Analysis of Combined Use of NN and MFCC for Speech Recognition

Safdar Tanweer, Abdul Mobin, Afshar Alam

***Abstract*—**The performance and analysis of speech recognition system is illustrated in this paper. An approach to recognize the English word corresponding to digit (0-9) spoken by 2 different speakers is captured in noise free environment. For feature extraction, speech Mel frequency cepstral coefficients (MFCC) has been used which gives a set of feature vectors from recorded speech samples. Neural network model is used to enhance the recognition performance. Feed forward neural network with back propagation algorithm model is used. However other speech recognition techniques such as HMM, DTW exist. All experiments are carried out on Matlab.

***Keywords*—**Speech Recognition, MFCC, Neural Network, classifier.

I. Introduction

IN the recent past, speech recognition by computer becomes an active field of research where words/digits spoken by human are made computer recognizable. Brief glimpse of previous work are initially reported in this paper. So far numbers of different methodologies have been proposed by different researcher for continuous and isolated word recognition. Recognition task falls broadly into two classes: speaker dependent and speaker independent [1]-[3]. However HMM has been assumed as reliable classifier for speech recognition as it widely used by many researcher and have got mix amount of success [4], [6]. In the recent past, ANN has also been widely acceptable and used as classifier for speech recognition. Various features have been used to model speech signal such as DTW, LPC, MFCC [5], [7], [8], [12] etc. singly or collectively shows improvement in recognition accuracies. Recognition task starts with signal captured by microphone in noise free environment. The high accuracy of recognition is the key thrust of this paper, since speech signal contains energy band of (0-5KHZ) their property varied with time (shown in Figs. 1-3) therefore time varying Fourier transform are used to study speech signals [5], [6] during very short interval of time some of the parameter of speech signals are constant such as energy, zero crossing correlation etc. Hence during that period of time, by introduction of hamming window, these constant signals are divided into blocks of small duration. The MFCC data are then evaluated. These data sets are used for training, validation and testing by using ANN [4]. The training data set is used to train the Feed Forward Neural Network where the weights are adjusted accordingly. According to the learning algorithms, Testing data sets are used to evaluate and analyse the efficiency of classifier. Testing data set is not used during the training session; these data sets are unknown to classifier.

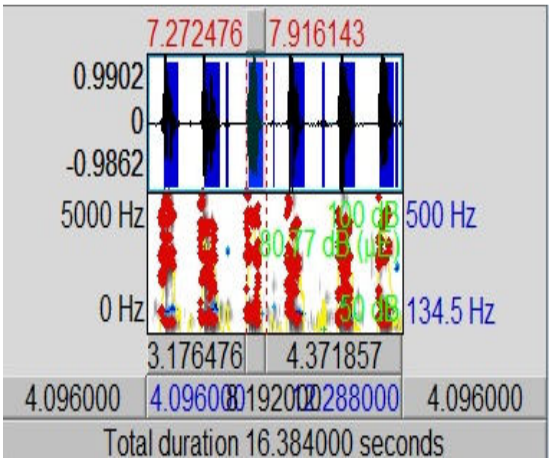

Fig. 1 Speech Sample

The utterances of speech samples are recorded from four different speakers for the isolated word zero to nine (0-9) keeping 16 bits sampling frequency by using PRAAT Software. We have computed here MFCC of the recorded utterances.

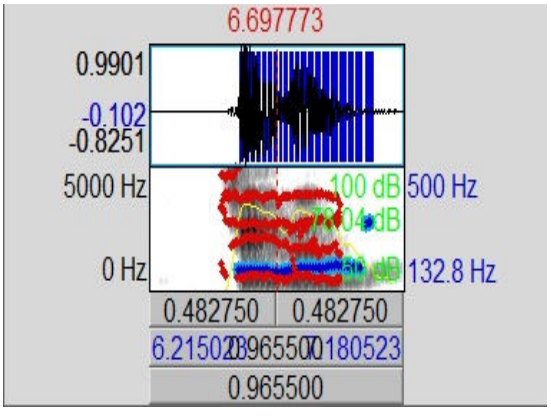

Fig. 2 Speech Sample of selected part

Safdar Tanweer is with the Department of Computer Science, Hamdard University, New Delhi, India (mobile: +919810465885; e-mail: safdartanweer@yahoo.com).

Dr. Abdul Mobin and Dr. Afshar Alam are with the Department Of computer science, Hamdard University, New Delhi, India.

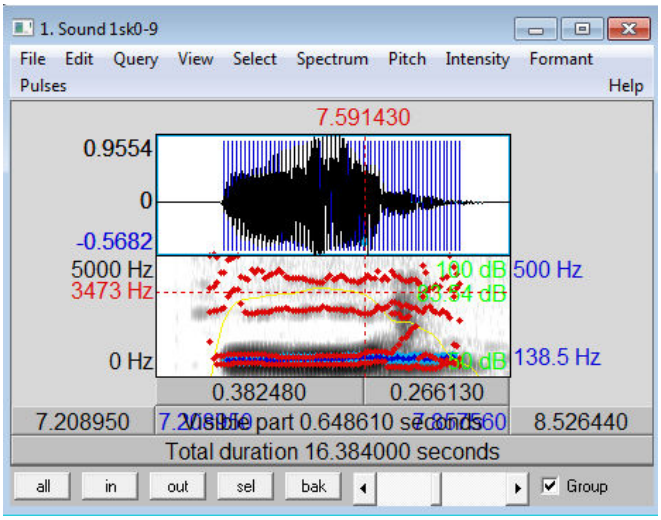

Fig. 3 Zoom in view of selected Speech Sample

A. Methodology [9], [10]

The recorded utterances have been segmented into frames. We have taken 20 ms window size and 15 ms frame size. The number of frames varied depending upon the amount of information available in speech sample. The steps involved are as below

i) Digitizing the speech that is to be recognized
ii) Capture the feature of the speech signal
iii) Reduce the feature set using self organized map
iv) Back propagation based phoneme classifier is used to classify each set of feature corresponding to the phoneme utterance of corresponding phoneme.

II. MFCC

Mel-scale frequency cepstral coefficients was developed by Stevens for feature extraction of front end input speech signal [3], [13] since MFCC is being categorized in frequency domain in nature. The key difference between MFCC and cepstral coefficients lies in the processing involved when characterizing the speech signals [6], [7].

A. MFCC Design Steps [3], [11]

1. Frame Blocking: Speech signal is a continuous signal their parameters are divided into frames for proper analysis and investigation.
2. Windowing: Since Speech signal is non stationery in nature, their parameters changes at every 15 ms approx. So we are windowing the frames at 15 ms for MFCC, these are logarithmic in nature.
3. FFT: Used for Transforming Time domain speech signal into Frequency domain. It is computationally easier to calculate DFT of any signal that in turn saves time and energy.
4. Mel-Filter Bank transformation: it is a logarithmic scale as the human auditory system which is also logarithmic in nature and is very robust for speech recognition

$$\text{Mel-}(f) = 2595 \log 10 \{1+f/1000\}$$

where f = actual frequency of speech.

5. DCT: Discrete cosine transform is applied for Mel filter Bank to obtain MFCC. It minimizes the distortion in frequency domain.

III. Back Propagation Neural Network

Neural Network [4], [14] is organized in layers which are made of inter connected nodes. Feature vectors of different samples (i.e. MFCC coefficients) are used as data sets for neural network. The feature vector data set is divided into three data sets. These are training data set, validation data set and testing data set. The weights are adjusted as per requirement. The hidden layer is linked to the output layer where the result is displayed. Fig. 4 shows input layer one or more hidden layers and output layer.

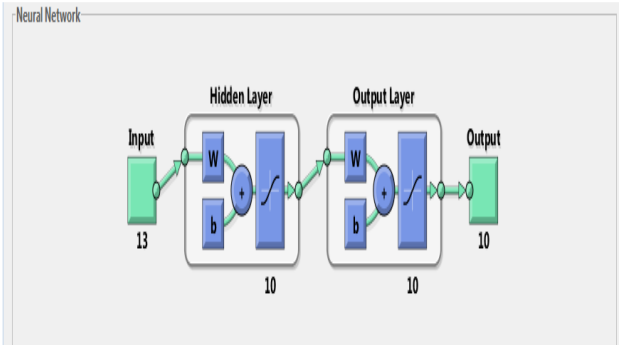

Fig. 4 Neural Network Architecture

In this paper we have used artificial neural network (ANN) for speech classification. The neural networks (NN) are trained in supervised manner using back propagation (BP) algorithm. There are many variations in training algorithm to classify new data and then the network is trained with this data set on the basis of test procedure.

IV. Speech Classification Using DTW and HMM

Researcher have also used DTW [5], [8] and HMM for speech recognition. DTW (Dynamic time warping) algorithm is widely used to measure similarities between two sequences. That may vary in time or speed as in case of speech recognition it allows computer to find a optimal match between two given patterns of speech sequence. The sequences are warped nonlinearly to match each other. The entire problem is divided into small number of states and needs a decision to be made. [7], [8] HMM is popular tool for modeling of time series Data of speech recognition and have got great success for speech pattern classification. The classification is based on combining digital signal processing technology (DSP) with pattern recognition methods that have been central to progress in automatic speech recognition (ASR) [11].

V. Experimental Result

On the basis of experiment carried out using MATLAB features are extracted in the form of 13 MFCC coefficients

from each frame. A database is formed for each isolated word using these MFCC coefficients. The database is divided into 10 classes corresponding to the isolated words spoken in the form of 0,1,2,3,4,5,6,7,8,9. Neural network pattern recognition tool of MATLAB is used for the classification of these MFCC pattern. The recognition efficiency of isolated word is found to be 100%.

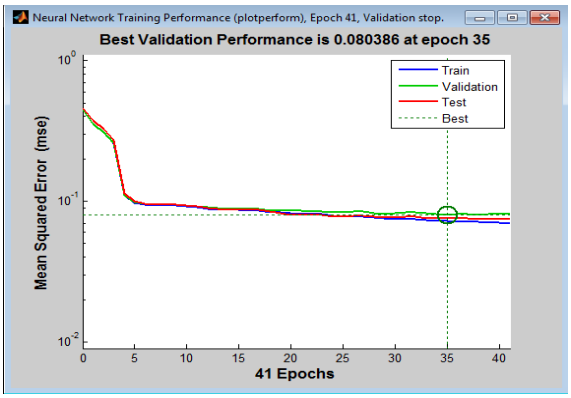

Fig. 5 Performance Curve

Fig. 5 shows the neural network performance in terms of mean squared error versus epochs curve.

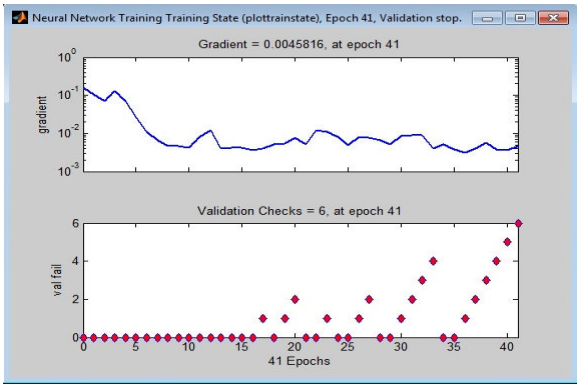

Fig. 6 Plot Training State

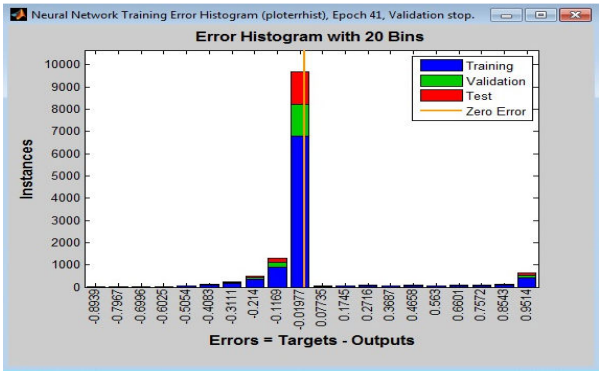

Fig. 7 Error Histogram

Fig. 6 shows gradient and validation during the training epochs. The result then is shown in Figs. 7 and 8 in the form of error histogram and confusion matrix. Fig. 7 shows the error histogram of training, validation and test phases. Error histogram shows the error is very close to the zero error between target and output of the pattern recognizer.

The confusion matrix consists of training, validation, test and all confusion matrixes. The diagonal of matrix shows the correct classification which is well enough for the recognition of isolated word in the form of a class.

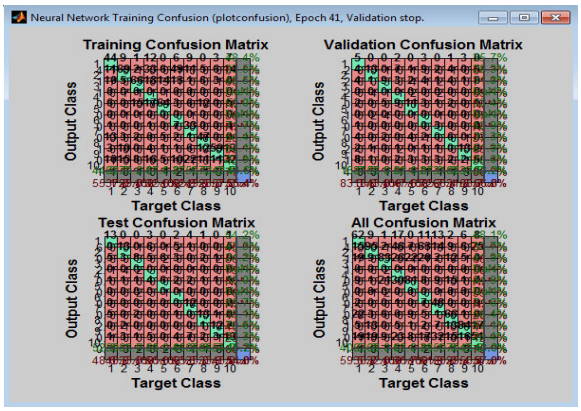

Fig. 8 Confusion Matrix

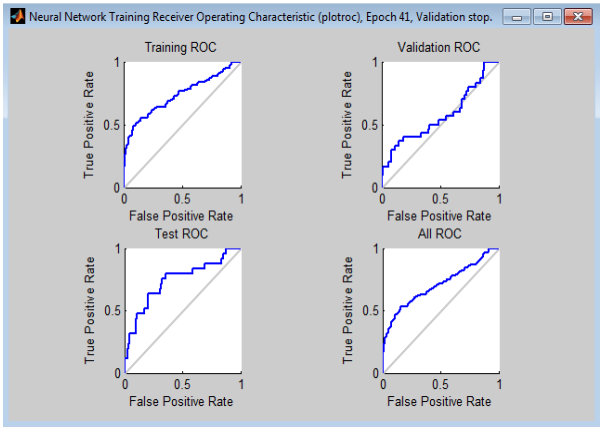

Fig. 9 Receiver Operating Characteristics (ROC)

Fig. 9 shows training, validation, testing and all receiver operating characteristics in terms of true and false positive rates.

VI. Conclusion

When features are classified using artificial neural network [ANN] pattern recognition tool with MATLAB 2010b, the testing result shows accuracy of 59.2 percent since accuracy of classification among ten classes [0-9] above 50 percent is considered as the recognition of one class is as 100 percent the recognition can be enhanced further by using more sample vectors of different groups of speaker.

Acknowledgement

We sincerely acknowledge the entire researcher for their wide contribution in the area of speech recognition technology that gives us ray of light to understand and work for further improvement in the above said field.

REFERENCES

[1] L.R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, Proc. IEEE, 77(2), 1989, 257-286.
[2] L.R Rabiner J. G. Wilpon, Speaker independent isolated word recognition for a moderate size voculabary, IEEE Transaction on Acoustics, Speech Signal Processing, ASSP-27, 1979, 583-587.
[3] Picheny M; Nahamou D, Goel V, Kingbusy B, Ramabhadran S.J Saon, G Trends and Advances in Speech recognition" IBM Journal of Research and Development, Vol no-5 PP-2:1-2:18 sept-oct-2011
[4] Haykin, S., "Neural Networks A Comprehensive Approach", Prentice Hall, 1999.
[5] L. Muda M. Begam, I. Elamvazuthi, Voice recognition algorithms using Mel Frequency Cepstral Coeffcient (MFCC) and Dynamic Time Warping (DTW) techniques, Journal of Computing, 2 (3), 2010, 138-143.
[6] Environmental Natural sound detection and classification using content based retrierval (CBB) and MFCC by Subarta Mandal, Institutional journal of engineering research and application (IJERA) ISSN:2248-9622, Vol:2 , issue-6 Nov-Dec 2012 PP-123-129.
[7] Chadawan Ittichaichareon, Siwat Sukasri and Tha-Weesak Yingthawornsuk" Speech recognition using MFCC, published in international conference on computer Graphics simulation and modeling (ICGSM 2012) July 28-29 2012 pattaya (Thailand).
[8] Stan Salvador and Pjilip Chan Fast DTW: Toward Accurate Dynamic time Warping in Linear time space, Florida Institute of Technology, Melbourne.
[9] Mohd Tamil, MOhd Yamani Idna Idris" Quarnic verse recitation feature extraction using MFCC AL-Quran & AL- Hadith Academy of Islamic Studeis of Malaya.
[10] 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.
[11] Performances Analysis of learning classifier for spoken digit under Noisy condition vol.4, No-3 March 2013 in Journal of emerging trends in computing and information science.
[12] W. Ghai, N Singh, Literature review on automatic speech recognition, International Journal of Computer Applications, 41 (8), 2012, 43-50.
[13] Pramod B. Patil" Multilayered Network for LPC Based Speech Recognition", IEEE 1998.
[14] Kung S, Digital Neural Network, Printice Hall 1993.