International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:2, No:11, 2008

Improved Closed Set Text-Independent Speaker Identification by combining MFCC with Evidence from Flipped Filter Banks

Sandipan Chakroborty*, Anindya Roy 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 SI applications. However, due to the structure of its filter bank, it captures vocal tract characteristics more effectively in the lower frequency regions. This paper proposes a new set of features using a complementary filter bank structure which improves distinguishability of speaker specific cues present in the higher frequency zone. Unlike high level features that are difficult to extract, the proposed feature set involves little computational burden during the extraction process. When combined with MFCC via a parallel implementation of speaker models, the proposed feature set outperforms baseline MFCC significantly. This proposition is validated by experiments conducted on two different kinds of public databases namely YOHO (microphone speech) and POLYCOST (telephone speech) with Gaussian Mixture Models (GMM) as a Classifier for various model orders.

***Keywords*—**Complementary Information, Filter Bank, GMM, IMFCC, MFCC, Speaker Identification, Speaker Recognition.

I. Introduction

ANY speaker Identification [1] system needs a robust acoustic feature extraction technique as a front-end block followed by an efficient modeling scheme for generalized representation of these features. MFCC [2], [3] has been widely accepted as such a front-end for a typical SI application as it is less vulnerable to noise perturbation, gives little session variability and is easy to extract. An illustrative SI system is shown in fig. 1.

Manuscript received November 5, 2006. This work was supported in part by the Indian Space Research Organization (ISRO).

*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).

A. Roy 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: anindya_7@yahoo.co.in).

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@ece.iitkgp.ernet.in).

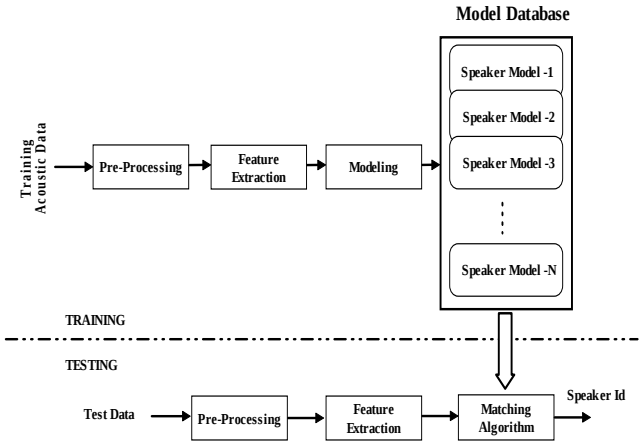

Fig. 1 A typical Speaker Identification System

But MFCC was first proposed for speech recognition [2] to identify monosyllabic words in continuously spoken sentences and not for SI. Also, calculation of MFCC is based on the human auditory system aiming for artificial implementation of the ear physiology [4] assuming that the human ear can be a good speaker recognizer too. However, no conclusive evidence exists to support the view that the ear is necessarily the best speaker recognizer.

Further, computation of MFCC involves averaging the low frequency region of the energy spectrum (approximately demarcated by the upper limit of 1 kHz) by closely spaced overlapping triangular filters while smaller number of less closely spaced filters with similar shape are used to average the high frequency zone. Thus MFCC can represent the low frequency region more accurately than the high frequency region and hence it can capture formants [5] which lie in the low frequency range and which characterize the vocal tract resonances [6]. However, other formants [6] can also lie above 1 kHz and these are not effectively captured by the larger spacing of filters in the higher frequency range.

All these facts suggest that any SI system based on MFCC can possibly be improved. In this work, we extract a new feature set from the speech signal which yields information that is complementary in nature to the human vocal tract characteristics described by MFCC. This makes it very suitable to be used with a parallel classifier [7] to yield higher accuracy in SI problem.

We propose to invert the entire filter bank structure [8], [9] such that the higher frequency range is averaged by more accurately spaced filters and a smaller number of widely spaced filters are used in the lower frequency range. We calculate a new feature set named Inverted Mel Frequency Cepstral Coefficients (IMFCC) following the same procedure as normal MFCC but using this reversed filter bank structure. This effectively captures those high frequency formants ignored by the original MFCC. Further, compared to high level features [10] used in [11]-[13], we will show that little extra computational burden is incurred in the calculation of this complementary feature set, which can be efficiently used if we go for parallel implementation with MFCC.

The importance of MFCC in SI cannot be understated. In order to exploit the best of both paradigms, we model two separate parallel classifiers using these two feature sets namely MFCC and IMFCC and fuse their scores to obtain the final classification decision. Viewed in another manner, we aim to reinforce the score generated by the MFCC based model by another score from a complementary source of information. A GMM [14] based classifier is developed which uses an unsupervised clustering technique to model the speakers. Since the classifiers are totally independent and modeled in parallel, the order of complexity is equal to that for a single MFCC based classifier. It is shown in the Result section that such parallel classifiers perform considerably better in all cases compared to a single classifier based on MFCC.

The rest of the paper is organized as follows: Section II briefly reviews the concept of MFCC. The proposed feature set is presented in Section III. Section IV outlines the GMM technique while 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

According to psychophysical studies, human perception of the frequency content of sounds follows a subjectively defined nonlinear scale called the Mel scale [15] (fig. 1). 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 [7] 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 triangular filters, uniformly spaced in the Mel scale (eqn. 1). The filter response $\Psi_i(k)$ of the ith filter in the bank (fig. 3) 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,

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

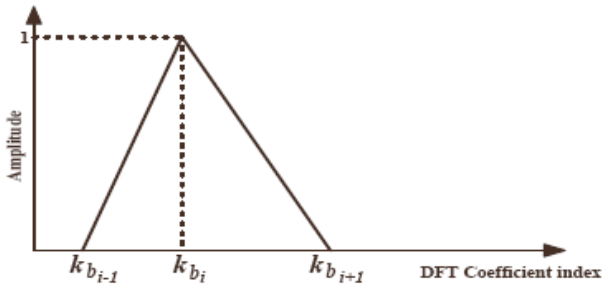

Fig. 2 Response $\Psi_i(k)$ of a typical Mel scale filter defined as in Eqn 3.

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 flow 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. 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 to model the individual speakers.

III. The Inverted Mel Frequency Cepstral Coefficient

Although MFCC presents a way to convert a physically measured spectrum of speech into a perceptually meaningful subjective spectrum based on the human auditory system [4], it is not certain that the human ear and hence MFCC is optimized for SI. Here we propose a new scale, the Inverted Mel Scale (fig. 2) 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,

$$\hat{\Psi}_i(k) = \Psi_{Q+1-i}\left(\frac{M_s}{2} + 1 - k\right) \quad (8)$$

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, $1 \leq i \leq Q$ 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 $\hat{\Psi}_i(k)_{i=1}^{Q}$ from eqn. 8 as follows,

$$\hat{\Psi}_i(k) = \begin{cases} 0 & for \quad k \leq \hat{k}_{b_{i-1}} \\ \dfrac{k - \hat{k}_{b_{i-1}}}{k_{b_i} - \hat{k}_{b_{i-1}}} & for \ \hat{k}_{b_{i-1}} \leq k \leq \hat{k}_{b_i} \\ \dfrac{\hat{k}_{b_{i+1}} - k}{\hat{k}_{b_{i+1}} - \hat{k}_{b_i}} & for \ \hat{k}_{b_i} \leq k \leq \hat{k}_{b_{i+1}} \\ 0 & for \ k \geq \hat{k}_{b_{i+1}} \end{cases} \quad (9)$$

where $1 \leq k \leq M_s$ and $\left\{\hat{k}_{b_i}\right\}_{i=0}^{Q+1}$, the boundary points of the Q filters are defined as,

$$\hat{k}_{b_i} = \left(\frac{Ms}{2}\right) + 1 - k_{b_{Q+1-i}} \quad (10)$$

TABLE I
BOUNDARY POINTS f_{b_i} AND $\hat{f}_{b_i}$ IN Hz FOR THE MFCC and IMFCC FILTER BANKS (with Q =20, flow = 31.25Hz and fhigh = 4kHz)

i	f_{b_i}	$\hat{f}_{b_i}$	i	f_{b_i}	$\hat{f}_{b_i}$
0	31.25	31.25 11	11	1237.9	2957.7
1	98.994	429.75 12	12	1417.4	3108.1
2	173.01	794.46 13	13	1613.5	3245.7
3	253.89	1128.2 14	14	1827.9	3371.7
4	342.26	1433.7 15	15	2062.1	3486.9
5	438.82	1713.3 16	16	2317.9	3592.4
6	544.32	1969.2 17	17	2597.5	3689
7	659.6	2203.4 18	18	2903	3777.4

Now, we can frame an equation analogous to eqn.4, linking $\left\{\hat{k}_{b_i}\right\}_{i=0}^{Q+1}$ to i, f_{low} and f_{high} as,

$$\hat{k}_{b_i} = \left(\frac{M_s}{F_s}\right) \cdot \hat{f}_{mel}^{-1}\left[\hat{f}_{mel}(\hat{f}_{low}) + \frac{i\left\{\hat{f}_{mel}(f_{high}) - \hat{f}_{mel}(f_{low})\right\}}{Q+1}\right] \quad (11)$$

Here, $\hat{f}_{mel}(f)$ is the subjective pitch in the proposed Inverted Mel Scale corresponding to f, the actual frequency in Hz. From eqns. 4, 10 and 11, it follows that,

$$\left(\frac{M_s}{F_s}\right) \cdot \hat{f}_{mel}^{-1}\left[\hat{f}_{mel}(f_{low}) + \frac{i\left\{\hat{f}_{mel}(f_{high}) - \hat{f}_{mel}(f_{low})\right\}}{Q+1}\right]$$
$$= \left(\frac{M_s}{2}\right) + 1 - \left(\frac{M_s}{F_s}\right) \cdot f_{mel}^{-1}\left[f_{mel}(f_{low}) + \frac{i\left\{f_{mel}(f_{high}) - f_{mel}(f_{low})\right\}}{Q+1}\right] \quad (12)$$

$$\hat{f}_{mel}^{-1}\left[f_{mel}(f_{low}) + \frac{i\left\{f_{mel}(f_{high}) - f_{mel}(f_{low})\right\}}{Q+1}\right]$$
$$= \left(\frac{F_s}{2}\right) + \left(\frac{F_s}{M_s}\right) - f_{mel}^{-1}\left[f_{mel}(f_{low}) + \frac{i\left\{f_{mel}(f_{high}) - f_{mel}(f_{low})\right\}}{Q+1}\right] \quad (13)$$

To maintain mathematical uniformity in the calculation of the DFT, we chose the new Inverted Mel Scale to share common boundary points with the actual Mel Scale, i.e., $\hat{f}_{mel}(f_{low}) = f_{mel}(f_{low})$ and $\hat{f}_{mel}(f_{high}) = f_{mel}(f_{high})$

Using this choice, we derive eqns. 12 and 13 from eqn. 11 by suitably choosing the integers Q and i, we can represent any frequency f in the linear (Hertz) scale as,

$$f = \hat{f}_{mel}^{-1}\left[f_{mel}(f_{low}) + \frac{i\{ f_{mel}(f_{high}) - f_{mel}(f_{low}) \}}{Q+1} \right] \quad (14)$$

From eqn. 13 it follows that,

$$f = \left(\frac{F_s}{2}\right) + \left(\frac{F_s}{M_s}\right) - f_{mel}^{-1}\left[f_{mel}(f_{high}) + f_{mel}(f_{low}) - f_{mel}(f) \right] \quad (15)$$

Finally, we obtain the equation,

$$f_{mel}(f) = f_{mel}(f_{high}) + f_{mel}(f_{low}) - f_{mel}\left[\frac{F_s}{2} + \frac{F_s}{M_s} - f \right] \quad (16)$$

which relates the proposed Inverted Mel Scale to the original Mel Scale [2]. For the current application, we have set (sec. II) F_s = 8 kHz, M_s = 256, $f_{low} = F_s/M_s$ = 31.25Hz and f_{high} = Fs/2 =4 kHz. Hence, using these values in eqn. 15, we define the proposed Inverted Mel Scale as,

$$\hat{f}_{mel}(f) = 2195.2860 - 2595 \log_{10}\left(1 + \frac{4031.25 - f}{700}\right) \quad (17)$$

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

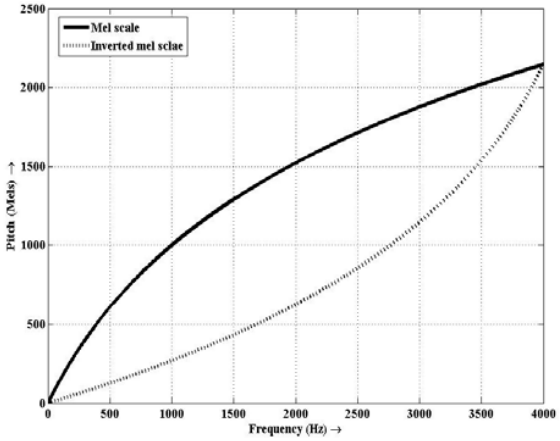

Fig. 3 Subjective Pitch vs Frequency. For Mel scale, corresponding to the human auditory system, pitch increases progressively less rapidly as the frequency increases, In direct contrast, it increases progressively more rapidly in the proposed Inverted Mel Scale.

We observe that, in this scale, pitch increases more and more rapidly (fig. 2) as the frequency increases. As we aimed, this is in direct contrast to the human auditory system (eqn. 1), where it increases less rapidly with rising frequency. Hence, the higher frequency zone coarsely approximated by normal MFCC can be represented more finely by this new scale. Hence it can capture the speaker-specific formant information present in this zone which could have been neglected by the original MFCC. These facts justify our choice of flipping the MFCC filter bank to obtain the new IMFCC feature set.

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

$$\hat{e}(i) = \sum_{k=1}^{\frac{Ms}{2}} |Y(k)|^2 \cdot \hat{\Psi}_i(k) \quad (18)$$

Computational burden is reduced since we do not need to recalculate the energy spectrum $|Y(k)|^2$ (fig. 4) when we go for parallel classifiers one using MFCC and the other using IMFCC.

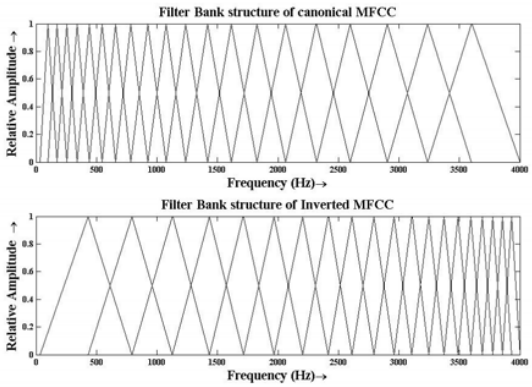

Fig. 4 Plot showing filter bank structures for the two systems

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

$$\hat{C}_m = \sqrt{\frac{2}{Q}} \sum_{l=0}^{Q-1} \log\left[\hat{e}(i+1)\right] \cdot \cos\left[m \cdot \left(\frac{2l-1}{2}\right) \cdot \frac{\pi}{Q} \right] \quad (19)$$

As with MFCC, we took *Q = 20, R = 20* and used the last 19 coefficients to model the individual speakers.

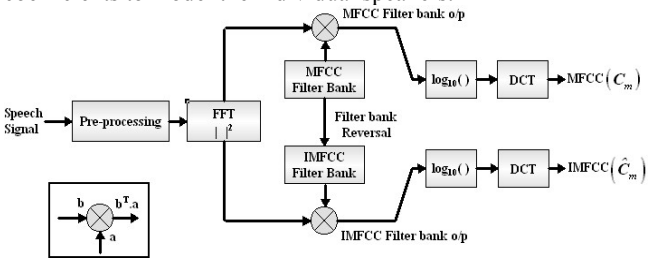

Fig. 5 Plot showing extraction of MFCC and IMFCC features.

IV. Theoretical Background on Gaussian Mixture Models (GMM)

A GMM [14] 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 $b_i^s(\mathbf{x})$, each parameterized by a mean vectors μ_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

$$p\left(x \mid \lambda_s\right) = \sum_{i=1}^{M} p_i^s b_i^s\left(x\right) \quad (20)$$

where,

$$b_i^s\left(x\right) = \frac{1}{\left(2\pi\right)^{\frac{D}{2}} \mid \Sigma_i^s \mid^{\frac{1}{2}}} e^{\left(-\frac{1}{2}\left(x-\mu_i^s\right)^t \left(\Sigma_i^s\right)^{-1}\left(x-\mu_i^s\right)\right)} \quad (21)$$

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

$$L_s\left(X\right) = log\ p\left(X \mid \lambda_s\right) = \sum_{t=1}^{T} log\ p\left(x_t \mid \lambda_s\right) \quad (22)$$

assuming the vectors to be independent for computational simplicity. For SI, the value of $L_s(X)$ is computed for all speaker models λ_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 λ, 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 K-means algorithm which was terminated after 5 iterations. This was followed by the E & M algorithm with 20 iterations. For all cases, diagonal covariance matrices were chosen because DCT has already decorrelated the features by (eqn. 7) and (eqn 19).

V. FUSION OF SPEAKER MODELS

Combining classifier decisions [16] to improve decision reliability has been successful in many pattern classification problems including SI. According to the available literature [12], [16], [17] the combination of two or more classifiers would perform better if they are 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. In this context, it should be noted that our computation of complementary information from IMFCC involves comparably lower computational complexity than higher level features [11]-[13].

During the training phase, two separate models were developed for each speaker from the MFCC and IMFCC feature sets respectively, using GMMM 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 [16], an uniform weighted sum rule [7], [12] 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} and S^i_{IMFCC} are the scores generated by the two models for the ith speaker then the combined score S^i_{com} is expressed as

$$S^i_{com} = \alpha S^i_{MFCC} + \left(1-\alpha\right) S^i_{IMFCC} \quad (23)$$

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

$$S^i_{com} = \alpha \sum_{t=1}^{T} log\ p\left(x_{tMFCC} \mid \lambda_{sMFCC}\right) + \left(1-\alpha\right) \sum_{t=1}^{T} log\ p\left(x_{tIMFCC} \mid \lambda_{sIMFCC}\right) \quad (24)$$

All the notations have their usual meanings. We have used a = 0.5 as the weight [11] for all combinations. However, more suitable weights can be investigated further to enhance the performance of the combined system. Finally, the identity of the true speaker i_{true} is given by:-

$$i_{true} = arg\ max_i\ S^i_{com} \quad (25)$$

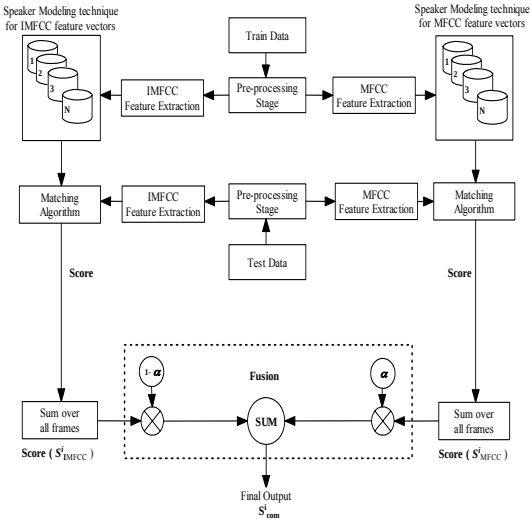

Fig. 6 Parallel classifier based SI system.

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

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 are calculated (ref. Sec II & III). The first coefficient ($\boldsymbol{C_0}$ and $\widehat{\boldsymbol{C}}_{\boldsymbol{0}}$) 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 voice verification corpus [18] 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 [19] 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 [20] 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 [14] and we have used the same:

$$\text{Percentage of identification accuracy (PIA)} = \frac{\text{No. of utterances correctly identified}}{\text{Total no. of utterances under test}} \quad (26)$$

D. Experimental Results

For each database, we evaluated the performance of an MFCC based classifier, an IMFCC based classifier and a parallel classifier fusing both models.

1) Results for YOHO Database

Table 2 describes identification results for various model orders of GMM. The last column in the table depicts the identification accuracies for the proposed combined scheme. The proposed 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.

To compare equal computation, we also implemented MFCC with higher no. of filters i.e. 39 from which a 38 dimensional cepstral vector is derived which is exactly twice as the no. of feature (i.e. 19) that has been already used for earlier MFCC implementation. Results show in fourth column of the table that the performances even after using higher no. of filter/cepstral order is worse than low dimensional MFCC. This is because increasing resolution neither captures spectral slope described by vocal tract filter nor approximate fine harmonic structure explained by glottal pulse. Further computational comparison can be made with single high dimensional MFCC in conjunction with *2x* no. of Gaussian with two lower dimensional streams (MFCC and IMFCC) each modeled with x no. of Gaussian where, $x \in \{4, 8, 16, 32, 64\}$. Here also the performance of the proposed combined system performs much better than the high dimensional MFCC.

TABLE II
RESULTS (PIA) OF GMM FOR YOHO DATABASE

No. of Mixtures	MFCC (19)	IFMCC (19)	MFCC (38)	**Combined System (19+19)**
2	75.05%	77.12%	66.18%	**83.04%**
4	84.96%	86.39%	71.18%	**90.56%**
8	91.87%	92.07%	79.86%	**94.89%**
16	94.09%	94.18%	87.01%	**95.72%**
32	96.20%	95.14%	91.07%	**97.19%**
64	97.19%	95.30%	94.42%	**97.74%**

2) Results for POLYCOST

Table shows the identification accuracies for the POLYCOST database. As with the YOHO database, it can be observed from these tables that our proposed combined scheme shows significant improvement over the baseline MFCC based system in all cases. Also, results improve as model order increases. We restrained ourselves to 4 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.

Higher order MFCC is also implemented with increased resolution for this database and results indicate that the proposed fused system outperforms both low dimensional and high dimensional MFCC based SI system.

TABLE III
RESULTS (PIA) OF GMM FOR POLYCOST DATABASE

No. of Mixtures	MFCC 19	IFMCC 19	MFCC 38	**Combined System (19+19)**
2	64.19%	58.49%	67.37%	**68.83%**
4	72.41%	67.50%	68.83%	**77.58%**
8	76.79%	74.54%	72.68%	**79.71%**
16	78.91%	77.32%	77.32%	**81.57%**

It is observed that the independent performance of IMFCC is not as good as MFCC for POLYCOST database as compared to YOHO. This is due to the fact that 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. Results also show that improvement of the proposed scheme is massive in case of lower order models especially in GMM and VQ based systems for both the databases. 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.

Further, a comparison with higher order of MFCC is also made for both the databases to ensure that the proposed system can also perform satisfactorily with lower order of Gaussian Mixtures.

VII. Conclusion

A new front-end acoustic feature set complementary to MFCC is proposed here that provides higher order speaker specific formant information usually ignored by MFCC. The proposed feature is extracted by flipping the triangular filter bank structure described by MFCC. Speaker models developed from this proposed feature when fused with existing MFCC based speaker models via weighted sum rule, gives significant improvements over the baseline system which can be attributed to availability of complementary information to two parallel models. The experiment is conducted with three different classifiers over different model orders on two kinds of databases, one based on microphone speech and the other on telephone speech. The results prove the superiority of our proposition irrespective of data type, amount of data and model orders. Further, the proposed scheme utilizes the same computational basis as MFCC unlike high level features that needs computationally expensive algorithms for extraction. The processing time could also be compared to a single-stream based system because of the inherent parallelism of the two feature sets. Performance could be further improved by choosing optimal weights to fuse the scores before the classification decision.

References

[1] J. P. Cambell, Jr., "Speaker Recognition:A Tutorial", *Proceedings of The IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept. 1997.

[2] 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.

[3] 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.

[4] Ben Gold and Nelson Morgan, *Speech and Audio Signal Processing*, Part- IV, Chap.14, pp. 189-203, John Willy & Sons,2002.

[5] U. G. Goldstein, "Speaker identifying features based on formant tracks", *J. Acoust. Soc. Am*, vol. 59, No. 1, pp. 176-182, Jan. 1976.

[6] Rabiner. L, Juang B. H, *Fundamentals of speech recognition*, Chap. 2, pp. 11-65, Pearson Education, First Indian Reprint, 2003.

[7] Daniel J. Mashao, Marshalleno Skosan, "Combining Classifier Decisions for Robust Speaker Identification", *Pattern Recog,*, vol. 39, pp. 147-155, 2006.

[8] 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.

[9] 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.

[10] 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.

[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*, Vol. 13, No. 4, pp. 575-582, July 2005.

[12] 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.

[13] 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.

[14] 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.

[15] D. O' Shaughnessy, *Speech Communication Human and Machine*,Addison-Wesley, New York, 1987.

[16] J. Kittler, M. Hatef, R. Duin, J. Mataz, "On combining classifiers", *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (1998) 226-239.

[17] Daniel Garcia-Romero, Julian Fierrez-Aguilar, Joaquin Gonzalez-Rodriguez, Javier Ortega-Garcia, "Using quality measures for multilevel peaker recognition", *Computer Speech and Language*, Vol. 20, Issue 2-3, pp. 192-209, Apr. 2006.

[18] J. Campbell, "Testing with the YOHO CDROM voice verification corpus", *ICASSP95*, 1995, vol.1 pp. 341-344.

[19] Petrovska, D., et al. "POLYCOST: A Telephone-Speech Database for Speaker Recognition", *RLA2C*, Avignon, France, April 20-23, 1998, pp. 211-214.

[20] 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, speech processing, speaker recognition, data fusion analysis. He is a student Member of IEEE.

Anindya Roy passed B.E in Electronics and Telecommunication Engineering with highest honours from Bengal Engineering and Science University, Shibpur, Howrah, India in 2005. Presently he is a seniour student of Masters of Technology (M. Tech) with Automation and Computer Vision as specialization in the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India. His research areas includes pattern recognition, neural networks, speech processing, speaker recognition, music information retrieval, data compression, cryptography etc.

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.