IJCSC International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 41-45

Performance Evaluation of MLP for Speech Recognition In Noisy Environments Using MFCC & Wavelets

P Phani Kumar[1], K S N Vardhan[2] & K Sri Rama Krishna[3]

[1]Assistant Professor, Dept of ECE, V R Siddhartha Engineering College, Vijayawada
[2]PG Student, Dept of ECE, V R Siddhartha Engineering College, Vijayawada
[3]Professor and Head, Dept of ECE, V R Siddhartha Engineering College, Vijayawada
E-mail: [1]polasi_phani@yahoo.co.in, [2]sivanagavardhan@yahoo.com, [3]srk_kalva@yahoo.com

ABSTRACT

Speech is a very powerful and fast tool for communication. That is the reason; the problem of automatic speech recognition has been fascinating the computer scientists. Two robust speech recognition systems in noisy environment have been implemented in this work. One system is based on Mel-Frequency Cepstrum Coefficients and the other is based on Wavelet Packet Filter bank. Major drawback of all Automatic Speech Recognition (ASR) systems is poor performance in the noisy environment. Hence, a speech segregation stage using Time-Frequency masking is employed as the front end of the systems implemented here. Other two stages are feature extraction stage and recognition stage as in conventional ASR systems. The accuracy of the entire system depends mainly on the performance of the segregation stage and feature extraction stage. By employing wavelet packet filter bank approximately matching to human cochlea, recognition accuracy is obtained for noisy speech. Then Multilayer Perceptron (MLP) neural network is for the train and test procedures.

Keywords: Multilayer Perceptron (MLP) Neural Network, Discrete Wavelet Transform (DWT), Mel's Scale Frequency Filter

1. INTRODUCTION

Speech is one of the most important tools for communication between human and his environment. Therefore manufacturing of Automatic System Recognition (ASR) is desire for him all the time. Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks in to a microphone. Automatic speech recognition technology has made enormous advances in the last 20 years and produced sufficiently good performance to be usefully employable in a variety of tasks but does not exhibit the robustness to environmental noise. Real world applications require that speech recognition systems be robust to interfering noise. Unfortunately, the performance of a speech recognition system drops dramatically when there is a mismatch between training and testing conditions.

In a speech recognition system, many parameters affect the accuracy of the Recognition System. Problems such as noisy environment, incompatibility between train and test conditions, dissimilar expressing of one word by two different speakers and different pronouncing of one word by one person in several times, is led to made system without complete recognition; So resolving each of these problems is a good step toward this aim. A speech recognition algorithm is consisted of several stages that the most significant of them are feature extraction and pattern recognition. In feature extraction category, best presented algorithms are Mel Frequency Cepstrum Coefficients (MFCC) and an auditory filter bank employing wavelet packets.

The performance gap between Automatic Speech Recognition and Human Speech Recognition (HSR) still remains large in the presence of noise. The performance of ASR system is almost similar with HSR performance in the absence of noise. The main reason for this performance gap is most of the developed ASR systems are performing speech segregation independently with recognition. A typical Speech Recognition system consists of a front end and back end. The front end gives the feature extraction of given input acoustic signal and back end is used for the recognition of speech by taking the feature vectors as input. Input signal is noisy speech, speech segregation unit separates the dominant speech from the noisy speech. From this speech, significant features are extracted for the recognition purpose. In the present work, the following methods are used for feature extraction, using Mel Frequency Cepstrum Coefficients (MFCC) and using an auditory filter bank employing wavelet packets. The performance of these two systems is compared. The purpose of speech feature extraction is to convert the speech waveform to some type of parametric representation for further analysis and processing.

Segregation of speech signals, when a representation of the sources exists such that the sources have disjoint support in that representation, it is possible to partition the support of the mixtures and obtain the original sources. One solution to the problem of demixing is thus to determine an appropriate disjoint representation of the sources and determine the partitions in this representation which demix. In this paper, we used the discrete short-time or windowed Fourier transform which is a good representation for demixing speech mixtures. Determining the partition blindly from one mixture is an open problem, but, given a second mixture, a method is described in, for partitioning the time-frequency lattice which separates the sources.

Our speech recognition process contains four main stages.

a. Acoustic processing that main task of this unit is filtering of the white noise from speech signals and consists of three parts, Fast Fourier Transform, Mel's Scale Bank pass Filtering and Cepstral Analysis.

b. W-Disjoint Orthogonality.

c. Feature extractions from MFCC and wavelet transform coefficients.

d. Classification and recognition using back propagation learning algorithm.

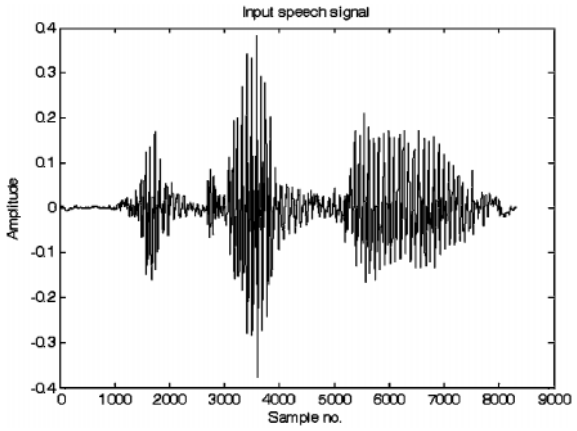

Fig. 1: Input Speech Signal

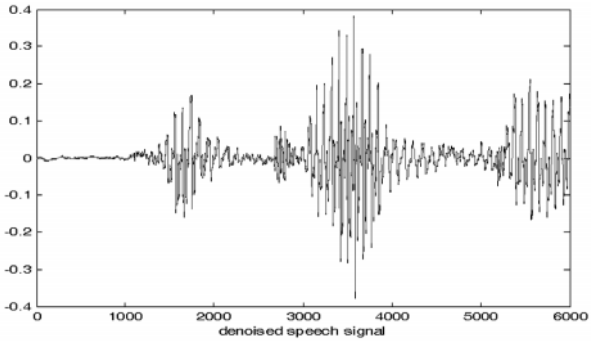

Fig. 2: Denoised Speech Signal

The digitized sound signal contains relevant, the data, and irrelevant information, such as white noise; therefore it requires a lot of storage space. Most frequency component of speech signal is below 5KHz and upper ranges almost include white noise that directly impact on system performance and training speed, because of its chromatic nature. So speech data must be pre-processed.

Fast Fourier Transform (FFT)

Fast Fourier Transform is that which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT) which is defined on the set of N samples $\{x_n\}$, as follow:

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N}, \; n = 0,1,2\ldots, N-1.$$

Mel Frequency Cepstral Coefficients System

The speech signal is a slowly timed varying signal called as quasi-stationary, when examined over a sufficiently short period of time between 5 and 100 msec, its characteristics are fairly stationary. However, over long periods of time where on the order of 1/5 seconds or more the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal.

A wide range of possibilities exist for parametrically representing the speech signal for the speech recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC is perhaps the best known and most popular, and is used in one of the systems implemented in this work. To simplify the subsequent processing of the signal, useful features must be extracted and the data should be compressed. The power spectrum of the speech signal is the most often used method of encoding.

Mel Frequency Cepstral Analysis is used to encode the speech signal. Mel scale frequencies are distributed linearly in the low range but logarithmically in the high range, which corresponds to the physiological characteristics of the human ear. Cepstral Analysis calculates the inverse Fourier transform of the logarithm of the power spectrum of the speech signal.

So first, speech signal was transferred to frequency domain by Fast Fourier Transform (FFT). Then the set of Mel scale filter banks is shown below was implemented on it and energy values of upper frequencies are decreased. Sub arrays are combined with each other and Inverse Fast Fourier Transform (IFFT) is performed.

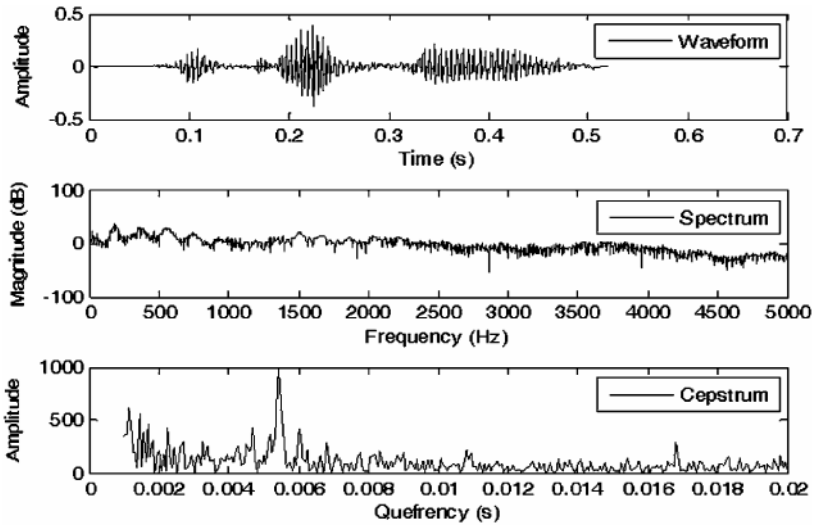

Fig. 3: Spectrum & Cepstrum for the Speech Signal

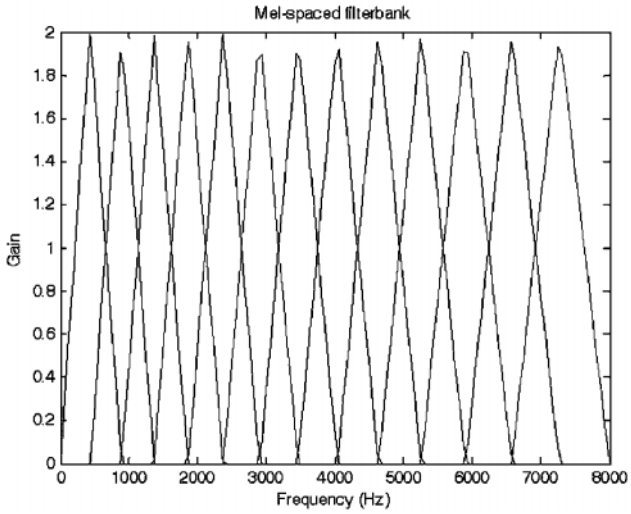

Fig. 4: Mel Frequency Filter Bank

The mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels.

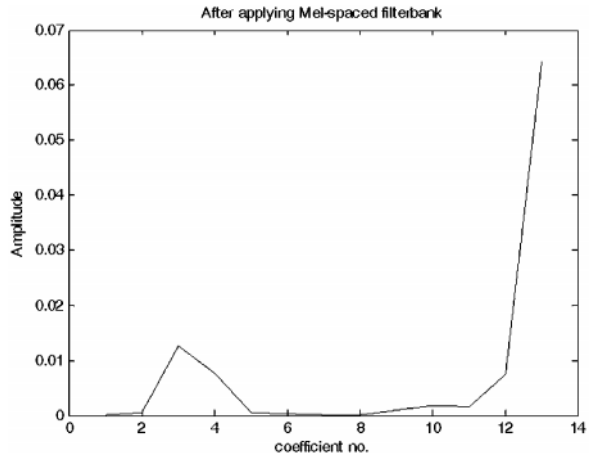

Fig. 5: After Applying Mel-spaced Filter Bank

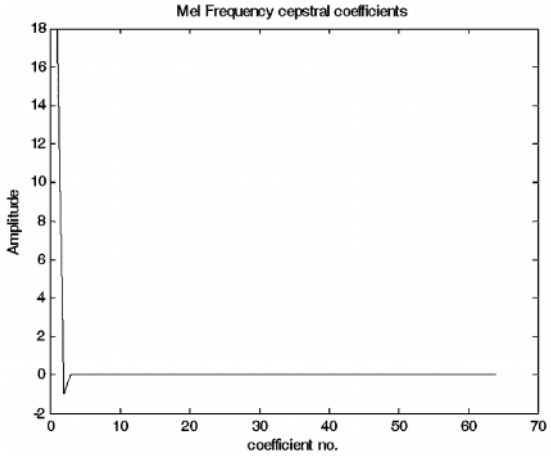

Fig. 6: Mel Frequency Cepstral Coefficients

W-Disjoint Orthogonality (W-DO)

Given a mixture $x_1(t) = \sum_{j=1}^{N} s_j(t)$ of sources $s_j(t)$, j = 1,……N, to recover the original sources. In order to accomplish this, we assume the sources are pair wise W-disjoint orthogonal. We call two functions s_1 and s_2 W-disjoint orthogonal (W-DO) if, for a given a window function **W**, the supports of the windowed Fourier transforms of s_1 and s_2 are disjoint.

The windowed Fourier transform of s_j is defined

$$F^W\left(s_j(\cdot)\right)(t,w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(\tau - t) s_j(\tau) e^{-iw\tau} d\tau$$

which we will refer to as $\hat{s}_j(t,\omega)$ where appropriate. Speech is sparse in that a small percentage of the time-frequency coefficients in the STFT expansion of speech capture a large percentage of the overall power. In other words, the magnitude of the time-frequency representation of speech is often small. Measure of

approximate W-disjoint orthogonality based on the demixing performance of time-frequency masks created using knowledge of the instantaneous source and interference time-frequency powers of speech mixtures. Experiments on speech mixtures reveal that speech is approximately W-DO. In order to measure W-disjoint orthogonality for a given mask, we combine two important performance criteria: how well the mask preserves the source of interest, and how well the mask suppresses the interfering sources.

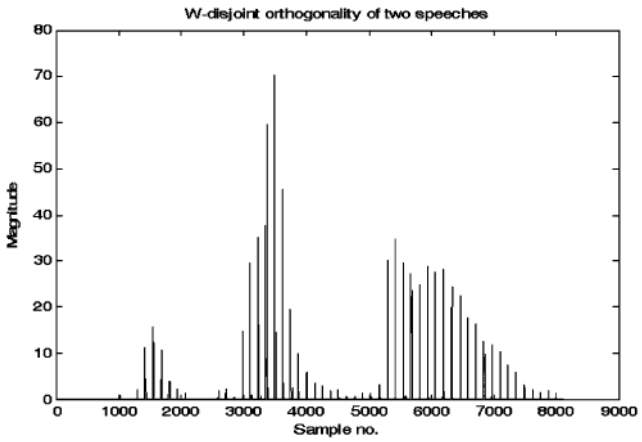

Fig. 7: W-disjoint Orthogonality of Two Speeches

Discrete Wavelet Transform

Wavelet analysis is a new development in the area of applied mathematics. Fourier analysis is ideal for studying stationary data, but is not well suited for studying data with transient events that cannot be statistically predicted from the data's past. Wavelets were designed with such nonstationary data in mind, and with their generality and strong results have quickly become useful to a number of disciplines.

Wavelet transform can be viewed as the projection of a signal into a set of basis functions named wavelets. Such basis functions offer localization in the frequency domain. Compare to STFT which has equally spaced time-frequency localization, wavelet transform provides high frequency resolution at low frequencies and high time resolution at high frequencies.

The discrete wavelet transform (DWT) of a signal $x[n]$ is defined based on so-called approximation coefficients, $w_{\varphi}[j_0,k]$, and detail coefficients, $w_{\psi}[j,k]$, as follows:

$$w_{\varphi}[j_0,k] = \frac{1}{\sqrt{M}}\sum_{n} x[n]\varphi_{j_0,k}[n] \qquad (1)$$

$$w_{\psi}[j,k] = \frac{1}{\sqrt{M}}\sum_{n} x[n]\psi_{j_0,k}[n] \text{ for } j \geq j_o$$

where $n = 0,1,2,\ldots, M-1$, $j = 0,1,2,\ldots, J-1$, $k = 0,2,\ldots, 2j-1$, and M denotes the number of samples to be transformed. The basis functions $\varphi_{j,k}[n]$, and $\psi_{j,k}[n]$ are defined as:

$$\varphi_{j,k}[n] = 2^{\frac{j}{2}}\varphi[2^j n - k] \qquad (2)$$

$$\psi_{j,k}[n] = 2^{\frac{j}{2}}\psi[2^j n - k]$$

$\varphi[n]$ is called scaling function and $\psi[n]$ wavelet function. For the implementation of DWT, the filter bank structure is often used. The approximation coefficients at a higher level are passed through a high pass and a low pass filter followed by a down sampling by two, to compute both the detail and approximation coefficients at a lower level. This tree structure is repeated for a multi-level decomposition. A wavelet packet filter bank matching to the critical bands of human cochlea is implemented for feature extraction. Human cochlea model consists of 24 critical bands with different bandwidths.

MLP Neural Network

A Multilayer Perceptron (MLP) network consists of an input layer, one or more hidden layers, and an output layer. Each layer consists of multiple neurons. An artificial neuron is the smallest unit that constitutes the artificial neural network. The actual computation and processing of the neural network happens inside the neuron. In this work, we use an architecture of the MLP networks which is the feed forward network with back propagation training algorithm (FFBP). In this type of network, the input is presented to the network and moves through the weights and nonlinear activation functions toward the output layer, and the error is corrected in a backward direction using the well-known error back propagation correction algorithm.

The number of neurons in each hidden layer has a direct impact on the performance of the network during training as well as during operation. Having more neurons than needed for a problem runs the network into an over fitting problem. Over fitting problem is a situation whereby the network memorizes the training examples. Networks that run into over fitting problem perform well on training examples and poorly on unseen examples. Also having less number of neurons than needed for a problem causes the network to run into under fitting problem. The under fitting problem happens when the network architecture does not cope with the complexity of the problem in hand. The under fitting problem results in an inadequate modeling and therefore poor performance of the network.

Unfortunately, coming up with the right number of hidden layers and neurons for each hidden layer can only be achieved by trial and error. Many experiments have been conducted to get the optimum number of hidden layers and neurons.

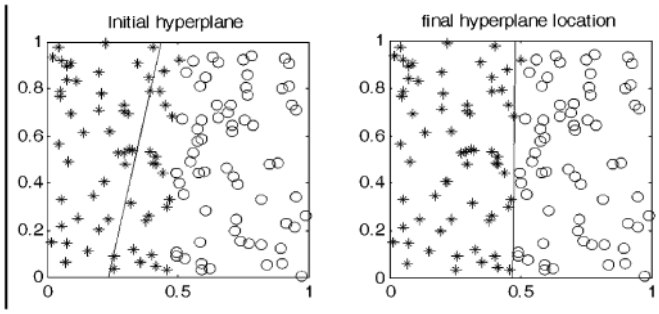

Fig. 8: Perceptron (Train and Test)

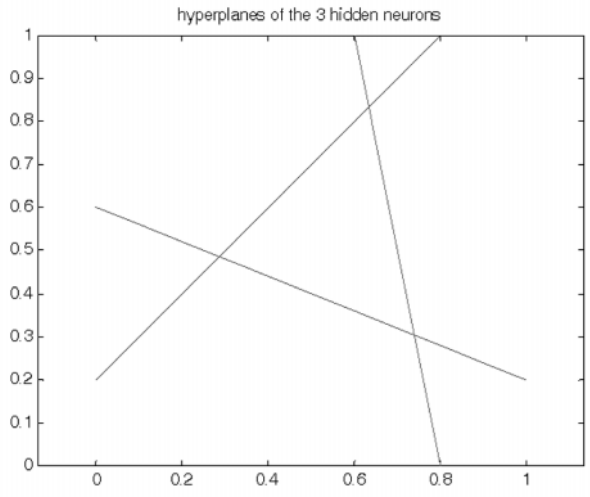

Fig. 9: Hyper Plane of Hidden Neurons

2. RESULT ANALYSIS

MLP neural network was successfully developed using MATLAB and has been selected for our network implementation. In MLP which is the feed forward network, the training and testing can be done with the thousand numbers of samples. Both train and test with 1000 samples.

3. CONCLUSION

In this paper, two robust speech recognition systems in noisy environment have been implemented. First system uses MFCC and the second system uses output of a wavelet packet filter bank closely matching to human cochlea, as features for the recognition stage. Performance of the wavelet packet filter bank based system is found to be better than that of MFCC based system. Performance of the speech segregation stage is also found to be better than that of related schemes. The efficiency of segregation stage depends on the degree of approximation of W-disjoint orthogonality of speech signals.

REFERENCES

[1] Abdul Ahad, Ahsan Fayyaz, Tariq Mehmood. "Speech Recognition using Multilayer Perceptron", *IEEE trans.* pp.103,2002.

[2] Song Yang, Meng Joo Er, and Yang Gao, "A High Performance Neural-Networks-Based Speech Recognition System", *IEEE trans*, pp.1527,2001.

[3] Ben Milner,Xu Shao: "Clean Speech Reconstruction from MFCC Vectors and Fundamental Frequency Using an Integrated Front End", *School of Computing Sciences*, University of East Anglia, Norwich NR4 7TJ, UK, *Speech Communication* 48(2006) pp.697-715

[4] I.Gavat, O.Dumitru, C. Iancu, Gostache, "Learning Strategies in Speech Recognition", Proc. Elmar 2005, pp.237-240, june 2005, Zadar, Croatia.

[5] Bahlmann. Haasdonk. Burkhardt, "Speech and Audio Recognition", *IEEE Trans*, **11**, May 2003.

[6] Tebelskis. J. "Speech Recognition Using Neural Networks", PhD. Dissertation, School of Computer Science, Carnegie Mellon University,1995.

[7] J. Tchorz, B. Kollmeier, "A Psychoacoustical Model of the Auditory Periphery as Front-end for ASR"; *ASAEAAiDEGA Joint Meeting on Acoustics*, Berlin, March 1999.

[8] R.P. Lippmann, "An Introduction to Computing with Meural Nets." *IEEE ASSP Mag.*, **4**, Apr.1997.

[9] MathWorks. Neural Network Toolbox User's Guide, 2004.

[10] S.M Peeling, R.K Moore and R.J.Tomlinson, "The Multi Layer Perceptron as a Tool for Speech Pattern Processing Research." *in Proc. IoA Autumn Conf.Speech Hearing*, 1986.

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