# Text Independent Speaker Identification with Finite Multivariate Generalized Gaussian Mixture Model and Hierarchical Clustering Algorithm

V Sailaja
Department of Electronics & Communication Engineering,
GIET affiliated to JNTUK, Rajahmundry, India

K. Srinivasa Rao
Department of statistics, Andhra University
Visakhapatnam,
India

K.V.V.S. Reddy
Department of Electronics & Communication Engineering,
Andhra University
Visakhapatnam

## ABSTRACT

In this paper we propose a Text Independent Speaker Identification with Finite Multivariate Generalized Gaussian Mixture Model with Hierarchical Clustering. Each speaker speech spectra are characterized with a mixture of Generalized Gaussian Distribution includes Gaussian and Laplacian distribution as a particular case. It also includes several of the platy, lepto and meso kurtic shapes of the speech spectra. The speech analysis is done with Mel Frequency Cepstral Coefficients extracted from front end process. Using the EM algorithm the model parameters are estimated. The numbers of acoustic classes associated with each speech spectra are determined through Hierarchical clustering. The performance of the proposed algorithm is studied through experimental evolution with 100 speaker's data base and found that this algorithm outperforms the existing speaker identification algorithm with GMM. It is also observed that this algorithm performs efficiently even heterogeneous population with small (less than 2 seconds utterances)

**Key Words:** Generalized Gaussian Mixture Model, Mel frequency cepstral coefficients, EM algorithm Hierarchical clustering.

## 1. INTRODUCTION

The development of efficient-Speaker Identification system has been a topic of active research during last two decades because they have a large number of potential applications in many fields that require accurate user identification such as shopping by telephone, bank transaction, accesses control and voicemail etc,. The Speaker Identification system is divided into two parts namely, Text Independent Speaker Identification and Text Dependent Speaker Identification. Among these two, Text Independent Speaker Identification is more complicated in open test. In Text Independent speaker recognition systems the model based methods are more efficient.

In both the systems the characterisation of the speaker speech is more important. Speech can't be merely characterised as a sequence of sound units. There are some characteristics that lend naturalness to speech. The variation of pitch provides recognised melodic properties to speech. This controlled modulation of pitch is referred as intonation. The sound units shortened or lengthened in accordance some under laying pattern giving rhythmic properties to speech. Some syllables are words may be made more prominent than others, resulting in linguistic stress[19]. This information gleaned from melody, timing and stress in speech increases the intelligibility of spoken message, enabling the listener to segment continuous speech in to phrases and words with ease [33]. It is also capable of conveying many more lexical and non lexical information such as lexical tone, prominence, accent and emotion. The characteristics that make us perceive that these effects collectively referred to as prosody. A prosodic cue includes stress, rhythm and intonation.

Prosodic characteristics such as rhythm, stress and intonation in speech conveys some important information regarding the identity of the spoken language. Results of perception studies on human language identification conforms that prosodic information. Specifically pitch and intensity are used for language identification under conditions where the acoustic of sound units and phonotactics are degraded [23], [17]. A study using resynthesis has revealed the importance of rhythm and intonation for language discrimination [26]. Since each speaker has unique physiological characteristics of speech production and speaking style. Speaker specific characteristics are also reflected in prosodic. Distinguishing the language - specific and speaker – specific aspect of prosodic using acoustic parameters is even more difficult. Therefore, it is a challenging task to extract and represents prosodic features for recognizing a language or a speaker.

To model the speaker-dependent acoustic features within the individual phonetic sounds that comprise the utterance is done comparing acoustic features from phonetic sounds in a test utterance with speaker-dependent acoustic features from similar phonetic sounds, the comparison measures speaker differences rather than textual difference. This approach can be accomplished using explicit or implicit segmentation of the speech into phonetic sound classes prior to speaker model training or recognition. In [39] and [28], explicit segmentation was performed using a hidden Markov model (HMM)-based continuous speech recognizer as a front-end segmented for text-independent speaker recognition systems. It was found in both studies that the front-end speech recognizer provided little or no improvement in speaker recognition performance compared to no front-end segmentation. Moreover, using a continuous speech recognizer front-end imposes a significant increase in computational complexity on both training and recognition.

Implicit segmentation, on the other hand, relies on some form of unsupervised clustering to provide implicit segmentation of the acoustic features during both training and recognition. The sound classes are not labeled, so separate training of a segmented is not required. Template based clustering, such as vector quantization [2], [11] and if-nearest neighbor with leader clustering [3], has proven to be very effective for this approach to speaker recognition. In the VQ approach, each speaker is represented by a codebook of spectral templates representing the phonetic sound clusters in his/her speech. While this technique has demonstrated good performance on limited vocabulary (digits) tasks, it is limited in its ability to model the possible variability's encountered in an unconstrained speech task. As has been shown in speech recognition, probabilistic models provide a better model of acoustic speech events and a framework for dealing with noise and channel degradations. HMM's, in a variety of forms, have been used as probabilistic speaker models for both text-independent and text-dependent speaker recognition [21], [35]. The HMM models are not only the underlying speech sounds, but also the temporal sequencing among these sounds. Although temporal structure modeling is advantageous for text-dependent tasks, for text-independent tasks the sequencing of sounds found in the training data does not necessarily reflect the sound sequences found in the testing data and contains little speaker-dependent information. This is supported by experimental results in [25] and [35] which found text-independent performance was unaffected by discarding transition probabilities in HMM speaker models.

Another important approach of text independent speaker identification method is using the Gaussian mixture speaker model which falls into the implicit segmentation approach to speaker recognition. It provides a probabilistic model of the underlying sounds of a person's voice, but unlike HMM's does not impose any Markovian constraints between the sound classes. The probabilistic framework also allows the application of newly developed noise and channel robustness techniques from the speech recognition area. In [30] a statistical background noise model is integrated with the Gaussian mixture speaker model for noise robustness using this framework. Furthermore, the new model is computationally efficient and can easily be implemented on a real-time digital signal processor [8], [7].

Recently the Mel frequency-cepstral coefficients have gained importance in Speaker Identification to describe the signal characteristics. According to psychophysical studies [10], human perception of the frequency content of sounds follows a subjectively defined nonlinear scale called the Mel scale [4]. This is defined as,

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (1.1)$$

where, fmel is the subjective pitch in Mels corresponding to f, the actual frequency in Hz. This leads to the definition of MFCC, a base line acoustic feature speech and speaker recognition applications. The computation steps for extracting the MFCC are as follows: (a) Take the Fourier transform of (a windowed excerpt of) a signal. (b) Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows. (c) Take the logs of the powers at each of the mel frequencies. (d) Take the discrete cosine transform of the list of mel log powers, as if it were a signal. (e) The MFCCs are the amplitudes of the resulting spectrum.

The Mel frequency cepstral coefficients provide more precise results. D.A Reynolds (1994) [27] has developed a Text Independent Speaker Identification using Gaussian Mixture Model with Mel frequency cepstral co-efficients as feature vectors for speaker identification. The main drawback of Gaussian mixture model is that the individual Gaussian components assigned for the feature vectors are symmetric and meso kurtic. In most of the voice frames the feature vector may be platy kurtic or lepto kurtic. Neglecting the realities of the kurtic nature, the MFC coefficients lead to a serious falsification of the model estimation. So to have a close approximation to the realistic situations it is needed to generalize the Speaker Identification method with a more general mixture distribution, which includes the Finite Gaussian Mixture model as a particular case.

The Generalised Gaussian Distribution includes the Gaussian distribution as a particular case and it can be parameterized in such a manner that its mean μ and variance σ2 coincide with the mean and variance of Gaussian distribution. In addition to local and scaling parameters, the Generalised Gaussian Distribution is having another parameter (shape parameter), 'ρ' also which is the measure of peakedness of the distribution

The Generalised Gaussian Distribution was used by Sharif .K et al [34] for modelling the atmospheric noise sub band encoding of Audio and Video signals, [6] has used this distribution for impulsive noise direction of arrival and independent component analysis. Wu.H.C.Y. Principe. J [38] has used the distribution for signal separation. Varanasi M. K. et al [36] discussed the parameter estimation for the Generalized Gaussian Distribution by using methods of moments and maximum Likelihood. Armando. J et al (2003) [5] developed a procedure to estimate the shape parameter in Generalized Gaussian Distribution.

Very little work has been reported regarding Speaker Identification based on Finite Multivariate Generalized Gaussian Mixture Distribution. The Hierarchical clustering is used to obtain the number of acoustic classes (say M) of the speech and to get the initial estimates of the model parameters of the EM algorithm. Hierarchical clustering algorithm preserves the neighboring information among the clustered classes. The model parameters are estimated by deriving the updated equation of EM algorithm. The performance of the developed model is evaluated by obtaining the percentage of correct identification through experimentation. This model also includes speaker Identification with Gaussian Mixture Model and speaker identification with Laplace mixture model as a particular case.

## 2. FINITE MULTIVARIATE GENERALIZED GAUSSIAN MIXTURE SPEAKER MODEL

Consider the Mel frequency cepstral coefficients (MFCC) to represent the features vectors for speaker identification. The Mel frequency cepstral coefficients of each are assumed to follow a Finite Multivariate Generalized Gaussian Mixture Distribution. The motivation for considering the mixture models is that the individual component densities of a multi model density like the mixture model may model some underlying set of acoustic processes. It is reasonable to assume the acoustic space corresponding to a speaker voice can be characterized by 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 its acoustic class can in turn be represented by the mean of its component density and the variation of the average spectral shape can be represented by the co-variance matrix. Therefore the entire speech spectra of the each individual speaker can be characterized as a M component Finite Multivariate Generalized Gaussian mixture distribution.

The probability density function of the each individual speaker speech spectra is

$$p(\vec{x}_t|\lambda) = \sum_{i=1}^{M} \alpha_i b_i(\vec{x}_t|\lambda) \qquad (2.1)$$

where, $\vec{x}_t = (x_{tij})$ j=1,2,…,D; i=1,2,3…,M; t=1,2,3,…,T is a D dimensional random vector representing the MFCC vector

λ is the parametric set such λ = (μ,ρ,∑ )

$\alpha_i$ is the component weight such that $\sum_{i=1}^{M} \alpha_i = 1$

$b_i(\vec{x}_t|\lambda)$ is the probability density of ith acoustic class represented by MFCC vectors of the speech data and the D-dimensional Generalized Gaussian (GG) distribution (M..Bicego et al (2008))[15] and is of the form

$$\mathbf{b_i}(\vec{\mathbf{x}}_\mathbf{t}|(\mu,\rho,\Sigma) = \frac{[\mathbf{det}(\Sigma)]^{-1/2}}{[\mathbf{z}(\rho)\mathbf{A}(\rho,\sigma)]^{\mathbf{D}}}\,\mathbf{exp}\left(-\left\|\frac{\Sigma^{-\frac{1}{2}}\,(\vec{\mathbf{x}}_\mathbf{t}-\vec{\mu}_\mathbf{i})}{\mathbf{A}(\rho,\sigma)}\right\|_{\rho}\right) \qquad (2.2)$$

where, $z(\rho) = \frac{2}{\rho}\Gamma\left(\frac{1}{\rho}\right)$ and $\mathrm{A}(\rho,\sigma) = \sqrt{\frac{\Gamma(1/\rho)}{\Gamma(3/\rho)}}$

$$(2.3)$$

and $\|x\|_\rho = \sum_{i=1}^{D}|x_i|^\rho$ stands for the $l_\rho$ norm of vector *x*, Σ is a symmetric positive definite matrix. The parameter $\vec{\mu}_i$ is the mean vector, the function A (ρ) is a scaling factor which allows the var(x) = $\sigma^2$ and ρ is the shape parameter when ρ=1, the Generalized Gaussian corresponds to a laplacian or double exponential Distribution. When ρ=2, the Generalized Gaussian corresponds to a Gaussian distribution. In limiting case ρ→ + ∞ Equation (2.2) Converges to a uniform distribution in (μ-√3σ, μ+√3σ) and when ρ →o +, the distribution becomes a degenerate one when x=μ.

The mean value of the univariate Generalized Gaussian distribution is

$$\mathrm{E}(x_{ij}) = \mu_{ij} \qquad (2.4)$$

The variance of the variate $x_{ij}$ is

$$\mathrm{var(X)} = \sigma_{ij} \qquad (2.5)$$

The model can have one covariance matrix per a Generalized Gaussian density of the acoustic class of each speaker. The covariance matrix ∑ can also be a full or diagonal. In this chapter the diagonal covariance matrix is used for speaker model. This choice is based on the initial experimental results. Therefore

$$\Sigma_i = \begin{bmatrix} \sigma_{i1}^2 & \dots & 0 \\ \dots & \sigma_{i2}^2 & \dots \\ 0 & \dots & \sigma_{iD}^2 \end{bmatrix}$$

As a result of diagonal covariance matrix for the feature vector, the features are independent and the probability density function of the feature vector is

$$\mathbf{b_i}(\vec{\mathbf{x}}_\mathbf{t}|(\lambda) = \prod_{\mathbf{j=1}}^{\mathbf{D}} \frac{\mathbf{exp}\left(-\left|\frac{(\mathbf{x_{ij}}-\mu_{\mathbf{ij}})}{\mathbf{A}(\rho_{\mathbf{ij}},\sigma_{\mathbf{ij}})}\right|^{\rho_{\mathbf{ij}}}\right)}{\frac{2}{\rho_{\mathbf{ij}}}\Gamma\left(1+\frac{1}{\rho_{\mathbf{ij}}}\right)\mathbf{A}(\rho_{\mathbf{ij}},\sigma_{\mathbf{ij}})} = \prod_{\mathbf{j=1}}^{\mathbf{D}} \mathbf{f_{ij}}(\mathbf{x_{tij}})$$

# 3. ESTIMATION OF THE MODEL PARAMETER THROUGH EXPECTATION MAXIMIZATION ALGORITHM

For developing the speaker identification model it is needed to estimate the parameters of the speaker model. For estimating the parameters in the model, consider the EM algorithm which maximizes the likelihood function of the model for a sequence of i training vectors $\overrightarrow{x_t} = (x_1, x_2, \dots\dots\, x_t)$ drawn from a speaker's speech spectrum which is characterized by the probability density function

$p(\vec{x}_t|\lambda) = \sum_{i=1}^{M} \alpha_i b_i(\vec{x})$, where, $b_i(\vec{x}_t)$ is as given in equation (2.2) is

$$L(\lambda) = \prod_{t=1}^{T}\left[\sum_{i=1}^{M} \alpha_i\, b_i(\vec{x}_t,\lambda)\right]$$

$$\mathbf{L}(\lambda) = \prod_{\mathbf{t=1}}^{\mathbf{T}}\left(\sum_{\mathbf{i=1}}^{\mathbf{M}} \alpha_\mathbf{i}\left(\prod_{\mathbf{j=1}}^{\mathbf{D}} \frac{\mathbf{exp}\left(-\left|\frac{(\mathbf{x_{tij}}-\mu_{\mathbf{ij}})}{\mathbf{A}(\rho_{\mathbf{ij}},\sigma_{\mathbf{ij}})}\right|^{\rho_{\mathbf{ij}}}\right)}{\frac{2}{\rho_{\mathbf{ij}}}\Gamma\left(1+\frac{1}{\rho_{\mathbf{ij}}}\right)\mathbf{A}(\rho_{\mathbf{ij}},\sigma_{\mathbf{ij}})}\right)\right)$$

where $\|x\|_\rho$ is same as given in equation (2.3). Since the variance matrix is considered to be diagonal we have

$$L(\lambda) = \prod_{t=1}^{T}\left(\sum_{i=1}^{M} \alpha_i\left(\prod_{j=1}^{D} \frac{exp\left(-\left|\frac{(x_{tij}-\mu_{ij})}{A(\rho_{ij},\sigma_{ij})}\right|^{\rho_{ij}}\right)}{\frac{2}{\rho_{ij}}\Gamma\left(1+\frac{1}{\rho_{ij}}\right)A(\rho_{ij},\sigma_{ij})}\right)\right) \qquad (3.1)$$

This implies

$$\mathrm{logL}(\lambda) = \log \prod_{t=1}^{T}\left[\sum_{i=1}^{M} \alpha_i\, b_i(\vec{x}_t,\lambda)\right]$$
$$= \sum_{t=1}^{T} \log\left[\sum_{i=1}^{M} \alpha_i\, b_i(\vec{x}_t,\lambda)\right]$$
$$= \sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{log}\left[\sum_{\mathbf{i=1}}^{\mathbf{M}} \alpha_\mathbf{i}\left(\prod_{\mathbf{j=1}}^{\mathbf{D}} \frac{\mathbf{exp}\left(-\left|\frac{(\mathbf{x_{tij}}-\mu_{\mathbf{ij}})}{\mathbf{A}(\rho_{\mathbf{ij}},\sigma_{\mathbf{ij}})}\right|^{\rho_{\mathbf{ij}}}\right)}{\frac{2}{\rho_{\mathbf{ij}}}\Gamma\left(1+\frac{1}{\rho_{\mathbf{ij}}}\right)\mathbf{A}(\rho_{\mathbf{ij}},\sigma_{\mathbf{ij}})}\right)\right] \qquad (3.2)$$

To find the estimate of the parameters $\alpha_i$ $\mu_{ij}$ and $\sigma_{ij}$ for i= 1,2,3 …,M, j=1,2,…,D, we maximize the expected value likelihood (or) log likelihood function. Here the shape parameters '$\rho_{ij}$' is estimated by the procedure given by Armando.J el at (2003) [5] for each acoustic class of each speech spectra.

The likelihood function contains the number of components M which can be determined from the Hierarchical clustering algorithm. Once M is obtained from the Hierarchical clustering, the EM algorithm can be applied for refining the parameters with up dated equations. The updated equations of the parameters for each Mel frequency cepstral coefficients are as follows

The updated equation for estimating $\alpha_i$ is

$$\boldsymbol{\alpha}_i^{(l+1)} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\alpha_i{}^{(l)}\ \mathbf{b_i}\left(\vec{x}_t,\boldsymbol{\lambda}^{(\mathbf{l})}\right)}{\sum_{i=1}^{M}\alpha_i{}^{(l)}\ \mathbf{b_i}\left(\vec{x}_t,\boldsymbol{\lambda}^{(\mathbf{l})}\right)}\right]$$

Where $\lambda^{(l)} = \left(\mu_{ij}{}^{(l)}, \sigma_{ij}{}^{(l)}\right)$ are the estimates obtained at the i^th iteration.

The updated equation for estimating $\mu_{ij}$ is

$$\boldsymbol{\mu}_{\mathbf{ij}}{}^{(\mathbf{l+1})} = \frac{\sum_{t=1}^{T}\mathbf{t_i}(\vec{\mathbf{x}}_{\mathbf{t}},\boldsymbol{\lambda})^{(\mathbf{l})}{}^{A(N,\rho_i)}\ (\mathbf{x_{tij}}-\boldsymbol{\mu}_{\mathbf{ij}})}{\sum_{t=1}^{T}\mathbf{t_i}(\vec{\mathbf{x}}_{\mathbf{t}},\boldsymbol{\lambda}^{(\mathbf{l})})^{A(N,\rho_{ij})}}$$

where, $A(N,\rho_{ij})$ is some function which must be equal to unity for $\rho_i = 2$ and must be equal to $\frac{1}{\rho_{ij}-1}$ for $\rho_i \neq 1$, in the case of N=2, we have also observed that $A(N,\rho_{ij})$ must be an increasing function of $\rho_{ij}$.

The updated equation for estimating $\boldsymbol{\sigma}_{ij}$ is

$$\boldsymbol{\sigma}_{\mathbf{ij}}{}^{(\mathbf{l+1})} = \left[\frac{\sum_{t=1}^{N}\mathbf{t_i}(\vec{\mathbf{x}}_{\mathbf{t}},\boldsymbol{\lambda}^{(\mathbf{l})})\left(\frac{\Gamma\left(\frac{3}{\rho_{ij}}\right)}{\rho_i\Gamma\left(\frac{1}{\rho_i}\right)}\right)\left|\mathbf{x_{tij}}-\boldsymbol{\mu}_{\mathbf{ij}}{}^{(\mathbf{l})}\right|^{\frac{1}{\rho_{ij}}}}{\sum_{t=1}^{T}\mathbf{t_i}(\vec{\mathbf{x}}_{\mathbf{t}},\boldsymbol{\lambda}^{(\mathbf{l})})}\right]^{\frac{1}{\rho_{ij}}}$$

## 4. INITILIZATION OF THE MODEL PARAMETERS THROUGH HIERARCHICAL CLUSTRING

The efficiency of the EM algorithm in estimating the parameters is heavily dependent on the number of acoustic classes of the speaker speech data (M) and the initial estimates of the model parameters $\mu_{ij}$, $\sigma_{ij}$, $\rho_{ij}$ and $\alpha_i$ ( i = 1,2,…,M ; j=1,2,…,D). Usually in EM algorithm, the mixing parameter $\alpha_i$ and the distribution parameters $\mu_{ij}$, $\sigma_{ij}$ are given with some initials values. A commonly used method in initialization is by drawing a random sample in the entire speech data. This method can be performed well when the sample size is large, but the computation is heavily increased. When the sample size is small it is likely that some small regions may not be sampled. The initial value of the $\alpha_i$ can be taken as $\alpha_i$=1/M, where, M is obtained from the Hierarchical clustering Algorithm. After obtaining the final value of M ,we obtained the initial estimates of $\sigma_{ij}$, $\mu_{ij}$ and $\alpha_i$ for i^th acoustic class of each speaker speech spectrum using the sample speech spectra ,Mel frequency cepstral coefficient values classified by Hierarchical clustering algorithm. After getting the initial estimates, the final refined estimates of the Model parameters are obtained through EM Algorithm given in section (3).

## 5. SPEAKER IDENTIFICATION 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 Finite Multivariate Generalized Gaussian Mixture Model.

**Step- 1:** Find feature vectors using front end process explained in section (1) for each individual speaker speech spectra with MFCCs

**Step- 2:** Divide the T samples into M groups by Hierarchical clustering algorithm. Find mean vector ($\mu_{ij}$) and variances vector ($\sigma_{ij}$) for each acoustic class of each speaker. Take $\alpha_i$ = 1/M, i = 1,2,3,4,5,. . . ,M.

Step- 3: Obtaining the refined estimates of $\mu_{ij,}\alpha_i$, and $\sigma_{ij}$ for each class of the $i_{th}$ speaker using the updated equations of the EM algorithm.

**Step-4:** Estimate the Speaker Model as $p(x|\lambda) = \sum_{i=1}^{M}\alpha_i b_i(x|\lambda_i)$
where, $\lambda_i$ ={ $\mu_{ij}$ , $\sigma_{ij}$ , $\alpha_i$} and $\lambda$ = {$\lambda_1$, $\lambda_2$, … $\lambda_M$ } from each speaker .

**Step- 5:** For Speaker identification, from a group of S Speakers S={1,2,…,S} each represented by Finite Multivariate Generalized Gaussian Mixture Model with parameters $\lambda'_1,\lambda'_2,\lambda'_3,\ldots,\lambda'_s$ we find the speaker model which has the maximum a posteriori probability for a given observation sequence such that is

$$\begin{aligned}\hat{s} &= \max_{1<k<S} p_r(\lambda_k|X)\\ &= arg\max_{1<k<S}[p(\lambda_k|X)p_r(\lambda_k)]\end{aligned}$$

where, the second equation is due to baye’s rule assuming equally likely speakers i,e $P_r(\lambda_k)$= 1/S and noting that p(X) is the same for all speaker models, the classification rule simplifies to

$$\begin{aligned}\hat{s} &= arg\max_{1<k<S} p_r(\lambda'_k)\\ &= arg\max_{1<k<S}\sum_{i=1}^{T}\log p\left(\vec{x}_t|\lambda'_k\right)\end{aligned}$$

in which $p\left(\vec{x}_t|\lambda'_k\right)$ is as given in section 3.

## 6. EXPERIMENTAL RESULT

To demonstrate the ability of the developed model, it is trained and evaluated by using a database of 100 speakers. For each speaker there are 10 conversations of approximately 2sec.each recorded in 10 separate sessions. Out of which four – five sessions are used for training data and the remaining sessions used for testing data. The speaker’s speech data was recorded locally by using high quality Microphone.

The test speech was first processed by front end analysis to produce a sequence of feature vectors (MFCCs) which are obtained for test sequence length 2 seconds. With the procedure given by given by D.A Reynolds (1995)[7]. The data set $\overrightarrow{\{x_1},\vec{x}_2,\vec{x}_3,\vec{x}_4,\ldots,\vec{x}_T\}$ is divided into a training set and a test set.

Using the classified data for each speaker the initial estimate of the parameters is obtained using the Hierarchical clustering algorithm and the moment estimators. With these initial estimates and the updated equations of the parameters given in section (3), the refined estimates of the parameters are obtained. With these estimates the global model for each speaker density is estimated. With the test data set, the efficiency of the developed model is studied by identifying the speaker with the Speaker identification algorithm given in section (4).

The percentage of correct identification is computed as

PCI = % correct identification =

$$\frac{\# correctly\ identified\ spea\ker s}{total\ \# of\ \ spea\ker s} X100$$

It is observed that this algorithm identifies the speaker correctly with 97.6%.

The variation of PCI is also computed by repeating the experiment over 10 sessions under different environment conditions and using binomial distribution the confidence interval for the correct identification is computed as

$$APCI \pm z_{\alpha}\sqrt{\frac{\frac{APCI}{100}\left(1-\frac{APCI}{100}\right)}{n}}$$

where APCI represents average percentage of correctness $\frac{\sum_{i=1}^{n} PCI_i}{n}$ and n is the number of sessions, $z_{\alpha}$ is the significant value computed from the binominal probabilities for the given level of significance α.

**Table 1. Average percentage of correct Identification versus for various Speaker Identification Models**

| Speaker Model | %Correct identification<br>(2 Sec test length) |
|---|---|
| GMM-nv | 94.6±1.8 |
| GMM-gv | 89.7±2.4 |
| TGMM | 80.2±3.1 |
| GC | 67.3±3.7 |
| FDTMGMM<br>(K-means) | 96.4±1.7 |
| FDTMGMM<br>(Hierarchical clustering) | 97.1±1.6 |
| FMGGMM<br>(Hierarchical clustering) | 97.8±1.4 |

A comparative study of the Performance of Finite Multivariate Generalized Gaussian Mixture Model is carried with reference to the speaker modeling techniques. Specially the other techniques are the unimodel Gaussian classifier given by H.Gish (1985)[12],Tied Gaussian Mixture model given by J. Oglesby and J. Mason,(1991)[14] and the Gaussian Mixture Model using nodal variance($GMM_{nv}$) and Gaussian Mixture Model using global variance ($GMM_{gv}$) by Douglas A Reynolds (1995)[8], and Finite Doubly Truncated Gaussian Mixture Model[37] using Mel frequency cepstral co-efficient as feature vectors. The average percentage of correct identification for 100 speakers utterances of the models are computed with their confidence intervals and are presented in Table 1.

From, Table 1, it is observed that the average percentage of correct identification for the developed model is 97.8% ± 1.4. The percentage correctness for the Gaussian Mixture Model with nodal variance is 94.6%±1.8. This clearly shows that the speaker identification model with multivariate Generalized Gaussian Mixture Model is having higher average percentage of correct identification than the other models.

The experiment is also repeated with respect to the size of the speakers population by considering the speaker size as S=20, S=40, S=60, S=80, and S=100 with the same experimental set up of locally recorded speakers speech data base with 2 seconds utterance repeated over 10 sessions, in different environmental conditions, The average percentage of correctness of identification is computed and given in Table 2 and the relationship between the speaker population size and average percentage of correct identification is shown in Fig. 1

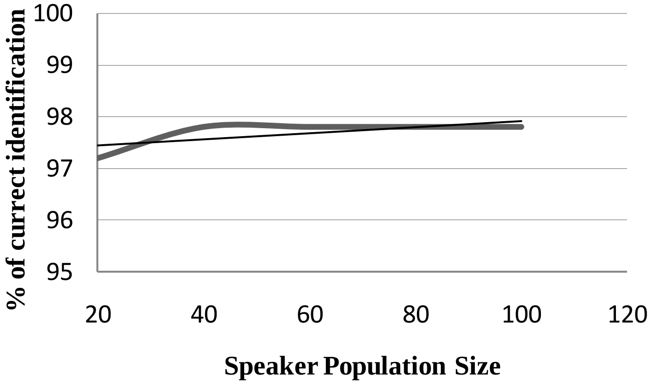

**Fig. 1 Speaker population size and average percentage of correct Identification**

A comparative study of the Performance of Finite Multivariate Generalized

From the Table 2 it is observed that the size of speaker population has an effect on percentage of correct identification. As the population size increases the average percentage of correctness is also increases and stabilizes after certain size.

**Table 2. Average percentage of correct identification versus speaker population size**

| Speaker population size | % Correct identification<br>(2 Sec test length) |
|---|---|
| 20 | 97.2±1.8 |
| 40 | 97.8±1 |
| 60 | 97.8±1.1 |
| 80 | 97.8±1.4 |
| 100 | 97.8±1.4 |

It is observed that the percentage of correctness for the developed model stabilizes when the speaker size approximately closer to 80. The proposed model is suitable for both homogeneous and heterogeneous speaker's population as it model each individual speaker uniquely.

Using the feature vectors derived from the test utterance, the evidence of different words at the output is noted. The evidence obtained for all the feature vectors in the test utterance are averaged to obtain the confidence scores for each word. Fig: 2 show the relationship between false alarm probability and miss probability of the proposed method. It is observed that this algorithm outperform the existence Text Independent Speaker Identification algorithms even in heterogeneous population with small utterance length.

## 7. CONCLUSIONS

In this paper a Text Independent Speaker Identification model is developed with the assumption that the feature vector associated with the speech spectra of each individual speaker follows a Finite Multivariate Generalized Gaussian Mixture Model. The Generalized Gaussian Mixture Model also includes Gaussian Mixture Model as a particular case.

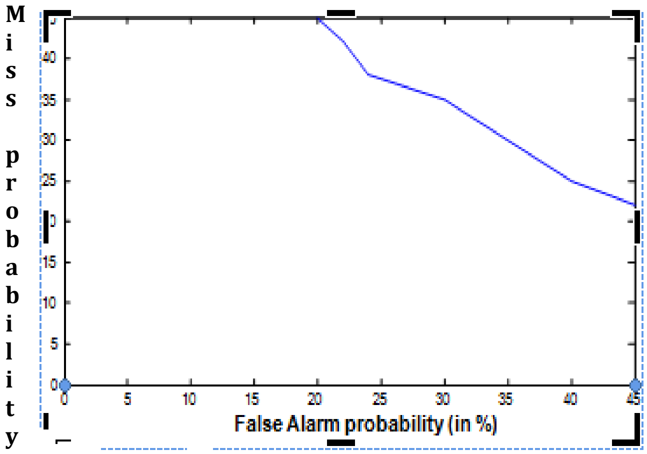

**Fig:2 Detection Error Tradeoff curve**

The Generalized Gaussian Mixture Model also includes the lepto kurtic or platy kurtic nature of the feature vector associated with each vocal class of individual speakers speech spectrum. It also includes Laplace mixture model. The Mel-frequency cepstral co-efficient of derived for the each speaker's speech data through front end procedure given by D A Reynolds (1995)[8]. The model parameters are obtained by deriving the up dated equations from the EM Algorithm associated with Finite Multivariate Generalized Gaussian Mixture Model. Using Hierarchical clustering algorithm and the diagonal nodal covariance matrix the initial estimates of the parameters are obtained. An experimentation with 100 speakers speech data revealed that this Text Independent Speaker Identification using Finite Multivariate Generalized Gaussian Mixture Model outperform the earlier existing Text Independent speaker identification models. It is also observed that this model perform much better even with large speaker data population sizes and independent of utterance (conversation) length.. The developed model can be applied for speaker identification like voice dialing, banking by telephone, telephone shopping, Forensic investigations, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers etc,.

## REFERENCES

[1] Akira kuremastu, Mariko Nakano-Mivatake, hector Perez-Meana, Eric Simancas Acevedo (2005), "performance analysis of Gaussian Mixture Model Speaker Recognition Systems with different speaker features," Electronic Journal Technical Acoustics, 14, ISSN 1819-2408.

[2] A. Higgins, L. Bahler, and J. Porter, (1993) "Voice identification using nearest-neighbor distance measure," in Proc. IEEE ICASSP,, pp. D-375-n-378.

[3] A. B. Poritz, (1982) "Linear predictive hidden Markov models and the speech signal," in Proc. IEEE ICASSP, , pp. 1291-1294.

[4] Ben Gold and Nelson Morgan (2002), "Speech and Audio Processing", Part IV , Chapter 14,pp 189 – 203 , John willy and sons.

[5] Armando. J et al (2003), "A practical procedure to estimate the shape Parameters in the Generalized Gaussian distribution".

[6] Choi S et al (2000), "Local Stability Analysis Of Flexible Independent Component Analysis Algorithm".Proceedings of 2000 IEEE international Conference on Acoustic speech and signal processing, ICA SSP 2000, PP. 3426- 3429.

[7] D. A. Reynolds, R. C. Rose, and M. J. T. Smith, (1992) "PC-based TMS320C30 implementation of the Gaussian mixture model text-independent speaker recognition system," in Proc. Int. Conf. Signal Processing Appl., Tech-no. l, pp. 967-973.

[8] Douglas A. Reynolds, and Richard C. Rose (1995), "Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Model," IEEE trans. Speech and Audio Processing, vol.3, pp. 72-83.

[9] Doddigntion.G. (2001) Speaker recognition based on idiolectic differences between speakers. In Proc. EUROSPEECH. Aalborg, Denmarks .pp 2521-2524.

[10] D O Shaaughnessy (1987), "Speech Communication Human and machine, Wesley publication, New York.

[11] F. Soong et al., "A vector quantization approach to speaker recognition," in Proc. IEEE ICASSP, 1985, pp. 387-390

[12] H. Gish et a (1985), "Investigation Of Text-dependent Speaker Identification Over Telephone Channels," in Proc. IEEE ICASSP, pp. 379-382.

[13] H. Gish et al.,(1986) "Methods and experiments for text-independent speaker cognition over telephone channels," in Proc. IEEE ICASSP, pp. 865-868.

[14] J . Oglesby and J. Mason (1991), "Radial basis function networks for Speaker Recognition," in Proceedings of IEEE ICASSP, pp. 393-396.

[15] JPool, J.A. du Preez. HF (1999) “Speaker Recognition. Thesis notes, Digital Signal Processing Group, Dept. of Electrical and Electronic Engineering, University of Stellenbosch. Acoustic Society of Japan (E), 20, 4, pp. 281-291.

[16] J. Market, B. Oshika, and A. Gray, Jr., (1977)"Long-term feature averaging for speaker recognition," IEEE Transaction Acoustic., Speech, Signal Processing, vol. ASSP-25, pp. 330- 337.

[17] Kometsu, M., Mori K., T. Arai, Murahara, Y., (2001) “ Human Language identification with reduced segmental information: Comparison between Monolinguals and bilinguals. In: Proc. EUROSPEECH, vol. 1, Scandinavia, pp.149-152.

[18] K.P.Markov, S.Nakagawa (1999), “Integrating pitch and LPC-residual Information with LPC- Cepstral for Text Independent Speaker Recognition.

[19] Leena Mary and B.Yegnanarayana, (2008) Extraction and representation of prosodic features for language and speaker recognition. speech communication 50 pp. 782, 796.

[20] L. Rudasi and S. A. Zahorian, (1991) "Text-Independent talker identification with neural networks," in Proc. IEEE ICASSP, pp. 389-392.

[21] L.Baum et al., (1970) "A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math Stat.*, vol.41, pp. 164-171.

[22] Md M. Bicego, D Gonzalez, E Grosso and Alba Castro (2008) “Generalized Gaussian distribution for sequential Data Classification” IEEE Trans. 978 -1- 4244-2175-6.

[23] Mori K. Toba N, Harada. T. Arai, Kometsu, M., Aoyagi, M., Murahara, Y., (1999) Human language identification with reduced spectral information. In Proc. EUROSPEECH. Vol.1. Budapest, Hungary. pp.391,394.

[24] Mclanchan G. and Krishan T (1997), “The EM Algorithm and Extensions”, John Wiley and Sons, New York – 2000.

[25] N. Z. Tishby, (1991)"On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 563-570.

[26] Ramus.F., Nespor, M., Mehler, J.,(1999). Correlates of linguistic rhythm in speech signal. Cognition 73(3), pp.265-292.

[27] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, (1994) "Integrated models of speech and background with application to speaker identification in noise," IEEE Trans. Speech Audio Processing, vol. 2, no. 2, pp. 245-257.

[28] R. E. Helms, (1981) "Speaker recognition using linear predictive vector code-books," Ph.D. thesis, Southern Methodist University.

[29] R. Rajeswara Rao, A. Nagesh, Kamakshi Prasad, K. Ephraim Babu (2007), “Text- Dependent Speaker Recognition System for Indian Languages”, IJCSNS International Journal of Computer Science and Network Security, vol.7 No.11, pp. 65 – 71.

[30] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds (1994), "Integrated models of speech and background with application to speaker identification in noise”,

[31] S. Furui (1981), "Cepstral analysis technique for automatic speaker verification, " IEEE Trans. Acoustic., Speech, Signal Processing, vol. ASSP-29, pp. 254- 272. IEEE Trans. Speech Audio Processing, vol. 2, no. 2, pp. 245-257.

[32] S. Furui, F. Itakura, and S. Saito, (1972) "Talker recognition by longtime averaged speech spectrum," Electron., Commun. in Japan, vol. 55-A, no. 10, pp. 54-61.

[33] Shriberg.E., Stolcke, A.,Hakkani- ur,D.,Tur,G.,(2000) “Prosody-based automatic segmentation of speech into sentences and topics” Speech Comm. Pp. 32,127-154.

[34] Sharif k etal(1995), “Estimation of shape parameters for generalized Gaussian Distribution in Sub band decomposition or video”, IEEE transaction on circuit systems vol.5 no.1 pp.52-56.

[35] T. Matsui and S. Furui, (1992) "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," in Proc.IEEE ICASSP, pp.n.157-11.164.

[36] Varanasi.MK et al(1989), “Parametric generalized Gaussian densit Estimation”, Journal Acoust. Soc.AM 86(4) pp. 1404.

[37] V.Sailaja (2010), “Some Studies on Text Independent Speaker Identification Models with Generalizations of Finite Gaussian Mixture Models”, unpublished Thesis notes Department of Electronics and Communication Engineering, Andhra University, Visakhapatnam.

[38] Wu.H.C.Y Principe J (1998), “Minimum entropy algorithm for source separation” proceedings of the Midwest symposium on system and circuits.

[39] Y. Kao, P. Rajasekaran, and J. Baras, (1992)"Free- Text speaker identification over long distance telephone channel using hypothesized phonetic segmentation," in Proc. IEEE ICASSP, pp. II. 177-11. 180.

[40] Y. Bennani and P. Gallinari, (1991) "On the use of TDNN-extracted features information in talker identification," in Proc. IEEE ICASSP, pp. 385-388.