Application Of Different Filters In Mel Frequency Cepstral Coefficients Feature Extraction And Fuzzy Vector Quantization Approach In Speaker Recognition

*Satyanand Singh
Associate Professor

**Dr. E.G. Rajan
Director

Abstract

Front-end or feature extractor is the first component in an automatic speaker recognition system. Feature extraction transforms the raw speech signal into a compact but effective representation that is more stable and discriminative than the original signal. Since the front-end is the first component in the chain, the quality of the later components is strongly determined by the quality of the front-end. Over the years, Mel-Frequency Cepstral Coefficients (MFCC) modeled on the human auditory system has been used as a standard acoustic feature set for speech related applications. In this paper it has been shown that the inverted Mel-Frequency Cepstral Coefficients is one of the performance enhancement parameters for speaker recognition, which contains high frequency region complementary information in it and also introduces the Gaussian shaped filter (GF) and Tukey while calculation MFCC and inverted MFCC in place of traditional triangular shaped bins. The main idea is to introduce a higher amount of correlation between subband outputs. The performance of both MFCC and inverted MFCC improve with GF and Tukey over traditional triangular filter (TF) based implementation, individually as well as in combination. In this study, Fuzzy Vector Quantization (FVQ) is used for speaker modeling. Fuzzy clustering methods allow objects to belong to several clusters simultaneously, with different degrees of membership. The performances of proposed GF and Tukey Filter based MFCC and IMFCC in individual and merged mode have been verified in two standard databases POLYCOST (Telephone Speech) and TIMIT each of which has more than 130 speakers as well as self voice collected from 90 speakers.

1.Introduction

A speaker recognition system mainly consists of two main module, speaker specific feature extractor as a front end followed by a speaker modelling technique for generalized representation of extracted features [1, 2]. Since long time MFCC is considered as a reliable front end for a speaker recognition application because it has coefficients that represents audio, based on perception [3, 4]. In MFCC the frequency bands are positioned logarithmically which approximated the human auditory systems response more closely than the linear spaced frequency bands of FFT or DCT. An illustrative speaker recognition system is shown in figure.1.The state of the art speaker recognition research primarily investigates speaker specific complementary information relative to MFCC. It has been observed that the performance of speaker recognition improved significantly when complementary information is merged with MFCC in feature level either by simple concatenation or by combining models scores. The main complementary information is pitch [5], residual phase [6], prosody [7], dialectical features [8] etc. These features are related with vocal chord vibration and it is very difficult to extract speaker specific information. It has been shown that complementary information can be captured easily from the high frequency part of the energy spectrum of a speech frame via reversed filter bank methodology [9]. There are some features of speaker which used to present at high frequency part of the spectrum and generally ignored by MFCC that can be captured by inverted MFCC is proposed in this paper. The complementary information captured by inverted MFCC is modelled by FVQ [10] technique. In this paper, Fuzzy Vector Quantization (FVQ) is used for speaker modelling. In many real situations, fuzzy clustering is more natural than hard clustering, as objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial memberships.So the present study was undertaken with the objective of to find out the speaker recognition efficiency improving components.

2. Methodology

In the present investigation GF and Tukey filter were used as the averaging bins instead of triangular for calculating MFCC as well as inverted MFCC in a typical speaker recognition application [11, 12]. There are three main inspiration of using GF and Tukey filter. First inspiration is both filter can provide much smoother transition from one subbands to other preserving most of the correlation between them. Second inspiring point is the means and variances of these can be independently chosen in order to have control over the amount of overlap with neighbouring subbands. Third inspiring point is the filter design parameters for GF and Tukey can be calculated very easily from mid as well as end-points located at the base of the original TF used for MFCC and inverted MFCC. In this investigation both MFCC and inverted MFCC filter bank are realized using a moderate variance where a GF's and Tukey coverage for a subbands and the correlation is to be balanced. Results show that GF and Tukey based MFCC and inverted MFCC perform better than the conventional TF based MFCC and inverted MFCC individually. Results are also better when GF and Tukey based MFCC & inverted MFCC is merged together their model scores in comparison to the results that are obtained by combining MFCC and inverted MFCC feature sets realized using traditional TF [13]. All the implementations have been done with FVQ [14]

3. Mel Frequency and Their Calculation

3.1 Mel-Frequency Cepstral Coefficients using triangular filters

According to psychophysical studies human perception of the frequency content of sounds follows a subjectively defined nonlinear scale called the Mel scale [15, 16]. MFCC is the most commonly used acoustic features for speech/speaker recognition. MFCC is the only acoustic approach that takes human perception (Physiology and behavioral aspects of the voice production organs) sensitivity with respect to frequencies into consideration, and therefore is best for speaker recognition. The acoustic model is defined as, voice production organs) sensitivity with respect to frequencies into consideration, and therefore is best for speaker recognition. The acoustic model is defined as,

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

Where f_{mel} is the subjective pitch in Mels corresponding transfer function and the actual frequency in Hz. This leads to the definition of MFCC, a baseline acoustic feature for speech and speaker recognition applications, which can be calculated as follows [17]

Let $\{y(n)\}_{n=1}^{Ns}$ represent a frame of speech that is

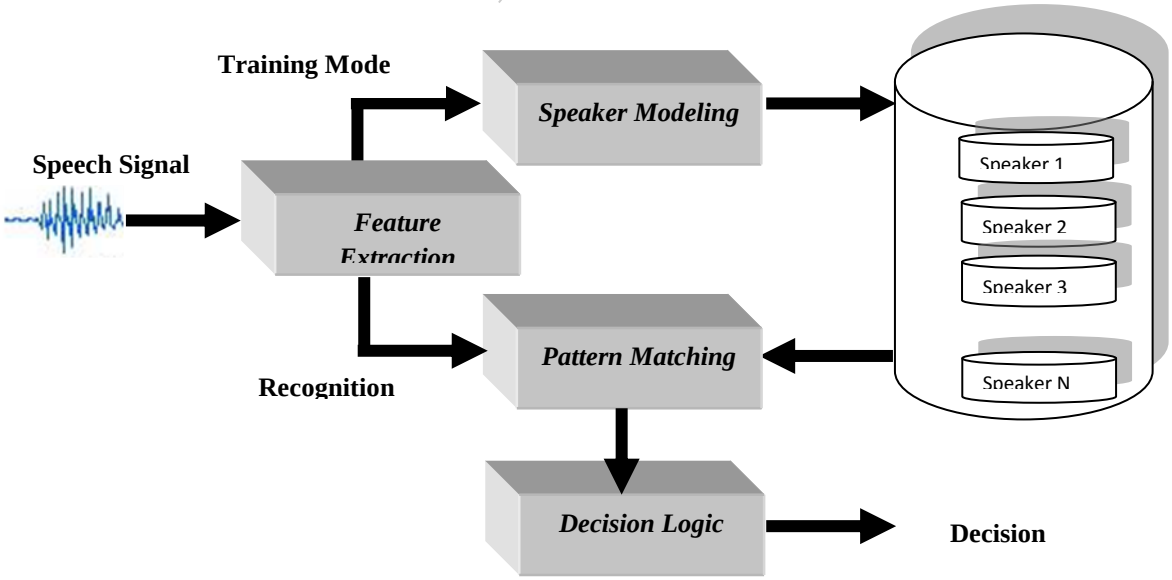

Figure 1: Speaker recognition system

preemphasized and Hamming-windowed.
First, y(n) is converted to the frequency domain by an M_s-point DFT which leads to the energy spectrum,

$$Y|k|^2 = \left|\sum_{n=l}^{Ms} y(n).e^{\left(-\frac{j2\pi nk}{Ms}\right)}\right|^2 \quad (2)$$

boundary points are equally spaced in the Mel scale which is satisfying the definition,

Where $1 \le k \le M_s$. this is followed by the construction of a filter bank with Q unity height TFs, uniformly spaced in the Mel scale eqn. (1). The filter response Ψi (k) of the ith filter in the bank (figure-2) is defined as,

$$\Psi_i(k) = \begin{cases} 0 & for\ k \le k_{b\,i-1} \\ \dfrac{k - k_{bi-1}}{k_{bi} - k_{bi-1}} & for k_{b\,i-1} \le k \le k_{bi} \\ \dfrac{k_{bi+1} - k}{k_{bi+1} - k_{bi}} & for\ k_{bi} \le k \le k_{b+1} \\ 0 & for\ k \ge k_{b\,i+1} \end{cases} \quad (3)$$

Where $1 \le i \le Q$, Q is the number of filters in the bank, $\{k_{bi}\}_{i=0}^{Q+1}$ are the boundary points of the filters and k denotes the coefficients index in the M_s point DFT. The filter bank

$$k_{bi} = \left(\frac{M_s}{F_s}\right) f_{mel}^{-1}\left[f_{mel}\,(f_{low}) + \frac{i\{f_{mel}\,(f_{high}) - f_{mel}\,(f_{low})\}}{Q+1}\right] \quad (4)$$

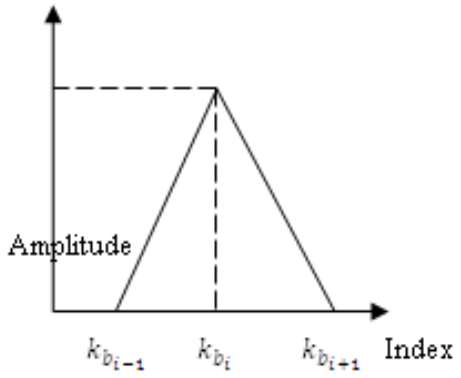

Figure 2: Response $\Psi_i(k)$ of a typical Mel scale filter

where the function fmel (•) is defined in eqn. (1), M_s is the number of points in the DFT eqn. (2), F_s is the sampling frequency, f_{low}, and f_{high} are the low and high frequency boundaries of the filter bank and f_{mel}^{-1} is the inverse of the transformation in eqn.(1) defined as,

$$f_{mel}^{-1}(f_{mel}) = 700.\left[10^{\frac{f_{mel}}{2595}} - 1\right] \quad (5)$$

The sampling frequency F_s and f_{low}, f_{high} frequencies are in Hz while f_{mel} is in Mels. In this work, F_s is 8 kHz. Ms is taken as 256 f_{low} = Fs/Ms = 31.25 Hz while f_{high} = F_s /2 = 4 kHz. Next, this filter bank is imposed on the spectrum calculated in eqn. (2). The outputs $e(i)_{i=1}$ Q of the Mel-scaled band-pass filters can be calculated by a weighted summation between respective filter response Ψi(k) and the energy spectrum $|Y(k)|^2$ as

$$e(i) = \sum_{k=1}^{\frac{M_s}{2}} |Y(K)|^2 \,.\Psi_i \quad (6)$$

Finally DCT is taken on the log filter bank energies $\{\log[e(i)]\}_{i=1}$ and the final MFCC coefficients C_m can be written as.

$$C_m = \sqrt{\frac{2}{Q}\sum_{l=0}^{Q-1} log[e(i+1)].cos\left[m\left(\frac{2l-1}{2}\right).\frac{\pi}{Q}\right]} \quad (7)$$

Where $0 \le m \le R-1$, R is the desired number of cepstral features.

3.2 Mel-Frequency Cepstrum Coefficients using Gaussian filters.

The transfer function of any filter is asymmetric, tapered and filter does not provide any weight outside the subband that it covers. As a result, the correlation between a subband and its nearby spectral components from adjacent subbands is lost. In this investigation a GF and Tukey is proposed, which produced gradually decaying weights at its both ends and symmetric for compensating possible loss of correlation. Referring to eqn. (3), the expression for GF can be written as [18]

$$\Psi_i^{g_{MFCC}} = e^{-\frac{(k-k_{bi})^2}{2\sigma_i^2}} \quad (8)$$

Where k_{bi} is a point between the ith transfer boundaries located at it base and it is considered here as a mean of the ith GF while the σ_i is the standard deviation and can be defined as,

$$\sigma_i = \frac{k_{bi+1} - k_{bi}}{\alpha} \quad (9)$$

Where α is the variance controlling parameter. However, in eqn. (8) the conventional denominator i.e $\sqrt{(2\pi)\sigma_i}$ is dropped, as its presence is only to ensure the area under Gaussian curve is unity [19]. Moreover, omitting the term helps a GF to achieve

unity as highest value at its mean,which is similar to unity height triangular shaped filter used for conventional MFCC. Note that, a TF become nonisosceles while they are mapped into from its two endsk_{bi} in base become unequal. For MFCCs' i_{th} filter, the relation becomes,

$$(k_{bi+1} - k_{bi}) > (k_{bi} - k_{bi-1}) \quad (10)$$

We took the maximum spread out of these two distances i.e. $(k_{bi+1} - k_{bi})$ to evaluate σ_i ensuring full coverage of the subband by the GF.

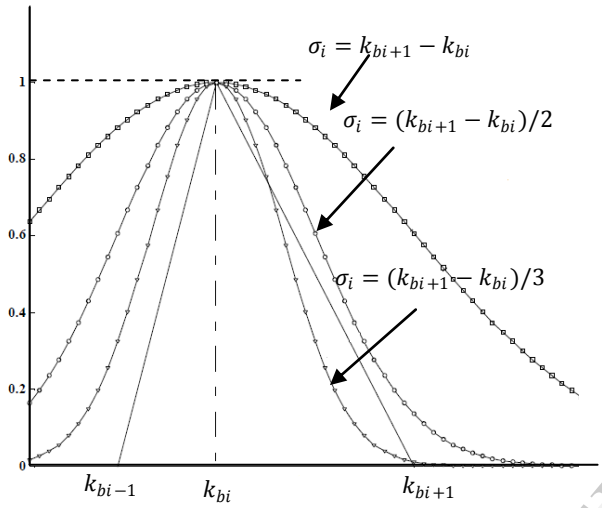

Figure. 3 Response of various shaped filters

Figure.3, shows the plot TF and GF for different values of sigma. The figure clearly depicts that a triangular window can give some sort of tapering at its both ends but lacks also in offering of no of weights outside its coverage.

In figure 4. shows the standard deviation for different values of α and k_{bi} is the centre point of al filters after that transfer functions are gradually decaying. However the Gaussian with higher variance shows larger correlation with nearby frequency component. Thus selection of α is a critical part for setting the variances of GF. In the present study the value of α = 2 then eqn. (9) can be written as,

$$2\sigma = k_{bi+1} - k_{bi} \quad (11)$$

95% of subband is covered once α = 2, is selected. Probability $[2\sigma \geq (k_{bi+1} - k_{bi})]$ = 0.95. Therefore, α = 2 can provide better correlation with nearby subbands in comparison to α = 3. In this study, we have chosen α = 2 to design filters for the MFCC filter bank. Thus, a balance is achieved where significant coverage of a particular subband is ensured while allowing moderate correlation between that subband and neighboring ones. The cepstral vector using GFs can be calculated from the filter's response eqn. (8) which is as follows

$$e^{g_{MFCC}}(i) = \sum_{k=1}^{\frac{Ms}{2}} |Y(K)|^2 \Psi^{g_{MFCC}} i(k) \quad (12)$$

and,

$$C^{g_{MFCC}} m = \sqrt{\frac{2}{Q}} \sum_{i=1}^{Q-1} log [e^{g_{MFCC}}(i+1)]. cos\left[m\left(\frac{2l-1}{2}\right).\frac{\pi}{Q}\right] \quad (13)$$

Here last 20 coefficient from both models are used and the value of Q=22 and R = 25 are taken.

Table I shows the different values of α and their coverage within the curve and outside the curve.

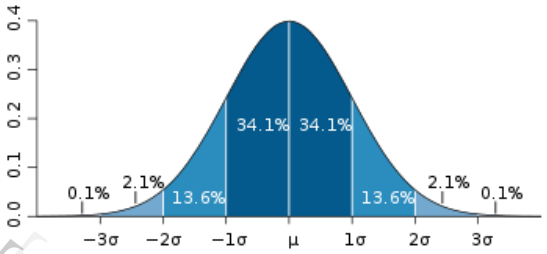

Figure 4: standard deviation

Table 1. Summary of sigma and its coverage

σ	% Within the curve	% Outside the curve
2σ	95.4499736%	4.5500264%
3σ	99.7300204%	0.2699796%

3.3 Mel-Frequency Cepstrum Coefficients using Tukey Filter

Based on evidence Tukey filter is a combination of the rectangular window and the Hann window [20]. In fact it is a cosine- tampered window and is defined as follows

$$\Psi_i(k) = \begin{cases} \frac{1}{2}\left(1 - cos\left(2\pi.\frac{k - k_{bi-1}}{N_{Hann}}\right)\right), \\ for\ k_{bi-1} \leq k_{bi-1} + 1 \leq k_{bi-1} + \frac{N_{Hann}}{2} \\ 1, \\ k_{bi-1} + \frac{N_{Hann}}{2} + 1 \leq k \leq k_{bi+1} - \frac{N_{Hann}}{2} \\ \frac{1}{2}\left(1 - cos\left(2\pi.\frac{k - N_{Rect}}{N_{Hann}}\right)\right), \\ k_{bi+1} - \frac{N_{Hann}}{2} \leq k \leq k_{bi+1} \\ 0, \quad for\ k \geq k_{b\,i+1} \end{cases} \quad (14)$$

Where $k = 0, 1 \ldots, N-1$ is discrete frequency, $0 \le k_{bi-1} < k_{bi} < k_{bi+1} < N$ are basic filter frequencies and unit amplitude of the filter and length of the filter is N sample. N_{Hann} is defined as

$$N_{Hann} = (1-\alpha).(k_{bi} - k_{bi-1} + 1) \quad (15)$$

Where α is the ratio of taper to constant section and $0 \le \alpha \le 1$. When $\alpha = 0$, then filter corresponds to a rectangular filter. When $\alpha = 1$, the filter corresponds to Hann filter. N_{Rect} in the Eq. (14) is a complement of the N_{Hann} and is defined as

$$N_{Rect} = \alpha.(k_{bi+1} - k_{bi-1} + 1) \quad (16)$$

Figure.5, shows the plot Tukey filter for different values of α. The figure clearly depicts that a Tukey window can give some sort of tapering at its both ends.

4. Inverted Mel Frequency Cepstral Coefficients Calculation

4.1 Inverted Mel-Frequency Cepstral Coefficients using triangular filters

The main objective is to capture that information which has been missed by original MFCC [21]. In this study the new filter bank structure is obtained simply by flipping the original filter bank around the point f = 2 kHz which is precisely the mid-point of the frequency range considered for speaker recognition applications. This flip-over is expressed mathematically as,

$$\hat{\Psi}_i(k) = \Psi_{Q+1-i}\left(\frac{M_s}{2} + 1 - k\right) \quad (17)$$

Where $\hat{\Psi}_i(k)$ is the inverted Mel Scale filter response while $\Psi_i(k)$ is the response of the original MFCC filter bank $1 \le i \le Q$ and Q is the number of filters in the bank. From eqn. (13) we can derive an expression for $\hat{\Psi}_i(k)$ with analogous to eqn. (3) $\hat{\Psi}_i(k)$ for the original MFCC filter bank.

$$\hat{\Psi}_i(k) = \begin{cases} 0 & for\ k \le \hat{k}_{b\,i-1} \\ \dfrac{k - \hat{k}_{b\,i-1}}{k_{bi} - k_{bi-1}} & for\ \hat{k}_{b\,i-1} \le k \le \hat{k}_{bi} \\ \dfrac{\hat{k}_{b_{i+1}} - k}{\hat{k}_{b_{i+1}} - \hat{k}_{bi}} & for\ \hat{k}_{bi} \le k \le \hat{k}_{b+1} \\ 0 & for\ k \ge \hat{k}_{b\,i+1} \end{cases} \quad (18)$$

Where $l \le k \le M_s$ and $\{\hat{k}_{bi}\}_{i=0}^{Q+1}$

Here inverted mel-scale is defined as follows

$$\hat{f}_{mel}(f) = 2195.2860 - 2595\, log_{10}\left(1 + \frac{4031.25 - f}{700}\right) \quad (19)$$

Where $\hat{f}_{mel}(f)$ is subjective pitch in the new scale corresponding to f, the actual frequency in Hz .The filter outputs $\{\hat{e}(i)\}_{i=1}^{Q}$ in the same way as MFCC from the same energy spectrum $|Y(K)|^2$ as

$$\hat{e}(i) = \sum_{k=1}^{\frac{Ms}{2}} |Y(K)|^2 . \hat{\Psi}_i(k) \quad (20)$$

DCT is taken on the log filter bank energies $\{\log_{10}[\hat{e}(i)]\}_{i=1}^{Q}$ and the final inverted MFCC coefficient $\{\hat{C}_m\}_{m=1}^{R}$ can be written as

$$\hat{C}_m = \sqrt{\frac{2}{Q}} . \sum_{l=0}^{Q-1} log[\hat{e}(i + 1)]\, cos\left[m.\left(\frac{2l-1}{2}\right)\frac{\pi}{Q}\right] \quad (21)$$

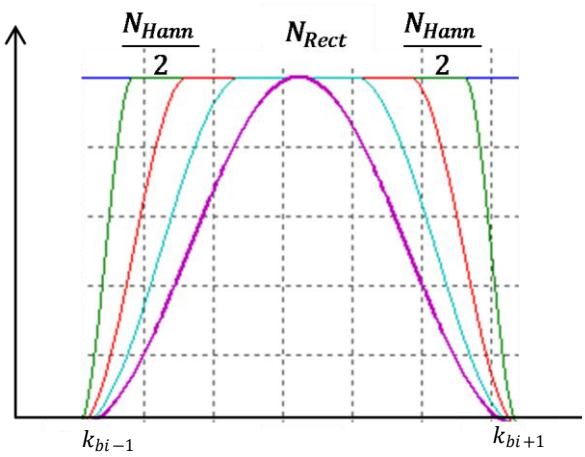

Figure 5: Response of Tukey Filter

4.2 Inverted Mel-Frequency Cepstral Coefficients using Gaussian filters

It is expected that introduction of correlation between subband outputs in inverted mel-scaled filter bank makes it more complementary than what was realized using TF. Flipping the original triangular filter bank, around 2 KHz inverts also the relation mentioned in eqn. (10) that gives

$$(\hat{k}_{bi} - \hat{k}_{bi-1}) > (\hat{k}_{bi+1} - \hat{k}_{bi}) \quad (22)$$

Here $\hat{k}_{bi}$ is the mean of the ith GF and standard deviation can be calculated as

$$\hat{\sigma}_i = \frac{\hat{k}_{bi} - \hat{k}_{bi-1}}{\alpha} \quad (23)$$

Here α value is chosen 2. The response of the GF for inverted MFCC filter bank and corresponding cepstral parameters can be calculated as follows;

$$\hat{\Psi}_i^{g_{IMFCC}} = e^{-\frac{(k-\hat{k}_{bi})^2}{2\hat{\sigma}_i^2}} \quad (24)$$

$$\hat{e}^{g_{IMFCC}}(i) = \sum_{k=1}^{\frac{Ms}{2}} |Y(k)|^2 . \hat{\Psi}_i^{g_{IMFCC}}(k) \quad (25)$$

And

$$\hat{C}_m^{g_{IMFCC}} = \sqrt{\frac{2}{Q}} \sum_{l=0}^{Q-1} log[\hat{e}^{g_{IMFCC}}(i+1)]. \cos\left[m.\left(\frac{2l-1}{2}\right).\frac{\pi}{Q}\right] \quad (26)$$

4.3 Inverted Mel-Frequency Cepstral Coefficients using Tukey Filter

The main objective is to capture that information which has been missed by original MFCC. In this study the new filter bank structure is obtained simply by flipping the original filter bank around the point f = 2 kHz which is precisely the mid-point of the frequency range considered for speaker recognition applications. This flip-over is expressed mathematically as

$$\hat{\Psi}_i(k) = \begin{cases} \frac{1}{2}\left(1 - \cos\left(2\pi.\frac{k - k_{bi-1}}{N_{Hann}}\right)\right), \\ for\ \hat{k}_{b\,i-1} \le \hat{k}_{b\,i-1} + 1 \le \hat{k}_{b\,i-1} + \frac{N_{Hann}}{2} \\ 1, \\ \hat{k}_{b\,i-1} + \frac{N_{Hann}}{2} + 1 \le k \le \hat{k}_{b\,i-1} - \frac{N_{Hann}}{2} \\ \frac{A}{2}\left(1 - \cos\left(2\pi.\frac{k - N_{Rect}}{N_{Hann}}\right)\right), \\ \hat{k}_{bi+1} - \frac{N_{Hann}}{2} \le k \le \hat{k}_{bi+1} \\ 0, \quad for\ k \ge \hat{k}_{b\,i+1} \end{cases} \quad (27)$$

5. Synthesis of MFCC and IMFCC

The idea of combining the classifier to enhance the decisions making process has been successful in many pattern classification problems including Speaker Identification. According to the available literature, the combination of two or more classifiers would perform better if they were supplied with information that are complementary in nature. Adopting this idea in our work, we supplied MFCC and IMFCC feature vectors, which are complementary in information content, to two classifiers respectively and finally fused their decisions in order to obtain improved identification accuracy. The same principle has been adopted for GF and Tukey based MFCC and IMFCC also. In this context, it should be noted that our computation of complementary information from IMFCC involves comparably lower computational complexity than higher-level features.

The MFCC and IMFCC features vectors, containing complementary information of speakers, were supplied to a given classifiers independently and the classification results for the MFCC features and the IMFCC features were fused in order to obtain optimum decision in the process of speaker recognition . A uniform weighted sum rule was adapted to fuse the core from two classifiers. If X_{MFCC} denotes the classification score based on the MFCC and X_{IMFCC} denotes the classification score based on the IMFCC, then the combined score for the m[th] speaker was given as

$$X_{com} = wX_{MFCC} + (1-w)X_{IMFCC} \quad (28)$$

The constant value of $\beta = 0.5$ was used in all cases. The speaker was determined as,

$$n_{class} = arm(maxX_{com}) \quad (29)$$

6. Theoretical Background of VQ

In VQ-based approach the speaker models are formed by clustering the speaker's feature vectors in K non-overlapping clusters. Each cluster is represented by a code vector C_i, which is the centroid [22]. The resulting set of code vectors $\{C_1, C_2, C_3, C_4, , \ldots\ldots\ldots\ldots\ldots C_k\}$is called a codebook, and it serves as the model of the speaker. The model size (number of code vectors) is significantly smaller than the training set. The distribution of the code vectors follows the same underlying distribution as the training vectors. Thus, the codebook effectively reduces the amount of data by preserving the essential information of the original distribution. K-means is an iterative approach; in each successive iteration it redistributes the vectors in order to minimize the distortion. The procedure is outlined below:

- Initialized the random centroids as the means of M clusters.
- Data points are associated with the nearest centroids.
- The centroids are moved to the centre of their respective clusters.
- Steps b and c were repeated until a suitable level of junction has been reached, i.e the distortion is minimized.

When the distortion is minimized, redistribution does not result in any movement of vectors among the clusters. This could be used as an indicator to terminate the algorithm. Upon the convergence, the total distortion does not change as a result of redistribution.

6.1 Linde, Buzo and Gray Clustering Technique

The acoustic vectors extracted from input speech of a speaker provide a set of training vectors. As described above, the next important step is to build a speaker-specific VQ codebook for this speaker using those training vectors. There is a well-known algorithm, namely LBG algorithm, for clustering a set of L training vectors into a set of M codebook vectors. The LBG VQ design algorithm is an iterative algorithm which alternatively solves the two optimality criteria. The algorithm requires an initial code C^0. This initial codebook is obtained by the splitting method. In this method, an initial code vector is set as the average of the entire training sequence. This code vector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two code vectors are split into four and the process is repeated until the desired number of code vectors is obtained.

6.2 K-means Clustering Technique

The standard k-means algorithm is a typical clustering algorithm used in data mining and which is widely used for clustering large sets of data. In 1967, Mac Queen firstly proposed the k-means algorithm; it was one of the most simple, non-supervised learning algorithms, which was applied to solve the problem of the well-known cluster. It is a partitioning clustering algorithm and this method are used to classify the given date objects into k different clusters iteratively, converging to a local minimum. So the results of generated clusters are compact and independent. The algorithm consists of two separate phases. The first phase selects k centers randomly, where the value k is fixed in advance. The next phase is to arrange each data object with the nearest centre. Euclidean distance is generally used to determine the distance between each data object and the cluster centers. When all the data objects are included in a cluster, the first step is completed and an early grouping is done. This process is repeated continues repeatedly until the criterion function becomes the minimum. Supposing that the target object is x, x_i indicates the average of cluster c_i . The criterion function is defined in eqn. (30).

$$E = \sum_{i=1}^{K} \|x - x_i\|^2 \qquad (30)$$

E is the sum of the squared error of all objects in database. The distance of the criterion function is Euclidean distance, which is used for determining the nearest distance between each data objects and cluster centre. The Euclidean distance between one vector $x = (x_1, x_2, \ldots\ldots\ x_n)$ and another vector $y = (y_1, y_2, \ldots\ldots\ y_n)$. The Euclidean distance $d = (x_i, y_i)$ can be obtained and shown as shown in eqn. (31)

$$d = (x_i, y_i = \left[\sum_{i=1}^{n} (x_i - y_i)^2\right]^{\frac{1}{2}} \qquad (31)$$

6.3 Fuzzy C-means Clustering

In speech-based pattern recognition, VQ is a widely used feature modeling and classification algorithm, since it is simple and computationally very efficient technique. FVQ reduces disadvantages of classical Vector Quantization. Unlike Linde-Buzo-Gray (LBG) and k-means algorithms, the FVQ technique follows the principle that a feature vector located between the clusters should not be assigned to only one cluster. Therefore, in FVQ each feature vector has an association with all clusters [23]. The discrete nature of hard partitioning also causes analytical and algorithmic intractability of algorithms based on analytic function values, since the function values are not differentiable. Fuzzy c-means is a clustering technique that permits one piece of data to belong to more than one cluster at the same time. It aims at minimizing the objective function defined by Equation (28).

$$J = \sum_{j=1}^{N} \sum_{i=1}^{C} u_{ij}^{m} \left(\left\|x_j^{(j)} - c_i\right\|\right)^2, 1 < m < \infty \qquad (32)$$

Where C is the number of clusters, N is the number of data elements, x_i is an column vector of X, and is defined as the centroid of the i_{th} cluster.u_{ij} is an element of U ,and denotes the membership of data element J to the i_{th} cluster, and S_i subject to the constraints $u_{ij} \in [0,1]$ and $\sum_{i=1}^{C} u_{ij} = 1$ for all j. m is a free parameter which plays a central role in adjusting the blending degree of different clusters. If m is set to 0, J is a sum of square error criterion, and u_{ij} a Boolean membership value (either 0 or 1). $\|*\|$ can be any norm expressing the similarity [24]. Fuzzy partitioning is carried out using an iterative optimization of the objective function with the update of membership function u_{ij} , an element of U, which denotes the membership of data element J to the I_{th} cluster. The cluster center C_j is derived using Equation (28) and (29).

$$u_{ij} = 1/\sum_{k=1}^{C} \left(\frac{\left\|x_i^{(j)} - C_J\right\|}{\left\|x_i^{(j)} - C_{Jk}\right\|}\right)^{\frac{2}{m-1}} \qquad (28)$$

$$C_j = \sum_{i}^{N} u_{ij}^{m} . x_i \Big/ \sum_{i}^{N} u_{ij}^{m} \qquad (29)$$

This iteration will stop when $\max_{ij} \left\{\left|u_{ij}^{K+1} - u_{ij}^{K}\right|\right\} < \varepsilon$, where ε is the termination criterion.

The algorithm for Fuzzy c-means clustering includes the steps:

I. Initialize C,N,m,U
II. Repeat
III. Minimize j, by computing :

$$u_{ij} = 1/\sum_{k=1}^{C}\left(\frac{\left\|x_i^{(j)} - C_j\right\|}{\left\|x_i^{(j)} - C_{jk}\right\|}\right)^{\frac{2}{m-1}}$$

IV. Normalize u_{ij} by $\sum_{i=1}^{C} u_{ij} = 1$
V. Compute Centroid C_j by using:

$$C_j = \sum_{i}^{N} u_{ij}^m . x_i \,/ \sum_{i}^{N} u_{ij}^m$$

VI. Until slightly change in U and V
VII. End

A schematic description of this scheme for parallel combination of classifiers is given in figure. 6.

7. Experimental setup

7.1 Pre-Processing Stag

In this work, each frame of speech is pre-processed as follows:

- Silence removal and end-point detection using an energy threshold criterion.
- Pre-emphasis with 0.97 pre-emphasis factor.
- Frame blocking with 20ms frame length, i.e N_s = 160 samples/frame 50 overlap, and finally Hamming-windowing.

The MFCC and IMFCC feature sets using triangular, GFs and Tukey are calculated.

7.2 POLYCON Database

The database was collected through the European telephone network. The recording has been performed with ISDN cards on two XTL SUN platforms with an 8 kHz sampling rate. In this work, a closed set text independent speaker identification problem is addressed where only the mother tongue (MOT) files are used. Specified guideline [25] for conducting closed set speaker identification experiments is adhered to, i.e. 'MOT02' files from first four sessions are used to build a speaker model while 'MOT01' files from session five onwards are taken for testing. In the POLYCOST case, the English prompts are fully annotated in terms of word boundaries. The mother tongue prompts are just labelled at the word level with no segmentation. In both case, the Speech Dat recommendations were used while performing the annotation.

7.3 Self Collected Voice Database

The voice corpus was collected in unhealthy environment by using Microsoft sound recorder. A good quality head phone belongs to different parts of India. The average duration of the training samples was 6 seconds per speaker and out of twenty utterances one is used for training purpose. For matching purposes remaining 19 voice corpus of the length 6 seconds, which was further divided into three different subsequence of the lengths 6 s (100%), 3 s (50%), 2s (33%) 1s (16 %) and 0.5s(8%) .Therefore, for 70 speakers we put 70X19X5 = 6650 utterance under test and valuated the identification efficiency.

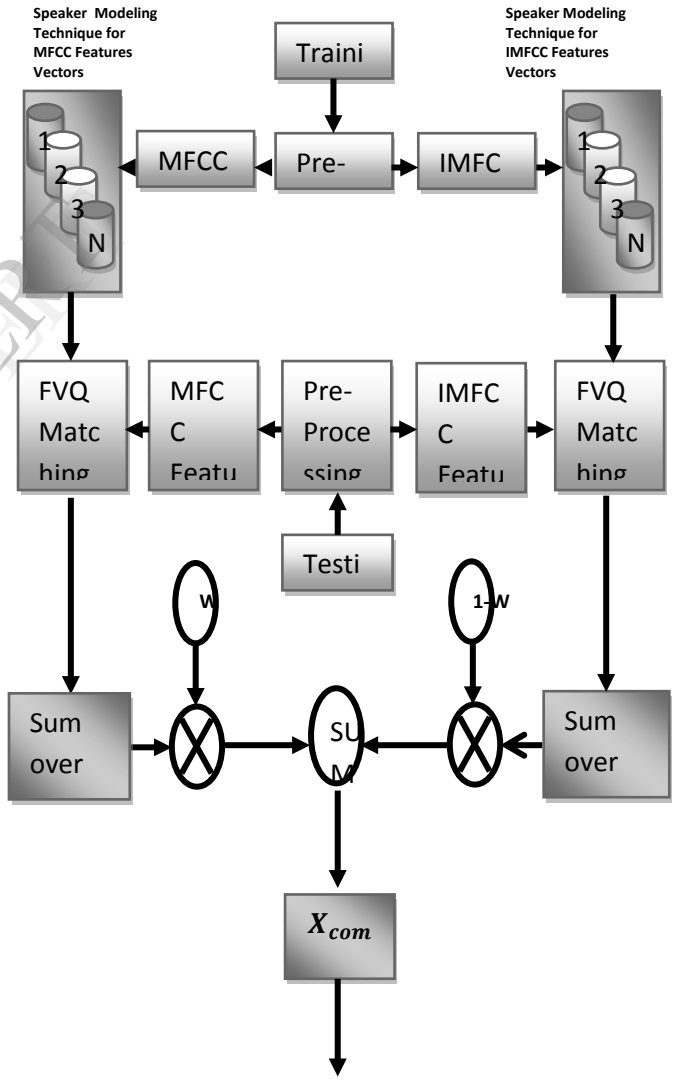

Fig. 6 Parallel classifier based SI system

7.4 TIMIT Database

The TIMIT speech corpus consists of 630 speakers (438 male and 192 female). For each speaker only one recording session was used. The speech data was recorded in a sound booth and contains fixed text sentences read by speakers and recorded over a fixed wideband channel. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. The speakers used American English. The main limitation of the TIMIT corpus is that the speech is recorded only during one session for each speaker, therefore the data does not reflect time related variations in speech characteristics. Moreover, the clean wideband speech environment in TIMIT has an ideal character and does not simulate the real world condition appearing in typical speaker recognition applications.

7.5 Score Calculation

For any closed-set speaker identification problem, speaker identification accuracy is defined as follows in and we have used the same:
Percentage of Identification Accuracy = No of utterance correctly identified /Total No of utterance under test.

8. Experimental Results

For each database, we evaluated the performance of an MFCC based classifier, an IMFCC based classifier where each feature set has been implemented using TF, GF as well as Tukey Filter.

8.1 Results for POLYCOST Database

Table II describes identification results for various model orders of fuzzy c-means clustering VQ with TF based MFCC and IMFCC features set. The last column in the table depicts the identification accuracies for the combined scheme. The combined scheme shows significant improvements over MFCC based SI system for different model orders. Further, even the independent performance of the IMFCC based classifier is comparable to that of the MFCC based classifier. Table III represents PIA of individual MFCC, IMFCC and fused scheme when GFs are used. It is evident from the table that individual performance of each feature set improves when compared against convention TF based MFCC and IMFCC. The fused scheme also outperforms GF based single streamed MFCC as well as earlier combined scheme using TFs, which in turn shows enhancement of complementary information applying GF for realizing the filter bank. Table IV represents PIA of individual MFCC, IMFCC andfused scheme when Tukey filter were used. It is evident from the table that individual performance of each feature set improves when compared against convention Tukey based MFCC and IMFCC. The fused scheme also outperforms Tukey based single streamed MFCC as well as earlier combined scheme using TFs, which in turn shows enhancement of complementary information applying Tukey for realizing the filter bank.

Table II
Results (PIA) for POLYCOST database using TF based MFCC & IMFCC

No. of Utterances	MFCC	IMFCC	**Combined Systems**
1300	77.4515	76.2599	**83.0345**
650	79.2349	78.0557	**84.1631**

Table III
Results (PIA) for POLYCOST database using GF based MFCC & IMFCC

No. of Utterances	MFCC	IMFCC	**Combined Systems**
1300	78.8472	77.6599	**84.0955**
650	80.9019	79.5862	**85.7586**

Table IV
Results (PIA) for POLYCOST database using Tukey filter based MFCC & IMFCC

No. of Utterances	MFCC	IMFCC	**Combined Systems**
1300	78.5472	77.4599	**83.3955**
650	79.8019	79.3862	**84.7586**

Results show that the complementary information supplied helps to improve the performance of MFCC in parallel classifier to a great extent for two types of filters. Thus it can be said that, compared to a single MFCC based classifier; a speaker can be modeled with the same accuracy but at a comparatively lower order model by an MFCC-IMFCC parallel classifier. It could be further concluded that GF based IMFCC provides better complementary information than TF and Turkey based IMFCC. Figure 7 shows the graphical presentation of percentage of identification accuracy of POLYCOST Database.

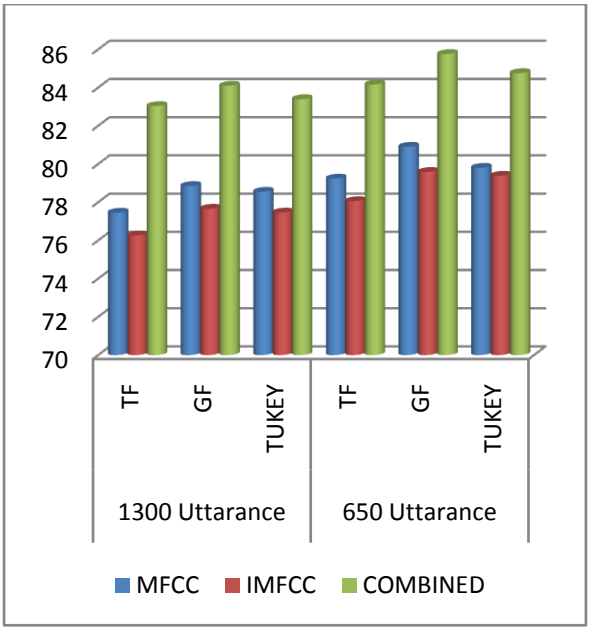

Figure 7. Graphical presentation of PIA for POLYCOST Database

8.2 Results for Self Collected Voice Database

Table V,VI and VII shows the identification accuracies for the self collected voice database for TF,GF and Tukey based filters respectively. PEA obtained using GF based filter bank improves in individual feature sets and combined scheme over various model orders. As with the result shows, it can be observed from these tables that combined scheme shows significant improvement over the baseline MFCC based system irrespective of the filter type.

Table V
Results (PIA) for SELF COLLECTED VOICE database using TF based MFCC & IMFCC

No. of Utterances	MFCC	IMFCC	**Combined Systems**
6650	81.9515	80.3259	**85.2345**
3325	83.8515	82.8557	**86.6631**

Table VI
Results (PIA) for SELF COLLECTED VOICE database using GF based MFCC & IMFCC

No. of Utterances	MFCC	IMFCC	**Combined Systems**
6650	82.9515	81.8734	**88.2445**
3325	84.8515	83.8557	**89.6731**

Table VII
Results (PIA) for SELF COLLECTED VOICE database using Tukey based MFCC & IMFCC

No. of Utterances	MFCC	IMFCC	**Combined Systems**
6650	82.5324	81.9934	**87.5443**
3325	84.3747	83.3145	**88.4534**

Figure 8. shows the graphical presentation of percentage of identification accuracy of self collected voice database. The GF is performing the best among all above mentioned filters.

8.3 Results for TIMIT

Table VIII,IX and X shows the identification accuracies for the TIMIT database for TF,GF and Tukey based filters respectively.

Table VIII
Results (PIA) for TIMIT database using TF based MFCC & IMFCC

No. of Utterances	MFCC	IMFCC	**Combined Systems**
6300	80.9545	79.2389	**83.6234**
3150	80.8978	79.7695	**83.6598**

Table IX
Results (PIA) for TIMIT database using GF based MFCC & IMFCC

No. of Utterances	MFCC	IMFCC	**Combined Systems**
6300	81.8976	79.9876	**85.2386**
3150	82.8734	81.5623	**85.2457**

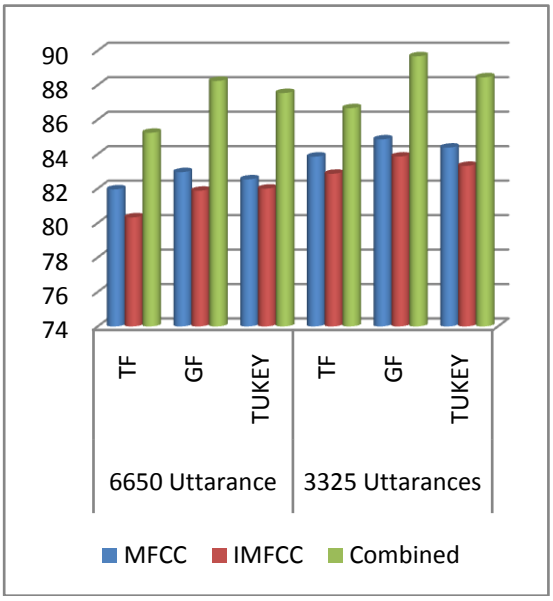

Figure 8. Graphical presentation of PIA for Self Collected Voice Database

Table X
Results (PIA) for TIMIT database using Tukey based MFCC & IMFCC

No. of Utterances	MFCC	IMFCC	**Combined Systems**
6300	80.9356	79.3563	**84.9823**
3150	81.9576	80.886	**84.8967**

It could be further concluded that GF based IMFCC provides better complementary information than TF and Tukey based IMFCC.

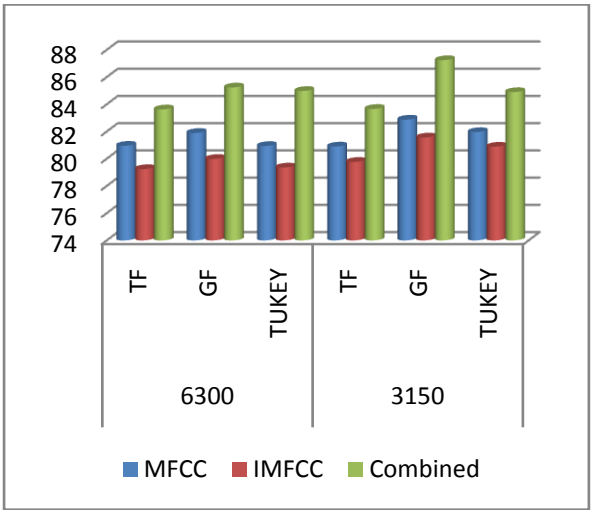

Figure 9. Graphical presentation of PIA for TIMIT Database

Figure 9. shows the graphical presentation of percentage of identification accuracy of TIMIT database. As mentioned above the GF is performing the best among all above mentioned filters.

9. Conclusion

Gaussian filter based mel and inverted mel scaled filter bank is proposed in this paper after getting promising accuracy by comparing result with TF and Tukey filter. An uniform variance is used to design the filter banks, which could maintain a good balance between a filter's coverage area and the amount of correlation. In both the scales, cepstral vectors are obtained and are modeled separately by fuzzy c-clustering VQ method. Performance is found to be superior when the individual performance of the each new proposed feature set is compared with its corresponding baseline. The result is shown for individual cases as well as for combined feature set for three speech databases each of which contains good number of speakers. The GF and Tukey filter show the better identification accuracy compare to TF.

10. References

[1] D. Gatica-Perez, G. Lathoud, J.-M. Odobez and I. Mc Cowan, "Audiovisual probabilistic tracking of multiple speakers in meetings" IEEE Transactions on Speech and Audio Processing, 2007, 15(2), pp. 601–616.
[2] J. P. Cambell, Jr, "Speaker Recognition A Tutorial Proceedings of the IEEE", 85(9), 1997, pp. 1437-1462.
[3] Faundez-Zanuy M. and Monte-Moreno E, "State-of-the-art in speaker recognition, Aerospace and Electronic Systems Magazine" IEEE, 20(5), 2005, pp. 7-12.
[4] K. Saeed and M. K. Nammous, "Heuristic method of Arabic speech recognition" in Proc. IEEE 7th Int. Conf. DSPA, Moscow, Russia, 2005, pp. 528–530
[5] D. Olguin, P.A.Goor, and A. Pentland, "Capturing individual and group behavior with wearable sensors, in Proceedings of AAAI Spring Symposium on Human Behavior Modeling" 2009.
[6] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences" IEEE Trans. On ASSP, 28(4), 1980, pp. 357-365.
[7] R. Vergin, B, O Shaughnessy and A. Farhat, "Generalized Mel frequency Cepstral coefficients for large-vocabulary speaker independent continuous-speech recognition" IEEE Trans. On ASSP,7(5), 1999 pp. 525-532.
[8] Chakroborty, S., Roy, A. and Saha, G, "Improved Closed set Text- Independent Speaker Identification by Combining MFCC with Evidence from Flipped Filter Bank" International Journal of Signal Processing, 4(2), 2007, pp. 114-122.
[9] S.Singh and Dr. E.G Rajan, "A Vector Quantization approach Using MFCC for Speaker Recognition" International conference Systemic, Cybernatics and Informatics ICSCI under the Aegis of Pentagram Research Centre Hyderabad, 2007, pp. 786-790.
[10] K. Sri Rama Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition" IEEE Signal Processing Letters, 13(1),2006, pp. 52-55.
[11] Yegnanarayana B., Prasanna S.R.M., Zachariah J.M. and Gupta C. S, "Combining evidence from source suprasegmental and spectral features for a fixed-text speaker verification system", IEEE Trans. Speech and Audio Processing, 13(4), 2005, pp. 575-582.
[12] J. Kittler, M. Hatef, R. Duin, J. Mataz, "On combining classifiers" IEEE Trans, Pattern Anal. Mach. Intell, 20(3), 1998, pp. 226-239.
[13] He, J., Liu, L., Palm, G, "A Discriminative Training Algorithm for VQ-based Speaker Identification" , IEEE Transactions on Speech and Audio Processing, 7(3), 1999, pp. 353-356.
[14] Laurent Besacier and Jean-Francois Bonastr, "Subband architecture for automatic speaker recognition" Signal Processing, 80, 2000, pp. 1245-1259.
[15] Zheng F., Zhang, G. and Song, Z, "Comparison of different implementations of MFCC" , J. Computer Science & Technology 16(6), 2001, pp. 582-589.
[16] Ganchev, T., Fakotakis, N., and Kokkinakis, G. "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task" Proc. of SPECOM Patras, Greece, 2005, pp. 1191-194.
[17] Zhen B., Wu X., Liu Z., Chi H, "On the use of band pass filtering in speaker recognition", Proc. 6th Int. Conf. of Spoken Lang. Processing (ICSLP), Beijing, China, 2000
[18] S. Singh, Dr. E.G Rajan, P.Sivakumar, M.Bhoopathy and V.Subha, "Text Dependent Speaker Recognition System in Presence Monitoring", International conference Systemic, Cybernatics and

Informatics ICSCI -under the Aegis of Pentagram Research Centre Hyderabad, 2008, pp. 550-554.
[19] A. Papoulis and S. U. Pillai, "Probability, Random variables and Stochastic Processes", Tata McGraw-Hill Edition, Fourth Edition, Chap. 4, 2002, pp. 72-122.
[20] Oppenheim, A.V., Schafer, R.W., Buck, J.R, "Discrete-Time Signal Processing", 2nd ed., Upper Saddle River,NJ, Prentice Hall, 1999
[21] Yegnanarayana B., Prasanna S.R.M., Zachariah J.M. and Gupta C. S, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system", IEEE Trans. Speech and Audio Processing, Vol. 13, No. 4, 2005, pp. 575-582.
[22] S.R. Mahadeva Prasanna, Cheedella S. Gupta, B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech", Speech Communication, 48(10), 2006, pp. 1243- 1261.
[23] H. S. Jayanna and S. R. M. Prasanna, "Fuzzy vector quantization for speaker recognition under limited data conditions" TENCON- IEEE Region 10 Conference ,2008, pp. 1 - 4.
[24] Haipeng Wang, Xiang Zhang, Hongbin Suo, Qingwei Zhao and Y. Yan, "A novel fuzzy-based automatic speaker clustering algorithm," ISNN, 2009, pp. 639–646.
[25] H. Melin and J. Lindberg, "Guidelines for experiments on the polycost database", In Proceedings of a COST 250 workshop on Application of Speaker Recognition Techniques in Telephony, 1996, pp. 59- 69, Vigo, Spain.