Text Independent Speaker Identification In a Distant Talking Multi-microphone Environment Using Generalized Gaussian Mixture Model

P. Soundarya Mala, Dr. V. Sailaja, Shuaib Akram

Abstract -- In speaker Identification System, the goal is to determine which one of the groups of an unknown voice which best matches with one of the input voices. The field of speaker identification has recently seen significant advancement, but improvements have tended on near field speech, ignoring the more realistic setting of far field instrumented speakers. In this paper, we use far field speech recorded with multi microphones for speaker identification. For this we develop the model for each speaker's speech. In developing the model, it is customary to consider that the voice of the individual speaker is characterized with Generalized Gaussian model. The model parameters are estimated using EM algorithm. Speaker identification is carried by maximizing the likelihood function of the individual speakers. The efficiency of the proposed model is studied through accuracy measure with experimentation of 25 speaker's database. This model performs much better than the existing earlier algorithms in Speaker Identification.

Keywords-- Generalized Gaussian model, EM Algorithm, and Mel Frequency Cepstral Coefficients.

INTRODUCTION

Speaker recognition is the process of recognizing who is speaking on the basis of information extracted from the speech signal. It has been number of applications such as verification of control access permission to corporate database search and voice mail, government lawful intercepts or forensics applications, government corrections, financial services, telecom & call centers, health care, transportation, security, distance learning, entertainment & consumer etc [2].

The growing need for automation in complex work environments and increased need for voice operated services in many commercial areas have motivated for recent efforts in reducing laboratory speech processing algorithms to practice. While many existing systems for speaker identification have demonstrated good performance and achieve high classification accuracy when close talking microphones are used. In adverse distant-talking environments, however the performance is significantly degraded due to a variety of factors such as the distance between the speaker and microphone, the location of the microphone or the noise source, the direction of the speaker and the quality of the microphone. To deal with these problems micro phone arrays based speaker recognizers have been successfully applied to improve the identification accuracy through speech enhancement [3][4][6].

In speaker identification since there is no identity claim, the system identifies the most likely speaker of the test speech signal. Speaker identification can be further classified into closed-set identification and open-set identification. Speaker identification can be further classified into closed-set identification and open-set identification. The task of identifying a speaker who is known a priori to be a member of the set of N enrolled speakers is known as closed-set speaker Identification. The limitation of this system is that the test speech signal from an unknown speaker will be identified to be one among the N enrolled speakers. Thus there is a risk of false identification. Therefore, closed set mode should be employed in applications where it is surely to be used always by the set of enrolled speakers. On the other hand, speaker identification system which is able to identify the speaker who may be from outside the set of

*Ms P.Soundarya mala Completed M.tech in Digital Electronics and Communication Engineering in GIET in Jawaharlal Nehru Technological University,Kakinada,INDIA in 2010,PH:9493493302.
Email:soundarya_palivela@yahoo.co.in
*Dr V.Sailaja received Ph.d degree in Speech Processing from Andhra University in 2010, INDIA,PH:9491444434,
Email:sailajagiet@gmail.com
*Mr Shuaib Akram studying IV B.Tech in Electronics and Communication Engineering, Jawaharlal Nehru Technological University , Kakinada,INDIA,PH:9703976497, Email:shuaib.akram@yahoo.com

N enrolled speakers is known as open-set speaker identification. In this case, first the closed-set speaker identification system identifies the speaker closest to the test speech data. The speaker identification system is divided into text independent speaker identification and text dependent speaker identification. Among these two, Text Independent Speaker Identification is more complicated in open test.

Speaker Identification:

Given different speech inputs X1,X2,……, Xc simultaneously recorded through C multiple microphones, whoever has pronounced X1,X2,…….Xc among registered speakers S={1,2,…..C} is identified by equation (1). Each speaker is modeled by GGMM λ_k.

$$\hat{S}= arg^{max}_{1\le k\le C}\ p(\lambda_k \mid X1,X2,\ldots.Xc).$$

$$= arg^{max}_{1\le k\le C}\ \frac{p(\lambda k \mid X1,X2,\ldots.Xc)\ .P(\lambda k)}{p(X1,X2,\ldots\ldots Xc)} \qquad (1)$$

By using Bayes's rule equal prior probability (ie. P(λk.)=1/C), and the conditional independency between different speech inputs X1,X2,….Xc given speaker model λp , and not in that p(X1,X2,…….Xc) is the same for all speakers, equation 1 can be simplified as,

$$\hat{S}= arg^{max}_{1\le k\le C}\ \prod_{c=1}^{C} p(Xc \mid \lambda_k) \qquad (2)$$

Taking the logarithm of equation 2, we obtain

$$\hat{S}= arg^{max}_{1\le k\le C}\ \sum_{c=1}^{C} p(Xc \mid \lambda_k) \qquad (3)$$

The identity of the speaker can be determined by the sum of hypothesis log likelihood scores obtained from C microphones. In a distance talking environment, however, the log likelihood score itself in equation(3) is expected to degraded, ie. Its reliability cannot be ensured. Furthermore, a variety of causes such as the location of the speaker or the noise, the direction of the speaker, and the distance can have a different effect on each microphone. Therefore the identification result obtained from a microphone can be better than the others. In such cases, the simple integration is greatly affected by the incorrect classification of single channel. Thus, we propose a new integration method to re-score the hypothesis scores, measure the distance between them and combine them.

2. FINITE MULTIVARIATE GENERALIZED GAUSSIAN MIXTURE SPEAKER MODEL

The Mel frequency cepstral coefficients (MFCC) are used to represent the features for speaker identification. In the set up used, the magnitude spectrum from a short frame is processed using a mel-scale filter bank. The log energy filter outputs are the cosine transformed to produce cepstral coefficients. The process is repeated every frame resulting in a series of feature vectors [1]. We assume that the Mel frequency cepstral coefficients of each are assumed to follow a Finite Multivariate Generalized Gaussian Mixture Distribution. 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

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

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 $\lambda = (\mu,\varrho,\sum)$ α_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)) [5] and is of the form

$$b_i(\vec{x}_t|(\mu,\rho,\sum) = \frac{[\det(\sum)]^{-1/2}}{[z(\rho)A(\rho,\sigma)]^D}\exp \qquad (5)$$

where,$z(\rho) = \frac{2}{\rho}\Gamma\left(\frac{1}{\rho}\right)$ and

$$A(\rho,\sigma) = \sqrt{\frac{\Gamma(1/\rho)}{\Gamma(3/\rho)}} \qquad (6)$$

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) = σ2 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 $\varrho \to +\infty$ Equation (5) Converges to a uniform distribution in ($\mu-\sqrt{3}\sigma$, $\mu+\sqrt{3}\sigma$) and when $\varrho \to +$, the distribution becomes a degenerate one when x=μ. The

generalized Gaussian distribution is symmetric with respect to μ,

The variance of the variate x_{ij} is

$$\text{var}(X) = \sigma_{ij} \quad (7)$$

The model can have one covariance matrix per a Generalized Gaussian density of the acoustic class of each speaker. The covariance matrix $\sum$ 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. 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

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

(8)

The model parameters are estimated and initialized by the EM algorithm with k-means [7].

The updated equation for estimating the model parameters are

The updated equation for estimating αi is

$$\alpha_i^{(l+1)} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\alpha_i^{(l)}\, b_i(\vec{x}_t, \lambda^{(l)})}{\sum_{i=1}^{M}\alpha_i^{(l)}\, b_i(\vec{x}_t, \lambda^{(l)})}\right] \quad (9)$$

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

The updated equation for estimating μ_{ij} is

$$\mu_{ij}^{(l+1)} = \frac{\sum_{t=1}^{T} t_i(\vec{x}_t,\lambda)^{(l)} A(N,\rho_i)(x_{tij}-\mu_{ij})}{\sum_{t=1}^{T} t_i(\vec{x}_t,\lambda^{(l)})^{A(N,\rho_{ij})}} \quad (10)$$

Where, $A(N,\rho_{ij})$ is some function which must be equal to unity for ϱi = 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 ρ_{ij}.

The updated equation for estimating σ_{ij} is

$$\sigma_{ij}^{(l+1)} = \left[\frac{\sum_{t=1}^{N} t_i(\vec{x}_t,\lambda^{(l)})\left(\frac{\Gamma\left(\frac{3}{\rho_{ij}}\right)}{\rho_i\Gamma\left(\frac{1}{\rho_i}\right)}\right)|x_{tij}-\mu_{ij}{}^{(l)}|^{\frac{1}{\rho_{ij}}}}{\sum_{t=1}^{T} t_i(\vec{x}_t,\lambda^{(l)})}\right]^{\frac{1}{\rho_{ij}}} \quad (11)$$

The number of mixture components is initially taken for K – Means algorithm by drawing the histogram of the first Mel frequency cepstral coefficient of the speech data.

3. EXPERIMENTAL SETUP

The proposed Generalized Gaussian Mixture Model in a multi microphone environment are evaluated with a database uttered by 25 speakers. For each speaker, there are 10 conversational sentences, which are recorded in single session. Each sample is of 2 seconds. Speech samples are recorded by using 8 micro phones. Each of which recorded at centre and diagonal. They were then re-recorded again by playing them back with a loud speaker placed at each position, means 1m, 3m and 5m by using 8 microphones which were placed at different locations. These are used to train GGM and to estimate the distribution of average log likelihood estimation.

4. EXPERIMENTAL RESULTS:

As the number of N-best hypotheses per channel (N- best classification results) employed for identification increases, the performance of the proposed method outperform the earlier existing methods.

Table1: Speaker identification accuracy

LOCAL / CH	1m		3m		5m	
	C	D	C	D	C	D
0	94.9	95.3	68.3	67.8	75.7	74.6
1	91.8	91.6	66.2	66.9	63.9	64.5
2	94.0	93.4	73.1	72.8	80.6	81.2
3	91.8	92.3	62.7	62.5	56.1	56.9
4	94.6	93.9	54.4	55.6	57.5	58.4
5	94.4	93.8	69.7	70.3	56.5	57.4
6	96.2	95.6	61.8	62.6	60.6	61.4
7	94.0	93.5	57.5	58.9	62.2	63.1
AVG	94.0	93.7	64.2	64.7	64.1	64.7
LS	93.0	93.9	67.4	80.4	70.9	78.3
AD	93.0	93.5	68.3	80.9	74.3	81.3
GGM	94.0	93.7	64.2	64.7	64.1	64.7

The figure represent the identification accuracies of the identification methods, as the number of N- best classification result increases to 25, which corresponds to total number of registered speakers .Table.1 corresponds to the performance of the identification methods in case of N=25. In terms of the base line method (LS), if we approximate k in Eq (3) to k ε N-best hypotheses, it is possible not to obtain the log likelihood score of the hypothesis from a certain channel depending on the order of hypotheses. Thus we considered only when N is equal to 25. Despite only using 2-best classification results per channel, the proposed method (GGM) is comparable to, or even better than, the baseline methods using all possible hypotheses, which consist of all the registered speakers. And the proposed identification method can achieve its best performance with less than eight best hypotheses per channel. Table 2 shows the average percentage of correct identification for various speaker identification models.

Table 2 Avg. percentage of correct identification vs. speaker identification models.

Speaker Model	% of Accuracy
LS	93.0
AD	93.0
GGM	94.0

*LS- Least Squared, AD-Adaptive Distribution,

GGM- Generalized Gaussian Model

A graph between False Alarm probability and Miss-probability called as DET curve is drawn below. If false alarm probability and Miss-probability are equal, the graph can be approximated as hyperbola.

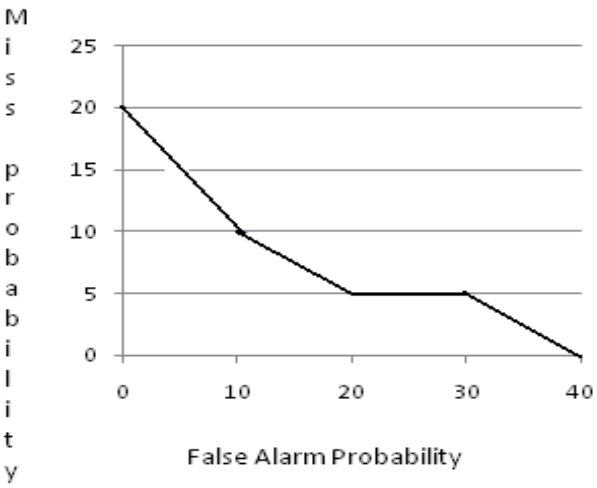

Figure 1: Plot of DET Curves for a speaker identification evaluation.

CONCLUSION

In this paper we propose Text Independent Speaker Identification model based on Finite Multivariate Generalized Gaussian Mixture Model, to improve identification accuracy in a Multi-microphone environment. In a relatively close-talking environment (1m), Generalized Gaussian model maintain high identification accuracy and they are superior to identification result of the best channel. As the distance between the speaker and microphone increases, the Generalized Gaussian Mixture Model shows more reliable identification performance than other existing models which are given in the table (1). For this, a text independent speaker identification model is developed with the assumption that the feature vector associated with speech spectra of each individual speaker follows a Finite Multivariate Generalized Gaussian Mixture Model. The model parameters are estimated and initialized by using EM algorithms with K-means. An experimentation with 25 speakers speech data revealed that this Text Independent Speaker Identification using Finite Multivariate Generalized Gaussian Mixture Model outperform the earlier existing Text Independent models.

References:

[1] B. Narayana Swamy and R. Gangadharaiah "Extracting Additional Information From GAUSSIAN MIXTURE MODEL PROBABILITIES for improved Text Independent Speaker Identification." Proc. ICASSP, vol.1, pp.621-624, Philadelphia, USA, March 2005.
[2] D.A.Reynolds, R.C.Rose. E.M. Hofstetter "Integrated Models of signal and background with application to speaker identification in noise" IEEE Trans. On speech and audio Processing, vol 2. No. 2. April 1994.
[3] D.A.Reynolds, R.C.Rose. "Robust Text Independent Speaker Identification using Gaussian Mixture Models" IEEE Trans. On speech and signal Processing, 3(1), January 1995, 72-83.
[4]. I.A. McCowan, J. Pelecanos and S. Sridharan, "Robust speaker recognition using microphone arrays" Proc. Speaker Odyssey, pp.101-106, Crete, Greece, June 2001.
[5] 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.
[6] Q. Lin, E. Jan, and J. Flanagan, "Microphone arrays and speaker identification", IEEE Trans. Speech Audio Process, vol.2, no.4, pp.622-628, Oct. 1994
[7] Sailaja, K Srinivasa Rao & K V V S Reddy (2010) "Text Independent Speaker identification with Multi Variate Gaussian Mixture Model" International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 475-480.