Language and Text-Independent Speaker Identification System Using GMM

S. SELVA NIDHYANANTHAN, R.SHANTHA SELVA KUMARI
Department of Electronics and Communication Engineering
Mepco Schlenk Engineering College
Sivakasi, Virudhunagar District, Tamil Nadu - 626005
INDIA
nidhyan@mepcoeng.ac.in, r_sskp@yahoo.com

Abstract: - This paper motivates the use of Dynamic Mel-Frequency Cepstral Coefficient (DMFCC) feature and combination of DMFCC and MFCC features for robust language and text-independent speaker identification. MFCC feature, modeled on the human auditory system has been the widely used feature for speaker recognition because of its less vulnerability to noise perturbation and little session variability. But the human auditory system also can sensitively perceive the pitch changes in the speech. Therefore adopting the algorithm which integrates the change in speaker specific pitch information in designing the Dynamic Mel scale filter bank exhibit improved effectiveness in speaker identification. The individual Gaussian component of Gaussian Mixture Model (GMM) represents vocal tract configurations that are effective for speaker identification. The performance of the speaker identification system is experimentally evaluated with microphone speech data base consisting of 120 speakers. The experiments examine the speaker Identification Error Rate (IDER) by testing using segments of different lengths and also using text-independent utterances in Tamil and English languages. In comparison with the identification error rate of 5.8% obtained with MFCC-based system and 2.9% with DMFCC system an error rate of 1.2% is obtained when DMFCC feature vectors are added with MFCC feature vectors to form the combined feature. Experimental results confirm that GMM is efficient for language and text – independent speaker identification.

Key-Words: - Speaker Identification, Mel- scale filter bank, Gaussian filters, Mel Frequency Cepstral Coefficient, Dynamic Mel Frequency Cepstral Coefficient, Gaussian Mixture Model.

1 Introduction

Speech signal is produced by exciting time varying vocal tract system with time varying excitation. Speech signal contains information about messages to be conveyed, speaker identity and language information. The speaker-specific characteristics of speech are due to differences in physiological and behavioral aspects of the speech production system in humans. The main physiological aspect of the human speech production system is the vocal tract shape [4]. The vocal tract modifies the spectral content of an acoustic wave as it passes through it, thereby producing speech. The vocal tract resonances vary based on the shape of the tract, which distinguishes one speaker from another. Speaker recognition is divided into two tasks: speaker identification and speaker verification. The goal of speaker identification is to determine the person by his or her voice. There are two types of speaker identification systems: Text-dependent, Text-independent [2]. In text-dependent, the speaker has to utter the same phrase during training and testing. In text-independent the phrase during testing may be independent of training phrase.

The earliest approach to speaker identification is to use long-term averages of acoustic features, such as spectrum representations [5] or pitch [6], [8]. The most frequently used parameters for speaker identification are MFCC, Linear Predictive Cepstral Coefficients (LPCC), pitch [34], formant frequency and bandwidth, Bark-Frequency Cepstral Coefficients (BFCC) and so on. Among these features MFCC is considered as an important characteristic parameter by researchers of speech and speaker recognition, because of the preferable simulation of the human hearing system's perception ability. MFCC, LPCC and BFCC features are based on the spectral information derived from a short time windowed segment of speech. They differ mainly in the detail of the power

spectrum representation. The formant, LPC and LPCC are related to vocal tract, and have good speaker identification characteristics with high SNR (signal to noise ratio). However, when the SNR is low, the differences between the vocal tract parameters estimated from noisy speech signal and those of the real vocal tract model are big. Thus, these characteristic parameters cannot correctly reflect speaker's vocal tract features. In the Gaussian filters in the filter bank based algorithm of MFCC, the number of Gaussian filters in the filter bank and the center frequency of each filter are fixed. At the same time the dynamic construction of Mel filter bank based on the speaker's pitch frequency, better represents the periodicity generated by vocal cords vibration and uniquely distinguishes the vocal characteristics of different people. Hence the feature derived from dynamic construction of Mel filter bank, the Dynamic MFCC feature is also considered to be equivalently important to MFCC features.

There are many modeling methods to model the speaker-dependent acoustic features within the individual phonetic sounds that comprise the utterance [7], [8] & [9]. These approaches can be accomplished using explicit or implicit segmentation of the speech into phonetic sound classes prior to speaker model training or recognition.

In the Hidden Markov Model (HMM) approach, the sequences of feature vectors extracted from speech waveforms are assumed to be a Markov process and modeled with HMM [10], [11]. The applicability of HMM-based automatic speech recognition is limited due to one critical issue: data-driven HMM-trained speech models do not generalize well from training to testing conditions. Such an inevitable mismatch is generally derived from 1) speaker effects, e.g., speech production, accent, dialect, and speaking rate differences and 2) speaking environment effects, e.g., interfering noise, transducers and transmission channel distortions [32]. HMM is not advantage over GMM in text-independent task [1].

In Vector Quantization (VQ) approach each speaker is represented by a codebook of spectral templates representing the phonetic sound clusters [14], [15]. However some speaker dependent temporal information is neglected in VQ. This technique provides good performance on limited vocabulary task [16], [17], while it limited its ability to model possible variabilities in an unconstrained speech.

The GMM falls into the implicit segmentation approach to speaker recognition. It provides a probabilistic model of the underlying sounds of a person's voice. The third approach to speaker recognition is the use of discriminative Neural Networks (NN). Rather than training individual models to represent particular speakers, discriminative NN's are trained to model the decision function which best discriminates speakers within a known set [12], [13]. The power and utility of the NN has been demonstrated in several applications including speech synthesis and pattern recognition [18], [19]. In the NN approach, each speaker has personalized NN that is trained to be activated only by those speakers' utterances. The testing waveforms are tested by the speakers personalized NNs to make speaker identification decisions. The major drawback of NN is that the complete network is retrained when a new speaker is added to the system [23].

In this paper GMM is used and evaluated for text- independent speaker identification. The individual component of GMM represents some vocal tract configurations that are speaker dependent for identifying the speaker. Gaussian mixture density provides smooth approximation to the sample distribution of observations obtained from utterance of a given speaker. In earlier works, LPCC and Reflection Coefficients have been used for speaker recognition; however they are affected by noise [29]. Recent studies [20] found that filter bank features are robust to noise in speech recognition. In this paper pitch based Mel-scale filter bank is used for speech analysis. In speaker identification a representation of the speech signal is obtained using digital signal processing techniques which preserve the features of the speech signal that are relevant to speaker identity. The resulting pattern of the speech signal is compared to previously prepared reference patterns and subsequent decision is made on the identity of the speaker.

This paper is organized as follows. System Model has been given in section 2. Preprocessing steps and Feature Extraction is explained in section 3 and section 4 respectively. Section 5 presents Gaussian Mixture Model. The experimental results are given in section 6 followed by concluding remarks in section 7.

2 System Model

Speech signal for training is preprocessed and feature vectors are obtained. Then Expectation Maximization (EM) algorithm is applied to feature vectors and the parameters of GMM are obtained. The test speech signal is preprocessed and with the feature vectors a posteriori probability is calculated along with training parameters in the parameter

database. The speaker with maximum a posteriori probability is the identified speaker. The block diagram used for speaker identification is given in Fig.1.

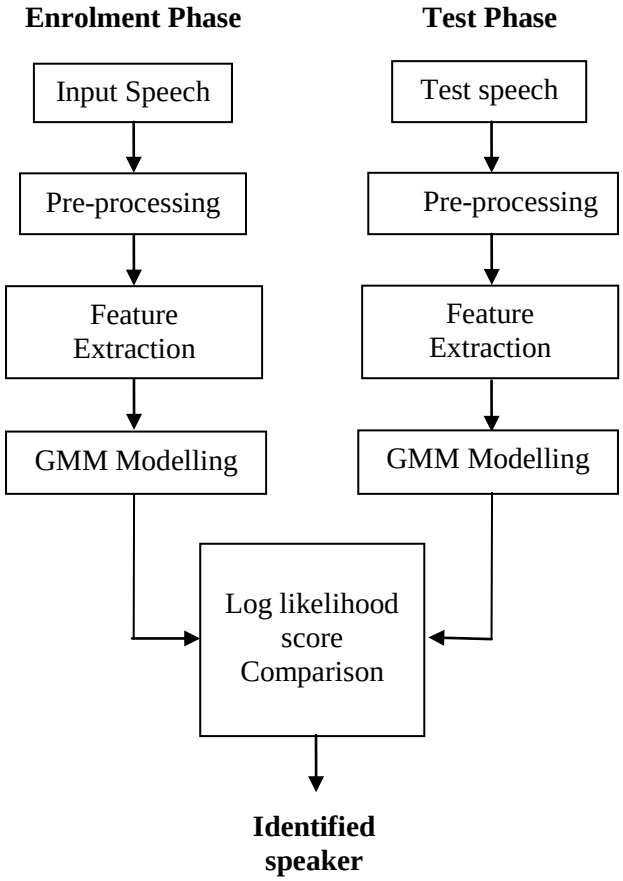


Fig.1. Speaker Identification system model

3 Preprocessing

Preprocessing of speech signal, involves segregating the voiced region from the silence/unvoiced portion of the captured signal, and is necessary in the development of a reliable speech or speaker recognition system. This is because most of the speech or speaker specific attributes are present in the voiced part of the speech signals. Moreover, extraction of the voiced part of the speech signal by marking or removing the silence and unvoiced region leads to substantial reduction in computational complexity at later stages.

3.1 Noise Removal

Noises are unnecessary signals which tend to degrade the performance of the speaker identification system. Noise removal is done by wavelet decomposition technique [13]. The steps followed in denoising are:

i) Compute Discrete Wavelet Transform (DWT) decomposition by choosing a wavelet at a decomposition level *N*.
ii) For each level 1 to *N*, select a threshold value and apply soft thresholding to both approximation and detailed coefficients.
iii) Wavelet reconstruction is computed based on the threshold approximation and detailed coefficients.

The wavelet decomposition is performed by choosing Daubechies wavelet of order 4 from the family of orthogonal wavelet. Decomposition is performed at level 1. In this both approximation and detailed coefficients are denoised using soft thresholding and the denoised signal is used for reconstruction.

3.2 Framing

The speech region has to be short enough so that it can reasonably be assumed to be stationary, for extracting the parameters. Thus to model dynamic parameters, the signal is divided into successive frames. Overlapping between frames is necessary. If the frames have no overlap, there may be loss of information, due to the presence of a small gap between adjacent frames. Framing is done with a frame size of 256 samples and overlap size of 156 samples. Good results are achieved with overlap size more than 50%. Typical value chosen for frame overlapping is 60%. Then Hamming windowing is done on each frame.

3.3 Windowing

Windowing is done to provide spectral smoothing [25]. It is done on each individual frame so as to taper the signal to zero at the beginning and at the end of frame. Windowing is also essential for capturing dynamic characteristics of vocal tract system in speech production mechanism [31]. The Hamming window is used because it has a wide main lobe and small side lobes, making it a smooth lowpass filter with less leakage [30].

Hamming window $w(n)$ has the form

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N-1 \tag{1}$$

where N represents the width, in samples, of a discrete-time window function. Typically it is an integer power-of-2, such as $2^{10} = 1024$.

n is an integer, with values $0 \le n \le N$-1.

4 Feature Extraction

In feature extraction, to the windowed signal $|FFT|^2$ is calculated [25] and preemphasis is done with a preemphasis factor of 0.97. Next the Mel-scale filter bank is constructed using Gaussian filters and filter response is obtained. In speech processing, MFCC is a representation of the short term power spectrum of a speech sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. This obtained by taking discrete cosine transform to the log energies. 20 MFCC coefficients are obtained from each frame.

4.1 Fast Fourier Transform

Features are the representative pattern vectors for speech signals. After windowing the speech signal Fast Fourier Transform (FFT) is applied to each frame and its squared magnitude is calculated. For sampled vector data Fourier analysis is performed using Discrete Fourier Transform (DFT). FFT is an efficient algorithm for computing DFT in the sequence.

4.2 Pre-emphasis

The higher frequency component of speech signal is generally weak, so high frequency energy may not be present to extract features in the upper frequency range. In many speech processing applications higher frequency components are necessary [3]. Preemphasis is used to boost the energy of high frequency signals. Thus preemphasis helps to equalize the spectral tilt in speech and the signal is spectrally flattened. The output of pre-emphasis [22] $\hat{s}(n)$ is related to input $s(n)$ by

$$\hat{s}(n) = s(n) - \alpha s(n-1) \tag{2}$$

where α is Preemphasis factor whose value varies from 0.9 to 1.

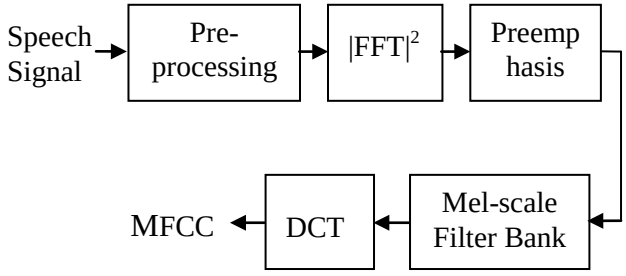


Fig.2.Mel-scale cepstral feature analysis

4.3 Mel Scale Filter Bank

Linear frequency scale is not unique for a speaker, so Mel frequency scale is used for speaker recognition [26], [27]. Mel is the unit of pitch. Mel-scale is linear below 1 kHz and logarithmic above 1 kHz [24]. Mel-scale filter bank is shown in Fig.3. The filters are equally spaced along Mel-scale

$$mel(f) = 2595\log_{10}\left(1+\frac{f}{700}\right) \tag{3}$$

mel(f) corresponding to mapping the actual frequency in Hertz to the mel frequency[36].

If triangular filters are used in filter bank, the correlation between a subband and adjacent subband is lost. In this paper Gaussian filters are used. The gaussian filters are chosen for many reasons. First, it is symmetric and high frequency components are involved. Second, gaussian shaped filters provide smooth transition from one subband to other preserving most of the correlation between them. The filters in the filter bank are arranged such that more number of filters are present in the low frequency range.

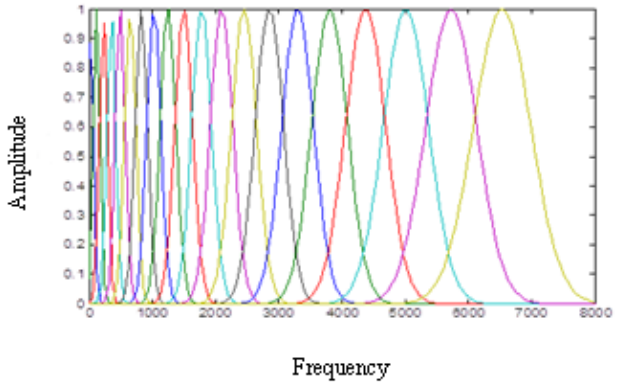


Fig.3. Mel scale Gaussian filter bank

The filter response is given by

$$H(k,m) = e^{-\frac{(f(k)-kb_i)^2}{2\sigma_i^2}} \tag{4}$$

$$where \qquad f(k) = \frac{kf_s}{Q} \quad k = 1,2,.....Q-1 \tag{5}$$

f_s is sampling rate which is chosen as 16 kHz

Q is the number of filters required to span the frequency range of speech.

Standard Deviation in (4) is given by

$$\sigma_i = \frac{kb_{i+1} - kb_i}{\alpha} \tag{6}$$

where kb_i are the boundary points of the filters and

K is the coefficient index in the *N* point DFT.
The term α in (6) controls variance. The value of α may be 1, 2 or 3.
The value of α is chosen to be 2, since it provides better correlation with adjacent subband.
f_{low} and f_{high} are the low and high frequency boundaries of filter bank, they are given as

$$f_{low} = \frac{fs}{N} = 62.5Hz \ ,$$
$$f_{high} = \frac{fs}{2} = 8\,kHz \quad (7)$$

The magnitude spectrum is scaled in both frequency and magnitude. First the frequency is scaled logarithmically using the so called mel-scale filter bank *H* (*k,m*) and then logarithm is taken.

$$X'(m) = \log_{10}\left(\sum_{k=1}^{N/2}|X(k)|^2 H(k,m)\right) \quad (8)$$

The value of *m* ranges from 1 to *Q*, where *Q* is the number of filters.

4.4 Discrete Cosine Transform

Discrete Cosine Transform (DCT) is applied to the log of the Mel Spectral Coefficients to obtain MFCCs. By applying DCT decorrelated coefficients are obtained. The zeroth coeffient has average log energy and hence it is discarded. In MFCC the frequency bands are logarithmically spaced. As the frequency bands are positioned logarithmically in MFCC it approximates the human response system more closely than any other system. These coefficients allow better processing of data. The Mel frequency cepstral coefficient is given by

$$c(l) = \sqrt{\frac{2}{Q}}\sum_{m=1}^{Q} X'(m)\cos\left(\frac{l\pi}{Q}\left(m-\frac{1}{2}\right)\right) \quad (9)$$

c(l) is the l^{th} Mel Frequency Cepstral Coefficient
where l = 1,2….Q , Q is number of filters.

4.5 Dynamic MFCC Using Pitch Frequency

The pitch frequency precisely represents the speakers' periodic characteristic of the vocal cords' vibration when speakers pronounce voiced sound [35]. The pitch information has been used in fields such as speech synthesis, and pronunciation defect correction, vocoder etc. It is also a main characteristic parameter in the speaker recognition, because it best exhibits the vocal cord characteristics of the speaker.

In the traditional algorithm based on MFCC [33], the number of Triangle or Gaussian filters, which constitute Mel filter, and the center frequency of each filter are fixed. This method does not fully consider the vocal cord characteristics of the speakers. Because the pitch frequency [34] represents the periodicity generated by vocal cords vibration when the speaker delivers a voiced sound, and portrays the vocal characteristics of different people, this paper combines MFCC and pitch frequency, and proposes design method of dynamic Mel filter based on speech signal variety. The steps involved in the design of dynamic Mel filter bank are given as follows:

(i) Get pitch freq f_p of each frame signal using autocorrelation method, then calculate

$$Mel(f) = 2595\log_{10}\left(1+{f_p}/{700}\right) \quad (10)$$

This mel(f) maps pitch frequency to mel frequency .

(ii) Dividing Mel frequency field into *N* sub-fields, and taking the corresponding frequency $f_p, f_{2p}, \ldots\ldots f_{NP}$ of each division point in actual frequency field as filter's center frequency, designs the Mel filter Bank $\{H_i(k), i = 1,2,\ldots\ldots N\}$.

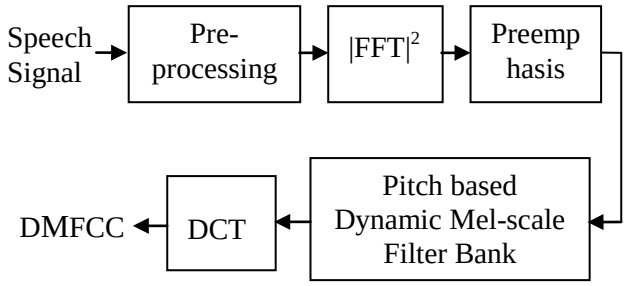


Fig.4. DMFCC feature extraction

This filter bank is applied on the short-time signal's energy spectrum of the speech signal followed by DCT computation to arrive at DMFCC features which is shown in Fig.4.

5 Gaussian Mixture Model

GMM can smoothly approximate the probability density function of arbitrary shape, portray distributed characteristic of different speaker's speech feature in the feature space. The GMM

probability density function may be expressed by the parameter set $\lambda_i = \{\vec{p}_i, \vec{\mu}_i, \vec{\Sigma}_i\}$ $i = 1,.....M$ which is called parameter model of the speaker y_i for example. This parameter model is dissimilar for different speakers. For each speaker, this model can describe the distribution of speech characteristics in the feature space. GMM belongs to the unsupervised classifiers category. This means that the training samples of a classifier are not labeled to show their category membership and the targets are not provided instead, during the training of the GMM classifier the underlying probability density functions of the observations are estimated. In the GMM classifier, the conditional-pdf of the observation vector with respect to the different classes is modeled as a linear combination of multivariate Gaussian pdf's. GMM has been successfully applied to speaker modeling in text-independent speaker identification because, the individual component Gaussian in a GMM represents some broad acoustic classes and a Gaussian mixture density provides a smooth approximation to the sample distribution of observations obtained from utterances of a given speaker.

Speech production is not deterministic. A particular sound is not produced by a speaker with exactly the same vocal tract shape, glottal flow, due to context, coarticulation, anatomical and fluid dynamical variations. One way to represent this variability is probabilistically through multi-dimensional Gaussian probability density function [28]. The use of GMMs for speaker recognition is described in [21]. A Gaussian probability density function is state dependent. A different Gaussian pdf is assigned for each acoustic class. The Gaussian probability density function of a feature vector for i^{th} state is given by

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} \exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)'\Sigma^{-1}{}_i(\vec{x} - \vec{\mu}_i)\} \quad (11)$$

where μ_i=mean vector,
$\sum_i$=covariance matrix and
D=dimension of the vector.

The probability of feature vector in any one of M acoustic class for a particular speaker model λ is represented by the union or mixture of different Gaussian pdf. This is represented as

$$p(\vec{x} / \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (12)$$

where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x})$, i=1……M are the component densities and p_i, i=1……M are the mixture weights.

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. The parameters are collectively represented by the notation

$$\lambda_i = \{\vec{p}_i, \vec{\mu}_i, \vec{\Sigma}_i\} \quad i = 1,.....M \quad (13)$$

For speaker identification, each speaker is represented by a GMM and is referred by his/her model λ.

5.1 Maximum Likelihood Parameter Estimation

The aim of ML estimation is to find the model parameters which maximize the likelihood of GMM.

For a sequence of T training vectors $X = \{\vec{x}_1, \vec{x}_2, ...\vec{x}_T\}$ the GMM likelihood can be written as

$$p(X / \lambda) = \prod_{t=1}^{T} p(\vec{x}_t / \lambda) \quad (14)$$

This expression is a nonlinear function of the parameters λ and so direct maximization is not possible. The ML parameter estimate is obtained iteratively using Expectation Maximization algorithm.

5.2 Expectation Maximization Algorithm

The most popular algorithm for GMM parameters estimation is the EM algorithm. This algorithm allows iterative optimization of the mixture parameters, under nondecreasing likelihood requirement. The EM algorithm begins with an initial model λ, to estimate a new model λ_1 .The new model then becomes the initial model and the process is repeated till convergence. The performance of this algorithm depends on its initialization due to its tendency to converge to local extrema. A proper initialization must be done for model parameter. On each EM iteration mixture weight, mean and variance are calculated using eqn. (15), (16) and (17) respectively.

Mixture weight: $$\vec{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i / \vec{x}_t, \lambda) \quad (15)$$

Mean: $$\vec{\mu}_i = \frac{\sum_{t=1}^{T} p(i / \vec{x}_t, \lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i / \vec{x}_t, \lambda)} \quad (16)$$

Variance: $$\vec{\sigma}_i = \frac{\sum_{t=1}^{T} p(i/\vec{x}_t,\lambda)x_t^{\,2}}{\sum_{t=1}^{T} p(i/\vec{x}_t,\lambda)} - \vec{\mu}_i^{\,2} \qquad (17)$$

The a posteriori probability for acoustic class i is given by

$$p(i/\vec{x}_t,\lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \qquad (18)$$

5.3 Speaker Identification

For speaker identification, a group of S speakers $S=\{1,2,.....S\}$ is represented by GMM's λ_1, λ_2, ……., λ_S. The objective is to find the speaker model which has the maximum a posteriori probability for a given observation.

$$\hat{S} = \arg\max_{1\le k\le S} P_r(\lambda_k / X) \qquad (19)$$

With reference to Baye's rule, the speaker identification system computes

$$\hat{S} = \arg\max_{1\le k\le s} \sum_{t=1}^{T} \log(p(\vec{x}_t/\lambda_k)) \qquad (20)$$

In which $p(\vec{x}_t/\lambda_k)$ is given in eqn. (12)

6 Experimental Results and Discussion

The Speaker Identification experiment is conducted on speech database created with a collection of 120 speakers. From each speaker 5 sessions, each of 20 sec duration is recorded in 16 kHz sampling rate. The speech is recorded using Gold Wave software in Tamil and English languages. The initial setting in this software is set to 16 kHz, 16 bit mono, then it is placed in recording mode and the speaker speaks using a condenser microphone. The recorded speech is saved with an extension of .wav file format.

Feature vectors are obtained for MFCC feature, DMFCC feature and combined MFCC and DMFCC feature. For every frame of the speech signal 20 DCT coefficients are obtained and the first DCT coefficient is discarded (since it yields the DC value). As a result every frame of the speech signal contributes 19 MFCC coefficients. When pitch frequency is used in Mel filter scale, every frame of the speech signal contributes 19 DMFCC coefficients. For the combined feature the 19 MFCC coefficients of every frame of speech signal is added with 19 DMFCC coefficients of every frame of speech signal. This addition of feature vectors result in combined MFCC and DMFCC feature. The feature vectors obtained are trained using EM algorithm in GMM modeling. The first three sessions of each speaker are taken for training and the remaining two sessions are taken for testing. The voice in English/Tamil of test speaker is compared to speakers in the trained data set and the number of incorrectly identified session trials is tabulated in Table 1. The Identification Error Rate is given by

$$IDER = \frac{Number\ of\ incorrect\ identification\ trials}{Number\ of\ identification\ trials} \times 100\,\%$$

Table 1. Speaker Identification Performance Using 1 sec Speech Utterance

Features	Number of mixture components	IDER (in %)
MFCC	4	60.2
	8	50.4
	16	80.2
	32	85
DMFCC	4	50.1
	8	30.6
	16	62
	32	80.2
MFCC + DMFCC	4	40.2
	8	22.3
	16	50
	32	71

The speaker identification performance is evaluated for two different lengths of speech utterances. In first case 1second speech is used. From Table 1, it is inferred that lesser identification error rate is achieved for 8 mixture components of GMM for all the three features.

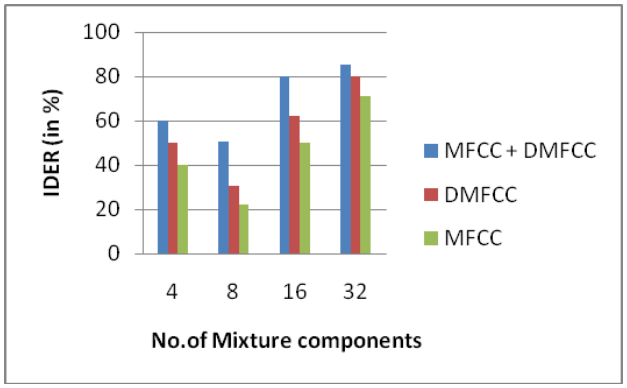


Fig.5. Identification Performance for 1 sec speech utterance

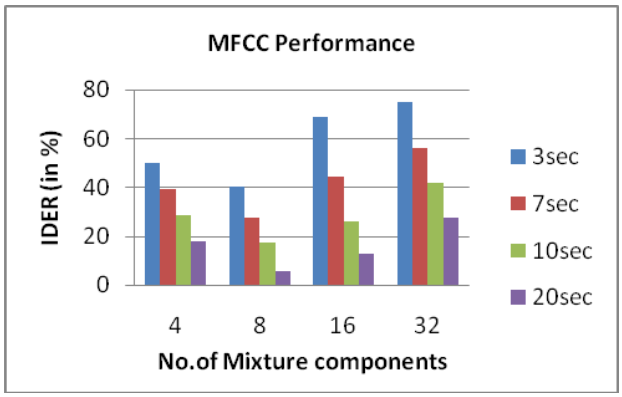
Fig.5 shows the Identification Performance for 1 sec speech utterance for MFCC feature, DMFCC feature and combined MFCC and DMFCC features.

Table 2. Speaker Identification Performance Using Segments of Different Lengths

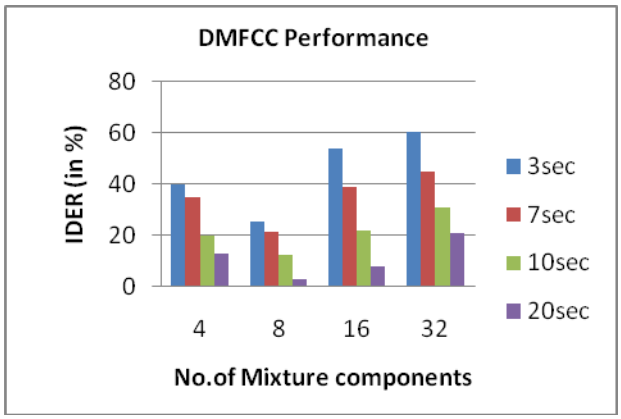
Features	Number of mixture Components	IDER (in %)			
		Test signal length			
		3sec	7 sec	10sec	20sec
MFCC	4	50	39	28.4	18
	8	40	27.3	17.2	5.8
	16	69	44.5	26	12.6
	32	75	56	42	27.5
DMFCC	4	40	35.2	20	13
	8	25.6	21.4	12.5	2.9
	16	54.2	39.2	22	8
	32	60.5	45	31	20.8
MFCC + DMFCC	4	35	31.13	13.75	8
	8	21.7	18.75	7.66	1.2
	16	37.27	30	15	2.5
	32	47.16	33.13	25.94	17.5

For 16 and 32 Gaussian mixture components the identification error rate increases beyond the identification error rate for 4 Gaussian mixture components.

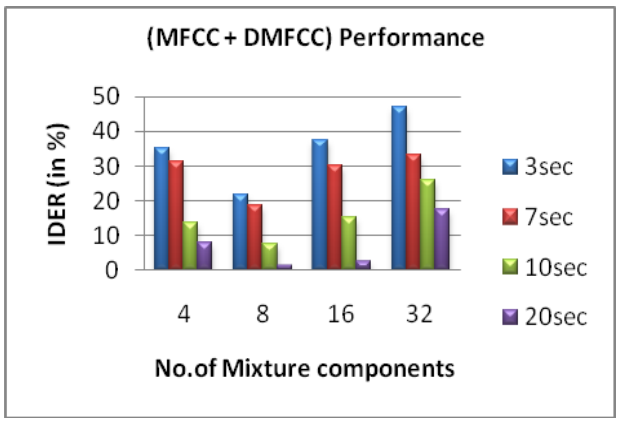
In the second case varying lengths of speech utterances is used. Comparing speaker identification performance of 1 second speech utterance in Table 1, and 3 seconds, 7 seconds, 10 seconds and 20 seconds speech utterances shown in Table 2 , it is observed that 1 second of speech utterance is not enough to obtain high identification performance, whereas for 20 second speech utterances, the identification performance increases for all the three features. It is also observed that there is a leveling off from 16 mixture components. This indicates that there is a lower limit on the number of mixture components necessary to adequately model the speakers. Models must contain at least this minimum number of components to maintain good identification performance. This limit seems to be 8 mixture components to this speaker identification. Fig.6 shows the increasing identification performance from 3 seconds to 20 seconds length segments for MFCC, DMFCC and combined features. The combined MFCC + DMFCC feature produces very low identification error rate compared to individual features.



(a)



(b)



(c)

Fig.6. Identification performance by (a) MFCC (b) DMFCC and (c) MFCC + DMFCC for different length of speech signals

The performance of the proposed combined MFCC, DMFCC feature based Speaker identification system is compared with the work on ‘features based on Cepstrum and Fourier – Bessel (FB)

Expansion' by K.Gopalan and the work on 'Speech Signal Image classification method for speaker identification' by Khalid Saeed. Gopalan used Greenflag database with 41 speakers and NATO database with 9 speakers, for the former he got maximum identification accuracy of 80.4% and for the later 88.0%. Khalid used 20 spaekers and obtained overall average success rate of 94.82%. Our combined MFCC and DMFCC based GMM model out performs these two works with 98.8% of identification accuracy(ie 1.2% IDER) for 120 speakers.

7 Conclusion

This paper uses MFCC feature, pitch based DMFCC feature and the combination of these two features in the experimental evaluation of text-independent, multilingual speaker identification performance. The performance of the speaker identification system is evaluated with microphone speech data base with 120 speakers. The experiments examined the speaker Identification Error Rate by testing using segments of different lengths and also using text-independent utterances in Tamil and English languages. In comparison with the identification error rate of 5.8% obtained with MFCC-based system and 2.9% with DMFCC system an error rate of 1.2% is obtained when DMFCC feature vectors are added with MFCC feature vectors to form the combined feature. This shows that combining the features modeled on the human vocal tract and auditory system yields better result than individual component model.

References:

[1] Bing Xiang,Toby Berger, Efficient Text-Independent Speaker Verification with Structural Gaussian Mixture Models and Neural Network, *IEEE Transactions on Speech And Audio Processing*, Vol.11, No.5, September 2003.

[2] Tomi Kinnunen and Haizhou Li, An overview of text-independent speaker recognition: From features to supervectors, *International journal of Speech Communication Speech Communication* 52 (2010), pp. 12- 40.

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

[4] Joseph P.Campell.JR, Speaker Recognition: A Tutorial. *IEEE Proc* Vol 85, No 9, September 1997.

[5] S. Furui, F. Itakura, and S. Saito, Talker recognition by longtime averaged speech spectrum, *Electron., Commun. in Japan*, Vol. %-A, No. 10, 1972, pp. 54-61.

[6] J. Markel, B. Oshika, and A. Gray, Jr., Long-term feature averaging for speaker recognition, *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-25, Aug. 1977, pp. 330-337.

[7] A. E. Rosenberg et al., Connected word talker verification using whole word hidden Markov models, in *Proc. ICASSP*, 1991, pp. 381-384.

[8] Tomoko Matsui and Sadaoki Furui, Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and DiscreteKontinuous HMM's, *IEEE transactions on speech and audio processing*, Vol. 2, No. 3, July 1994.

[9] Chang-Hoon Lee and Soo-Young Lee, Noise-Robust Speech Recognition Using Top-Down Selective Attention With an HMM Classifier, *IEEE Signal Processing Letters*, Vol. 14, No. 7, July 2007.

[10] Janne Pylkkönen and Mikko Kurimo, Analysis of Extended Baum–Welch and Constrained Optimization for Discriminative Training of HMMs, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 9, November 2012.

[11] L.R.Rabiner, A Tutorial on Hidden Markov Models and selected applications in speech recognition, *IEEE signal Process. Lett.*, Vol. 8, No. 1, July 2001, pp. 196-199.

[12] Muzhir Shaban Al-Ani, Thabit Sultan Mohammed and Karim M. Aljebory, Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform, *Journal of Computer Science* Vol. 3, No. 5, 2007, pp. 304-309.

[13] F. Phan, M. T. Evangelia, and S. Sideman, Speaker identification using neural networks and wavelets, *IEEE Engineering in Medicine and Biology Magazine*, Vol. 191, 2000, pp. 92-101.

[14] Burton, D., Text-dependent speaker verification using vector quantization source coding, *IEEE Trans. Acoustics, Speech, Signal Process*, Vol. 35. No. 2, 1987, pp. 133–143.

[15] Karpov,E., Kinnunen,T., Fra¨nti,P., Symmetric distortion measure for speaker recognition. In: *Proc. Ninth Internat. Conf. on Speech and Computer* (*SPECOM 2004*), St. Petersburg, Russia, September 2004, pp. 366–370.

[16] Louradour, J., Daoudi, K., Andre´-Obrecht, R., Discriminative power of transient frames in

speaker recognition. In: *Proc. Internat Conf. on Acoustics, Speech, and Signal Processing* (*ICASSP 2005*), Vol. 1, Philadelphia, USA, 2005, pp. 613–616.

[17] Hautama, ki .V., Kinnunen, T., Ka¨rkka¨inen, I., Tuononen, M., Saastamoinen, J., Fra¨nti, P., Maximum a posteriori estimation of the centroid model for speaker verification. *IEEE Signal Process. Lett*. 15, 2008, pp. 162–165.

[18] J.Hertz,A. Krogh, and R.J.Palmer, *Introduction to the Theory of Neural Computation, Santa Fe Institute Studies in the Sciences of Complexity*, Addison-Wesley, Reading, Mass, USA, 1991.

[19] J. Oglesby and J. S. Mason, Optimization of neural models for speaker identification, in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing* (*ICASSP'90*), Albuquerque, NM,USA, Vol. 1, April1990, pp. 261–264.

[20] Nirmalya Sen and T.K. Basu, Features Extracted Using Frequency-Time Analysis Approach from Nyquist Filter Bank and Gaussian Filter Bank for Text-Independent Speaker Identification, *BioID 2011, LNCS* 6583, pp. 125–136, Springer-Verlag Berlin Heidelberg 2011

[21] Douglas A. Reynolds and Richard C. Rose, Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, *IEEE transactions on Speech and Audio processing*, Vol. 3, No. 1. January 1995.

[22] Jun Xu,Aladdin Ariyaeeinia,Reza Sotudeh , Zaki Ahmad, Preprocessing Speech Signals in FPGAs. School of Electronic Communication and Electrical Engineering, *IEEE* 2005.

[23] Oglesby.J , Mason.J, Radial Basis Function Networks for speaker Recognition. *Proc. IEEE ICASSP*, May 1991, pp.393-396.

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

[25] Ryo Mukai, Hiroshi Sawada, Shoko Araki, Shoji Makino, Frequency Domain Blind Source Separation Of Many Speech Signals Using Near-Field and Far-Field Model. ,*EURASIP Journal on Applied Signal Processing*, Article ID 83683, 2006.

[26] Senthil Raja.G, Dandapat.S, Performance of Selective Speech Features for Speaker Identification. *Journal of the Institution of Engineers* (*India*), Vol. 89, May 29, 2008.

[27] Sharma.A, Shukla.P, Meghwani.M, Secure Speech Recognition, *Journal of the Institution of Engineers* (*India*), Vol. 89, Nov 17, 2008.

[28] Thomas F.Quatieri, *Discrete Time Speech Signal Processing Principles and Practice*. Pearson Education, New Delhi, 2007.

[29] Tierney.J, A study Of LPC analysis of speech in additive noise, *IEEE Trans. Acoust., Speech,Signal Processing*, Vol. ASSP-28, Aug 1980, pp.389-397.

[30] Yanmei Zhan, Henry Leung, Keun-Chang Kwak, Automated Speaker Recognition for Homeservice Robots Using Genetic Algorithm And Dempster–Shafer Fusion Technique, *IEEE Transactions On Instrumentation And Measurement*, Vol. 58, No. 9, September 2009.

[31] Yeganarayana.B , Satyanarayana Murthy.P, Source System Windowing For Speech Analysis And Synthesis, *IEEE Transactions on Speech And Audio Processing*, Vol.4, No.2, March 1996.

[32] Yu Tsao, Chin-Hui Lee, An Ensemble Speaker and Speaking Environment Modeling Approach to Robust Speech Recognition, *IEEE Transactions On Audio, Speech, And Language Processing*, Vol. 17, No. 5, July 2009.

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

[34] K. Sreenivasa Rao, Voice Conversion by Mapping the Speaker-specific features using Pitch Synchronous Approach, *Computer Speech and Language, Elsevier*, Vol. 24, 2010, pp. 474-494.

[35] Harrag A. Mohamadi., Serignat J.F., (2005). LDA Combination of Pitch and MFCC Features in Speaker Recognition. *Proceedings of INDICON 2005*, IIT Chennai, Indian, pp. 237-240.

[36] Sahidullah, Goutam Saha, Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition, *Speech Communication* 54, 2012, pp.543–565.