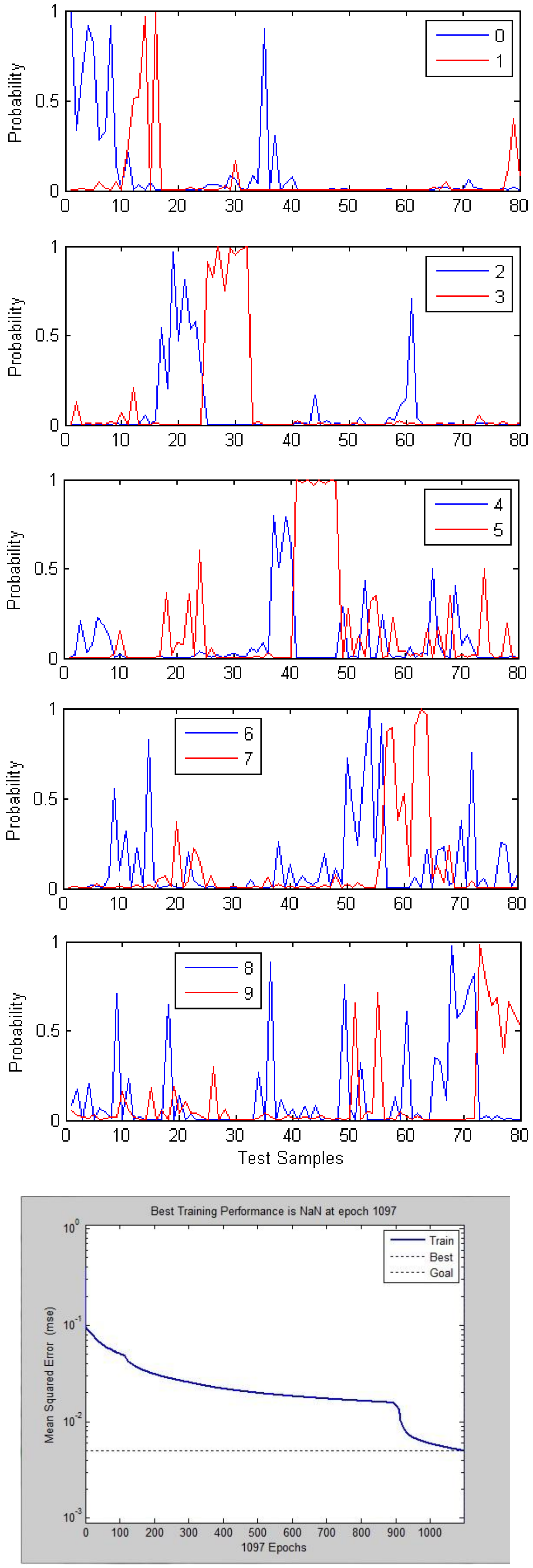

Fig. 4: NN classification of digits 0 to 9 (a) output plots (b) convergence plot

The output plots indicate the probability of the test samples as belonging to one of the 10 classes. Since the test samples are fed to ANN sequentially, samples 1-8 belong to class ‘0’, 9-16 to class ‘1’, 17-24 to class ‘2’, 25-32 to class ‘3’, 33-40 to class ‘4’, 41-48 to class ‘5’, 49-56 to class ‘6’, 57-64 to class ‘7’, 65-72 to class ‘8’, 73-80 to class ‘9’. The class memberships are indicated by a peak in the probability values.

V. ANALYSIS

Table 1 indicates the accuracies obtained by implementing the proposed algorithm on a dataset consisting of speech samples by 28 different speakers (both male and female). In each of the cases, the accuracy obtained is compared with those outlined in [9] and [14] consisting of single features of LPC and MFCC, on the same data set. Classification done using ANNs are also compared with results obtained by the Euclidean metric. The time required for calculating four features of all 280 speech samples is 24 seconds on a system with 3GB RAM and Intel Core2Duo Processor. To put the results in perspective with the state of the art, the system described in [8] achieves 94% accuracy with isolated digit recognition. Error rates of 23.1% are detected if only MFCC features and PLP features are considered separately in [16]. An accuracy of 91.4% is reported in [9] and 79.5% in [14].

Table 1 Recognition Accuracies

Classifier	Only LPC	Only MFCC	LPC + MFCC + ZCR + STE
ANN	37.5%	51.25%	85%
Euclidean Distance	23.75%	30%	57.5%

VI. CONCLUSIONS AND FUTURE SCOPES

This paper outlines a system to recognize English words corresponding to digits zero to nine, spoken by a set of 28 speakers. Words are classified using a combination of features based on LPC, MFCC, ZCR and STE. The recognition accuracy is seen to be better than achieved using these features individually, as has been done in some of the previous works, and is comparable to those reported in extant literature. The overall accuracy can be enhanced by combining more features of the speech samples. Also different windows like Hamming, Hanning or Blackman windows can be considered for filtering the speech samples.

REFERENCES

[1] M. B. Herscher, R. B. Cox, An adaptive isolated word speech recognition system, *Proc. Conf. on Speech Communication and Processing*, Newton, MA, 1972, 89-92.

[2] F. Itakura, Minimum prediction residual principle applied to speech recognition, *IEEE Transaction on Acoustics, Speech and Signal Processing, ASSP-23*, 1975, 67-72.

[3] V. N. Gupta, J. K. Bryan, and J. N. Gowdy, A speaker-independent speech recognition system based on linear prediction, *IEEE Transactions on Acoustics, Speech, Signal Processing, ASSP-26*, 1978, 27-33.

[4] L. R. Rabiner, J. G. Wilpon, Speaker independent isolated word recognition for a moderate size vocabulary, *IEEE Transactions on Acoustics, Speech, Signal Processing, ASSP-27*, 1979, 583-587.

[5] L. R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, *Proc. IEEE, 77(2)*, 1989, 257-286.

[6] A. Betkowska, K. Shinoda, S. Furui, Robust speech recognition using factorial HMMs for home environments, *EURASIP Journal on Advances in Signal Processing*, Article ID 20593, 2007, 1-10.

[7] M. S. Rafiee, A. A. Khazaei, A novel model characteristics for noise-robust automatic speech recognition based on HMM, *Proc. IEEE Int. Conf. on Wireless Communications, Networking and Information Security (WCNIS)*, 2010, 215-218.

[8] R. Low, R. Togneri, Speech recognition using the probabilistic neural network, *Proc. 5th Int. Conf. on Spoken Language Processing*, Australia, 1998.

[9] Thiang, S. Wijoyo, Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile robot, *International Conference on Information and Electronics Engineering IPCSIT vol.6*, 2011, 179-183.

[10] D. Paul, R. Parekh, Automated speech recognition of isolated words using neural networks, *International Journal of Engineering Science and Technology (IJEST), 3(6)*, 2011, 4993-5000.

[11] J. P. Sendra, D. M. Iglesias, F. D. Maria, Support vector machines for continuous speech recognition, *Proc. 14th European Signal Processing Conference*, Italy, 2006.

[12] L. Muda, M. Begam, I. Elamvazuthi, Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques, *Journal of Computing, 2(3)*, 2010, 138-143.

[13] J. L. Ostrander, T. D. Hopmann, E. J. Delp, Speech recognition using LPC analysis, *Technical Report RSD-TR-1-82*, University of Michigan, 1982.

[14] A. A. M. Abushariah, T. S. Gunawan, O. O. Khalifa, English digits speech recognition system based on Hidden Markov Models, *Int. Conf. on Computer and Communication Engineering*, 11-13 May 2010, 1-5.

[15] B. Kotnik, D. Vlaj, Z. Kacic, B. Horvat, Robust MFCC feature extraction algorithm using efficient addictive and convolutional noise reduction procedures, *ICSLP'02 Proceedings*, 2002, 445-448.

[16] A. Zolnay, R. Schlueter, H. Ney, Acoustic feature combination for robust speech recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 2005, 457-460.

[17] P. Ambalathody, Speech recognition using wavelet transform, unpublished (www.scribd.com/doc/ 36950981/ Main-Project-Speech-Recognition-using-Wavelet-Transform)

[18] W. Ghai, N. Singh, Literature review on automatic speech recognition, *International Journal of Computer Applications, 41(8)*, 2012, 43-50.