# Speech Enhancement Using MEL Frequency Cepstral Coefficients

**M. Deebalakshmi(PG scholar)[1], B. Kirubagari( Assistant professor)[2]**
[1, 2] Department of Computer Science and Engineering
[1, 2] Annamalai University, Chidambaram, India

***Abstract-*** *This paper proposes a novel method of speech enhancement that moves away from conventional filtering-based methods. It aims to reconstruct clean speech from a set of speech features. Speech enhancement based on Gaussian Mixture Model (GMM) using Mel-frequency cepstral coefficients (MFCC) is well established in the fields of speech processing. A maximum a posterior approach (MAP) is proposed to estimate the clean MFCC vector is made from noisy MFCC vector. The enhanced signal is reconstructed by reducing noise. Analysis is done on NOIZEUS database which consists of speech signals corrupted by eight different real world noises at different SNR levels. The performance of the proposed algorithm is evaluated using different measures such as log-likelihood ratio, signal-to-noise ratio.*

***Keywords***- Speech enhancement, Gaussian Mixture Model, MAP, MFCC, Speech Processing.

## I. INTRODUCTION

Speech is the ability to express thoughts by articulate sounds. Speech communication is ten times faster than written communication. Speech is commonly used method of transferring information from one person to another. Speech signal is produced as a result of time varying excitation of the time varying vocal tract system [1, 2]. Speech production mechanism essentially consists of a vibrating source of sound coupled to a resonating system [3]. Speech enhancement involves processing speech signal for human listening or as preparation for further processing before listening. The enhancement process aims to improve the overall quality of degraded speech signal, to increase the speech intelligibility in order to reduce the listener fatigue, ambiguity etc depending on specific application. The enhancement system may be designed only to achieve one of these aims or several.

Speech enhancement is closely related to speech restoration. When speech is degraded, its restoration to the original speech signal often leads to speech enhancement [4]. There are, however some important differences between enhancement and restoration. In the speech restoration, an ideal speech signal is degraded and the objective is to make the processed speech signal as close as possible to the original. On the other hand the objective of speech enhancement is to make the processed signal sound better than the unprocessed signal. In substantiation of this it can be said that an original un-degraded signal cannot be further restored, but can be enhanced by making it sound clearer. Speech enhancement system is shown in figure 1.

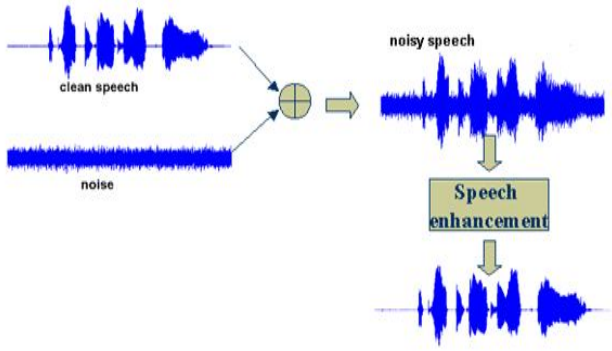


Figure1 : Speech Enhancement System

This paper is organized as follows: Section 2 describes MFCC feature extraction. Section 3 explain about GMM. Section 4 explains MAP estimation. Section 5 gives the proposed methodology. In Section 6 the experimental results are given. Performance analysis is presented in section 7 and the conculation is given in section 9.

## II. MEL FREQUENCY CEPSTRAL COEFFICIENTS EXTRACTION

MFCC is commonly used for automatic speech recognition (ASR) [6]. It is to represent the spectrum of an audio frame [5]. MFCC is a representation of the short term power spectrum of a sound based on a non linear mel scale of frequency. MFCC feature extraction is shown in Figure 2.

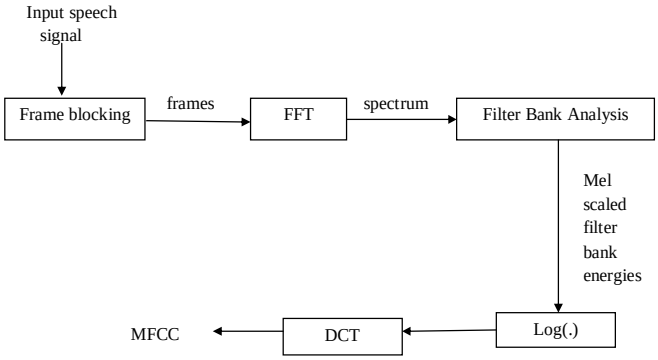


Figure 2: Block Diagram of MFCC Feature Extraction

MFCCs are commonly derived as follows [6,7]:

- Frame blocking: The speech signal is decomposed into short frames of size 20ms with a frame shift of 10ms. The signal is windowed using hamming window.
- Frequency Domain Transformation: The signal is transformed using Fast Fourier Transform.
- Mel Filters are applied.
- Log is applied to compress the value
- The values are transformed using Discrete Cosine Transformation.

Human hearing is not equally sensitive at all frequency bands. It is less sensitive at higher frequencies roughly above 1000 Hz. The mapping of frequency in mel scale is linear below 1000Hz and logarithmic above 1000 Hz. So the band edges and centre frequencies of the filters are linear for low frequency and logarithmically increase with increasing frequency. We call these filters as mel-scale filters and collectively a mel-scale filter bank [8].

As can be seen, the filters used are triangular and they are equally spaced along the mel-scale which is defined by

$$\text{Mel}(f) = 2595 \log_{10} (1 + \frac{f}{700}) \quad (1)$$

## III. GAUSSIAN MIXTURE MODEL

Gaussian is a characteristic symmetric bell curve shape that quickly falls off towards 0 (practically) mixture model is a probabilistic model which assumes the underlying data to belong to a mixture distribution [9]. A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system [10]. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm.

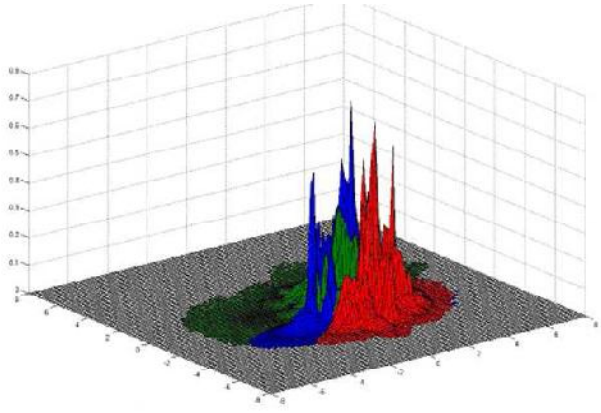

Figure 3 : Gaussian Mixture model

Mathematical Description of GMM is define by

$$p(x) = w1\ p1\ (x) + w2p2\ (x) + \ldots\ldots\ldots + wn\ pn\ (x) \quad (2)$$

where p(x) = mixture component
w1, w2 ….. wn = mixture weight or mixture coefficient pi (x) = Density functions

Gaussian mixture model (GMM) is a mixture of several Gaussian distributions, represent different subclasses inside one class. The probability density function is defined as a weighted sum of Gaussians. Gaussian mixture models are formed by combining multivariate normal density components. Gaussian mixture models are often used for data clustering. It is shown in Figure 3.

The most common mixture distribution is the Gaussian (Normal) density function, in which each of the mixture components are Gaussian distributions, each with their own mean and variance parameters.

$$p(x) = w1N(\ x\ |\mu 1 \sum 1\ ) + \ldots + w1N(\ x\ |\ \mu n \sum n\ ) \quad (3)$$

μi - mean and ∑i - covariance-matrix of individual components (probability density function)

## IV. MAP ESTIMATION

To remove the effects of noise, a MAP estimate of the clean MFCC vector is made from the noisy MFCC vector [11]. This approach is similar to SPLICE [12], which makes clean feature estimates from noisy feature for a robust speech recognition system. A model is first made of the joint density of the clean and noisy MFCC vectors using a Gaussian mixture model(GMM). Training begins by first creating a joint feature vector, $z_i$

$$z_i = [c_i, n_i] \quad (4)$$

Where ci and ni are clean and noisy MFCC vector representing frame i. From a set of training vector expectation – maximization (EM) clustering is applied to create a GMM, $\phi$, to model the joint density

$$\mu_k^z = \begin{bmatrix} \mu_k^c \\ \mu_k^n \end{bmatrix} \quad \sum_k^z = \begin{bmatrix} \sum_k^{cc} & \sum_k^{cn} \\ \sum_k^{nc} & \sum_k^{nn} \end{bmatrix} \qquad (5)$$

The GMM comprises a set of K Gaussian probability density function ($PDF_s$), $\phi_k$ , that localise the joint density of the clean and noisy MFCC vectors. $\alpha_k$ represents the prior probability of the $k^{th}$ cluster and $\sum_k$ represent the mean covariance of the join vector within the $k^{th}$ Gaussian distribution

The mean vector comprises clean and noisy MFCC mean vector, $\mu_k^c$and $\mu_k^n$ The covariance matrix consist of clean and noisy covariance matrices,$\sum_k^{cc}$ and $\sum_k^{nn}$and cross-covariance of the clean and noisy MFCCs, $\sum_k^{cn}$ and $\sum_k^{nc}$

A MAP estimate of the clean MFCC vector can be made from the noisy MFCC vector and their joint density. For the $k^{th}$ clustering in the GMM, $\phi_k$, the MAP estimate of the clean MFCC vector from the noisy MFCC vector, $n_i$ is given

$$\hat{c}_i^k = \arg\max(\Pr(c_i| n_i, \phi_k)) \qquad (6)$$

The estimates from each cluster in joint density can be combined by weighting by the posterior probability of the noisy MFCC vector belonging to the $k^{th}$ cluster, to give a weighted estimate of the clean MFCC vector

$$\hat{c}i=\sum_{k=1}^{k} h_k(n_i) \arg\max (\Pr (c_i|n_i, \phi_k)) \qquad (7)$$

$h_k(n_i)$ represents the posterior probability of the noisy MFCC vector is defined

$$h_k(n_i)=\frac{\alpha_k \Pr (n_i|\phi_k )}{\sum_{k=1}^{K} \alpha_k \Pr (n_i|\phi_k )} \qquad (8)$$

$pr(ni|\phi_k)$ is the marginalized distribution of the noisy MFCC vector. Finally, the estimate of the clean speech MFCC vector can be calculated.

$$\hat{c}i=\sum_{k=1}^{k} h_k (n_i)\left(\mu_k^c + \sum_k^{cn}(\sum_k^{nn})^{-1}(n_i - \mu_k^n)\right) \qquad (9)$$

Finally the effects of noise a MAP estimate of the clean MFCC vector is made from the noisy MFCC vector.

## V. THE PROPOSED METHOD

Speech enhancement improves the overall perceptual quality of the degraded speech signal that moves away from conventional filtering-based methods and instead aims to reconstruct clean speech from a set of speech features. Speech enhancement based on Gaussian Mixture Model using Mel-frequency cepstral coefficients (MFCCs) is well established in the fields of speech processing. The noisy signal is decomposed into short frames and Mel-frequency cepstral coefficients (MFCCs) are obtained and then these coefficients are trained and tested with Gaussian Mixture Model (GMM). A maximum a posterior (MAP) approach is proposed estimating the clean MFCC vector from noisy MFCC vector. A set of subjective tests, measuring speech quality, noise intrusiveness and overall quality is used, which is shown in Figure 4.

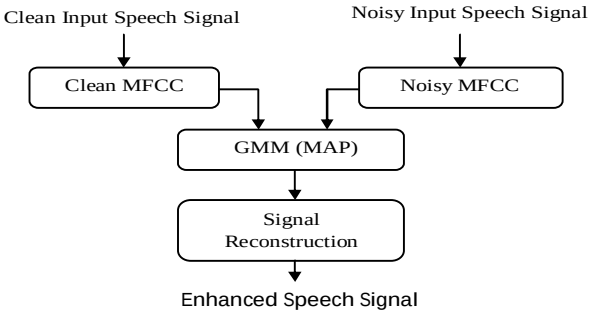


Figure 4: Architectural Diagram

## VI. EXPERIMENTAL AND RESULT

Experiments are done on noisy speech signal database (NOIZEUS) which is corrupted by eight different noises at 0dB, 5dB, 10dB, and 15dB SNR levels. The noisy database produced by three male and three female. The noise was taken from the AURORA database and includes suburban train noise, babble, car, exhibition hall, restaurant, street, and airport and railway-station noise [13]. The feature vectors are calculated for the given input noisy speech signal and clean speech signal. Further GMM is used to calculate the probability density function after that noise is removed from the noisy speech signal.

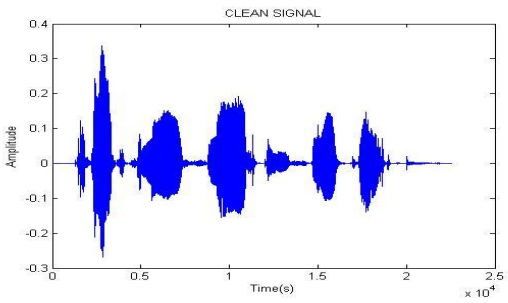


Figure 5: Clean Speech Signal

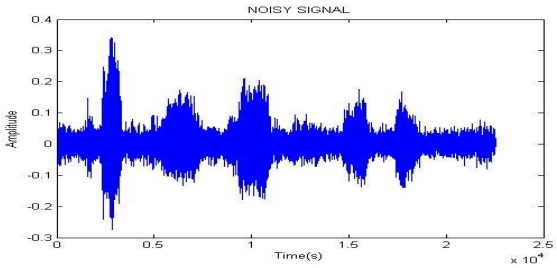


Figure 5: Noisy Speech Signal

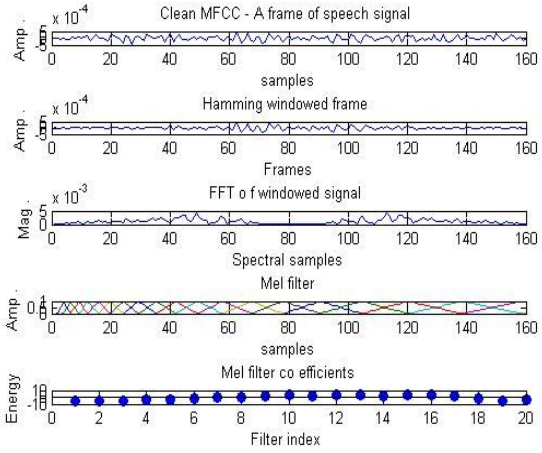


Figure 6: Clean MFCC vector

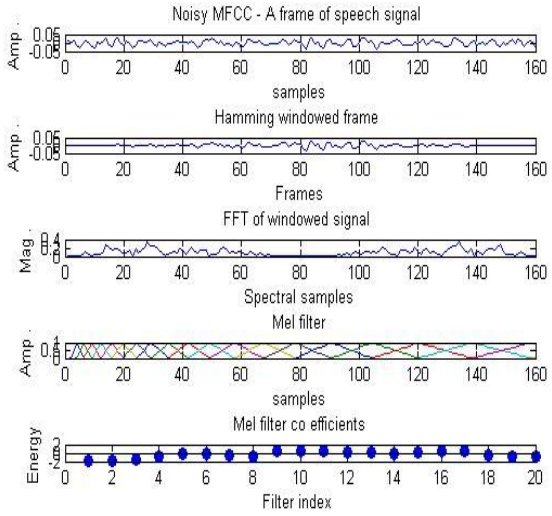


Figure 7: Noisy MFCC vector

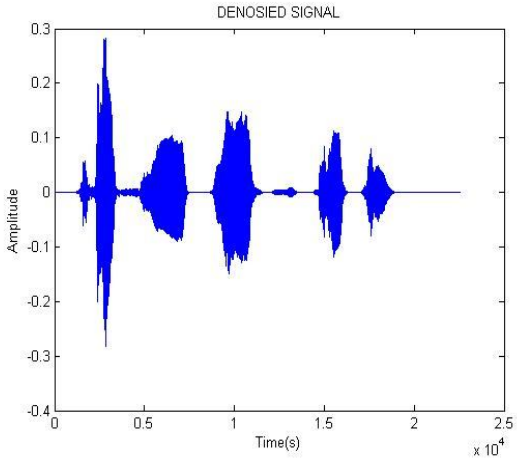


Figure 8: Enhanced Speech signal

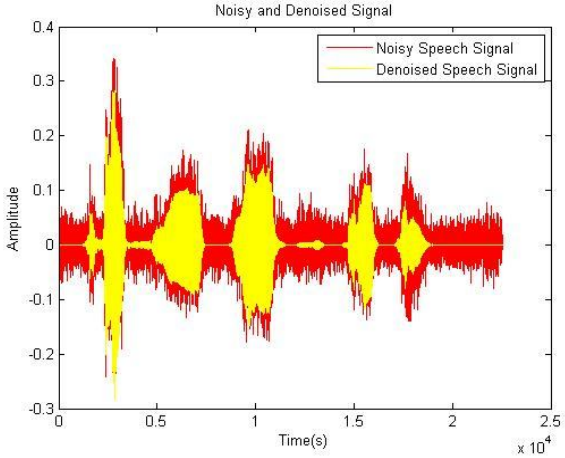


Figure 9: Clean and Enhanced Speech Signal

## VII. PERFORMANCE ANALYSIS

The performance of enhanced speech signal is measured by Signal to Noise Ratio (SNR) and Log Likelihood Ratio (LLR).

**Signal to Noise Ratio**

Computationally, this is the simplest test, but the most un-reliable one. Let, s(t), z(t), $\hat{S}(t)$ be the clean, corrupted and enhanced speech signal, respectively, and T the sample size [14]

Define by,

$$SNR_{in} = 10\log10\frac{\Sigma^{T}S^{2}(t)}{\Sigma^{T}_{i=1}[s(t)-Z(t)]^{2}} \quad (10)$$

$$SNR_{out} = 10\log10\frac{\Sigma^{T}S^{2}(t)}{\Sigma^{T}_{i=1}[s(t)-Z(t)]^{2}} \quad (11)$$

The SNR levels in the input and in the output of the evaluated enhancer. Define by the difference

$$G=SNR_{out} - SNR_{in} \quad (12)$$

The SNR improvement is achieved by the enhancement algorithm (in dB). It is important to emphasize that the improvement in SNR generally does not translate into improvement in speech quality and/or intelligibility. The performance is calculated by defining the SNR values for the relevant speech signals, the speech signals are taken in various areas like airport, babble, car, exhibition, train, restaurant and station. The noise are added to the clean signal and processed. The experimental SNR values given in the below Table 1 shows the SNR of enhanced signal. The noise removed from 0dB and 5dB gives good result compared with 10dB and 15 dB. The performance chart of enhanced signal is shown in figure 10.

Table 1: SNR values for Enhanced Speech Signal

| SNR | 0dB | 5dB | 10dB | 15dB |
|---|---|---|---|---|
| Airport | 2.32 | 7.14 | 8.47 | 8.20 |
| Babble | 2.04 | 6.53 | 8.54 | 8.73 |
| Car | 4.55 | 8.40 | 8.66 | 8.80 |
| Exhibition | 6.14 | 5.68 | 8.79 | 8.69 |
| Restaurant | 1.48 | 7.31 | 8.38 | 8.55 |
| Station | 1.50 | 8.66 | 8.80 | 8.72 |
| Street | 6.85 | 7.72 | 8.39 | 8.65 |
| Train | 5.12 | 8.31 | 8.69 | 8.59 |

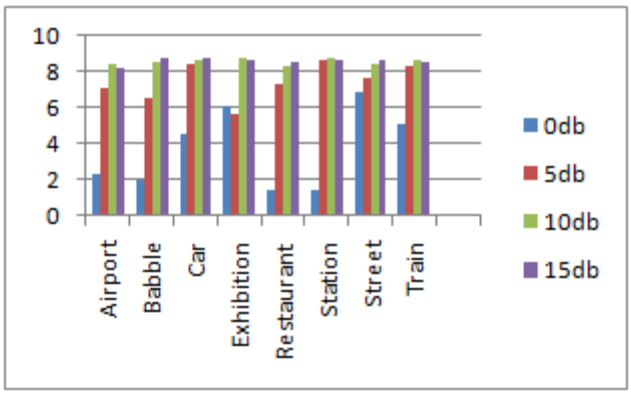


Figure 10 : SNR values of Enhanced Speech Signal

**Log-likelihood ratio**

The Log-likelihood ratio (LLR) measure is one of the three types of LPC-based objective measures to evaluate the quality of based on linear predictive coding (LPC) techniques [15]. LLR measure define by,

$$d_{LLR}(A,B) = \ln\left[\frac{A.V.A^T}{B.V.B^T}\right] \quad (13)$$

where $A = [1 - a_1 - a_2A - a_N]^T$ is the LPC vector of the original speech signal frame,

$B = [1 - b_1 - b_2A - b_N]^T$ is the LPC vector of the enhanced speech frame and $V$ is the autocorrelation matrix of the original speech signal. A larger value of $d_{LLR}$ indicates the better quality of enhanced speech. The experimental LRR values given in the below Table 2 shows the enhanced signal. The noise removed from 0dB and 5dB gives good result compared with 10dB and 15 dB. Figure 11 shows the performance chart of enhanced signal.

Table 2 : LLR values for Enhanced Speech Signal

| | 0dB | 5dB | 10dB | 15dB |
|---|---|---|---|---|
| Airport | 7.69 | 8.73 | 8.01 | 8.13 |
| Babble | 7.61 | 9.34 | 8.52 | 8.06 |
| Car | 7.77 | 8.48 | 8.54 | 7.54 |
| Exhibition | 8.48 | 10.37 | 8.85 | 8.02 |
| Restaurant | 7.40 | 8.92 | 7.69 | 7.87 |
| Station | 8.75 | 8.89 | 8.42 | 8.08 |
| Street | 8.76 | 8.20 | 8.56 | 7.86 |
| Train | 7.60 | 8.34 | 10.36 | 7.81 |

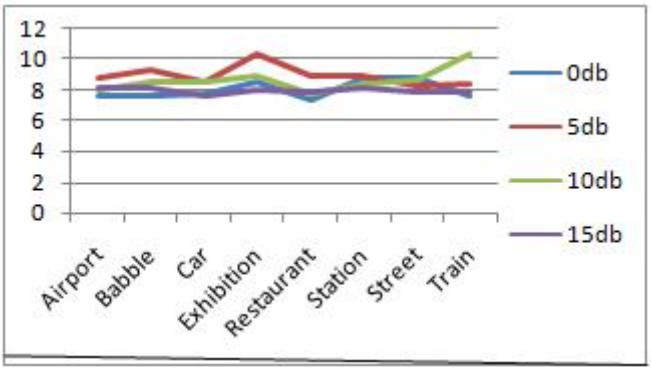


Figure 11: LLR values of Enhanced Speech Signal

## VIII. CONCLUSION

Speech enhancement improves the overall perceptual quality of degraded speech signal. This project shows that speech enhancement can be achieved by reconstructing speech from a set of clean speech features extracted from noisy speech. Speech enhancement is based on Gaussian Mixture Model using Mel-frequency cepstral coefficients (MFCCs) is proposed. A maximum a posterior (MAP) approach is proposed estimating the clean MFCC vector is made from noisy MFCC vector. Analysis is done on NOIZEUS database which consists of speech signals corrupted by eight different real world noises at different SNR levels. The performance of the proposed algorithm is evaluated using different measures such as log-likelihood ratio, signal-to-noise ratio.

## REFERENCES

[1] L. Rabiner and B.H. Juang, 'Fundamentals of Speech Recognition', Pearson Education, Singapore, 2003.

[2] D. O. Shaughnessy, 'Speech Communications-Human and Machine', University Press, India, 2001.

[3] S. Palanivel, 'Person authentication using speech, face and visual speech', Ph.D. thesis, IIT, Madras, September 2004.

[4] J. Darch and B. Milner, 'A comparison of estimated and MAP predicted formants and fundamental frequencies with a speech reconstruction application', in ICSLP, pp. 542–545, August 2007.

[5] L. Rabiner and R.W. Schafer, 'Digital Processing of Speech Signals', Pearson Education, 2005

[6] Jesper Jensen and Zheng-Hua Tan, 'Minimum Mean-Square Error Estimation Of Mel-Frequency Cepstral Features–A Theoretically Consistent Approach', IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, November 2015.

[7] K. T. Deepak and S. R. Mahadeva Prasanna, 'Foreground Speech Segmentation and Enhancement Using Glottal Closure Instants and Mel Cepstral Coefficients', IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 24(7), July 2016.

[8] Z.Tufekci, John N. Gowdy, Sabri Gurbuz and Eric Patterson,' Applied mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition',Speech Communication, 2006.

[9] M. J. Alam, P. Kenny, P. Dumouchel and D. O'Shaughnessy, 'Noise Spectrum Estimation using Gaussian Mixture Model-based Speech Presence Probability for Robust Speech Recognition', 2014.

[10] L. Deng, A. Acero, M. Plumpe, and X. Huang, ,Large-vocabulary speech recognition under adverse acoustic environments,' in ICSLP, vol. 3, 2000, pp, 806-809.

[11] Tahira Mahboob, Memoona Khanum, Malik Sikandar Hayat Khiyal and Ruqia Bibi, 'Speaker Identification Using GMM with MFCC', International Journal of Computer Science Issues, Volume 12, Issue 2, March 2015.

[12] Cheang Soo Yee and Abdul Manan Ahmad,' Mel Frequency Cepstral Coefficients for Speaker Recognition Using Gaussian Mixture Model-Artificial Neural Network Model',2003.

[13] Hu, Y. And Loizou, p. 'Subjective evaluation and compression of speech enhancement algorithms', speech communication, vol. 49, no. 7, pp. 588-601, 2007.

[14] Ashraf M. Aziz, 'Subband Coding of Speech Signals Using Decimation and Interpolation', International Confrence on Aerospace Sciences & Aviation Technology", pp.1-16, May 2009.

[15] B. Kirubagari and S. Palanivel, ' Three-stage hybrid system for speech signal enhancement,, Int. J. Signal and Imaging Systems Engineering, Vol. 8, No 1/2, 2015.