Speaker Identification using MFCC and DTW Technique on the Enhanced Speech Signal in a Noisy Environment

S. Malini,
PG Student, Dept. of ECE,
Sri Venkateswara College of Engineering,
Sriperumbudur,

R. Kousalya
Assistant Professor, Dept. of ECE,
Sri Venkateswara College of Engineering,
Sriperumbudur,

***Abstract*— Speech processing is an emerging technology of signal processing. The trending research area under speech processing is Speaker Identification (SI), which identifies the original speaker with respect to the extracted features from the speech signal. SI is mainly used in forensic analysis; voice activated home control systems and database access services. SI is implemented by feature extraction and feature matching. The most efficient feature extraction technique at present is Mel Frequency Cepstral Coefficient (MFCC) but it fails to perform well under a noisy condition. In order to overcome such a drawback, we introduce a different technique which involves filtering the speech signal before extracting features from it. The speech signals were collected from the NOIZEUS database and the feature extraction technique used is Dynamic Time Warping (DTW). The entire project results were simulated and verified using MATLAB 2010a. The results of the proposed technique are found to be more efficient than the results produced by the existing technique which uses MFCC alone for feature extraction.**

Keywords: MFCC, Speaker Identification, DTW, Filter.

I. INTRODUCTION

Speech processing is the study of speech signals and its processing. The general aspects of speech processing include the acquisition, manipulation, storage, transfer and output of speech signals. Hence, in order to achieve easy manipulation of the speech signals, it is necessary to convert them into digital domain. And therefore, speech processing falls under the broad category called digital signal processing. Some of the research areas of speech processing include speech synthesis, recognition of the speech and Speaker Identification (SI). Among those, speaker identification is a difficult task and is still a hot research area of speech processing.

Speaker Identification process is generally implemented by means of two processes, one after the other. The first process (or) the front end module is called feature extraction, which involves extraction of the speech parameters from the clean speech signals. This phase is also called as the training phase. Following the training phase is the second process (or) the back end module, called as the feature matching which tries to match the extracted parameters of speech samples in both test and the train phase. Usually the feature matching technique makes use of a classifier to identify the level of match between the signals.

The structure of this research paper proceeds as follows. Section II reviews the literature survey part of the existing speech recognition technology. Section III presents the proposed project work, the associated blocks and its functionality. Section IV presents the simulated results and the tabulation of identification rates. Section V concludes the scope of the research work and section VI discusses about the references.

II. LITERATURE SURVEY

G.V.P. Chandra Sekhar Yadav, Et. all, [1] proposed a paper which considers some portion of the existing work and also performs some additional noise cancellation on it. They have considered step size as the main factor for the convergence speed and mean square error. Wiener filter showed better performance but it involves large number of computations, so adaptive filter is considered as the alternate approach for removal of noise with moderate complexity and cost. The simulation result showed that wiener filter gives the better performance but at the cost of high expense, hence adaptive filter is the choice in many applications.

S. Nithyananthan and R. Shantha Selva Kumari [2] worked a paper that combines the two best processes to yield RASTA-MFCC feature which is robust to noise and also contributes speaker dependent information to identify the speaker efficiently.

I. Nancy Catherine and S. Dhanapani [3] proposed a model for Active Noise Control (ANC) by using Voice Activity Detection (VAD) and Weiner filtering. This method attenuates the effect of multiple noise source environments and produces clear signal at the output. The background noises are reduced by the iterative adaptive filtering method.

Palden Lama and Mounika Namburu [4] presented the various speech recognition technologies and more specifically, it gave explanation on the steps involved in designing an effective speech recognition system. The paper focuses on the pre-processing technique used to extract features from the speech signal and a Dynamic Time Warping (DTW) technique used to efficiently

compare the feature vectors of speech signals. The system was developed and tested using MATLAB software.

III. PROPOSED SYSTEM

On analyzing the existing methods, we have seen that MFCC technique performs exceptional speaker identification in a noiseless environment. But, the same fails to perform well in the case of a noisy environment. Hence, a method is proposed in order to improve the speaker identification rate when compared to the original technique. The proposed technique differs from the actual MFCC technique by means of introducing a filtering block. By doing so, the proposed technique can be used for speaker identification in the real world noisy environment as well as in a clean environment.

The proposed Speaker Identification (SI) system is organized into two modules for better understanding. The first module which is called as the Feature Extraction, involves extraction of small amount of valuable information from the available audio wave signal. This module explains about some of the preprocessing techniques implemented over the noisy speech signal along with the filtering technique. It also deals with one of the most commonly used feature extraction technique called Mel Frequency Cepstral Coefficient (MFCC) [11] technique. The second module is called as the Feature Matching module. In this module, the noisy speech is classified by means of a classifier. The proposed work has included Dynamic Time Warping (DTW) as its feature matching technique.

A brief explanation on the preprocessing, feature extraction and feature matching technique are explained below. The modules present in the SI system are shown in Figure 1.

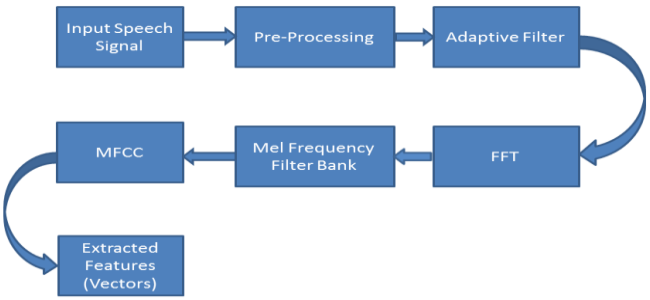

Fig. 1 SI System Model

The input speech signal which was collected from the database is initially read using the MATLAB command in order to use for further processing. The plot containing one of the input speech signals is shown below in Figure 2.

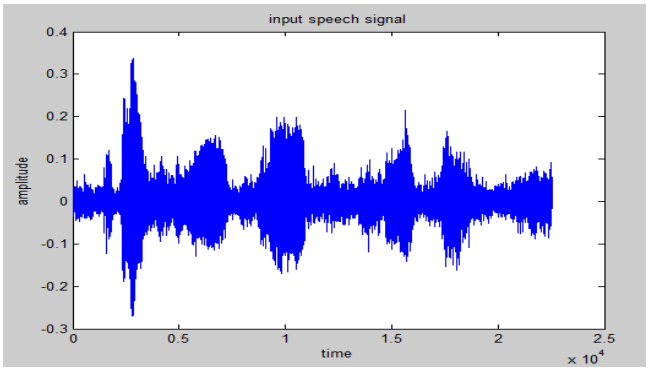

Fig.2 Plot of Input Speech Signal

An audio wave signal is not stationary by nature and hence applying processing steps on it tends to be tedious. Therefore, the signal is converted into multiple overlapping frames of shorter duration to perform various processing steps on it I such a way that there occurs no loss of information. Usually, the frames of 25ms duration are considered and the sampling rate is 16000 Hz. These frames are further passed through a Hamming window to cancel out the correlated samples present in each overlapping frame.

$$w(n) = \alpha - \beta \cos\left(\frac{2*pi*n}{N-1}\right) \text{ with } \alpha=0.54 \ \& \ \beta=0.46 \quad (1)$$

Equation (1) explains the construction of Hamming window function. Framing and windowing steps complete the pre-processing procedure. Figure 3 shows the Hamming window along with one of the windowed frame. These windowed frames are then passed through the filter in order to filter out the excess noise.

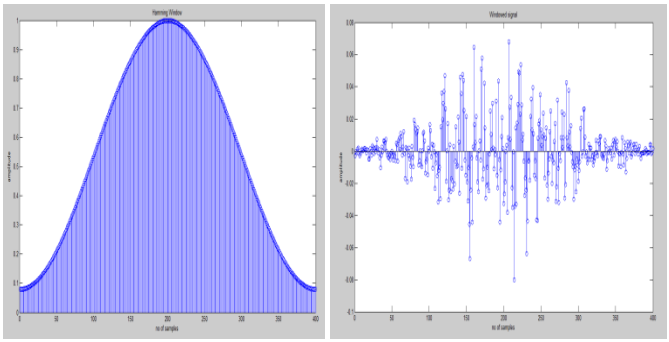

Fig.3 Hamming Window and One Windowed Frame

The windowed frame is passed on through the adaptive LMS filter block to remove the noise present along with the valuable signal. The adaptive filter offers good convergence rate and hence produces better results. Following the filtering operation, the signals are converted to frequency domain by applying Fast Fourier Transform (FFT) on it. FFT transforms the signal from time to frequency domain with lesser number of computations and at a faster rate. The general transformation function is shown in equation (2).

$$S_i(k) = \sum_{n-1}^{N} s(n) * w(n) * e^{\frac{-j*2*\pi*k*n}{N}} \quad (2)$$

Once the signals are converted to frequency domain, they are to be passed through the Mel Frequency Filter Bank to extract features from them. The Mel Frequency Filter Bank for feature extraction from a speech signal was designed with 20 filters from 0Hz to 4500Hz, which is the audible range of human beings. This is used to extract the power spectrum at each filter bank output, which helps in identifying the features present in the speech signal. The structure of the Mel Frequency Filter Bank is shown in Figure 4 and its equation is given in (3).

$$Mel(f) = 2595 * \log\left(1 + \frac{f}{700}\right) \quad (3)$$

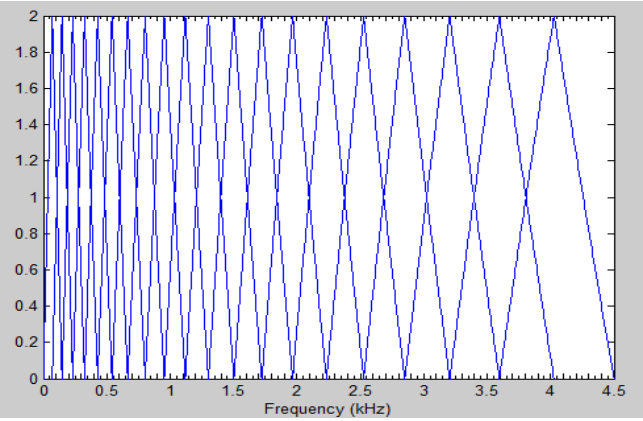

Fig. 4 Mel Frequency Filter Bank Structure

The Mel Frequency Cepstral Coefficients [8]-[9] were obtained for the input speech signal by passing it through the Mel space Filter Bank with 20 band pass filters. In order to easily understand the concept of MFCC, the plot was depicted using a spectrogram approach, where the horizontal axis denotes the power spectral estimate in time and the vertical axis denotes the Mel space Filter Bank number. The MFCC for the input speech signal, which was depicted in the form of a spectrogram, is shown in Figure 5. In the Feature Matching module, the noisy speech is classified by means of a classifier. The proposed technique includes Dynamic Time Warping (DTW) [4]-[7], [10] as its classifier.

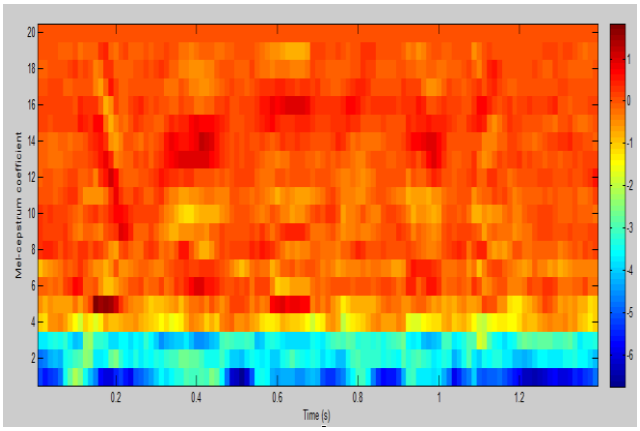

Fig. 5 MFCC Spectrogram

IV. RESULTS AND DISCUSSIONS

Feature Matching is performed using DTW technique over the various speech samples. The dataset was collected from the NOIZEUS database. The proposed project was carried out considering 6 speakers, with each speaker speaking 2 different sentences in 3 different types of noisy environment. The 3 noisy environments considered were airport, bubble and car. DTW was carried out by comparing each noisy speech signal with the speech signal produced by different users in different noisy environment. The identification rates were tabulated for different levels of noise measured by SNR values such as 5dB, 10dB and 15dB. The tabulation showing the various identification rates are shown in Table 1.

TABLE 1
ACHIEVED IDENTIFICATION RATES IN DIFFERENT NOISY ENVIRONMENTS

SNR (Feature Extraction)	Airport (%)	Bubble (%)	Car (%)	Average (%)
5dB	77.08	77.08	79.16	77.77
10dB	91.66	100	91.66	95.83
15dB	100	100	100	100

V. CONCLUSION

Due to the environmental noise scenario, the identification rates of the traditional MFCC technique were very low and hence a new technique was proposed to offer good identification rates. The proposed SI system was designed using an adaptive LMS filter to effectively reduce the noise level present in the audio wave signal. By filtering out the noise from the signal, better identification rates of around 91.2% were achieved by using the proposed MFCC technique. The project was simulated using MATLAB software and DTW technique was used for Feature Matching.

VI. REFERENCES

[1] G.V.P.Chandra Sekhar Yadav, Et. all "Performance of Wiener Filter and Adaptive Filter for Noise Cancellation in Real-Time Environment" in *Int. Journal on Computer Applications*, Vol. 97–No.15, July 2014.

[2] S.Selva Nidhyananthan and R.Shantha Selva Kumari, "Text Independent Voice Based Students Attendance System under Noisy Environment using RASTA-MFCC Feature" in *IEEE International Conf. on Communication and Network Tech.*, 2014.

[3] I. Nancy Catherine and S. Dhandapani, "Noise Reduction In Speech Processing Using Improved Active Noise Control (ANC) Technique" in *IJRET* Vol. 3 Issue: 03, Mar-2014.

[4] Palden Lama, Mounika Namburu, "Speech Recognition with Dynamic Time Warping using MATLAB" in *SPRING* 2010.

[5] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques" in *Journal Of Computing*, Vol. 2, Issue 3, March 2010.

[6] Sinith, M.S, Gowri Sankar, K, "A novel method for Text-Independent speaker identification using MFCC and GMM" in *International Conference on Audio Language and Image* Processing, 2010.

[7] Zeinali, H., Sameti, H. "A fast Speaker Identification method using nearest neighbor distance" in *IEEE 11th International Conference on Signal Processing (ICSP)*, 2012.

[8] M.G.Sumithra Et. all, "A Study on Feature Extraction Techniques for Text Independent Speaker Identification" in *IEEE* 2012.

[9] Sarangi, S.K., Saha, G., "A novel approach in feature level for robust text-independent speaker identification system" in *4th International Conference on Intelligent Human Computer Interaction (IHCI)*, 2012.

[10] Dalmiya C.P, Dr. Dharun V.S, Rajesh K.P, "An Efficient Method for Tamil Speech Recognition using MFCC and DTW for Mobile Applications in *IEEE Conference on Information and Communication Technologies*, 2013.

[11] Martinez, J. Et. all "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques" in *Electrical Communications and Computers (CONIELECOMP), IEEE*, 2012.