Text Independent Speaker Identification with Finite Multivariate Generalized Gaussian Mixture Model with Distant Microphone Speech

K.Jyothi
Department of Electronics & Communication Engineering,
GIET affiliated to JNTUK,
Rajahmundry, India

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

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

ABSTRACT
An effective and efficient speaker Identification (SI) system requires a robust feature extraction module followed by a speaker modeling scheme for generalized representation of these features. In recent, years Speaker Identification has seen significant advancement, but improvements have tended to be bench marked on the near field speech, ignoring the more realistic setting of far field instrumented speaker. A novel speaker model is developed by using Finite Multivariate Generalized Gaussian Mixture Model, Minimum Variance Distortion less Response Cepstral coefficients as feature Vectors. The performance of the developed model is studied through experimental evaluation with 45 speaker's data base and identification accuracy.

Key Words: Generalized Gaussian Mixture Model, Minimum Variance Distortion less Response Cepstral coefficients, EM algorithm

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 authentation 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.

Any speaker identification system needs a robust acoustic feature extraction technique as a front end block followed by an effective and efficient modeling scheme for generalized representation of these features. MFCC[4] [10] has been widely accepted front end block for typical speaker identification applications as it is less vulnerable to noise perturbation gives less session variability and is copy to extract.

Speaker identification (SI) methods can be divided into text independent and text dependent methods. In a text independent system, speaker models capture characteristic of speaker's speech which show up irrespective of what one is saying. In a text dependent system, on the other hand, the recognition of the speaker's identity is based on his or her speaking one or more specific phrases like, passwords, card numbers, PIN codes etc .Whether text independent or text dependent, each has its own advantages and disadvantages and any require different modeling techniques. The choice of which technologies to use is application specific [4].

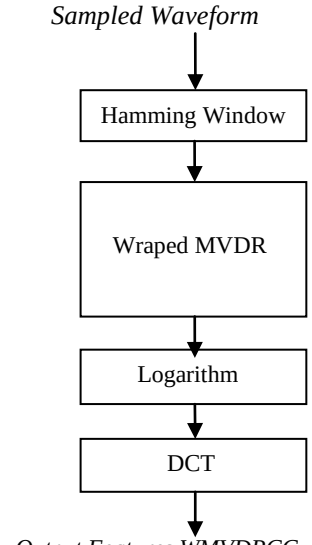
But MFCC was first proposed for speech recognition [13] to identify monosyllabic words in continuously spoken sentences and not for Speaker Identification SI. Also, calculation of MFCC is based on the human auditory system aiming for artificial implementation of the ear physiology [9] assuming that the human ear can be a good speaker recognizer too. The MFCC techniques make use of two types of filters namely linearly special filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz.

Thus MFCC can represents the low frequency region more accurately than the high frequency region and hence it can capture formant [5] which lie in the low frequency range and which characterize vocal tract resonances [6] .However, other formant can also lie above 1KHz and these are not effectively captured by the longer spacing of filters in the higher frequency range. Recently MVDR Cepstral coefficients as features used for speaker identification.

1.1 Speech 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. LPC & MFCC are commonly used feature extraction techniques. But in adverse conditions such as noisy environment and for medium and high pitched noise the above two will not provide good spectral estimate [4]. This is because LP based all – pole filters tend envelop the spectrum as tightly as possible and will under certain conditions descend down to the level of residual noise in the gap between harmonic partials, which will happen when the space between partials is large, as in high pitched sounds, and when the order is high enough (when there are large number of poles to cover every partial peak).Therefore filter which yields spectral estimate with smoother contour than all pole filters are desired. A best alternative to above techniques is MVDR, minimum variance distortion less response cepstral

coefficients which provides superior modeling for medium and high pitched voices.

The MVDR coefficients can be computed by using steps shown in Fig: 1



2. THE GENERALISED GAUSSIAN MIXTURE MODEL

This section describes the form of the Generalized Gaussian Mixture Model (GGMM) and motivates its use as a representation of speaker for text-independent speaker identification. Previously Douglas A.Reynold (1995) et al

[3],Sandipan Chakraborty(2006) et al[6] and Qin Jin, RUnix Li (2010) et al [8] developed speaker identification models using GMM using MFCC as feature vectors.

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 feature vectors leads 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 σ^2 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 [12] for modelling the atmospheric noise sub band encoding of Audio and Video signals, Wu.H.C.Y. Principe. J [15] has used the distribution for signal separation. Varanasi M. K. et al [14] discussed the parameter estimation for the Generalized Gaussian Distribution by using methods of moments and maximum Likelihood. Armando. J et al (2003) [1] 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 K-Means algorithm is used to obtain the number of acoustic classes of the speech and to get the initial estimates of the model parameters of the EM algorithm. K-Means 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.

We use the Minimum variance Distortion less Response cepstral coefficients (MVDR) to represent the features vectors for speaker identification. The MVDR Cepstral coefficients of each are assumed to follow a Finite Generalized Gaussian Mixture Distribution.

3. MODEL DESCRIPTION

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

$$p(\vec{x}_t|\lambda) = \sum_{i=1}^{M} \alpha_i b_i(\vec{x}_t|\lambda) \qquad (3.1)$$

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 MVDR vector λ is the parametric

set such λ={ μ ,σ,α}

$b_i(\vec{x}_t|\lambda)$ is the probability density of i^{th} acoustic class represented by MVDR vectors of the speech data and the *D*-dimensional Generalized Gaussian (GG) distribution (M..Bicego et al (2008)) [16] and is of the form

$$\mathbf{b_i(\vec{x}_t|(\lambda) = \prod_{j=1}^{D} \frac{\exp\left(-\left|\frac{(x_{ij}-\mu_{ij})}{A(\rho_{ij},\sigma_{ij})}\right|^{\rho_{ij}}\right)}{\frac{2}{\rho_{ij}}\Gamma\left(1+\frac{1}{\rho_{ij}}\right)A(\rho_{ij},\sigma_{ij})} = \prod_{j=1}^{D} f_{ij}(x_{tij})} \qquad (3.2)$$

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

and $\|x\|_\rho = \sum_{i=1}^{D}|x_i|^\rho$ stands for the l_ρ 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) = σ^2 and ρ is the shape parameter.

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all

component densities. These parameters are collectively represented by the notation

$\lambda_i = \{ \mu_{ij}, \sigma_{ij}, \alpha_i \}$

The α_i are the mixture weights satisfying stochastic constraints $\sum_{i=1}^{M} \alpha_i = 1$

4. 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 $\vec{x}_t = (x_1, x_2, \ldots \ldots \; x_t)$ 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 (3.1) is

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

$$\mathbf{L(\lambda)} = \prod_{\mathbf{t=1}}^{\mathbf{T}} \left(\sum_{\mathbf{i=1}}^{\mathbf{M}} \boldsymbol{\alpha}_{\mathbf{i}} \left(\prod_{\mathbf{j=1}}^{\mathbf{D}} \frac{\exp\left(-\left| \frac{(\mathbf{x_{tij}} - \boldsymbol{\mu}_{\mathbf{ij}})}{\mathbf{A}(\boldsymbol{\rho}_{\mathbf{ij}}, \boldsymbol{\sigma}_{\mathbf{ij}})} \right|^{\boldsymbol{\rho}_{\mathbf{ij}}} \right)}{\frac{2}{\boldsymbol{\rho}_{\mathbf{ij}}} \boldsymbol{\Gamma}\left(1 + \frac{1}{\boldsymbol{\rho}_{\mathbf{ij}}}\right) \mathbf{A}(\boldsymbol{\rho}_{\mathbf{ij}}, \boldsymbol{\sigma}_{\mathbf{ij}})} \right) \right)$$

where $\|x\|_\rho$ is same as given in section 3. Since the variance matrix is considered to be diagonal we have

$$\boldsymbol{L(\lambda)} = \prod_{\boldsymbol{t=1}}^{\boldsymbol{T}} \left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{M}} \boldsymbol{\alpha}_{\boldsymbol{i}} \left(\prod_{\boldsymbol{j=1}}^{\boldsymbol{D}} \frac{\boldsymbol{exp}\left(-\left| \frac{(\boldsymbol{x}_{tij} - \boldsymbol{\mu}_{ij})}{\boldsymbol{A}(\boldsymbol{\rho}_{ij}, \boldsymbol{\sigma}_{ij})} \right|^{\boldsymbol{\rho}_{ij}} \right)}{\frac{\mathbf{2}}{\boldsymbol{\rho}_{ij}} \boldsymbol{\Gamma}\left(\mathbf{1} + \frac{\mathbf{1}}{\boldsymbol{\rho}_{ij}}\right) \boldsymbol{A}(\boldsymbol{\rho}_{ij}, \boldsymbol{\sigma}_{ij})} \right) \right)$$

This implies

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

$$= \sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{log} \left[\sum_{\mathbf{i=1}}^{\mathbf{M}} \boldsymbol{\alpha}_{\mathbf{i}} \left(\prod_{\mathbf{j=1}}^{\mathbf{D}} \frac{\exp\left(-\left| \frac{(\mathbf{x_{tij}} - \boldsymbol{\mu}_{\mathbf{ij}})}{\mathbf{A}(\boldsymbol{\rho}_{\mathbf{ij}}, \boldsymbol{\sigma}_{\mathbf{ij}})} \right|^{\boldsymbol{\rho}_{\mathbf{ij}}} \right)}{\frac{2}{\boldsymbol{\rho}_{\mathbf{ij}}} \boldsymbol{\Gamma}\left(1 + \frac{1}{\boldsymbol{\rho}_{\mathbf{ij}}}\right) \mathbf{A}(\boldsymbol{\rho}_{\mathbf{ij}}, \boldsymbol{\sigma}_{\mathbf{ij}})} \right) \right]$$

To find the estimate of the parameters α_i μ_{ij} and σ_{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 'ρ_{ij}' is estimated by the procedure given by Armando.J el at (2003) [1] for each acoustic class of each speech spectra.

The likelihood function contains the number of components M which can be determined from the histogram of the MVDR Cepstral coefficients associated with the speaker and counting the number of peaks. Once M is obtained from the K- means clustering, the EM algorithm can be applied for refining the parameters with updated equations. The updated equations of the parameters for each MVDR Cepstral coefficients are as follows

The updated equation for estimating α_i is

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

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 μ_{ij} is

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

where, $\mathrm{A}(N, \rho_{ij})$ 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 $\mathrm{A}(N, \rho_{ij})$ must be an increasing function of ρ_{ij}.

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

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

This technique involves iterative update of each of the parameter s μ_{ij} , σ_{ij} , α_i.In the present work initialization of feature vectors were done by the K-Means algorithm which was terminated after 5 iterations. This was followed by the E&M algorithm with 20 iterations. For all cases diagonal covariance matrices were chosen.

5. SPEAKER IDENTIFICATION

Once the speech spectrum of a speaker is observed the main purpose is to identify the speaker from the group of S speakers. For speaker identification a group of S speakers S={1,2,…,S} is represented by FMGGMM's $\lambda_1, \lambda_2, \lambda_3, \ldots \lambda_S$[3].

The objective is to find the speaker model which has the maximum a posterior probability for a given observation sequence. Formally

$\hat{s} = \max_{1<k<S} p_r(\lambda_k|X)$

$\hat{s} = arg \max_{1<k<S} \frac{p_r(X|\lambda_k) p_r(\lambda_k)}{p(X)}$

Where the second equation is due to Baye's rule. Assuming equally likely speaker (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

$$\hat{s} = \max_{1<k<S} p_r(X|\lambda_k)$$

Using logarithms and the independence between observations, the speaker identification system computes

$$\hat{s} = arg\max_{1<k<S} \sum_{i=1}^{T} \log p\left(\vec{x}_t|\lambda_k'\right)$$

in which $p(\vec{x}_t|\lambda_k)$ is as given in section (3)

which has the maximum a posteriori probability for a given observation sequence such that is

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

6. EXPERIMENTAL RESULTS

To demonstrate the ability of the developed model, it is trained and evaluated by using a database of 45 speakers. For each speaker 10 conversations in 6 sessions were recorded at several distances from the speaker location. Out of which four – five sessions are used for training data and the remaining sessions used for testing data.

We define two train-test conditions:

Long-Long: 90-sec of training and 30-sec of test.

Short-Short: 30-sec of training and 10-sec of test

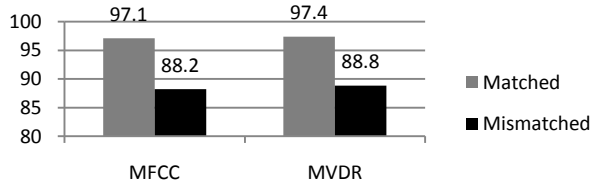


Fig 2: SID accuracy (Long-long)

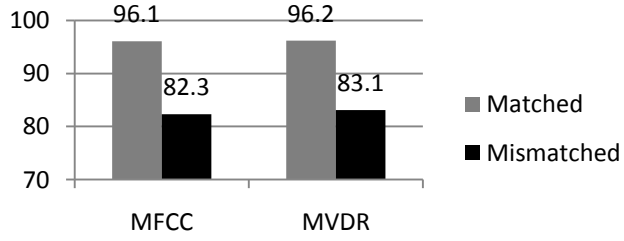


Fig 3: SID accuracy (Short-short)

The results under the Long-Long train and test is shown in Fig 2.From Fig 2 we can see that MVDR features achieves

better performance than MFCC features under mismatched condition. Fig 3 show the results under the short-short train-test condition where less audio was used for train and test has and somewhat less accuracy than long-long

The test speech was first processed by front end analysis to produce a sequence of feature vectors (MVDR) Cepstral Coefficients. The data set (feature vectors) $\{\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4, \ldots, \vec{x}_T\}$ is divided into a training set and a test set.

With the test data set, the efficiency of the developed model is studied by identifying the speaker with the Speaker identification process given in section (5).

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 model identifies the speaker correctly with 97.4%.

Table 1

SID accuracy (Long-Long and Matched)

	Baseline	FMGGMM(K-means)
MFCC	96.03%	97.1%
MVDR	95.73%	97.4%

Table 2

SID accuracy (Long-Long and Mismatched)

	Baseline	FMGGMM(K-means)
MFCC	85.81%	88.2%
MVDR	87.72%	88.8%

Tables 1 and 2 compare the speaker identification performance under Long-Long train-test condition of the baseline system with GMM/UBM speaker modeling and the system with FMGGM is applied. We can see that FMGGM provides significant improvements over the baseline GGM/UBM speaker modeling.

7. CONCLUSIONS

This paper has introduced and evaluated the use of Finite Multivariate Generalized Gaussian Mixture Model for text-independent speaker identification. The FMGGM model was evaluated for identification tasks with speech data base of 45 speakers using short duration and long duration utterances from unconstrained conversational speech. MVDR Cepstral coefficients as feature for speaker feature extraction. Experimental results shows that the developed model out perform the traditional speaker identification model. The primary focus of this work was on a task domain for real application, such as employee facing and customer or public facing. The dominant employee facing

application is password reset. Other applications includes speaker identification for field service reporting, physical and data access, wire transfer ,access to internal/secured internal telephone calls. The dominant customer or public facing applications are account access and related transactions.

REFERENCES

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

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

[3] Douglas A. Reynolds, member, IEEE and Richard C.Rose, Member, IEEE 'Robust text- Independent Speaker Identification Using Gaussian Mixture Speaker Models'

[4] EL-Jaroudi,A. and Makhoul,J (1991)" Discrete all-pole modelling" .IEEE rans. Speech Processing, vol.39:pp.411-423.

[5] J.P. Cambell. Jr(1997)."Speaker Recognition: A Tutorial", Processing of the IEEE, vol.85, no9, pp. 1437-1462..

[6] 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.

[7] Md.Rashidul Hasan, Mustafa Jamil, Md Golam Rabbani Md.Saifar Rahman (2004) "Speaker identification using Mel frequency cepstral coefficients" ICECE 2004, 28-30 December, ISBN 984- 32-1804-4, pp 565 to568.

[8] Qin Jin ,Ruxxin Li,Qian Yang,Kornel Laskowski, Tanja Schultz (2010). 'Speaker Identification with Distant Microphone Speech' IEEE978-1-4244-4296-6/10.

[9] Rabiner .Ljuang B.H(2003) "fundamentals of speech reconition",Chap.2,pp.11-65, Pearson Education, First Indian reprint.

[10] R.Vergin,B.OShaughnessy and A. Farhat (Sep.1999) "Generalize Mel frequency cepstral coefficient for large-vocabulary speaker independent continuous speech recognition, IEEE Trans. On ASSP, vol.7, no.5, pp.525-532.

[11] Sandipan Chakraborthy, Anindya Roy and Goutam Saha (2006) "Improved Closed Set text-Independent Speaker identification by combining MFcc with Evidence from Flipped Filter banks'IJSp vol 4,no 2

[12] 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.

[13] U.G.Goldstein(1976.) "speaker Identifying features based on formant tracks" J.Acoust, Soc.am, vol.59, No.1,pp.176-182.

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

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