International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:3, No:11, 2009

# Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter

Sandipan Chakroborty* and Goutam Saha

***Abstract*—**A state of the art Speaker Identification (SI) system requires a robust feature extraction unit followed by a speaker modeling scheme for generalized representation of these features. 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. On a recent contribution by authors, it has been shown that the Inverted Mel-Frequency Cepstral Coefficients (IMFCC) is useful feature set for SI, which contains complementary information present in high frequency region. This paper introduces the Gaussian shaped filter (GF) while calculating MFCC and IMFCC in place of typical triangular shaped bins. The objective is to introduce a higher amount of correlation between subband outputs. The performances of both MFCC & IMFCC improve with GF over conventional triangular filter (TF) based implementation, individually as well as in combination. With GMM as speaker modeling paradigm, the performances of proposed GF based MFCC and IMFCC in individual and fused mode have been verified in two standard databases YOHO, (Microphone Speech) and POLYCOST (Telephone Speech) each of which has more than 130 speakers.

***Keywords*—**Gaussian Filter, Triangular Filter, Subbands, Correlation, MFCC, IMFCC, GMM.

## I. Introduction

ANY speaker identification [1],[2] system consists of a speaker specific feature extractor as a front-end module followed by a robust speaker modeling technique for generalized representation of extracted features. MFCC [3], [4] is considered as a reliable front-end for a typical SI application (Fig. 1) as it can describes the vocal tract characteristics and easy to extract. An illustrative SI system is shown in fig. 1.

The state of the art speaker recognition research primarily investigates speaker specific complementary information relative to MFCC. The speaker identification performance improves significantly when this complementary information is fused with MFCC in feature level either by simple concatenation or by combining models' scores. The examples of such complementary information are pitch [5], residual phase [6], prosody [7], dialectical features [8] etc. However, these features are related with vocal chord vibration from which faithful extraction of speaker specific information are quite difficult. In a recently contributed article [7] by present authors, it has been shown that complementary information [9] can be captured easily from the high frequency part of the energy spectrum of a speech frame via reversed filter bank methodology. The work has proposed a new feature set called IMFCC to capture speaker specific information lying in higher frequency part of the spectrum and is usually ignored by MFCC. The complementary information captured by IMFCC is modeled by standard Gaussian Mixture Modeling (GMM) [10] technique. It is seen that when fused with MFCC based speaker model, the performance of a SI system outperforms baseline MFCC significantly.

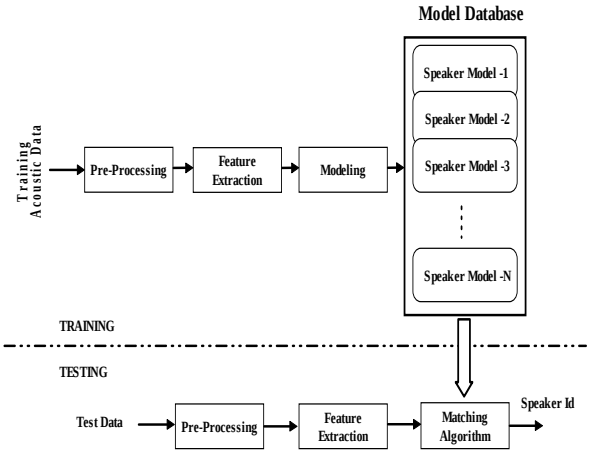


Fig. 1 A typical Speaker Identification System

Nevertheless, both MFCC and IMFCC use TF to get subband outputs from energy spectrum. A triangular filter (TF) provides crisp partitions in an energy spectrum by providing non-zero weights to the portion covered by it while giving zero weight outside it. The phenomena cause loss of correlations between a subband output and the adjacent spectral components that are present in the other

*S. Chakroborty is with the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, Kharagpur-721 302, Kharagpur, India; Phone: +91-3222-281470; fax: +91-3222-255303; (email: mail2sandi@gmail.com).

G. Saha is with the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, Kharagpur-721 302, Kharagpur, India; Phone: +91-3222-281470; fax: +91-3222-255303; (email:gsaha.iitkgp@gmail.com ).

subbands. Note that, in [11], a parallel redundant architecture was proposed that could keep most of the correlation between subbands for improving SI performance. Lippmann [12] showed that the redundancy between subbands might be a source of human robustness to speech degradation. However, no attempts have been yet made to extract features to introduce correlation in a systematic way that is missed by TFs for their crisp division from one subband to other.

In this work, we use Gaussian Shaped filters (GF) as the averaging bins instead of triangular for calculating MFCC as well as IMFCC in a typical SI application. The motivation of using GF is threefold. First, Gaussian shaped filters can provide much smoother transition from one subband to other preserving most of the correlation between them. Second, the means and variances of these GFs can be independently chosen in order to have control over the amount of overlap with neighboring subbands. Third, the filter design parameters for GF can be calculated very easily from mid as well as end-points located at the base of the original TF used for MFCC and IMFCC. Both MFCC and IMFCC filter bank [13], [14] are realized using a moderate variance where a GF's coverage for a subband and the correlation is to be balanced.

Results show that GF based MFCC and IMFCC perform better than the conventional TF based MFCC and IMFCC individually. Results are also better when GF based MFCC & IMFCC is fused by their model scores in comparison to the results that are obtained by combining MFCC and IMFCC feature sets realized using traditional TF. All the implementations have been done with GMM as speaker modeling paradigm with different model orders in two standard databases (YOHO [15], Microphone speech & POLYCOST [16], telephone speech), each containing more than 130 speakers.

The rest of the paper is organized as follows: Section II briefly reviews the concept of MFCC, its implementation using proposed GF. Next, IMFCC and GF based IMFCC are discussed in Section III. Section IV outlines the GMM based speaker model used for SI task. Section V explains the scheme for the fusion of classifiers. Section VI reports the experimental results. Finally, Section VII draws the principal conclusions of the paper.

## II. MEL FREQUENCY CEPSTRAL COEFFICIENTS AND THEIR CALCULATION BY GAUSSIAN FILTERS

### *A. Mel-Frequency Cepstral Coefficients using triangular filters*

According to psychophysical studies [17], human perception of the frequency content of sounds follows a subjectively defined nonlinear scale called the Mel scale [18] (Fig. 5, Solid curve). This 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 to $f$, the actual frequency in Hz. This leads to the definition of MFCC, a baseline acoustic feature [19] for Speech and Speaker Recognition applications, which can be calculated as follows.

Let $\{y(n)\}_{n=1}^{N_s}$ represent a frame of speech that is pre-emphasized and Hamming-windowed. First, $y(n)$ is converted to the frequency domain by an $M_s$-point DFT which leads to the energy spectrum,

$$|Y(k)|^2 = \left|\sum_{n=1}^{M_s} y(n)\cdot e^{\left(-\frac{j2\pi nk}{M_s}\right)}\right|^2 \quad (2)$$

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 $\Psi_i(k)$ of the $i$th filter in the bank (fig. 2) is defined as,

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

where $1 \le i \le Q$, $Q$ is the number of filters in the bank, $\{k_{b_i}\}_{i=0}^{Q+1}$ are the boundary points of the filters and $k$ denotes the coefficient index in the $M_s$-point DFT. The filter bank boundary points, $\{k_{b_i}\}_{i=0}^{Q+1}$ are equally spaced in the Mel scale which is satisfied by the definition,

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

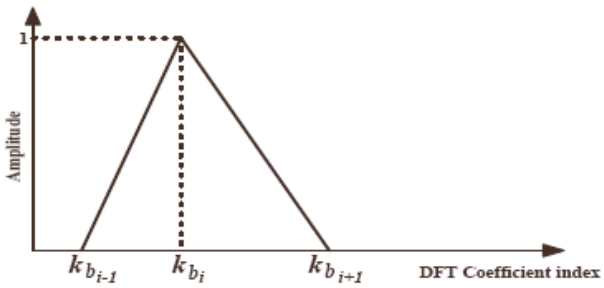


Fig. 2 Response Ψi(k) of a typical Mel scale filter defined as in eqn. 3

where the function $f_{mel}(\bullet)$ is defined in eqn. 1, $M_s$ is the number of points in the DFT (eqn. 2), $Fs$ 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,

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

The sampling frequency $F_s$ and the frequencies $f_{low}$ and $f_{high}$ are in Hz while $f_{mel}$ is in Mels. For both the databases considered in this work, Fs are 8 kHz. $M_s$ was taken as 256, $f_{low} = F_s/M_s = 31.25$ 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 $\Psi_i(k)$ and the energy spectrum $|Y(k)|^2$ as

$$e(i) = \sum_{k=1}^{\frac{Ms}{2}} |Y(k)|^2 \cdot \Psi_i(k) \quad (6)$$

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

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

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

### B. Mel-Frequency Cepstral Coefficients using Gaussian filters

A TF is asymmetric, tapered but 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. We propose here to use of GF, which is symmetric and provides gradually decaying weights at its both ends for compensating possible loss of correlation. Referring to eqn. 3, the expression for GF can be written as

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

where, $k_{b_i}$ is a point between the $i$th TF's boundaries located at its base and considered as the mean of $i$th GF while the $\sigma_i$ is the standard deviation or square root variance of and can be defined as,

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

where, $\alpha$ is the parameter that controls the variance. However, in the eqn. 8, the conventional denominator i.e. $\sqrt{(2\pi)\sigma_i}$ is dropped, as its presence is only to ensure the area under a Gaussian curve [20] is unity. 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 non-isosceles while they are mapped into normal frequency domain (Ref. eqn 5) for which the two distances calculated from its two ends to $k_{b_i}$ in base become unequal. For MFCCs' $i$th filter, the relation becomes,

$$\left(k_{b_{i+1}} - k_{b_i}\right) > \left(k_{bi} - k_{b_{i-1}}\right) \quad (10)$$

We took the maximum spread out of these two distances i.e. $\left(k_{b_{i+1}} - k_{b_i}\right)$ to evaluate $\sigma_i$ ensuring full coverage of the subband by the GF.

In Fig.3, we plot TF and GF for different $\sigma$ values. 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.

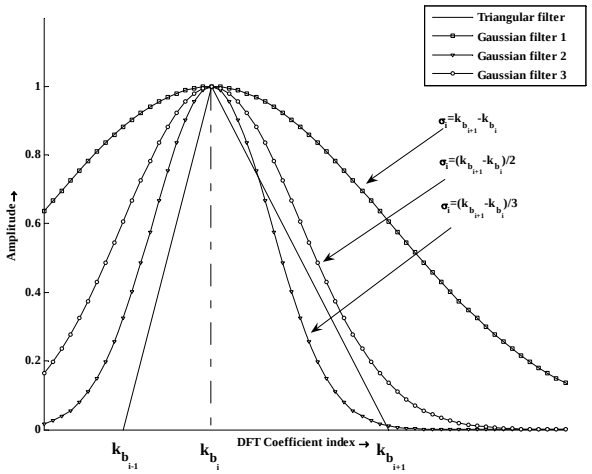


Fig. 3 Response of various shaped filters

All three GFs are centered around $k_{b_i}$ and offer gradual decaying weights to the portion of the spectrum away from center. However, the Gaussian with higher variance (Gaussian Filter 1 in fig. 3) shows larger correlation with adjacent frequency components. Thus the choice of $\alpha$ is crucial for setting the variances of GF. If $\alpha$ =3 then by eqn. 9 we can write,

$$k_{b_{i+1}} - k_{b_i} = 3\sigma \quad (10)$$

which in turn signifies that Pr [$(k_{b_{i+1}} - k_{b_i}) \leq 3\sigma_i$ ] = 0.997, where Pr[•] denotes the prior probability [17] of an event. The above relation depicts 99.7% coverage of a GF for a particular subband. Similarly, for $\alpha$ =2, a GF guarantees 95% coverage within a subband, since Pr[$(k_{b_{i+1}} - k_{b_i}) \leq 2\sigma_i$ ] = 0.95. Therefore, $\alpha$ =2 can provide better correlation with adjacent subbands in comparison to $\alpha$ =3 which sets a Gaussian window to deliver only 0.3% of its total weights to the frequencies other than its own subband. One could have chosen $\alpha$ =1 for which the variance will be, too high with 68% (because Pr [$(k_{b_{i+1}} - k_{b_i}) \leq \sigma_i$ ] = 0.68) of a filter's total weights to the subband of interest. Therefore, a trade-off occurs between the coverage by A GF for a particular subband and the portion lying outside of it. We have chosen $\alpha$ =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. Figure 4. shows the structure of MFCC filter bank using triangular and Gaussian bins. The cepstral vector using GFs can be calculated from the filter's response (eqn. 8) which is as follows;

$$e^{g_{MFCC}}(i)=\sum_{k=1}^{\frac{Ms}{2}}\left|Y(k)\right|^2\cdot\Psi^{g_{MFCC}}{}_i(k) \tag{11}$$

and,

$$C^{g_{MFCC}}{}_m=\sqrt{\frac{2}{Q}}\sum_{l=0}^{Q-1}log\left[e^{g_{MFCC}}(i+1)\right]\cdot cos\left[m\cdot\left(\frac{2l-1}{2}\right)\cdot\frac{\pi}{Q}\right] \tag{12}$$

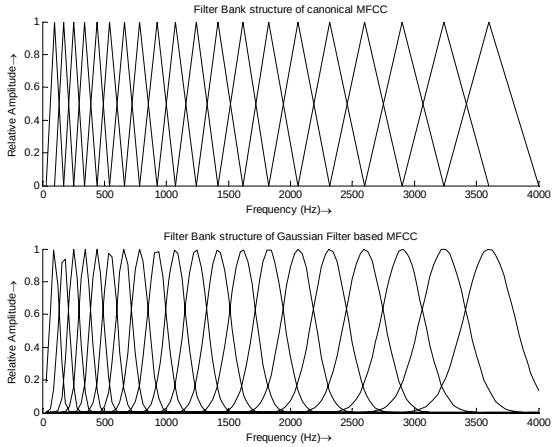


Fig. 4 Filter Bank Structure for canonical MFCC and Gaussian Filter based MFCC

Typically, $Q$ = 20 and 10 to 30 cepstral coefficients are taken for speech processing applications. Here we took $Q$ = 20, $R$ = 20 and used the last 19 coefficients from both normal MFCC and GF based MFCC to model the individual speakers.

## III. INVERTED MEL FREQUENCY CEPSTRAL COEFFICIENTS CALCULATION BY GAUSSIAN FILTERS

### *A. Inverted Mel-Frequency Cepstral Coefficients using triangular filters*

The Inverted Mel Scale (fig. 5, dotted line) proposed by present authors [7] is defined by a competing filter bank structure which is indicative of a hypothetical auditory system which has followed a diametrically opposite path of evolution than the human auditory system. The idea is to capture those information which otherwise could have been missed by original MFCC.

We obtain the new filter bank structure 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 SI applications, i.e. (0 to 4 kHz (sec. II)). This flip-over is expressed mathematically as,

$$\widehat{\Psi}_i(k)=\Psi_{Q+1-i}\left(\frac{M_s}{2}+1-k\right) \tag{13}$$

where $\widehat{\Psi}_i(k)$ is the Inverted Mel Scale filter response while $\Psi_i(k)$ is the response of the original MFCC filter bank, $1\leq i\leq Q$ and $Q$ is the number of filters in the bank. Analogous to eqn. 3 for the original MFCC filter bank, we can derive an expression for $\widehat{\Psi}_{i(k)\,i=1}{}^{Q}$ from eqn. (13) as follows,

$$\widehat{\Psi}_i(k)=\begin{cases}0 & for\quad k\leq\hat{k}_{b_{i-1}}\\ \dfrac{k-\hat{k}_{b_{i-1}}}{k_{b_i}-\hat{k}_{b_{i-1}}} & for\ \hat{k}_{b_{i-1}}\ \leq\ k\leq\hat{k}_{b_i}\\ \dfrac{\hat{k}_{b_{i+1}}-k}{\hat{k}_{b_{i+1}}-\hat{k}_{b_i}} & for\ \hat{k}_{b_i}\ \leq\ k\leq\hat{k}_{b_{i+1}}\\ 0 & for\ k\ \geq\ \hat{k}_{b_{i+1}}\end{cases} \tag{14}$$

where $1\leq k\leq M_s$ and $\left\{\hat{k}_{b_i}\right\}_{i=0}^{Q+1}$

The inverted mel-scale is then defined by the following expression.

$$\hat{f}_{mel}(f)=2195.2860-2595\,log_{10}\left(1+\frac{4031.25-f}{700}\right) \tag{15}$$

Note that, detailed calculations for deriving the inverted mel scale can be found in [7];

where $\hat{f}_{mel}(f)$ is the subjective pitch in the new scale corresponding to $f$, the actual frequency in Hz.

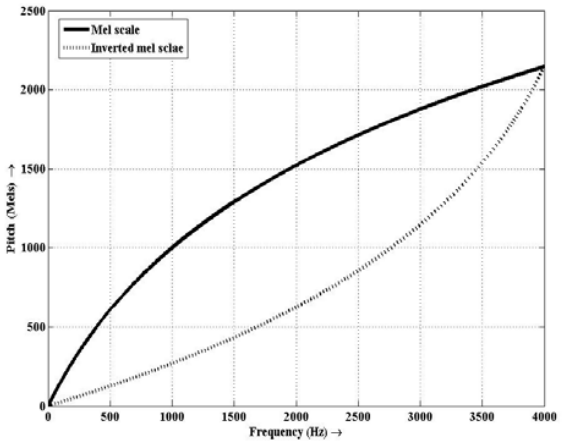


Fig. 5 Subjective Pitch vs Frequency. For Mel scale, corresponding to the human auditory system, pitch increases progressively less rapidly as the frequency increases, In direct contrast, it increases progressively more rapidly in the proposed Inverted Mel Scale

We find the filter outputs $\left\{\hat{e}(i)\right\}_{i=1}^{Q}$ in the same way as MFCC from the same energy spectrum $|Y(k)|^2$ as,

$$\hat{e}(i)=\sum_{k=1}^{\frac{Ms}{2}}\left|Y(k)\right|^2\cdot\widehat{\Psi}_i(k) \tag{16}$$

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

$$\hat{C}_m=\sqrt{\frac{2}{Q}}\sum_{l=0}^{Q-1}log\left[\hat{e}(i+1)\right]\cdot cos\left[m\cdot\left(\frac{2l-1}{2}\right)\cdot\frac{\pi}{Q}\right] \tag{17}$$

### *B. Inverted Mel-Frequency Cepstral Coefficients using Gaussian filters*

In this, work IMFCC filter bank is also realized using Gaussian bin. 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

$$\left(\hat{k}_{bi}-\hat{k}_{b_{i-1}}\right)>\left(\hat{k}_{b_{i+1}}-\hat{k}_{b_i}\right) \tag{18}$$

for inverted mel scale. Here, $\hat{k}_{b_i}$ is taken as mean of *i*th GF while standard deviation $\widehat{\sigma}_i$ can be estimated by the following relation;

$$\widehat{\sigma}_i=\frac{\hat{k}_{b_i}-\hat{k}_{b_{i-1}}}{\alpha} \tag{19}$$

Note that $\alpha$ is chosen also here as 2 for the same reason mentioned in Sec. II B. Therefore, response of the GF for IMFCC filter bank and corresponding cepstral parameters can be calculated as follows;

$$\widehat{\Psi}_i^{g_{IMFCC}}=e^{-\frac{(k-\hat{k}_{b_i})^2}{2\hat{\sigma}_i^2}} \tag{20}$$

$$\hat{e}^{g_{IMFCC}}(i)=\sum_{k=1}^{\frac{Ms}{2}}\left|Y(k)\right|^2\cdot\widehat{\Psi}_i^{g_{IMFCC}}(k) \tag{21}$$

and finally,

$$\hat{C}_m^{g_{IMFCC}}=\sqrt{\frac{2}{Q}}\sum_{l=0}^{Q-1}log\left[\hat{e}^{g_{IMFCC}}(i+1)\right]\cdot cos\left[m\cdot\left(\frac{2l-1}{2}\right)\cdot\frac{\pi}{Q}\right] \tag{22}$$

As with MFCC, we took *Q = 20, R = 20* and used the last 19 coefficients to model the individual speakers. Figure 5 shows the typical structure of IMFCC realized using triangular and Gaussian window and figure 6 describes extraction of the feature sets.

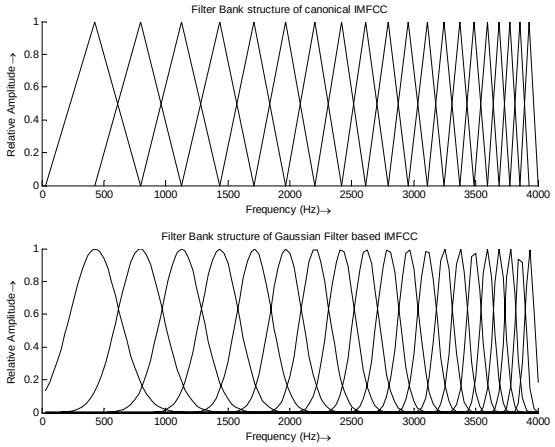


Fig. 6 Filter Bank Structure for canonical IMFCC and Gaussian Filter based IMFCC

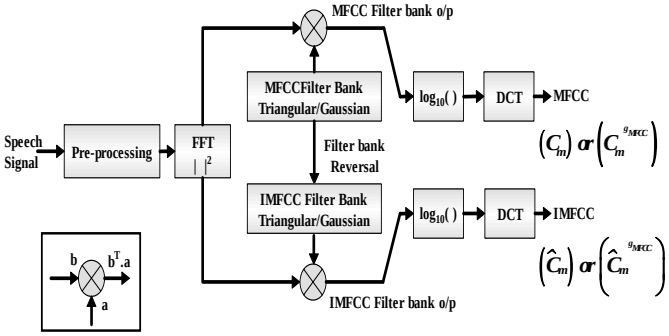


Fig. 7 Plot showing extraction of Triangular /Gaussian MFCC and Triangular/Gaussian IMFCC features

## IV. Theoretical Background on Gaussian Mixture Models (GMM)

A GMM [10] can be viewed as a non-parametric, multivariate probability distribution model that is capable of modeling arbitrary distributions and is currently one of the principal methods of modeling speakers for SI systems. The GMM of the distribution of feature vectors for speaker s is a weighted linear combination of *M* unimodal Gaussian densities $b_i^s(\boldsymbol{x})$, each parameterized by a mean vectors $\mu_i^s$ with a diagonal covariance matrix $\sum_i^s$. These parameters, which collectively constitute the speaker model, are represented by the notation $\left\{p_i^s,\mu_i^s\Sigma_i^s\right\}_{i=1}^{M}$. The $p_i^s$ are the mixture weights satisfying stochastic constraint $\sum_{i=1}^{M}p_i^s=1$.

For a feature vector $\boldsymbol{x}$ the mixture density for a speaker *s* is computed as

$$p\left(x\mid\lambda_s\right)=\sum_{i=1}^{M}p_i^s b_i^s\left(x\right) \tag{23}$$

where,

$$b_i^s(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i^s|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(x-\mu_i^s)^t (\Sigma_i^s)^{-1}(x-\mu_i^s)\right)} \quad (24)$$

and $D$ is the dimension of the feature-space.

Given a sequence of feature vectors $X=\{x_1, x_2 \ldots x_T\}$ for an utterance with $T$ frames, the log-likelihood of a speaker model $s$ is

$$L_s(X) = log\ p(X \mid \lambda_s) = \sum_{t=1}^{T} log\ p(x_t \mid \lambda_s) \quad (25)$$

assuming the vectors to be independent for computational simplicity. For SI, the value of $L_s(X)$ is computed for all speaker models $\lambda_s$ enrolled in the system and the owner of the model that generates the highest value is returned as the identified speaker. During training, feature vectors collected from a speaker's utterances are trained using the Expectation and Maximization (E & M) algorithm. This technique involves an iterative update of each of the parameters in $\lambda$, with a consequent increase in the log-likelihood at each step. Usually, within a few iterations (10 to 25) the model parameters converge to stable values. In the present work, initialization of seed vectors for Gaussian centers was done by the split Vector Quantization [21] algorithm This was followed by the E & M algorithm with 10 iterations. For all cases, diagonal covariance matrices were chosen because DCT has already uncorrelated the features (eqn. 7, 12, 17, and 22) in respective cases.

## V. Fusion of Speaker Models

Combining classifier [9] decisions to improve decision reliability has been successful in many pattern classification [6], [9], [19], [22] problems including SI. 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 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 [6-8] [23].

During the training phase, two separate models were developed for each speaker from the MFCC and IMFCC feature sets respectively, using GMM technique (Sec. IV). During the test phase, MFCC and IMFCC features were extracted in a similar way from an incoming speech utterance as done in the training phase and were sent to their respective models. For each speaker, two scores were generated, one each from the MFCC and IMFCC models. Since sum rule outperforms other combination strategies due to its lesser sensitivity to estimation errors, an uniform weighted sum rule was adopted to fuse the scores from the two classifiers.

Further, since in each case we fused the scores of two classifiers of the same type (GMM-GMM), no score adaptation or normalization was necessary before combination.

If $S^i_{MFCC}$ (or $S^{ig}_{MFCC}$) and $S^i_{IMFCC}$ (or $S^{ig}_{IMFCC}$) are the scores generated by the two models for the $i$th speaker then the combined score $S^i_{com}$ (or $S^{ig}_{com}$) is expressed as

$$S^i_{com} = wS^i_{MFCC} + (1-w) S^i_{IMFCC} \quad (26\ a)$$

Or, $$S^{i_g}_{com} = wS^{i_g}_{MFCC} + (1-w) S^{i_g}_{IMFCC} \quad (26\ b)$$

A governing equation is given below which describes fusing outputs of parallel classifiers methodology via weighted sum rule.

$$S^i_{com} = w\sum_{t=1}^{T} log\ p(x_{tMFCC} \mid \lambda_{sMFCC}) + (1-w)\sum_{t=1}^{T} log\ p(x_{tIMFCC} \mid \lambda_{sIMFCC}) \quad (27\ a)$$

Or,

$$S^{i_g}_{com} = w\sum_{t=1}^{T} log\ p(x^g_{tMFCC} \mid \lambda^g_{sMFCC}) + (1-w)\sum_{t=1}^{T} log\ p(x^g_{tIMFCC} \mid \lambda^g_{sIMFCC}) \quad (27\ b)$$

All the notations have their usual meanings. We have used $w$ = 0.5 as the weight for all combinations. However, other weighting schemes that are more suitable can be investigated further to enhance the performance of the combined system. Finally, the identity of the true speaker $i_{true}$ is given by-

$$i_{true} = arg\ max_i\ S^i_{com} \quad (28\ a)$$

Or, $$i^g{}_{true} = arg\ max_i\ S^{i_g}_{com} \quad (28\ b)$$

Note that, for eqns. (26b, 27b, and 28b) have superscripted **'g'** symbol, which signify GF based MFCC and IMFCC filter bank.

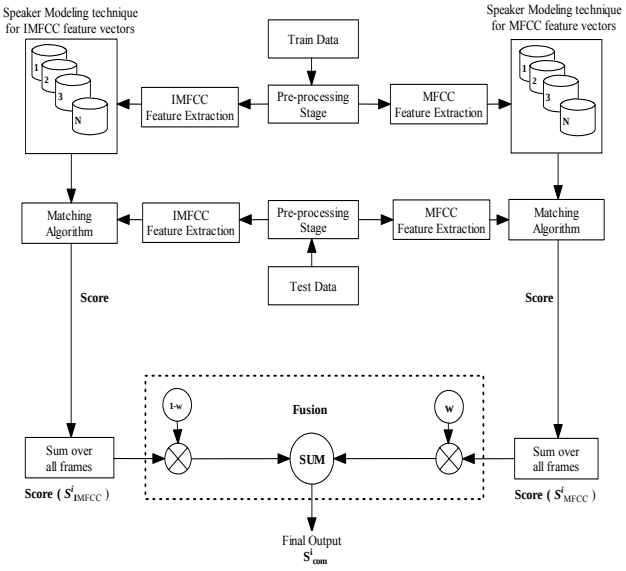


Fig. 8 Parallel classifier based SI system

A schematic description of this scheme for parallel combination of classifiers is given in fig. 8.

## VI. Experimental Evaluation

### A. Pre-processing stage

In this work, each frame of speech is pre-processed by i) silence removal and end-point detection using an energy threshold criterion, followed by ii) pre-emphasis with 0.97 pre-emphasis factor, iii) frame blocking with 20ms frame length, i.e $N_s$ = 160 samples/frame (Sec. II) & 50 overlap, and finally iv) Hamming-windowing. Next, the MFCC and IMFCC feature sets using both triangular and GFs are calculated (ref. Sec II & III). The first coefficient ($C_0$, $C^g{}_0$, $\widehat{C}_0$, and $\widehat{C}^g{}_0$) is discarded since it contains only the energy of the spectrum and the resulting 19 dimensional vector is used.

### B. Databases for experiments

#### a) YOHO Database

The YOHO [1], [12] voice verification corpus was collected while testing ITT's prototype speaker verification system in an office environment. Most subjects were from the New York City area, although there were many exceptions, including some nonnative English speakers. A high-quality telephone handset (Shure XTH-383) was used to collect the speech; however, the speech was not passed through a telephone channel. There are 138 speakers (106 males and 32 females); for each speaker, there are 4 enrollment sessions of 24 utterances each and 10 test sessions of 4 utterances each. In this work, a closed set text-independent speaker identification problem is attempted where we consider all 138 speakers as client speakers. For a speaker, all the 96 (4 X 24 utterances) utterances are used for developing the speaker model while for testing, 40 (10 sessions X 4 utterances) utterances are put under test. Therefore, for 138 speakers we put 138 X 40 =5520 utterances under test and evaluated the identification accuracies.

#### b) POLYCOST Database

The POLYCOST database [20] was recorded as a common initiative within the COST 250 action during January- March 1996. It contains around 10 sessions recorded by 134 subjects from 14 countries. Each session consists of 14 items, two of which (MOT01 & MOT02 files) contain speech in the subject's mother tongue. 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 [24] 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. Unlike YOHO database all the speakers do not have the same number of sessions. Further, three speakers (M042, M045 & F035) are not included in our experiments as they provide sessions which are lower than 4. A total 754 'MOT01' utterances are put under test. As with YOHO database, all speakers (131 after deletion of three speakers) in the database were registered as clients.

### C. Score Calculation

For any closed-set speaker identification problem, identification accuracy is defined as follows in [10] and we have used the same:

$$\text{Percentage of identification accuracy (PIA)} = \frac{\text{No. of utterances correctly identified}}{\text{Total no. of utterances under test}} \quad (29)$$

### D. 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 as well as GF and a parallel classifier fusing both models.

#### 1) Results for YOHO Database

Table I describes identification results for various model orders of GMM 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. Note that, identification accuracies increase with increase in model order.

Table II 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. Here also, PIA increases with increase in model order.

TABLE I
Results (PIA) for YOHO Database using TF based MFCC & IMFCC

| No. of Mixtures | MFCC | IMFCC | **Combined System** |
|---|---|---|---|
| 16 | 94.2029 | 94.1486 | **96.2500** |
| 32 | 95.6703 | 95.2174 | **97.2645** |

TABLE II
Results (PEA) for YOHO Database using GF based MFCC & IMFC**C**

| No. of Mixtures | MFCC | IMFCC | **Combined System** |
|---|---|---|---|
| 16 | 95.4891 | 94.2572 | **96.5036** |
| 32 | 96.8279 | 95.2355 | **97.4275** |

*2) Results for POLYCOST*

Table III & IV show the identification accuracies for the POLYCOST database for TF and GF .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 YOHO database, it can be observed from these tables that combined scheme shows significant improvement over the baseline MFCC based system irrespective of the filter type. In addition, results improve as model order increases. We restrained ourselves to 2 different sized mixtures for GMM. This is because less number of feature vectors is obtained from the POLYCOST database that prevents development of meaningful higher order GMMs.

TABLE III
RESULTS (PIA) FOR POLYCOST DATABASE USING TF BASED MFCC & IMFCC

| No. of Mixtures | MFCC | IMFCC | **Combined System** |
|---|---|---|---|
| 8 | 77.8515 | 76.2599 | **81.0345** |
| 16 | 77.8515 | 77.0557 | **81.1631** |

TABLE IV
RESULTS (PIA) FOR POLYCOST DATABASE USING GF BASED MFCC & IMFCC

| No. of Mixtures | MFCC | IMFCC | **Combined System** |
|---|---|---|---|
| 8 | 78.6472 | 76.2599 | **82.0955** |
| 16 | 80.9019 | 77.5862 | **82.7586** |

It is observed that the independent performance of IMFCC is not as good as MFCC for POLYCOST database as compared to YOHO. This is because the data in POLYCOST is based on telephone speech where higher frequency information used by IMFCC are somewhat distorted. Nevertheless, results show that the complementary information supplied by it 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 based IMFCC..

## VII. CONCLUSION

Gaussian filter based mel and inverted mel scaled filter bank is proposed in this paper. 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 GMM. 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 two speech databases YOHO, microphone speech, & POLYCOST, telephone speech, each of which contains more than 130 speakers. The increment of speaker identification accuracy in combined method using GF over TF based feature sets also suggest the possible enhancement of higher frequency complementary information relative to mel-scaled filters. The performance can further be improved by proper choice of mixing proportion of two streams in combined model.

## REFERENCES

[1] J. P. Cambell, Jr., "Speaker Recognition:A Tutorial", *Proceedings of The IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept. 1997.
[2] Faundez-Zanuy M. and Monte-Moreno E., "State-of-the-art in speaker recognition", *Aerospace and Electronic Systems Magazine, IEEE*, vol. 20, No. 5, pp. 7-12, Mar. 2005
[3] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. On ASSP*, vol. ASSP 28, no. 4, pp. 357-365, Aug. 1980.
[4] R. Vergin, B. O' Shaughnessy and A. Farhat, "Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-idenpendent continuous-speech recognition, *IEEE Trans. On ASSP*, vol. 7, no. 5, pp. 525-532, Sept. 1999.
[5] Harrag A. Mohamadi T., Serignat J.F., "LDA Combination of Pitch and MFCC Features in Speaker Recognition", *Proceedings of INDICON 2005*, pp. 237-240, 11-13 Dec., IIT Chennai, India, 2005.
[6] K. Sri Rama Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition", *IEEE Signal Processing Letters*, vol 13, no. 1, pp. 52-55, Jan. 2006.
[7] 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, pp. 575-582, July 2005.
[8] Chakroborty, S., Roy, A. and Saha, G., "Improved Closed set Text-Independent Speaker Identification by Combining MFCC with Evidence from Flipped Filter Banks". *International Journal of Signal Processing*, Vol. 4, No. 2, Page(s):114-122, 2007.
[9] J. Kittler, M. Hatef, R. Duin, J. Mataz, "On combining classifiers", *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (1998) 226-239.
[10] D. Reynolds, R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models", *IEEE Trans. Speech Audio Process.*, vol. 3, no.1, pp. 72-83, Jan. 1995.
[11] Laurent Besacier and Jean-Francois Bonastre, "Subband architechute for automatic speaker recognition", *Signal Processing*, vol-80, pp. 1245-1259, 2000.
[12] R. P. Lippmann, ``Speech recognition by machines and humans", *Speech Communication*, vol. 22, No. 1, pp. 1-15, 1997.
[13] Zheng F., Zhang, G. and Song, Z., "Comparison of different implementations of MFCC", *J. Computer Science & Technology, vol.* 16 no. 6, pp. 582-589, Sept. 2001.
[14] Ganchev, T., Fakotakis, N., and Kokkinakis, G. "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task", *Proc. of SPECOM 2005*, October 17-19, 2005. Patras, Greece, vol. 1, pp.191-194.
[15] J. Campbell, "Testing with the YOHO CDROM voice verification corpus", *ICASSP95*, 1995, vol.1 pp. 341-344.
[16] Petrovska, D., et al. "POLYCOST: A Telephone-Speech Database for Speaker Recognition", *RLA2C*, Avignon, France, April 20-23, 1998, pp. 211-214.
[17] D. O' Shaughnessy, *Speech Communication Human and Machine*, Addison-Wesley, New York, 1987.
[18] Ben Gold and Nelson Morgan, *Speech and Audio Signal Processing*, Part- IV, Chap.14, pp. 189-203, John Willy & Sons ,2002.
[19] Daniel J. Mashao, Marshalleno Skosan, "Combining Classifier Decisions for Robust Speaker Identification", *Pattern Recog,*, vol. 39, pp. 147-155, 2006.

[20] A. Papoulis and S. U. Pillai, "Probability, Random variables and Stochastic Processes", Tata McGraw-Hill Edition, Fourth Edition, Chap. 4, pp. 72-122, 2002.
[21] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design", IEEE Trans. Commun., vol. 28, no. 1, pp. 84-95, 1980.
[22] Daniel Garcia-Romero, Julian Fierrez-Aguilar, Joaquin Gonzalez-Rodriguez, Javier Ortega-Garcia, "Using quality measures for multilevel speaker recognition", Computer Speech and Language, Vol. 20, Issue 2-3, pp. 192-209, Apr. 2006,
[23] S.R. Mahadeva Prasanna, Cheedella S. Gupta b, B. Yegnanarayana, Extraction of speaker-specific excitation information from linear prediction residual of speech", Speech Communication, Vol. 48, Issue 10, pp. 1243- 1261, October 2006.
[24] 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*, pp. 59- 69, Vigo, Spain, November 1996.

**Sandipan Chakroborty** passed B.E in Electronics from Nagpur University, India in 2001 and passed Masters of Engineering (M.E) having specialization in Digital System and Instrumentation with highest honours from Bengal Engineering and Science University, Shibpur, Howrah, India in 2003. Presently he is a seniour research scholar in the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India. His current area of research includes pattern recognition, neural networks, speaker recognition and data fusion strategies. He is also a student Member of IEEE.

**Goutam Saha** graduated in 1990 from Dept. of Electronics & Electrical Communication Engineering, Indian Institute of Technology (IIT), Kharagpur, India. The author worked in Tata Steel, India in the period 1990-1994, joined IIT Kharagpur as CSIR research Fellow in 1994 and completed Ph. D work in 1999. He worked in Institute of Engineering & Management, Salt Lake, Kolkata as a faculty member during 1999-2002 and since 2002 serving IIT Kharagpur as Assistant Professor till date. An active researcher in the field of speech processing, biomedical signal processing, modeling and prediction he has published papers in reputed journals like Physical Review E, IEEE Trans. on Systems, Man & Cybernetics, IEEE Trans. on Biomedical Engineering etc.