International Journal of Future Innovative Science and Engineering Research (IJFISER)
Volume - 2, Issue - II ISSN (Online): 2454- 1966

Research Manuscript Title

Automatic Speech Recognition (ASR) in Noisy Environment using Adaptive Filter for Speech Enhancement

S. Malini
PG student, Dept. of ECE,
Sri Venkateswara College of Engineering,
Sriperumbudur,
smalini28@live.com

R. Kousalya
Assistant Professor, Dept. of ECE,
Sri Venkateswara College of Engineering,
Sriperumbudur,
kousi@svce.ac.in

JUNE – 2016

www.istpublications.com

International Journal of Future Innovative Science and Engineering Research (IJFISER), Volume - 2, Issue – II,
ISSN (Online): 2454- 1966 www.istpublications.com.

Automatic Speech Recognition (ASR) in Noisy Environment using Adaptive Filter for Speech Enhancement

S. Malini
PG student, Dept. of ECE,
Sri Venkateswara College of Engineering,
Sriperumbudur,
smalini28@live.com

R. Kousalya
Assistant Professor, Dept. of ECE,
Sri Venkateswara College of Engineering,
Sriperumbudur,
kousi@svce.ac.in

ABSTRACT

Speaker Identification (SI) is the current research area under Speech Processing domain. Existing techniques made use of Mel Frequency Cepstral Coefficients (MFCC) as the feature extraction technique in order to extract the useful parameters from the audio wave signal. MFCC is an effective feature extraction technique used for the purpose of Speaker Identification. But, the efficiency of the existing MFCC technique fails to extract features effectively when it is used in a noise prone environment. Hence, a new technique has been proposed in order to achieve a better Identification Rate when compared to the existing module. The proposed technique makes use of adaptive filter to enhance the quality of the input audio wave signal produced in a noisy environment followed by MFCC technique to extract the features from it. Vector Quantization is used for feature matching between the test and the train signals. The proposed work has attained a better Identification Rate when compared to the existing MFCC technique.

Keywords*: Speaker Identification, Adaptive Filter, Mel Frequency Cepstral Coefficient (MFCC), Vector Quantization***.**

I. INTRODUCTION

Speech is the general means of communication among humans. Since the rapid growth in technology, speech has also developed as a means of communication with machines. This rapid growth has contributed to the development of a new domain called Speech Processing, which deals with the study about the behavior of speech signals along with the various methods to process them effectively. General aspects of Speech Processing cover various functions like acquisition of the speech signal, manipulating them, applying different speech enhancement techniques over them and finally storing and transferring the processed speech signals. Speech Processing encloses vast research areas like speech synthesis, recognition of the speech, speech to text conversion and speaker identification. Among these, Speaker Identification (SI) is a developing technology in the current scenario, but its efficiency is often degraded by the presence of the environmental noise. This has paved a way to perform efficient speaker identification in noisy environmental scenario by using various speech enhancement techniques, generally filtering.

Speaker Identification in general can be classified into two broad categories, which is Text Dependent speaker identification and Text Independent speaker identification. In the first category, speaker identification is carried out for a predefined set of sentences or words. Also, the word or the sentences used in the test and the train module should be one and the same. The other category doesn't impose any such constraints over the utterances of word or sentences. From which it is clear that the sentences or words uttered in the test and the train module may or may not be the same.

An Automatic Speech Recognition (ASR) system can be implemented by allowing the speech signal to get processed in two different modules. The Front-End module consists of a feature extraction technique which tries to extract the speech features from the input speech signal and develop a reference database containing the codeword entries of N different speakers. The Front-End module is usually called as the training phase and is depicted in Fig. 1.

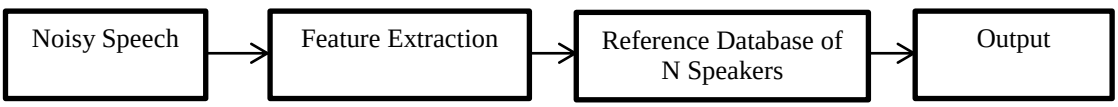

Fig. 1 Front-End Module

The Back-End module tries to match the input speech signal produced in a noisy environment with any one of the appropriate speech signals that are already stored in the database. The matching among the speech signal in clean and noisy environment can be implemented by using various types of classifiers. The Back-End module of ASR system is usually called as the testing phase and is shown in Fig. 2.

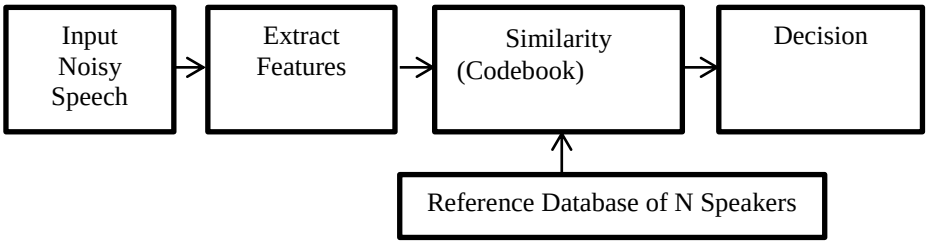

Fig. 2 Back-End Module

Speaker identification is dependent on the fact that each user tends to exhibit unique speech characteristics. But, there may occur a mismatch between the speech samples in train and test phase due to change in voice characteristics, speaking rates and noise background. Hence, the main aim of the project is to develop a module to enhance the quality of the speech signal before extracting features from it. The contents of this paper are organized in the following manner: Section II covers the literature survey part, Section III portrays a detailed explanation about the proposed speaker identification system, Section IV discusses about the achieved results and finally section V concludes the efficiency of the proposed system design.

II. LITERATURE SURVEY

Jesper Jensen, Zheng-Hua Tan [1] proposed a means of enhancing the quality of the speech signal by estimating the minimum mean square error value (MMSE). They have considered well-established and consistent assumptions for the purpose of estimating the minimum MMSE of the MFCC features. And ultimately the performance of the proposed model was closer enough or better than the existing MFCC technique, though it efficiently enhanced the quality of the speech signal.

S.Selva Nidhyananthan, R.Shantha Selva Kumari [2] has suggested a new technique called RelAtive SpecTrA-MFCC (RASTA) processing of speech in order to improve the performance of the identification system in the presence of additive and convolution noise by performing band pass filtering over each frequency channel. This paper combines two efficient techniques to contribute RASTA-MFCC which performs well in a noisy environment and also identifies the speaker effectively. The outcomes from the proposed project were found to achieve an accuracy of 93.2%.

Dalmiya C.P, Dr. Dharun V.S, Rajesh K.P [3] has proposed a different speaker identification system which made use of MFCC as the feature extraction technique followed by Dynamic Time Warping (DTW) as the feature matching technique. Also they have compared the performance of the DTW algorithm with the other existing algorithms like ANN, HMM, etc. Finally, they have concluded that the performance of the DTW feature matching technique was better when it was used along with MFCC feature extraction technique.

Paresh M. Chauhan and Nikita P. Desai [4] have undergone a research by using wiener filter for the purpose of filtering out the environmental noise form the input noisy speech signal, followed by MFCC to extract the speech features from the filtered speech signal. They have considered input speech signals with varying levels of noise i.e. with different Signal to Noise Ratio (SNR) values for the purpose of implementing a speaker identification system. The proposed work has achieved better results by using Neural Networks as the feature matching technique.

III. PROPOSED METHODOLOGY

The prime motive of Automatic Speech Recognition (ASR) system is to perform speaker identification i.e. to acquire the speech signal from a noisy environment, extract the speech features from it and finally try to match it with that of the actual speaker. As inferred from the literature survey part, it was clear that the MFCC is an efficient feature extraction feature in a noiseless environment but, fails to achieve a similar performance when used in noisy environmental scenario. Hence, the proposed SI system makes use of an additional speech enhancement (adaptive filter) block along with the existing MFCC technique in order to achieve a better performance. The proposed SI system is depicted in Fig.3.

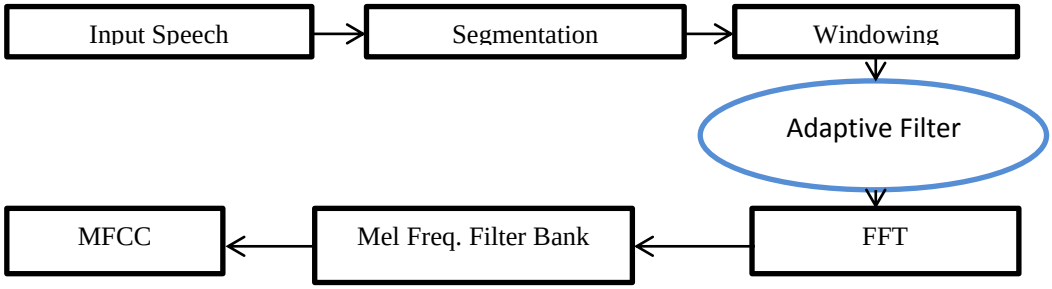

Fig.3 Proposed SI system for noisy environment

Generally when we consider an audio speech signal, it tends to change constantly i.e. it is quasi-stationary [9] - [10]. Extracting feature from an audio wave signal as such is not much efficient and hence certain amount of pre-processing has to be applied over it. So, the first step to perform speech processing is to segment the entire audio wave signal into several small frames of duration 20 to 40ms each in order to make them look stationary. An important constraint to note here is that the segmented frames should be overlapping because considering non-overlapping frames might lead to loss of information. In order to make calculations simpler, the number of frames are usually considered to be in powers of 2 (generally 256 frames are considered).The input speech signal and one its segmented frame can be observed in Fig. 4.

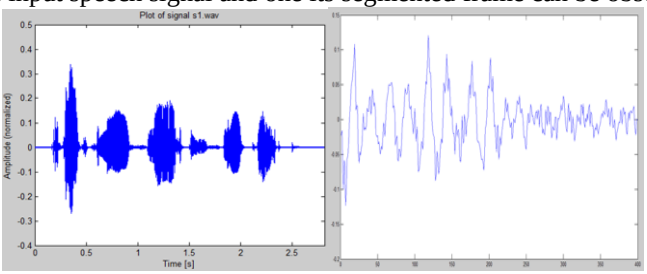

Fig.4 Input signal and one Segmented Frame

After segmenting the audio wave signal into overlapping frames, a window has to be multiplied with each frame in order to reduce the amount of spectral distortions at the start and end of each overlapping frame. The efficient window technique used for the purpose of speech processing is the hamming window.

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2*\pi*N}{N-1}\right); \quad 0<n<N-1 \qquad (1)$$

A hamming window, w (n) is developed in such a way that it reduces the discontinuity at the start and end of each frame to zero; thereby removing the overlapping components between the adjacent frames and at the same time restores the valuable speech information. The design of the hamming window function is denoted by equation (1) and its resultant shape is denoted in Fig. 5.

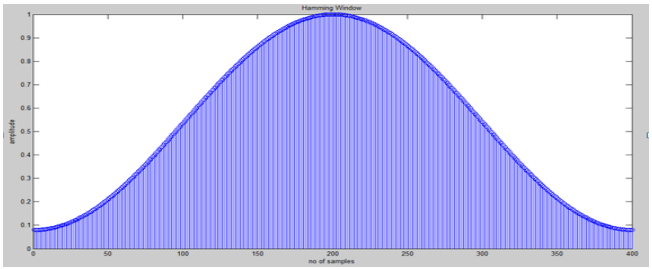

Fig.5 Hamming window function

Once the frames are windowed, the resultant frames are enhanced by means of an adaptive filter [12]. An adaptive LMS filter changes its filter parameters i.e. filter coefficients on an adaptive manner, so as to adapt to the varying signal characteristics. The filter tries to minimize the Mean Square Error (MSE) [5]-[7] present in the speech signal acquired from the noisy environment and hence produces enhanced speech as its output. The effect of adaptive filtering on the noisy speech signal is displayed in Fig.6.

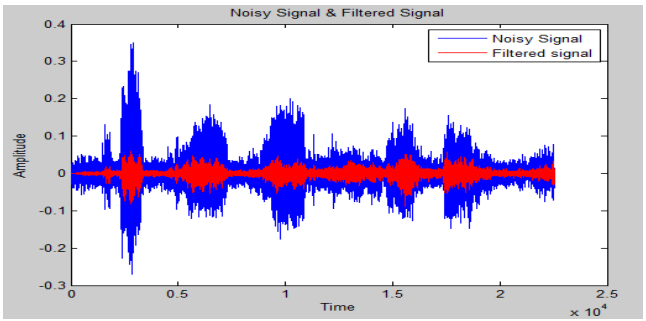

Fig.6 Effect of adaptive filtering on the noisy speech signal

The filtered speech signals in time domain are next passed through the Fast Fourier Transform (FFT) block for the purpose of converting the time domain frames into frequency domain. The FFT technique given by equation (2) is capable of

performing rapid transformations from one domain to another, i.e. from time domain to frequency domain or vice versa. Whenever a speech signal has to be processed, analyzing it time domain is highly tedious because the samples will be highly correlated with the adjacent ones. Hence, the signals are converted are converted to frequency domain, which clearly describes about the speaker.

$$S_i(k) = \sum_{n-1}^{N} s(n) * w(n) * e^{\frac{-j*2*\pi*k*n}{N}} \; ; k=0\ldots N-1 \qquad (2)$$

The nature of human ear is well defined to be linear till 1 KHz frequency range and as the frequency rises its behavior tends to be logarithmic. Hence, a transformation model has to be followed and its performance should match the tendency of human ear. A well-known and efficient model for the purpose of transforming the speech signal is called Mel-Frequency Transformation function [11] and the inverse Mel Frequency transformation function is given by equation (3) and (4) respectively. Using the above transformation function, a filter bank is generated which depicts the tendency of the human ear and tries to retrieve the pitch information of individual speaker. The filter bank is called as the Mel-scale filter bank and it comprises of 20 triangular band pass filter structures which are designed based on the Mel Frequency transformation function.

$$M(f) = 1125 * \ln\left(1 + \frac{f}{700}\right) \qquad (3)$$

$$f = 700 * e^{\left(\frac{M(f)}{700} - 1\right)} \qquad (4)$$

The structure of the Mel-scale filter bank in the human audible range (0 Hz to 4000 Hz) is depicted in Fig. 7. Each frame is made to pass through the filter bank one followed by the other and the corresponding energies at the output of each filter bank are summed up, thus leading to 20 different energy vectors for each frame. The power spectrum is shown in Fig.8.

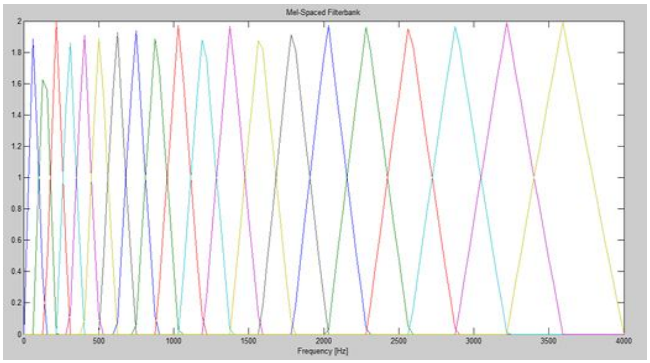

Fig.7 Mel-scale filter bank

These vectors are converted to logarithmic scale, because humans perceive loudness in a logarithmic scale. Finally, the logarithmic energies are passed through a Discrete Cosine Transform (DCT) block for the purpose of de-correlating the signal and to achieve maximum energy compaction. The final vectors obtained after performing DCT operation are referred to as the Mel-Frequency Cepstral Coefficients (MFCC) which terminates the feature extraction part.

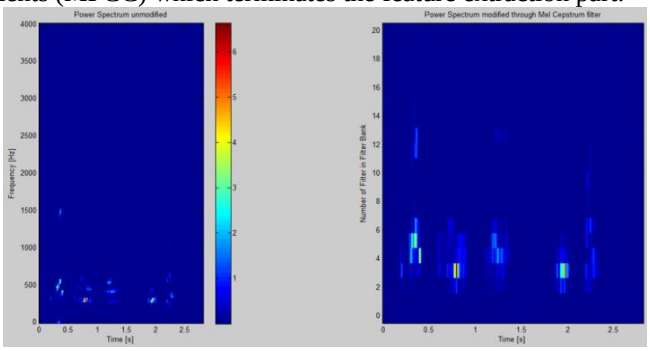

Fig.8 Power spectrum obtained from Mel-scale filter bank

Finally, the extracted features are used for the purpose of matching the input speech signals to its corresponding original speaker. The feature extraction technique used in the proposed SI system is called Vector Quantization technique. This technique is implemented using Linde-Buzo Gray algorithm [8] in order to store the codebook entries pertaining to each speaker in both the test and the train phase.

IV. RESULTS AND DISCUSSION

The proposed SI system involves comparing the codebook entry of the filtered version of the incoming noisy speech signals with the codebook entries of different trained speakers. The test and the train database consist of 8 different speakers uttering sentences in different types of noisy environment with varying noise levels as denoted by their Signal to Noise Ratio (SNR) values. It was inferred that as the SNR value of the input speech signal increases, good identification rates were achieved i.e. higher the value of SNR implies that the level of noise gets decreased. The tabulation for the Identification Rates achieved using the proposed SI system is shown in table 1.

TABLE 1- IDENTIFICATION RATES ACHIEVED USING THE PROPOSED SPEAKER IDENTIFICATION SYSTEM

SNR (dB)	Airport (%)	Babble (%)	Car (%)	Station (%)	Street (%)	Average (%)
5dB	62.5	75	62.5	62.5	50	62.5
10dB	75	87.5	62.5	87.5	75	77.5
15dB	100	100	87.5	100	87.5	95

V. CONCLUSION

Through this research work, we have a proposed a new speaker identification system which is a variant of all the existing MFCC techniques used in noisy environment. In experiments where the input speech signals are corrupted by various sources of noise, the proposed method has succeeded in attenuating the level of noise present in it. The proposed method uses Vector Quantization as the classifier in a noise mismatch condition and has achieved an identification rate of approximately 80% with 8 different speakers. Also, we have analyzed that the identification rates were improved as the SNR increases i.e. as the signal strength increases.

VI. REFERENCES

[1] Jesper Jensen, Zheng-Hua Tan, "Minimum Mean-Square Error Estimation of Mel-Frequency Cepstral Features- A Theoretically Consistent Approach" in *IEEE/ACM TRANSACTIONS on Audio, Speech and Lang. Processing*, Vol. 23, No. 1, Jan 2015.

[2] S.Selva Nidhyananthan, 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] 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 Proceedings of 2013 *IEEE Conference on Information and Communication Tech. (ICT 2013)*, pp. 1263-1268.

[4] Paresh M. Chauhan and Nikita P. Desai, "Mel Frequency Cepstral Coefficients (MFCC) based Speaker Identification in Noisy Environment using Weiner Filter", *IEEE* 2014.

[5] Martinez. J, Et. all, "Speaker Recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) techniques" in *Electrical Communications and Computers, IEEE* 2012

[6] Anuja N. Untwale and Kishori S. Degaonkar, "Survey on Noise Cancellation Techniques of Speech Signals by Adaptive Filtering", in 2015 *IEEE International Conference on Pervasive Computing*.

[7] Sivaranjan Goswami, Et. all, "A Novel Approach for Design of a Speech Enhancement System using NLMS Adaptive Filter and ZCR based Pattern Identification" in *ICETACS*, 2013.

[8] Arup Kumar Pal and Anup Sar, "An Efficient Codebook Initialization Approach for LGB Algorithm", in 2011 *International Journal of Computer Science, Engg. And Applications (IJCSEA),* Vol. 1, No. 4, pp. 72-80.

[9] Ian McLoughlin, "*Applied Speech and Audio Processing: With MATLAB_ Examples*", Cambridge University Press- BOOK, 2009.

[10] Patrick A. Naylor, "*Image, Video Processing and Analysis, Hardware, Audio, Acoustic and Speech Processing- Introduction to Audio Signal Processing*", Vol. 4- BOOK, 2014.

[11] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/- Web Link.

[12] G.V.P Chandra Sekhar Yadav, Et. all, "Performance of Weiner Filter and Adaptive Filter for Noise Cancellation in Real Time Environment" in *International Journal on Computer Applications*, 2014.