International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 475-480

TEXT INDEPENDENT SPEAKER IDENTIFICATION USING FINITE DOUBLY TRUNCATED GAUSSIAN MIXTURE MODEL

V Sailaja[1], K Srinivasa Rao[2] & K V V S Reddy[3]

Speaker identification is the problem of deciding who is speaking in a given utterance. When the actual sequence of words that are spoken is known then the problem is test dependent and when it is unknown the problem is test independent. Several Test independent speaker identification models are proposed by using Gaussian mixture model (GMM).They assumed that the feature vector associated with individual speaker is having infinite range and symmetric. However, many of the feature vectors associated with individual speaker identification are having finite range and asymmetrically distributed. In this paper Test Independent Speaker Identification model is developed using Finite Doubly Truncated Gaussian Mixture Distribution. The speech analysis is done with Mel frequency cepstral coefficients for the voice spectrum. Using the EM algorithm the model parameters are estimated. The speaker identification algorithm is developed. The performance of the model is studied through experimental evaluation with 50 speaker's data base and identification accuracy. This model performance is much better than the earlier speaker identification model with GMM.

Keywords: Gaussian Mixture Model, Mel Frequency Cepstral Coefficients, EM Algorithm

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 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. Several authors have developed different text independent speaker identification systems Sadaoki Furi (1981), Douglas A.Reynolds and Richard C.Rose (1995) A.Kirankurematsu, etal(2005),Nemat.S and Abdel Khader (2008),Leena Marry and B.Yegnanarayana (2008) A.Revathi et al (2009) Md.Rabiul Islam and Md.fayzur Rahman (2009) Sandipan Chakroborty and Goutam Saha (2009) were developed and analysed Text Independent Speaker Identification with Gaussian Mixture Models.

One of the widely used stochastic methods for speaker recognition task is the Gaussian mixture model which is similar to be HMM with the difference that the GMM omits the temporal information implicit in HMM (K P Markov (1999) and Pool J etal (1999)). The Gaussian mixture model has only one state and does not meet the transition time from one state to another state as required in the HMM and also it does not impose Morkovian constrains. The Gaussian mixture model uses unique Gaussian mixture distribution to represent each speaker. In addition in most cases it is not necessary to use all co-variance matrix components because all Gaussian components are acting together to model the overall probability density function. Then the full co-variance is not necessary even if the features are not statistically independent. This is because to take all components of the co-variance of the matrix is equivalent to take only the main diagonal of the co-variance matrix from each speaker model (Akira kurematsu(2005)). Thus the GMM has been widely used in text independent speaker recognition system because it decides its desirable features mentioned above. It has the capacity of representing broad acoustic classes with its individual Gaussian components. In Gaussian mixture model the performance strongly depends on the characterization of the feature vector that allows unambiguous representation of the pattern under analysis. Usually the speaker characteristic is estimated through linear prediction of the speech signal. The reason for this is the structure of the vocal tract can be satisfactorily represented by using these parameters. However, it has been reported that better performance can be obtained with cepstral analysis which allows to get a robust speaker characterization with low sensitivity to the distortion introduced in the signal transmitted through communication channel (D.A Reynolds (1995)).

Recently the mel frequency -cepstral coefficients used in speaker identification to describe the speech

[1]Department of Electronics & Communication Engineering, GIET, Rajahmundry, India

[2]Department of statistics, Andhra University, Visakhapatnam, India

[3]Department of Electronics & Communication Engineering, A U, Visakhapatnam, India

Email: sailajagiet@gmail.com, ksraoau@gmail.com

characteristics. According to psychophysical studies (D.O' Shaughnessy,(1987)),human perception of the frequency content of sounds follows a subjectively defined nonlinear scale called the Mel scale (Ben Gold and Nelson Morgan (2002)). This is defined as,

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

where f_{mel} is the subjective pitch in Mels corresponding to f, the actual frequency in Hz.

D.A.Reynolds(1994), D.Hard and K.Feiibaum and Daniel j mashao(2006) have used mel frequency cepstral coefficients as base line Acoustic feature for text independent Speaker Identification. The feature vectors associated with the Text independent Speaker Identification can be computed by using the steps shown in Fig.1

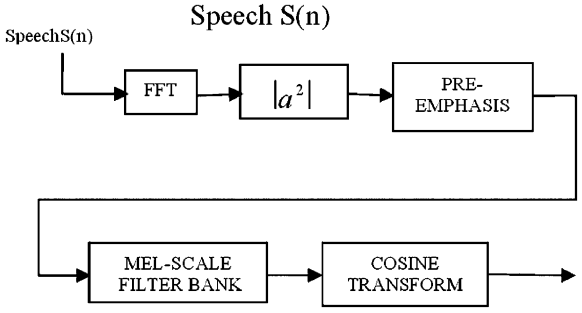

Fig.1: Feature Vector Extraction

In all these methods they consider that the Mel frequency cepstral co-efficient associated with each Speaker follows a Gaussian or finite Gaussian Mixture Model. These Speaker Identification models serve well when the Mel frequency cepstral co-efficient associated with the Speakers are having infinite range and symmetrically distributed misokurtosis. But in some of the voice frames the elements in the feature lies between two finite values and in some cases the distribution of the feature vector may be asymmetric and skewed. Neglecting the realities of the finite range to the MFC coefficient leads to a serious falsification of the model estimation. So to have a robust model for the MFC coefficients. It is needed to consider a Finite doubly truncated Gaussian distribution which characterizes the features distribution for each speaker as finite doubly truncated Gaussian mixture Model.

The FDTGMM includes the GMM also as a limiting case. It also includes the skewed nature of the component in the model. The effect of truncation in Gaussian distribution has been discussed by several researchers in other application areas (Cohen A.C. (1950) Johnson A.C.(1996) Matto R.S (2000)). But, no serious attempt is made to develop and analyse text independent speaker identification model with finite doubly truncated Gaussian Mixture Model. Using the k-means algorithm the number of components in recognising the speech spectrum of the individual speaker is given. The model parameters are estimated using EM algorithm. The efficiency of the model is compared with that of the text independent speaker identification model with GMM given by (D A Reynold (2006)) through experimental results and rate of accuracy.

2. DOUBLY TRUNCATED GAUSSIAN MIXTURE SPEAKER MODEL

In this section we briefly describe the DTGMM 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 Mel frequency cepstral coefficients of each speaker spectrum as the features for speaker identification. The Mel frequency cepstral coefficients are assumed to follow a DTGMM. 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 co-variance 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 DTGMM 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.

Gaussian full co-variance is not necessary even is the features are not statistically independent.

The probabilities density functions of the finite M component Doubly Truncated Gaussian Mixture distribution is

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

The Doubly truncated D variate Gaussian density is

$$b_i(\vec{x}) = \frac{1}{(B-A)(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x_i}-\vec{\mu_i})\Sigma_i^{-1}(\vec{x_i}-\vec{\mu_i})\right\} \qquad (2)$$

where, $\vec{x}$ is a D dimensional random vector ($\vec{x}_t = (x_1, x_2, \ldots\ldots x_t)$) is the future vector, $\vec{\mu}_i$ is the i^{th} component feature mean vector, Σ_i is the i^{th} component of variance co- variance matrix and

$$A = \int_{-\infty}^{xL} \ldots \int_{-\infty}^{xL} b_i(\vec{x}_t)\vec{d}x_t$$

$$B = \int_{-\infty}^{xM} \ldots \int_{-\infty}^{xM} b_i(\vec{x}_t)\vec{d}x_t$$

$$E(\vec{x}) = \mu_i + \sigma_i^2 \left[\frac{f(\vec{x}_L) - f(\vec{x}_M)}{\emptyset(\vec{x}_L) - \emptyset(\vec{x}_M)} \right] \quad (3)$$

where, $\emptyset(\vec{x}_L)$ are $\emptyset(\vec{x}_M)$ the standard normal areas and $\vec{x}_L, \vec{x}_M$ are the lower and upper truncated points of the feature vectors.

$b_i(\vec{x})$, , i = 1 ……. M are the component densities and $\alpha_i(\vec{x})$, i = 1 ……. M are the mixture weights, with mean vector.

The variance of each feature vector (Mel frequency cepstral coefficients) is Σ with diagonal elements as

$$V_i^2 = \left[1 + \frac{\left(\frac{\vec{x}_L - \vec{\mu}_i}{\sigma_i}\right)\vec{x}_L - \left(\left(\frac{\vec{x}_L - \vec{x}_i}{\sigma_i}\right)\vec{x}_M\right)}{B - A} \right] \sigma_i^2 \quad (4)$$

The mixture weights satisfy the constraints $\Sigma_{i=1}^{M} \alpha_i = 1$

Then the FDTGMM is parameterized by the 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 = 1,2, ...M.

For speaker identification each speaker is represented by FDTGMM and is referred to by his /her model parameter λ. The FDTGMM can represent different forms depending on the choice of the co-variance matrices and truncation parameters. One co-variance matrix for all Gaussian component (grand co-variance) or a single co-variance matrix shared by all speakers models (global covariance) used in FDTGMM. The covariance matrix can also be full or diagonal .In the present work the diagonal covariance matrix as primarily used for speaker model. This choice is based on the works given by (D.A Renold, Richard rose (1995)) and initial experimental results indicating better identification performance and hence Σ can be represented as

$$\sum = \begin{bmatrix} V_1^2 & 0 & 0 & 0 \\ 0 & V_2^2 & 0 & 0 \\ - & - & - & - \\ 0 & 0 & 0 & V_M^2 \end{bmatrix} \quad (5)$$

This simplifies the computational complexities.

3. ESTIMATION OF THE MODEL PARAMETERS

For developing the speaker identification model it is needed to estimate the parameters of the speaker model. For estimating the parameters in the model we consider the EM algorithm which maximizes the likelihood function of the model for a sequence of i training vectors $\vec{x}_t = (x_1, x_2, \ldots\ldots x_t)$.

The likelihood function of the DTGMM is

$$p(\vec{x}; \lambda_i) = \prod_{i=1}^{T} p_i(\vec{x}; \lambda_i) \quad (6)$$

where, $p(\vec{x}; \lambda_i)$ is given in equation (2).

The likelihood function contains the number of components M which can be determined from the k-means algorithm. The k-means algorithm requires the initial number of components which can be taken by plotting the histogram of Mel frequency cepstral coefficients associated with the speaker and counting the number of peaks. Once k-is assigned 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:

$$\alpha_k^{l+1} = \frac{1}{T} \sum_{i=1}^{T} p(i \mid \vec{x}_t, \lambda^1) \quad (7)$$

$$\mu_k^{l+1} = \frac{\Sigma_{i=1}^{T} \overrightarrow{x_{tp}}(i \mid \vec{x}_t, \lambda^1) + \Sigma_{i=1}^{T} \frac{f(x_M) - f(x_L)}{B-A} \sigma_k^2 p(i \mid \vec{x}_t, \lambda^1)}{\Sigma_{i=1}^{T} p(i \mid \vec{x}_t, \lambda^1)} \quad (8)$$

$$\sigma_k^{l+1} = \frac{\Sigma_{i=1}^{T} p(i \mid \vec{x}_t, \lambda^1)(\vec{x}_t - \mu_i^{(l+1)})^2}{c\Sigma_{i=1}^{T} p(i \mid \vec{x}_t, \lambda^1)} \quad (9)$$

where C is given by

$$C = \frac{1}{(B-A)}(1 + \mu_k^{l+1})[(f(\vec{x}_M) - f(\vec{x}_L)) + (x_M(f(\vec{x}_L) - x_L f(\vec{x}_M))]$$

and $f(x_{M'}) = \int_{-\infty}^{x_M} b_i(x_t)dx_t, \qquad f(x_{L'}) = \int_{-\infty}^{x_L} b_i(x_t)dx_t$

The a posterior probability for acoustic class i is given by

$$p(i \mid \vec{x}_t, \lambda^1) = \frac{\alpha_i b_i(\vec{x}_t)}{\Sigma_{i=1}^{k} \alpha_i b_i(\vec{x}_t)} \quad (10)$$

4. INITIALIZATION OF THE MODEL PARAMETERS

To utilize the EM algorithm we have to initialize the parameters μ_i, σ_i and α_i i = (1 ….. M) X_M and X_L obtained can be estimated with the values of the maximum and the minimum values of each feature vector respectively. The initial values of α_i can be taken α_i = 1/M where M is obtained from the histogram of the mel frequency cepstral

coefficients. The initial estimates of μ_i, σ_i and α_i of the i^{th} component is obtained using the method given by A.C. Cohen (1950).

k-means Algorithm: Step.1: Begin with a initial value of M (number of components) from the sample histogram of the first Mel frequency cepstral coefficients of a speaker.

Step. 2: Put any initial partition that classifies the Mel frequency cepstral co-efficient into M–clusters, we can arrange the training samples randomly, or systematically as follow.

a. Take the first M–training samples of Mel frequency cepstral coefficients of i^{th} speaker.

b. Assign each of the remaining (T – M) training samples to the components with the nearest centroid. Let there be exactly M components (C_1, C_2 – CM) and T coefficients to be classified such that each pattern is classified in to exactly one component. Let each component C_M and T_M coefficients, the mean vector or the centre of the cluster C_M is defined as the centroid of the cluster and given by

$$m^M = \frac{1}{T_M} \Sigma_{i=1}^{T_M} x_i^{(M)}$$

where $x_i^{(M)}$ is the i^{th} pattern belonging to the cluster C_k. After each assignment, recomputed the centriod of the gaining component.

Step. 3: Take each sample in sequence and compute its distance from the centriod of each component. If the sample is not currently in the cluster with the closest centriod switch this sample to that component and update the centrioi of the component gaining the new sample and cluster losing the sample.

Step. 4: Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments.

After obtaining the final values of the number of clusters M, we obtain the initial estimates of μ_i, σ_i, and α_i for i^{th} component of each speaker using the method given by Cohen A.C. (1950), for Doubly Truncated Gaussian Distribution.

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 Doubly Truncated Gaussian Mixture Model.

1. Find feature vectors using front end process explained in section 1.
2. Divide the T samples into 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 obtaining the refined estimates of μ_i, α_i, and σ_i for each component of the i^{th} speaker.
6. Write the speaker Model as $p(x/\lambda) = \Sigma_{i=1}^{M} \alpha_i p_i((x/\lambda_s)$ where $\lambda_i = \{\mu_i, \sigma_i, \alpha_i\}$ put $\lambda = \lambda_1, \lambda_2, \ldots \lambda_6$. from each speaker.
7. For Speaker identification, from a group of S Speakers S = {1, 2,.....S} each represented by FDTGMM's with parameters , $\lambda_1, \lambda_2, \lambda_3, \ldots \ldots \lambda_s$ we find the speaker model which has the maximum a posteriori probability for a given observation sequence that is

$$\hat{s} = \max_{1<k<s} p_r(\lambda_k \mid x)$$

$$= \arg\max_{1<k<s} [p_r((\lambda_k \mid X) p_r(\lambda_k)]$$

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

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

$$= \arg\max_{1<k<S} \Sigma_{i=1}^{T} \log p_r(\vec{x}_t \mid \lambda_k)$$

in which $p(\vec{x}_t \mid \lambda_k)$ is given in equation (2).

6. EXPERIMENTAL RESULT

The DTGM for 50 speaker model was trained and evaluated by using a data base of 25 speakers for each speaker there twenty words of approximately 2sec. each recorded in a single session. The speech was recorded for on a high quality Micro phone locally.

The feature vectors are mel frequency-ceptral co-efficient which are obtained for test sequence length 2 secs. Corresponds to different number of vectors based on Front end process given by D.A Reyonolds (1995).

The data set $\{\vec{x}_1, \vec{x}_2, \vec{x}_3 \ldots\ldots\ldots \vec{x}_t\}$ is divided into a training set and a test set. Using the training set of the sample histogram of individual speaker the first Mel frequency cepstral coefficients is the re drawn and shown in fig (2).

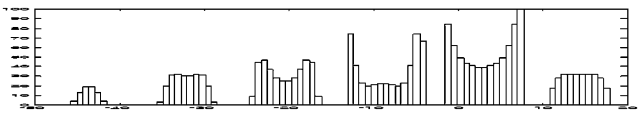

Fig. 2: Histogram of the Feature Vector

From the fig (2) It is observed that the initial number of components in each speaker can be considered as M = 6 using the k-means algorithm for each speaker training data set the feature vector classification is done in to 6 components. From the six components data from each speech data the initial estimate of the parameters are obtained using the K-means algorithm and the moment estimators given by A.C. Cohen (1950). With these initial estimates and the up dated 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. Using 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

% correct identification =

$$\frac{\#\text{correctly identified segments}}{\text{total \# of segments}} \times 100$$

It is observed that this algorithm identifies the 96% speaker correctly. The variation of this is also computed by repeating the experiment over 20 words.

A comparative study of the Performance of DTGMM is carried with reference to the non truncated GMM with Mel frequency cepstral co-efficient of the feature vectors. The percentage correct identification for 50 speaker utterance length of both the models are computed with their confidence intervals and presented in table 1.

Table 1
Speaker Identification Performance for Speaker Models Discussed in Text

Speaker Model	% Correct Identification (2 Sec test length)
GMM-nv	94.5±1.8
GMM-gv	89.5±2.4
TGMM	80.1±3.1
GC	67.1±3.7
FDTGMM	96±1.9

From table 1. It is observed, that the developed speaker identification model with DTGM distribution performs much better than the earlier speaker identification model with GM distribution. The improvement is highly significant with respect to the quality of identification and authentication.

CONCLUSION

In this paper the proposed a text independent speaker identification model based on Finite Doubly truncated GMM with EM and K-means algorithm. The model parameters are estimated through EM algorithm after identifying the number of component densities in each speaker voice spectrum using mel frequency cepstral co-efficient as feature vectors with the component maximum likelihood. The speaker identification algorithm is developed. The DTGM feature vector components is a generalization of the Finite Gaussian Mixture Model. This also includes the Gaussian mixture model as limiting case when the truncated points tend to infinite. 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,

REFERENCES

[1] S.Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Trans. Acoustic, Speech, Signal Processing, Assp-29, pp.254-272, Apr.1981.

[2] Douglas A.Reynolds, and Richard C. Rose, "Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Model," IEEE Trans. Speech and Audio Processing, 3, pp.72-83, Jan1995.

[3] Akira Kurwmastsu, Mariko Nakano-Miyatake, Hecto-perez-Meana, Eric simancas Acevedo, "Performance Analysis of Gaussian Mixture Model Speaker Recognition Systems with Different Speaker Features," Electronic Journal Technical Acoustics, May.2005.

[4] K.P.Markov, S.Nakagawa, "Integrating Pitch and LPC-residual Information with LPC-Cepstral for Text Independent Speaker Recognition". J.Acoustic Society of Japan (E), 1999.

[5] Sandipan Chakroborty, Anindya Roy and Goutam Saha "Improved Closed Set Text- Independent Speaker Identification by Combining MFCC with Evidence from Flipped Filter Banks" International Journal of Signal Processing, 4, pp.114-121, Nov 2006.

[6] K.Sri Rama Murthy and Yegnanarayana B, "Combining Evidence from Residual Phase and MFCC Features for Speaker Recognition", IEEE Signal Processing Letters, 13, Jan, 2006.

[7] 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.

[8] Mattos, R. S. et al., "Estimating Kings Ecological Inference Normal Model via EM Algorithm". In The 2000 Midwest Political Science Association Meeting, 27–30, 2000.

[9] Leena Marry and B.Yegnanarayana "Extraction and Representation of Prosodic Features for Language and Speaker Recognition" Speech Communication, April 2008.

[10] Nemat.S and Abdel Khader "Effect of GSM System on Text Indenpendent Speaker Recognition Performance" Journal of Theoretical and Applied Information Technology.

[11] A.Revathi R Genapathy and Y Venkatrramani, "Text Indenpendent Speaker Recognition and Indenpendent Speech Recognition using Iterative Clustering Approach" Journalof Computer Science and Information Technology, 1, No. 2., November 2009.

[12] Md.Rabiul Islam and Md. Fayzur Rahman " Improvement of Text Dependent Speaker Identification System using Neuro-genetic Hybrid Algorithm in Office Environmental Conditions" International Journal of Computer Science, 1, 2009.

[13] Sandipan Chakroborty and Goutam Saha "Improvemed Text Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter" International Journal of Signal Processing Winter, 2009.