International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 3– No.1, July 2012 – www.ijais.org

Improving Speaker Identification Performance by Combining Vocal Tract Features

S.Selva Nidhyananthan
Assistant Professor
Mepco Schlenk Engineering College
Sivakasi, India

R.Shantha Selva Kumari
Professor
Mepco Schlenk Engineering College
Sivakasi, India

G.Jaffino
PG Student
Mepco Schlenk Engineering College
Sivakasi, India

.

ABSTRACT

This paper proposes fusion and addition techniques of vocal tract features such as Mel Frequency Cepstral Coefficients (MFCC) and Dynamic Mel Frequency Cepstral Coefficients (DMFCC) in speaker identification. Feature extraction plays an important role as a front end processing block in Speaker Identification (SI) process. Mel frequency features are used to extract the spectral characteristics of the speech such as formant frequency and the bandwidth of formant frequency. This feature estimation method leads to robust recognition performance. The Dynamic Mel frequency features are used to extract the dynamic behavior of the human vocal tract using pitch frequency. This work is focused to increase the identification accuracy with databases containing short length speech signal. Experimental evaluation is carried out on TIMIT database with 630 speakers using Gaussian Mixture Model (GMM).

Keywords

DMFCC, MFCC, GMM, Feature extraction, Speaker identification.

1. INTRODUCTION

In speaker identification task the unknown speech signal is provided to the system as input and the system finds out who the speaker is? By matching the input speech signal with the speech signals present in the known database [1]. Speaker recognition system task is categorized into two main phases, Enrollment and verification phases. In enrollment phase, the speakers' voice is recorded and typically a number of features are extracted to form a model. In verification phase, a speech sample is compared against a previously created voice print. In the modern world technology, speech based techniques are being used in many applications such as, voice recognition as entry point into the secured information access, Biometric applications and Forensic applications. Speech may be categorized into voiced and unvoiced speech. Voiced speech is produced when air flow produced by the lungs are forced through the glottis on to the vocal cords. Voiced sounds are distinguished by the presence of periodicity in the corresponding acoustic waveform. Voiced sounds are having more energy content. Unvoiced speeches, do not have any periodicity, and do not have a distinct association with pitch [2]. The main difference between voiced and unvoiced sounds falls in the significant vibration on the vocal cords.

In this paper Mel Frequency Cepstral Coefficients (MFCC) and Dynamic Mel Frequency Cepstral Coefficients (DMFCC) feature extraction techniques are discussed and then both these features are added. The MFCC are coefficients that are derived from a sort of cepstral representation of the signal. These feature extraction techniques extract both linear and non-linear properties. The frequency bands are evenly spread out on the Mel scale and this estimates the human auditory system's response closer than linearly spaced frequency bands would.

Text independent speaker identification system can be modeled using any of the following techniques: Hidden Markov Model (HMM) [3], Gaussian Mixture Model (GMM) [4] and Vector Quantization (VQ) [5]. In this paper Gaussian Mixture Model is used for speaker identification. Gaussian mixture density provides smooth approximation to the sample distribution of observations obtained from utterances of a given speaker. Gaussian Mixture Model is a parametric probability density function represented as a weighted sum of Gaussian component densities [6]. The parameter of GMM is unique for a speaker and it is necessary for speaker identification. Preprocessing is the process of making the speech signal is more suitable for further processing. This step includes silence removal, framing and windowing.

The remainder of this paper is organized as follows. Section 2 describes the procedure for MFCC feature extraction. Section 3 describes the DMFCC feature extraction method. Gaussian Mixture Modeling is given in section 4 and the experimental evaluation is explained in section 5. Section 6 describes the conclusion and suggestions for future work.

2. MFCC FEATURE EXTRACTION

Feature extraction involves simplifying the amount of resources requires to describe a large set of data accurately. The goal of feature selection is to find a transformation to a relatively low-dimensional feature space that preserves the information pertinent to the application while enabling meaningful comparisons to be performed using simple measures of similarity [6]. The block diagram of proposed Fusion based speaker identification is shown in Figure 2.1 and the proposed addition based Speaker Identification is shown in Figure 2.2.

2.1 Fast Fourier Transform

After preprocessing, Fast Fourier Transform and its squared magnitude are calculated to each frame of the speech signal. Fast Fourier Transform is an efficient algorithm for computing DFT in sequence. The DFT pair equation is given by,

$$x(n) = \frac{1}{N}\sum_{k=1}^{N-1} X(k) e^{j\frac{2\pi}{N}kn}, 0 \le n \le N-1 \qquad (1)$$

The sequence $x(n)$ and the transform $X(k)$ specified the interval $[0, N-1]$.

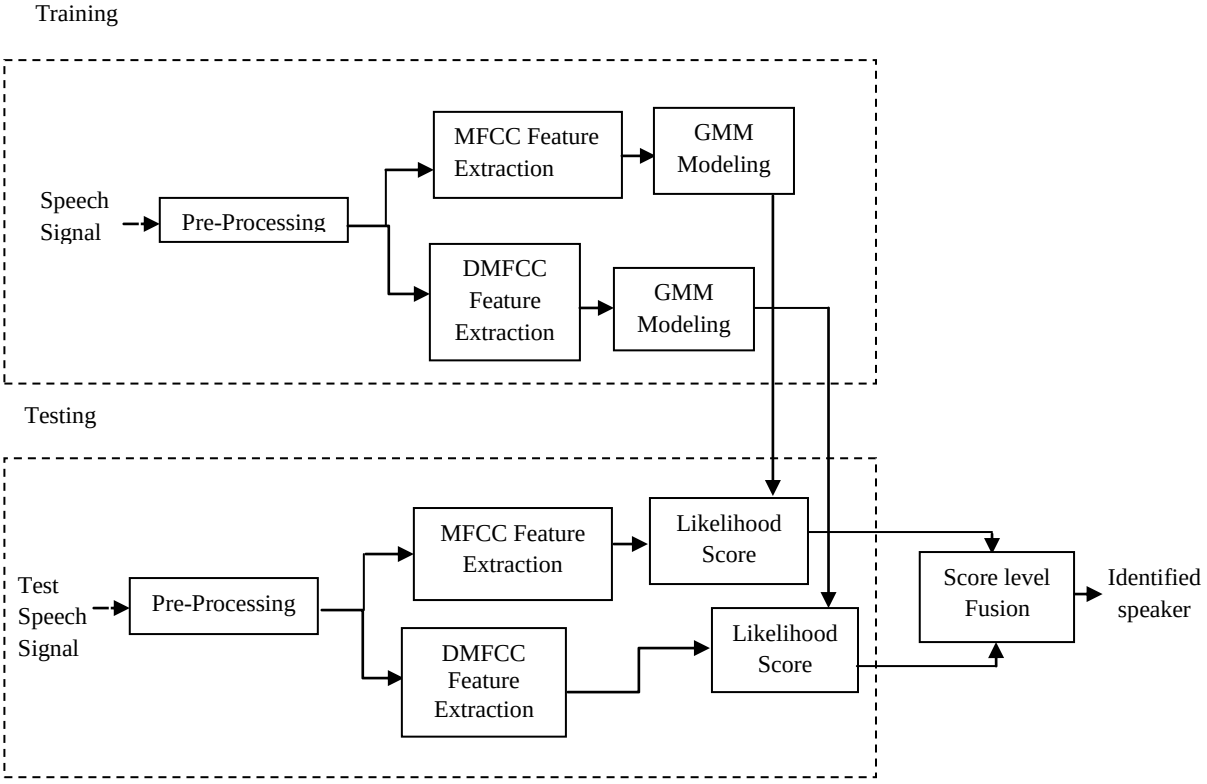

Figure 2.1 Block diagram of proposed Fusion based Speaker Identification

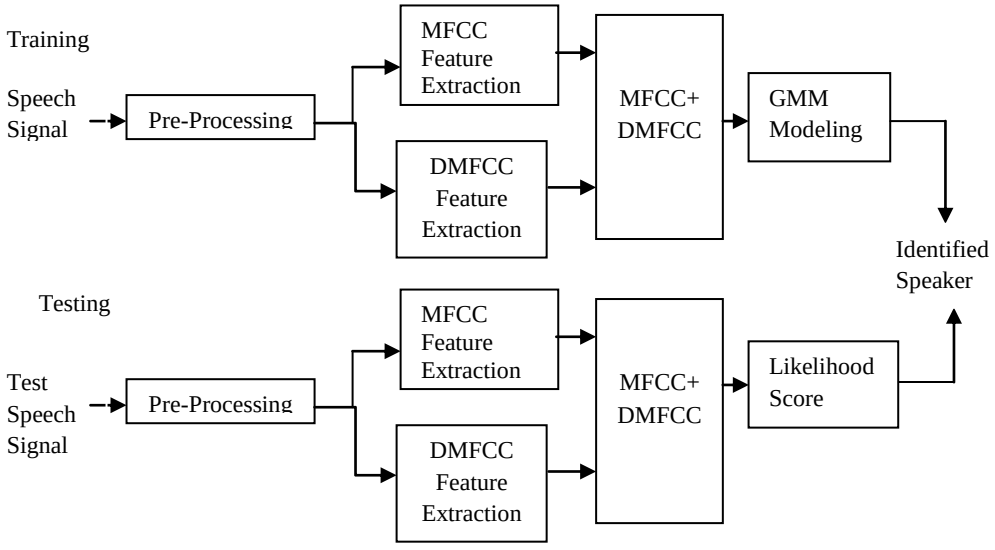

Figure 2.2 Block diagram of proposed addition based Speaker

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}, 0 \le k \le N-1 \qquad (2)$$

2.2 Pre-emphasis

Pre-emphasis refers to a system process designed to increase within a band of frequencies, the magnitude of some higher frequencies with respect to the magnitude of some lower frequencies [7]. Pre-emphasis helps to equalize the spectral the spectral tilt in speech and the signal is spectrally flattened. The output of pre-emphasis is given by,

$$y(n) = x(n) - \alpha x(n-1) \tag{3}$$

Where $x(n)$ is the speech signal, α is pre-emphasis factor. In this paper, the pre-emphasized factor is chosen as 0.97.

2.3. Mel-Scale Filter bank

There are two types of filter bank. Uniform and non-uniform filter bank. Since speech signal contains more information in the low frequency range; there is a need of more number of filters in this low frequency range, this prompts the use of non uniform filter bank [8]. So in this project, non-uniform filter bank is used.

For a non-uniform filter bank the bandwidth b_i and center frequency f_i is given by,

$$b_1 = C \tag{4}$$

$$b_i = \alpha b_{i-1} \tag{5}$$

$$f_i = f_1 + \sum_{j=1}^{i-1} b_j + \left(\frac{b_i - b_1}{2} \right), 2 \le i \le Q \tag{6}$$

Where

C *and* f_1 are the arbitrary bandwidth and center frequency of the first filter and α is the logarithmic growth factor. The mel scale filter bank equation is given by,

$$mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{7}$$

Where $mel(f)$ is the subjective pitch in Mels corresponding to the actual frequencies in hertz.

Conventionally used filter bank in speech feature extraction is Triangular filter bank. This paper explains the Gaussian filter bank and its advantages over Triangular filter bank. Generally Triangular filter is asymmetric, tapered but does not provide any weight outside the sub band. The response of triangular filter is shown in Figure 2.1.

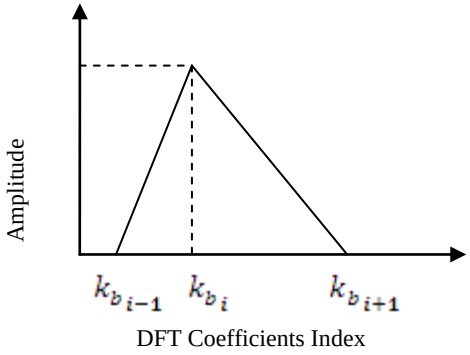

Figure 2.1 Response of Triangular filter

In triangular filter bank, if there is a high energy area in the left part of the filter bank, this might be partially suppressed as a result of filtering. To overcome this, Gaussian filters are used to better capture the energies in that frequency bin [9]. The response of Gaussian filter is shown in Figure 2.2.

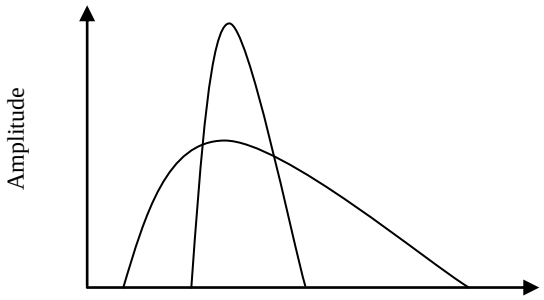

Figure 2.2 Response of Gaussian filter

The filter response is given by,

$$H(k,m) = e^{-\frac{(f(k)-f_c(m))^2}{2\sigma_i^2}} \tag{8}$$

Where

$$f(k) = \frac{kf_s}{N}, k = 1,2,...N-1 \tag{9}$$

2.4 Discrete Cosine Transform

A discrete cosine transform expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies. The expression of DCT is given as,

$$y(k) = w(k) \sum_{n=1}^{N} x(n) \cos\left(\frac{\pi(2n-1)(k-1)}{2N} \right) \tag{10}$$

Where $k = 1,2,...N$

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 1 \\ \sqrt{\frac{2}{N}}, & 2 \le k \le N \end{cases} \quad (11)$$

These DCT coefficients are taken as the MFCC features.

3. DYNAMIC MEL FREQUENCY CEPSTRAL COEFFICIENTS

The Mel frequency cepstral coefficient is the most widely used feature in speech and speaker identification. As the human auditory system can sensitively perceive the pitch changes in the speech, the algorithm, which combines the speaker information obtained by the MFCC with the pitch, can dynamically construct a set of Mel-filters according to the result of pitch detection. The Mel-filters are then used to extract the dynamic MFCC parameter, which represents the speaker's identity characteristics, and enhance accuracy of speaker recognition [10].

After pre-emphasis, the speech signal is multiplied with the windowed signal and the equation is given by,

$$S_w(n) = y(n) * w(n) \quad (12)$$

where

$y(n)$ is the pre-emphasized output

$w(n)$ is the windowed output signal.

The pitch frequency f_p is calculated by using autocorrelation method.

This pitch frequency is given to the input of the Mel scale filter bank. This equation is given by,

$$mel(f) = 2595 \log(1 + f_p / 700) \quad (13)$$

Next step is to calculate the energy spectrum of the signal. An energy spectrum is a distribution of energy among a large assemblage of particles. It is a statistical representation of wave energy, and an empirical estimator of the spectral function. The energy spectrum is calculated by using,

$$X(k) = \left| DFT(S_w(n)) \right| \quad (14)$$

where

$S_w(n)$ is the windowed and pre-emphasized multiplied signal.

The last step is, discrete cosine transform. The discrete cosine transform attempts to decorrelate the data [11]. After decorrelation, each transform coefficient can be encoded independently without losing compression efficiency. DCT is also used for good energy compaction.

4. GAUSSIAN MIXTURE MODELING

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 [12]. For $X_1, X_2, ...X_N \in R^D$ are a given set of *N* points in *D*-dimensions, and a family *F* of probability density functions on R^D. The functions *F* could be mixtures of Gaussian functions then,

$$f(x;\theta) = \sum_{k=1}^{K} p_k g(x; m_k, \sigma_k) \quad (15)$$

Where

$$g(x; m_k, \sigma_k) = \frac{1}{\left(\sqrt{2\pi}\sigma_k\right)^D} e^{-\frac{1}{2}\left(\frac{\|x - m_k\|}{\sigma_k}\right)^2}$$

is a *D*-dimensional isotropic Gaussian function $\theta = \theta_1, \theta_2 ...., \theta_K = ((p_1, m_1, \sigma_1), ..., (p_{K,} m_K, \sigma_K))$ is a $K(D+2)$-dimensional vector containing the mixing probabilities p_k as well as the means m_k and standard deviations σ_k of the *K* Gaussian functions in the mixture. Each Gaussian function integrates to one.

$$\int_{R^D} g(x; m_k, \sigma_k) dx = 1 \quad (16)$$

Since *f* is a density function, it must be non-negative and integrate to one [13].

$$1 = \int_{R^D} f(x;\theta) dx = \int_{R^D} \sum_{k=1}^{K} p_k g(x; m_k, \sigma_k) dx$$
$$= \sum_{k=1}^{K} p_k \int_{R^D} g(x; m_k, \sigma_k) dx = \sum_{k=1}^{K} p_k \quad (17)$$

The numbers p_k must be nonnegative, $f(x)$ takes on negative values must be added up to one.

$$p_k \ge 0 \text{ and } \sum_{k=1}^{K} p_k = 1 \quad (18)$$

This is why the numbers p_k are called mixing probabilities. The likelihood functions can be defined as,

$$\wedge(X;\theta) = \prod_{n=1}^{N} f(x_n;\theta) \quad (19)$$

For mixtures of Gaussian functions can be defined as,

$$\wedge(X;\theta) = \prod_{n=1}^{N}\sum_{k=1}^{K} p_k g(x_n;m_k,\sigma_k) \quad (20)$$

Parametric density function problem can be defined as,

$$\hat{\theta} = \arg\max_{\theta} \wedge(X;\theta) \quad (21)$$

The conditional probability of selected component *k* given that data point x_n is,

$$p(k \mid n) = \frac{q(k,n)}{\sum_{m=1}^{K} q(m,n)} \quad (22)$$

The logarithm of likelihood function can be defined as,

$$\lambda(X;\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} p_k g(x_n;m_k,\sigma_k) \quad (23)$$

On EM iteration, the following re-estimation formulas are used.

Mean $$m_k = \frac{\sum_{n=1}^{N} p(k \mid n) x_n}{\sum_{n=1}^{N} p(k \mid n)} \quad (24)$$

Standard deviation

$$\sigma_k = \sqrt{\frac{1}{D} \frac{\sum_{n=1}^{N} p(k \mid n) \left\| x_n - m_k \right\|^2}{\sum_{n=1}^{N} p(k \mid n)}} \quad (25)$$

Mixture weight $$p_k = \frac{1}{N} \sum_{n=1}^{N} p(k \mid n) \quad (26)$$

These are the parameters in Gaussian Mixture Modeling [14].

The score level fusion of features would perform better if they are provided with speaker specific information that is complementary in nature. A governing equation that describes the fusion of outputs parallel classifiers via weighted sum rule is given as,

$$S_{com}^{i} = w\sum_{t=1}^{T} \log p(x_{tMFCC} \mid \lambda_{sMFCC}) + (1-w)\sum_{t=1}^{T} \log p(x_{tDMFCC} \mid \lambda_{sDMFCC}) \quad (27)$$

Where S_{com}^{i} denotes the combined score for GMM based system, x_k denotes the feature vector corresponds to MFCC/DMFCC and λ_n denotes the object corresponding to MFCC/DMFCC. The identity of the true speaker i_{true} is given by,

$$i_{true} = \arg \underbrace{\max}_{i} S_{com}^{i} \quad (28)$$

The speaker corresponding to the object giving the maximum likelihood is identified as the currect speaker. All the signals in the database are tested for identification and the percentage of identification for each of these models are computed. Next the MFCC and DMFCC features are added and then modeled using Gaussian Mixture Model.

5. EXPERIMENTAL EVALUATION

This experimental work uses TIMIT speech corpus for speaker identification task. Texas Instruments-Massachusetts Institute of Technology (TIMIT) is a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects. The speech was recorded at TI, transcribed at MIT, and has been maintained, verified, and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST) [15]. The TIMIT database consists of 630 speakers, out of which 70% are male and 30% are female from 10 different dialect regions in America. Each speaker contains ten speech signals. Out of ten speech signals of every speaker six signals are used for training and the remaining four signals used for testing.

The identification performance of DMFCC feature for 1024 frame size is tabulated in Table 1.

Table 1 Identification Performance of DMFCC feature using 1024 frame size

Sl. No	Total number of speakers	GMM	
		Number of Mixtures	Percentage of Correct Identification (%)
1	630	8	55.45
2		16	59.58
3		32	**63.45**

For this 1024 frame size, mixture weight 32 gives the maximum identification accuracy of 63.45%. The identification performance of DMFCC feature using 512 frame size is tabulated in Table 2.

Table 2 Identification Performance of DMFCC feature using 512 frame size

Sl.No	Total number of speakers	GMM	
		Number of Mixtures	Percentage of Correct Identification (%)
1	630	8	69.39
2		16	**80.12**
3		32	74.85

It can be seen from these observations that the maximum percentage of identification achieved is 80.12% which is achieved for the mixture weight of 16. The frame size for this analysis is taken as 512 samples with 256 samples as overlapping. The testing is done for different mixture components and the results are analyzed. The identification performance of DMFCC feature using 256 Frame size are tabulated in Table 3.

Table 3 Identification Performance of DMFCC feature using 256 frame size

Sl.No	Total number of speakers	GMM	
		Number of Mixtures	Percentage of Correct Identification (%)
1	630	8	83.23
2		16	**89.78**
3		32	86.39

The overall performances of DMFCC feature extraction, the maximum identification accuracy as 89.78%. The identification performance for MFCC feature extraction using 1024 frame size is tabulated in Table 4.

Table 4 Identification Performance of MFCC feature using 1024 frame size

Sl.No	Total number of speakers	GMM	
		Number of Mixtures	Percentage of Correct Identification (%)
1	630	8	63.15
2		16	**71**
3		32	66.2

For this performance of MFCC feature extraction technique the maximum identification accuracy is 71%. The identification performance for MFCC feature extraction using 512 frame size is tabulated in Table 5.

Table 5 Identification Performance of MFCC feature using 512 frame size

Sl.No	Total number of speakers	GMM	
		Number of Mixtures	Percentage of Correct Identification (%)
1	630	8	79.67
2		16	81
3		32	**86.12**

The maximum identification accuracy is 86.12% for the mixture weight of 32. The identification performance for MFCC feature extraction using 256 frame size is tabulated in Table 6.

Table 6 Identification Performance of MFCC feature using 256 frame size

Sl.No	Total number of speakers	GMM	
		Number of Mixtures	Percentage of Correct Identification (%)
1	630	8	88.72
2		16	**92.53**
3		32	91.01

The fusion technique is used to fuse MFCC and DMFCC with the weight of w=0.77 and it is tabulated in Table 7.

Table 7 Identification performance of Fusion Technique

	Frame size		
Mixture weight	1024	512	256
8	61.43	75.57	91.56
16	67.32	88.67	**98.02**
32	69.04	79	90.19

For this fusion technique the maximum identification accuracy is 98.02%. The identification of MFCC+DMFCC added results are tabulated in Table 8.

Table 8 Identification performance for MFCC+DMFCC added feature

	Frame size		
Mixture weight	1024	512	256
8	49.52	51	63.02
16	55.18	57.68	**76.12**
32	52.35	68.15	69

6. CONCLUSION

In this project, GMM modeling technique is evaluated for speaker identification using TIMIT speech database under fusion of model scores and addition of features cases. MFCC feature extraction is a better technique for speaker identification system, by adding complementary DMFCC feature along with it. An efficient fusion technique to fuse the scores from the GMM model of MFCC and DMFCC is used. For a TIMIT database with 630 speakers a maximum percentage of 89.78% is achieved for DMFCC feature and 92.53% is achieved for MFCC feature. In this project, good identification accuracy of 98.02% is achieved with fusion of both the features which outperforms the result obtained by adding the complementary features. This work can be further extended to mismatched and noisy conditions of speech signal.

7. REFERENCES

[1] Douglas O' Shaughnessy,"Speech Communication Human and Machines," II nd edition, Universities press (India) Limited (2001).

[2] S.Selva Nidhyananthan, R.Shantha Selva Kumari and G.Jaffino," Text-Independent speaker identification using residual feature extraction Technique," CiiT International Journal of Digital signal processing, march 2012.

[3] A. E. Rosenberg et al., "Connected word talker verification using whole word Hidden Markov Models," in Proc. ICASSP, 1991, pp. 381-384.

[4] D. A. Reynolds and R.C.Rose published a paper, "Robust test-independent speaker identification using Gaussian mixture Speaker models."IEEE Transaction on Speech Audio Processing, vol.3, 1995, pp 72-83.

[5] Tomoko Matsui and Sadaoki Furui, "Comparison of Text Independent Speaker Recognition Methods Using VQ Distortion and Discrete Continuous HMM's," IEEE transactions on speech and audio processing, vol. 2, no. 3, July 1994.

[6] Md.Rashidul Hasan, Mustafa Jamil Md.Golam Rabbani,Md.Saifur Rahman, "Speaker Identification using Mel Frequency Cepstral Coefficients", 3rd International conference on Electrical and computer engineering ICECE 2004,Dec 2004.

[7] Douglas O' Shaughnessy, "Speech communication Human and Machines*", IInd edition , Universities Press(India) Limited(2001).*

[8] Prodesy and speech recognition by Alex Waibel, vol. 1, Nos.1-2, 2007.

[9] Sandipan Chakroborty, Goutam Saha, "Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter" , International Journal of Signal Processing 5:1, 2009.

[10] Wang Yutai, Li Bo, Jiang Xiaoqing, Liu Feng, Wang Lihao," Speaker Recognition Based on Dynamic MFCC Parameters" IEEE proceedings 2009.

[11] Tomi Kinnunen, Haizhou Li,"An Overview of Text-Independent Speaker Recognition: From Features to Super vectors", august 2009.

[12] Miyajima, Y.Hattori, K.Tokuda, T.Kabayashi and T.Kitamura " Text-Independent Speaker Identification using Gaussian Mixture Models based on multispace probability distribution," IEEE Transactions on information and system, vol.E84-B,2001,pp.847-855.

[13] C.Miyajima, Y.Hattori, K.Tokuda, T.Kabayashi and T.Kitamura," Text-Independent speaker identification using Gaussian mixture models based on multispace probability distribution," IEEE Transactions on information and system, vol.E84-B, 2001, pp.847-855.

[14] Murthy.K and Yegnanarayana.B," Combining evidence from residual phase and MFCC features for speaker Recognition," Signal Processing Letters; IEEE, vol.13, no.1, pp.52- 55, Jan 2006.

[15] Victor Zue, Stephanie Seneff, James Glass,"Speech database development at MIT: Timit and beyond", Speech Communication, Volume 9, Issue 4, August 1990, Pages 351–356.