Inverted Mel Feature Set based Text-Independent Speaker Identification using Finite Doubly Truncated Gaussian Mixture Model

V. Sailaja
Dept.of ECE
Pragati Engineering College
Surampalem

P. Sunitha
Dept.of ECE
Pragati Engineering College
Surampalem

B. Vasantha Lakshmi
Dept.of ECE
Pragati Engineering College
Surampalem

ABSTRACT
This paper provides an efficient approach for text-independent speaker identification using the Inverted Mel-frequency Cepstral Coefficients as feature set and Finite Doubly Truncated Gaussian Mixture as Model (FDTGMM). 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. Furthermore, it has been shown that the Inverted Mel-frequency Cepstral Coefficients (IMFCC) is also a useful feature set for Speaker identification, which contains information complementary to MFCC as, it covers high frequency region more closely. The performance of the developed model is studied through experimental evaluation with 45 speaker's data base and identification accuracy.

Keywords
Speaker Identification, IMFCC, FDTGMM, Identification accuracy.

1. INTRODUCTION
Speech is one of the natural forms of communication. It conveys the information regarding identity of the speaker. Recent developments have made it possible to use this in the security and authentication systems. In speaker identification the task is to use a speech sample to select the identity of the person that produced the speech from among a population of speaker. Text dependent and text independent speaker identification are two main categories in speaker identification. In text dependent speaker identification, the speaker has to utter same phrase or word during enrollment and verification process. In text independent speaker identification, the speaker has to utter different phrases or words during verification process [1] [2]. The text-independent Speaker Identification systems are most commonly used for speaker identification because they require very little cooperation by the speaker. Thus it gives an efficient secure authentication mechanism. In this paper, a feature set called IMFCC is used to capture speaker specific information lying in higher along with capturing speaker specific information lying in lower frequency part of the spectrum by MFCC. In this paper, a feature set called IMFCC is used to capture speaker specific information lying in higher along with capturing speaker specific information lying in lower frequency part of the spectrum by MFCC.

This paper is organized as follows. In section 2, feature extraction is described. In section 3, a brief overview of Finite Doubly Truncated Gaussian Mixture Model (FDTGMM) is given. The implementation and results are discussed in section 4 and concluding remarks are given in section 5.

2. FEATURE EXTRACTION
The purpose of this module is to convert speech signal to some type of parametric representation. Since speech is a slowly varying signal, when examined over short period of time (eg 5 & 50 ms), its characteristics are fairly constant. During long course of time order of 0.3sec and more the signal characteristics changes which reflect the different speech sounds spoken by the speaker. MFCC are commonly used feature extraction techniques.

MFCC, thus, represents the low frequency region more accurately than the high frequency region and hence, can capture formants efficiently, which lie in the low frequency range and which characterize the vocal tract resonances. However, other formants that lie above 1 kHz are not effectively captured by the larger spacing of filters in the higher frequency range. To increase the frequency resolution in the high frequency range, the inverted Mel wrapping function (for sampling frequency of 8 kHz) the empirical relation (1) have been used .

Cepstral coefficients are calculated using the inverted Mel filter bank as shown in fig.1 The detailed procedure is given in publication [3].

$$f_{inverted\ mel} = 2146.1 - 2595 \log 10\left(1 + \left(\frac{400 - f}{700}\right)\right)$$

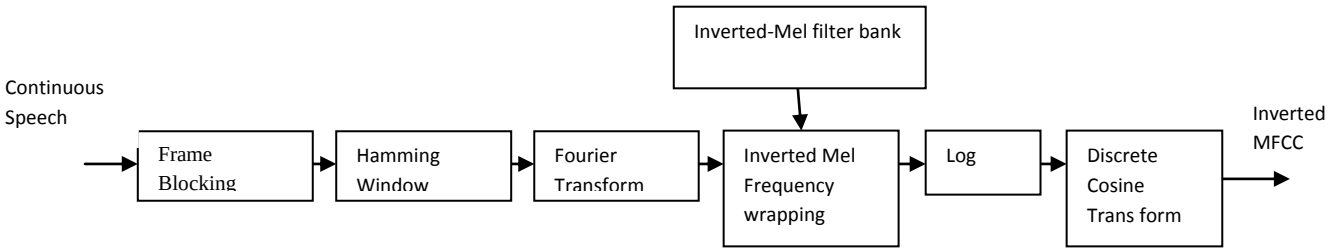

Fig1.Block diagram for inverted-Mel scale Cepstral Coefficients

3. FINITE DOUBLY TRUNCATED GAUSSIAN MIXTURE SPEAKER MODEL

In this section we briefly describe the FDTGMM and motivate its use as a representative of the speaker identity for test independent Speaker identification. The choice of the probability density function is largely dependent on the features being used. Consider the inverted Mel frequency cepstral coefficients of each speaker spectrum as the features for speaker identification. The inverted Mel frequency cepstral coefficients are assumed to follow a FDTGMM. The motivation of this assumption is that the individual component densities of a multi model density, model the underlying set of acoustic process of the speaker.

It is reasonable to assume the acoustic space corresponding to a speaker voice can be characterized by a acoustic classes representing some broad phonetic events such as vowels nasals or fricatives. These acoustic classes reflect some general speaker dependent vocal tract configurations that are useful for characterizing speaker identity. The spectral shape of the its acoustic class can intern be represented by the mean of the its component density and the variation of the average spectral shape can be represented by the covariance matrix. Assuming the independent feature vectors, the observation density of the feature vectors drawn from these acoustic classes is a Doubly Truncated Gaussian Mixture. Also it is given that a linear combination of Gaussian basis function is capable of representing a large class of sample distributions. The FDTGMM is a generalization of the GMM and also as in the case of a Gaussian full co-variance is not necessary even is the features are not statistically independent.

The Finite doubly truncated D variate Gaussian density is

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

$$A = \int_{-\infty}^{x_L} \dots \int_{-\infty}^{x_L} b_i(\vec{x}_t) d\vec{x}_t$$

$$\text{B} = \int_{-\infty}^{x_M} \dots \int_{-\infty}^{x_M} b_i(\vec{x}_t) d\vec{x}_t \quad (3)$$

$\vec{x}_t = (x_1, x_2, \dots x_t)$ is the feature vector $\vec{\mu}_i$ is the ith component of feature mean vector, $\sum$I is the ith component of variance co-variance matix.

The probability density function of the finite M component doubly truncated Gaussian mixture distribution is

$$p(\vec{x}/\lambda) = \sum_{i=1}^{M} \alpha_i \, b_i(\vec{x}) \quad (4)$$

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

The mixture weights satisfy the constraints $\sum_{i=1}^{M} \alpha_i = 1$. Then the FDTGMM is parameterized by mean vector, co-variance matrix and mixture weights from all components densities. The parameters are collectively represented by $\lambda_i = \{\alpha_i, \mu_i, \sigma_i\}$ i=1…M. For speaker identification each speaker is represented by FDTGMM and is referred to by his/her model parameter λ.

3.1 Maximum Likelihood Parameter Estimation

The aim of ML estimation [1] is to obtain the model parameters which maximize the likelihood of FDTGMM. For a sequence of training vector $\vec{x}_t = (x_1, x_2, \dots x_t)$, The FDTGMM likelihood cam be written as

$$p(\vec{x}_t/\lambda_i) = \prod_{i=1}^{T} p_i(\vec{x}_t/\lambda_i) \quad (5)$$

This expression is a non linear function of the parameter λ and so direct maximization is not possible. The ML parameter estimate is obtained iteratively using Expectation maximization algorithm[4].

3.2 Expectation Maximization (Em) Algorithm

The EM algorithm begins with an initial model λ to estimate a new model λI .The new model then becomes the initial model and the process is repeated till convergence. A proper initialization must be done for model parameters. The EM algorithm can be applied for refining the parameters with up dated equations.

The updated equations of the parameters for each Inverted Mel frequency cepstral coefficients are as follows

$$\alpha_k^{l+1} = \frac{1}{T}\sum_{i=1}^{T} p(i|\vec{x}_t, \lambda^l) \quad (6)$$

$$\mu_k^{l+1} = \frac{\sum_{i=1}^{T} p(i|\vec{x}_t,\lambda^l) + \sum_{i=1}^{T} \frac{f(x_M)-f(x_L)}{B-A}\sigma_k^2 p(i|\vec{x}_t,\lambda^l)}{\sum_{i=1}^{T} p(i|\vec{x}_t,\lambda^l)} \quad (7)$$

$$\sigma_k^{l+1} = \frac{\sum_{i=1}^{T} p(i|\vec{x}_t,\lambda^l)(\vec{x}_t - \mu_i^{l+1})^2}{c\sum_{i=1}^{T} p(i|\vec{x}_t,\lambda^l)} \quad (8)$$

Where c is given by

$$C = \frac{1}{(B-A)}\left(1 + \mu_i^{l+1}\right)\left[\left(f(x_M) - f(x_L)\right) + \left(x_M f(x_L)\right) - x_L f(x_M)\right]$$

$$f(x_M) = \int_{-\infty}^{x_M} b_i \vec{x}_t \, d\vec{x}_t \;\; , \;\; f(x_L) = \int_{-\infty}^{x_l} b_i \vec{x}_t \, d\vec{x}_t$$

$$f(x_M) = \int_{-\infty}^{x_M} b_i \vec{x}_t \, d\vec{x}_t \;\; , \;\; f(x_L) = \int_{-\infty}^{x_l} b_i \vec{x}_t \, d\vec{x}_t$$

The a posterior probability for acoustic class i is given by

$$p(i|\vec{x}_t, \lambda^l) = \frac{\alpha_i b_i(\vec{x}_t)}{\sum_{i=1}^{k} \alpha_i b_i(\vec{x}_t)} \quad (9)$$

3.3 Initialization of the Model Parameters

To utilize the EM algorithm we have to initialize the parameters μi, σi, and αi i=(1 …..M) and XM and XL obtained can be estimated with the values of the maximum and the minimum values of each feature vector respectively. The initial estimates of μi, σi, and αi of the ith component is obtained using method given by A.C Cohen(1950)[5]

3.4 Speakeridentification Algorithm

Once the speech spectrum of a speaker is observed the main purpose is to identify the speaker from the group of S speakers. The following algorithm can be adopted for speaker identification using Doubly Truncated Gaussian Mixture Model.[6]

1. Find feature vectors using front end process explained in section 1.
2. Divide the T samples in to M groups by K-means algorithm.
3. Find mean vector (μ_i), and variance vector(σ_i) for each group
4. Take α_i=1/M, I=1,2,3,4,5,…M initially
5. Use EM algorithm for obtained the refined estimates of μ_i, σ_i, and α_i for each component of i^{th} speaker.

6. Write the speaker Model as $p(x/\lambda) = \sum_{i=1}^{M} \alpha_i \, p_i(x/\lambda_s)$ where $\lambda_i = \mu_i, \sigma_i, \alpha_i$ put

   $\lambda_i = \lambda_1, \lambda_2, \lambda_3, \lambda_4 \ldots \lambda_6$ From each speaker.

7. For a speaker identification from a group of S speakers S={1,2,3,…S} each is represented by FDTGMM's with parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \ldots \lambda_s$. We find the speaker model which has the maximum a posterior probability for a given observation sequence that is

$$\hat{s} = max_{1<k<s} p_r(\lambda_k|x)$$
$$= arg\ max_{1<k<s} [p_r(\lambda_k|x) p_r(\lambda_k)]$$

4. IMPLEMENTATION AND RESULT

To demonstrate the ability of the developed model it is trained and evaluated by using a database of 45 speakers. For each speaker there are 10 conversations of approximately 2sec. each recorded in 10 separate sessions under different environmental conditions locally by using high quality Microphone. The test speech was first processed by front end analysis to produce a sequence of feature vectors (IMFCCs) which are obtained for test sequence length 2 seconds, with the procedure given by D.A Reynolds (1995).

The data set $\{\vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots \vec{x}_t\}$ is divided into a training set and a test set. Using the Hierarchical clustering algorithm for each speaker the sampled speech data is classified and using the moment estimators the initial estimate of the parameters are obtained.

With these initial estimates and the updated equations of the parameters given in section (3.2), the refined estimates of the parameters are obtained using these estimates. The global model for each speech spectrum is estimated. The efficiency of the developed model is studied by identifying the speaker with the Speaker identification algorithm given in section (3.4) with the test data set.

The percentage of correct identification is computed as

$$PCI = \%correct\ identification = \frac{\#correctly\ identified\ speakers}{total\#of\ speakers} X100$$

It is observed that this algorithm identifies the speaker with 97.4%±1.7 correctly

A comparative study of the Performance of FDTMGMM is carried with reference to the speaker modeling techniques. The other techniques are the uni modal Gaussian classifier given by H.Gish (1985), Tied Gaussian Mixture model given by J. Oglesby and J.Mason, (1985) and the Douglas A Reynolds(1995) with Gaussian Mixture Model using nodalvariance (GMMnv) and Gaussian Mixture Model using global variance (GMMgv) using Mel frequency cepstral co-efficient as feature vectors. The average percentage of correct identification for 45 speakers utterances of the models are computed with their confidence intervals and are presented in Table 1.

Table 1:Speaker Identification Performance forVarious Speaker Models

Speaker Model	% Correct identification (2 Sec test length)
GMM-nv(MFCC)	94.6 ±1.8
GMM-gv(MFCC)	89.7±2.4
TGMM	80.2±3.1
GC	67.3±3.7
FDTGMM(IMFCC)	97.4±1.7

5. CONCLUSION

In this paper the proposed text independent speaker identification model based on Finite Doubly truncated GMM with EM algorithm. The model parameters are estimated through EM algorithm after identifying the number of component densities in each speaker voice spectrum using inverted mel frequency cepstral co-efficient as feature vectors with the component maximum likelihood. The speaker identification algorithm is developed. Experimental results shown that the proposed model as better identification capabilities compared to the finite Gaussian mixture speaker model. This is also validated through a comparative study using speaker identification quality metrics, % of correct identification and its confidence interval. The developed model is much useful for robust text independent speaker identification in at places like banking by telephone, telephone shopping, Data Base access services, information services, voice interactive system, Security Control for confidential information areas and Remote access etc,..

6. REFERENCES

[1] Douglas A. Reynolds, Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Models" , IEEE Trans. on Speech and Audio Processing, January 1995, Vol. 3, No. 1, pp72-82.

[2] Herbert Gish, Michael Schmidt, Text Independent Speaker Identification, IEEE Signal Processing Magazine, Oct 1994.

[3] Ruchi Chaudhary "Performance study of Text-independent Speaker identification system using MFCC & IMFCC for Telephone and Microphone Speeches" International Journal of Computing and Business Research (IJCBR) ,ISSN (Online) : 2229-6166, Volume 3 Issue 2 May 2012 pp.1-11.

[4]R.Shantha Selva Kumari , S. Selva Nidhyananthan , Anand.G "Fused Mel Feature sets based Text-Independent Speaker Identification using Gaussian Mixture Mode" International Conference on Communication Technology and System Design 2011, Procedia Engineering 30 (2012) pp. 319 – 326

[5] Cohen A.C. Jr. "Estimating the Mean and Variance of Normal Populations from Singly and Doubly Truncated Samples", Ann.Maths. Statist., 21, pp 557-569.1950.

[6] V Sailaja , K Srinivasa Rao & K V V S Reddy "Text Independent Speaker Identification Using Finite Doubly Truncated Gaussian Mixture Model", International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 475-480.