International Journal of Electronics Signals and Systems

Volume 3 | Issue 1 Article 11

July 2013

MFCC AND CMN BASED SPEAKER RECOGNITION IN NOISY ENVIRONMENT

DEBASHISH DEV MISHRA
Royal School of Engineering and Technology, Guwahati, Assam, India., debashish.dm@gmail.com

MFCC AND CMN BASED SPEAKER RECOGNITION IN NOISY ENVIRONMENT

DEBASHISH DEV MISHRA[1], UTPAL BHATTACHARJEE[2] & SHIKHAR KUMAR SARMA[3]

[1]Royal School of Engineering and Technology, Guwahati, Assam, India.
[2]Department of Computer Science, Rajiv Gandhi University, Itanagar, Arunachal Pradesh, India.
[3]Department of Computer and Information Technology, IST, Gauhati University, Guwahati, Assam, India
Email-debashish.dm@gmail.com

Abstract- The performance of automatic speaker recognition (ASR) system degrades drastically in the presence of noise and other distortions, especially when there is a noise level mismatch between the training and testing environments. This paper explores the problem of speaker recognition in noisy conditions, assuming that speech signals are corrupted by noise. A major problem of most speaker recognition systems is their unsatisfactory performance in noisy environments. In this experimental research, we have studied a combination of Mel Frequency Cepstral Coefficients (MFCC) for feature extraction and Cepstral Mean Normalization (CMN) techniques for speech enhancement. Our system uses a Gaussian Mixture Models (GMM) classifier and is implemented under MATLAB®7 programming environment. The process involves the use of speaker data for both training and testing. The data used for testing is matched up against a speaker model, which is trained with the training data using GMM modeling. Finally, experiments are carried out to test the new model for ASR given limited training data and with differing levels and types of realistic background noise. The results have demonstrated the robustness of the new system.

Keywords-*ASR, MFCC, CMN, GMN, MLLR*

I. INTRODUCTION

Automatic Speaker Recognition (ASR) systems are generally provided with the ability to take decision in a range of environment where noise maybe a common ingredient. To develop such capabilities, ASR s are trained using data which are acquired in noise free environments [1]. In real world operating conditions, several factors can degrade the quality of the speech signal and therefore reduce the performance of a ASR system [2]. Some of the factors are distortion due to the nature of the environment, stationary or non-stationary ambient noise such as fan or people talking respectively. Also the acoustics of the room can introduce reverberations and echo in the signal. The quality of microphones can cause linear and non-linear distortion. The distance from the speaker to the microphone can vary. This will result in a distortion of the amplitude of the signal.

The transmission channel such as (GSM) or (PSTN) networks can introduce perturbations to the signal. Additive noise such as electrical perturbations of the transmission lines, impostors such as mimicry by human, speakers bad pronunciation, emotion states such as anger, sickness, tiredness and aging all together are factors that distort the signal. Speaker recognition system must be robust to these real environment perturbations [3].

A major problem of most ASR systems is their unsatisfactory performance in noisy environments [4]. In this experimental research, we have studied a combination of Mel Frequency Cepstral Coefficients (MFCC) for feature extraction and Cepstral Mean Normalization (CMN) techniques for speech enhancement. Our system uses a Gaussian Mixture Models (GMM) classifier and is implemented under MATLAB®7 programming environment. The process involves the use of speaker data for both training and testing. The data used for testing is matched up against a speaker model, which is trained with the training data using GMM modeling. Finally, experiments are carried out to test the new model for ASR given limited training data and with differing levels and types of realistic background noise. The results have demonstrated the robustness of the new system. Some of the relevant literature are [5]-[9].

II. THEORETICAL BACKGROUND

The detailed block diagram of a general ASR system is shown in Figure 1 [4]. It mainly consists of digital speech data acquisition, feature extraction, pattern matching, making an accept/reject decision, and enrolment to generate speaker reference models [4]. A general feature extraction block diagram is shown in figure 2 [6].

Feature extraction is the estimation of variables (feature vector) from the observation of a speech signal which contains different information such as dialect, context, speaking style and speaker emotion.

The aim is to transform the speech signal into a collection of variables that can preserve the signal information and that can be used to make comparisons.

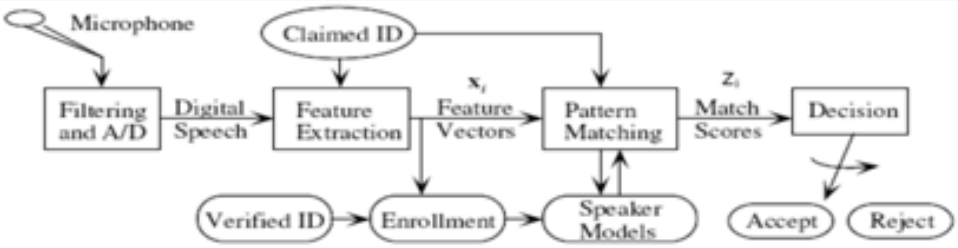

Figure 1: Generic ASR

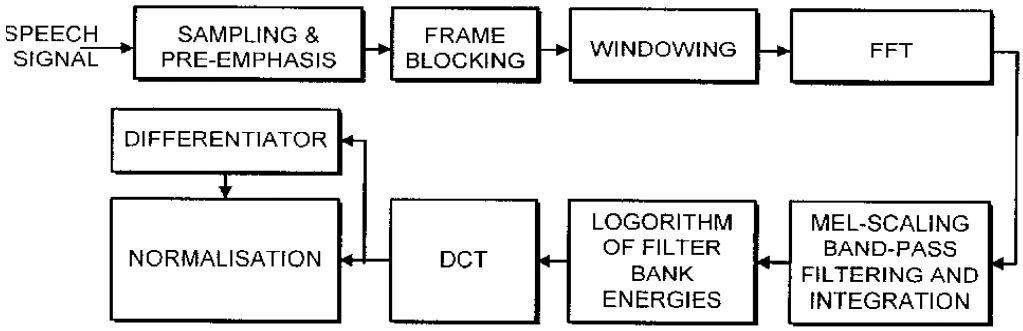

Figure 2: General feature extraction of a ASR

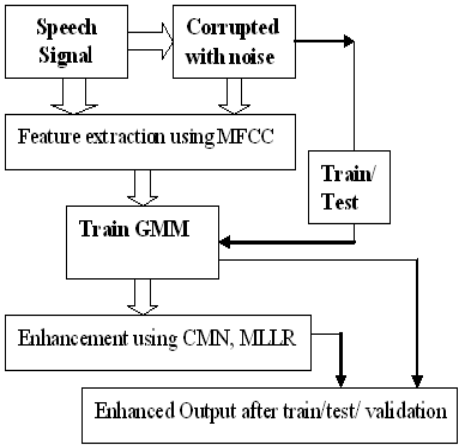

Figure 3: System Model

III. SYSTEM MODEL AND EXPERIMENTAL DETAILS

The system model is shown in Figure 3.
Initially we record certain number of speech samples out of which some are retained in clean form and a few are corrupted. Next, we extract features using MFCC. The feature set contains sample which are clean and nose corrupted. These are next modeled using GMM. There is a training phase during which - the GMM learns the clean and nose corrupted samples. Next, test and validation processes are performed during which the GMM demonstrates the decision making role as part of the ASR. The speech samples derived from the GMM are further enhanced by the CMN and MLLR approaches which contribute to the performance of the system. In order to evaluate the clean speech in real environment condition, the clean speech is deteriorated -

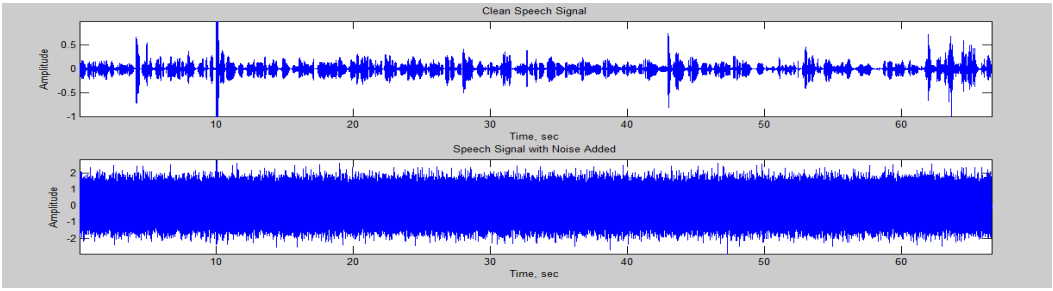

Figure 4: Clean speech signal and speech signal with white Gaussian noise added.

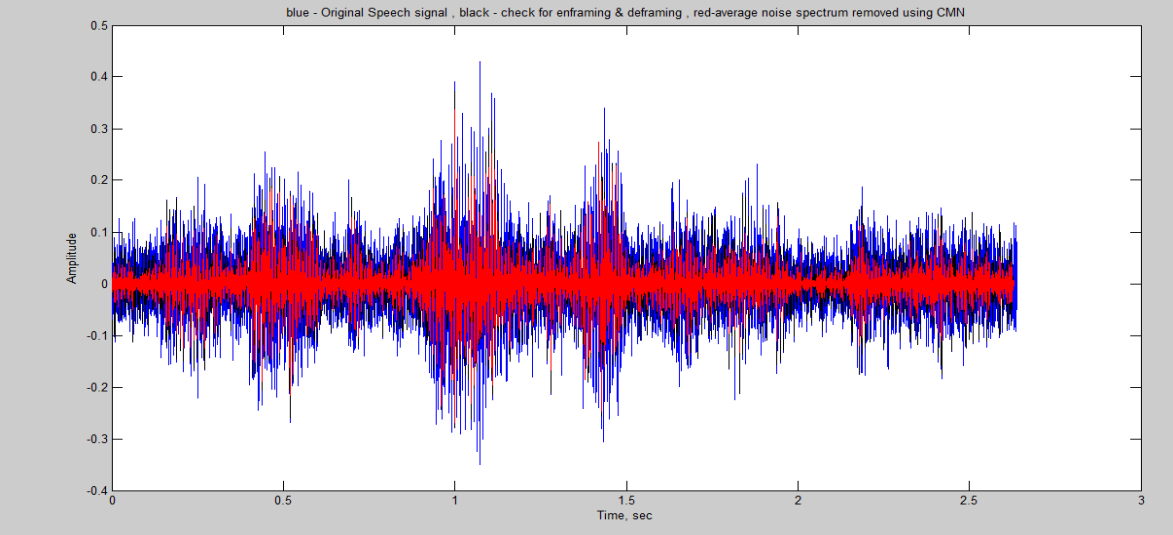

Figure 5: Clean speech signal in blue, check signal in black and enhanced signal in red.

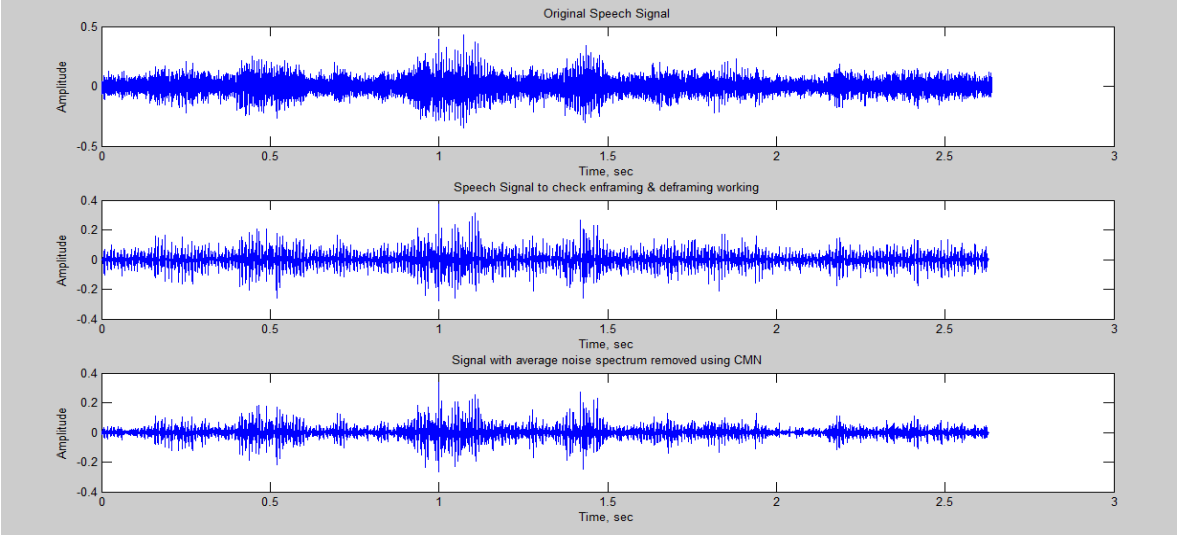

Figure 6: Clean speech signal, check signal and enhanced signal derived using CMN

Table I: Speaker Dependent Threshold Values

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
-4.257851	-4.2128	-4.31461	-4.40416	-4.0026	-3.48019	-3.47964	-4.3755	-4.01796	**-4.592466**
-4.9951	-5.0494	-5.0238	-5.0230	-5.3953	-5.3630	-5.4144	-5.1321	-4.9343	**-4.5289**

Table II: Speaker Dependent Threshold Values with Gaussian noise added

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
-3.1564	-3.1605	-3.13352	-3.14865	-3.1424	-3.13914	-3.13421	-3.1443	-3.14916	**-3.156431**
-3.1556	-3.1637	-3.1644	--3.1619	-3.1584	-3.1603	-3.1603	-3.1650	-3.1627	**-3.1564**

Table III- Speaker Dependent Threshold Values with Gaussian Noise Removed

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
-5.5456	-5.5129	-5.5036	-5.5614	-5.3898	-5.44945	-5.46972	-5.5814	-5.48780	**-5.570331**
-5.6090	-5.6185	-5.6322	-5.5934	-5.6287	-5.6055	-5.6557	-5.5950	-5.6490	**-5.5685**

by adding the white Gaussian noise to the clean speech. The assumptions made are that the noise is additive and not correlated with the speech signal. The speech enhancement techniques identified to overcome the noisy conditions are CMN and MLLR. The noisy signal due to the addition of the white Gaussian noise is shown in the second plot of the below Figure 4. The signal-to- noise ratio (SNR) is set at 5dB. After applying CMN to the noisy signal, the enhanced speech signal is shown below in Figure 5. The original speech signal is plotted in blue, the signal checked for enframing and deframing is shown in black and the average noised removed signal using CMN is shown in red.

A similar set of results are generated using a combination of CMN and MLLR.

IV. RESULTS AND DISCUSSION

Initially a speaker recognition system of 10 speakers has been created. The initial convergence likelihood received has been listed below. For clean speech, a speaker will be accepted if the distance between convergence likelihood and calculated likelihood is 0.2. The value of 0.2 is experimentally calculated by making False acceptance and False Rejection rate equal. Table 1 shows speaker dependent threshold values which necessary during the decision are making process of the ASR. The speakers are named S1 to S10 and the corresponding threshold values are shown.

A similar set of values are derived for noise corrupted samples which are shown in Table II while Table III shows another set of values derived from clean samples obtained from the system. From the above (Tables I-III) we see that the ASR properly recognizes the speaker S10 in all the conditions considered. This result is obtained after performing over twenty numbers of trials with varied input sequences. The outcome thus establishes the effective of the proposed ASR approach for speaker recognition under noisy environment.

V. CONCLUSION

This paper has examined the an ASR approach designed MFCC features and GMM aided by CMN and MLLR enhancement techniques. Experimental results show that the system is robust under a range of noise environments.

REFERENCES

[1]. S. Furui, “50 years of progress in Speech and Speaker Recognition Research”, ECTI Transactions on computer and information Technology, Vol. 1, no.2, Nov 2005.

[2]. D. R. Reddy, “An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave.” Tech . Report No. C549, Computer Science Dept., Stanford Univ., 1966

[3]. J.P. Campbell, “Speaker recognition: A tutorial,” Proceedings of the IEEE, vol 85, pp. 1437 – 1462, September 1997.

[4]. P. C. Loizou, “ Speech Enhancement Theory and Practise”, CRC Press, 2007.

[5]. T. Azetsu, E. Uchino and N. Suetake, “Blind separation and sound localization by using frequency-domain ICA,” *Soft Computing*, vol.11, no.2, pp.185–192 2007.

[6]. J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, Robust Speaker Recognition in Noisy Conditions, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 5, JULY 2007.

[7]. J. H. L. Hansen, “Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,” *Speech Commun.*, vol. 20, pp. 151–173, 1996.

[8]. D. Giuliani, M. Omologo, and P. Svaizer, “Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation,” in *Proc. ICSLP’96*, Trento, Italy, pp. 1329–1332, 1996.

[9]. J. Ming, P. Jancovic, and F. J. Smith, “Robust speech recognition using probabilistic union models,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 403–414, Sep. 2002.