[1] **Jayanthi Kumari T.R**

[2] **Anita R**

[3] **Suraj Duncan T**

Speaker Verification Comparison between GMM and GMM-UBM Under Limited Data Condition

Abstract: **-** This work demonstrates the verification of speakers with the Constraint of Limited data (<15 sec). The existing techniques for speaker verification work well for sufficient data (>1 minutes). Developing techniques for verifying the speakers for limited data condition is a challenging issue. In this paper, a comparision study is made using Gaussian Mixture Model (GMM) and GMM-Universal background model (GMM-UBM) with mel-frequency cepstral coefficients (MFCC) as a feature is given. The NIST-2003 database is used to carry-out the experiments. The experiments are conducted using different amount of training and testing data. The experimental results show that GMM-UBM gives a lower equal error rate (EER) compared to GMM.

Keywords: GMM, Speaker, GMM-UBM, data condition, system.

I. INTRODUCTION

The objective of speaker recognition is to recognize the speakers using their voice [1]. Speaker recognition involves speaker verification and speaker identification [2]. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. In speaker identification since there is no identity claim, the system identifies the most likely speaker of the test speech signal [3]. Speaker recognition can be classified into closed-set and open-set recognition. The task of recognizing 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. On the other hand, the speaker recognition system which is able to identify the speaker who may be from outside the set of N enrolled speakers is known as open-set Speaker identification [3] [4]. Depending on the mode of operation, the speaker recognition system can be either text-dependent or text-independent [5]. The same text is used for both training and testing in the text-dependent case while no restrictions on text are made in the text-independent case.

Speaker recognition in limited data condition aims at recognizing speaker with the constraint that both training and testing data are limited. Limited data symbolizes the case of having speech data of few seconds (less than or equal 15 seconds) [6]. Since the amount of data available is less in limited data condition, the number of feature vectors is insufficient to model and discriminate the speaker well. Therefore, it is a challenging task to improve the speaker recognition in such situations.

Speaker recognition under limited data conditions could be used in the following applications [7] [8]:

1) Controlled access and authentication like banking operations through telephone.
2) Criminal and forensic investigations.
3) Person authentication using voice as a Biometric.

All the above mentioned applications face the constraint of limited data. Therefore, it is essential to build a system which will be able to recognize speaker with the constraint of limited data.

The most widely used classifier for speaker recognition is GMM which was proposed by Reynolds in 1995 [9]. In GMM, the underlying probability density function of the feature vectors of each speaker is captured using Gaussian mixtures [10]. The complete GMM is parameterized by the mean vectors, covariance matrices and mixture weight. In GMM, Log likelihood ratio test is used for testing. The advantage of using GMM is, it's more economical and is based on a well-understood statistical model for text-independent speaker verification [9]. The disadvantage of GMM is that it requires sufficient data to model well the speaker parameters [9] [11]. To overcome this problem, GMM-UBM modelling technique is used for speaker recognition task [12].

In GMM-UBM, speech data from a large pool of speakers were used to design a speaker independent model. UBM is trained, which acts as a speaker-independent model. The speaker-dependent model is then created from the UBM by performing maximum a posterior (MAP) adaptation technique using speaker-specific training speech. As a result GMM- UBM gives better results than the GMM. The advantage of the UBM-based modelling technique is

[1] Professor, Department of ECE, EPCET Bangalore, India. trjayanthikumari.ece@eastpoint.ac.in
[2] Professor, Department of ECE, EPCET Bangalore, India. anitar.ece@eastpoint.ac.in
[3] Senior Salesforce Developer, Point Orgn Technologies India Pvt. Ltd. suraj98d@gmail.com

that it provides good performance even through the speaker-dependent data is minimal. The disadvantage is that a gender-balanced large speaker set is required for UBM training [11].

The paper is organised as follows: Section II describes the database used for the experiments. Feature extraction using MFCC and speaker modelling using GMM and GMM-UBM are presented in Section III. In Section IV comparison of experimental result is presented. Section V gives summary of the present work and scope for the future work.

II. Database For The Study

The NIST-SRE-2003 database consists of speech data from 356 speakers (149 male and 207 female). The spontaneous speech of speakers was collected over cellular phone, sampled at 8 kHz and stored with 16 bits/sample resolution for use as training and testing data. The range of speech data varies from few seconds to few minutes. Since the database is not meant for limited data condition, we have taken four, five and six seconds of each speaker data to create the database for the present work. A detailed description of the database can be found in the NIST-SRE-2003 plan (NIST 2003) [13]

III. Feature Extraction And Modeling

A. *Mel-frequency cepstral coefficients (MFCC)*

The purpose of feature extraction is to extract the speaker- specific information in the form of feature vectors at a reduced data rate [14]. In this work, features are extracted using MFCC technique. The state-of-the-art speaker verification system uses MFCC as a feature for recognizing speakers [15]. Fig.1 shows the block diagram representation of the MFCC method. Speech signals were sampled at the rate of 8 kHz. Frame duration of 20 msec and overlapping duration of 10 msec (160 and 80 samples respectively) are considered. Upon framing, we advance to the Windowing process. Windowing (Hamming) method is carried out to minimize the spectral distortion. The mathematical expression for the Hamming window is as follows:

$$h(n) = 0.54 - 0.46\cos(2\pi n/N - 1) \quad (1)$$

Fourier transform is then applied on the windowed frame signal to obtain the magnitude frequency response. A magnitude spectrum is computed. The resulting spectrum is passed through a set of triangular band pass filters. We have considered 22 filters. These filters are equally spaced along the Mel-frequency scale. The Mel scale is a mapping between the real frequency scale (Hz) and the perceived frequency scale (Mels). The mapping from linear scale to Mel scale is given in equation 2

$$fmel = 2595\log10(1 + f/700) \quad (2)$$

In order to get the cepstral coefficients, Discrete cosine transform (DCT) is applied. In this work, 13 coefficients are considered as feature vectors. Since the 0th coefficient can be regarded as a collection of average energies of each frequency bands, it is unreliable [11].

B. *Gaussian mixture model (GMM)*

The Gaussian mixture model (GMM) is a density estimator and is one of the most commonly used types of classifier for speaker recognition [16]. In GMM, the distribution of the feature vector is modelled clearly using a mixture of M

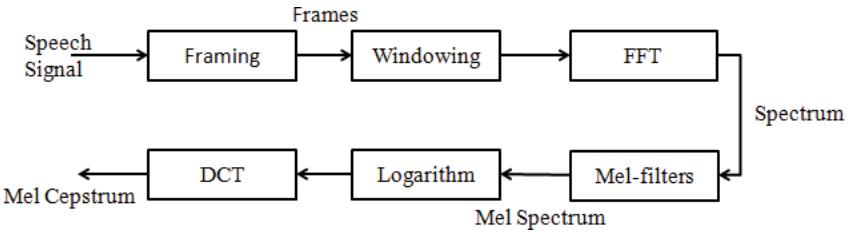

Fig. 1: Block diagram of MFCC Technique

Gaussians. Given a collection of training vectors, maximum likelihood model parameters are estimated using iterative expectation-maximization (EM) algorithm [10].The EM algo- rithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for the observed feature vectors. The complete GMM is parameterized by the mean vectors, covariance matrices and mixture weights from all components densities. These parameters are collectively represented by the notation,

$$\lambda = \{\omega i, \mu i, \Sigma i\} \quad (3)$$

During testing the speaker recognition system uses the matching and decision logic [11]. Test features vectors are matched with the reference models, matching gives a score which represents how well the test feature vectors are close to the reference models. Decision will be taken based on a final set of matched scores, which depends on the threshold value. In GMM, Log likelihood ratio test is used for testing.

C. GMM-Universal background model (GMM-UBM)

UBM is a large GMM which represents the speaker inde- pendent distribution of features. UBM is generally built using large population of speech. UBM is the core part of GMM- UBM speaker verification system. UBM should be balanced with respect to male and female speakers. To train a UBM, the simplest approach is to merely pool all the data and use it to train the UBM via the EM algorithm. Maximum a posteriori (MAP) adaptation integrates coupled target and background speaker model components is an effective way of performing speaker recognition [17].During the testing stage, log likelihood ratio test is used for testing.

IV. EXPEREMENTAL RESULTS

The present work focuses on speaker verification using two different modelling methods. A perfect speaker verification system should accept all the true claims and reject all the false claims. But in practice, some true trials may be rejected and some false trials may be accepted by speaker verification system. The speaker verification performance is measured in terms of false rejection rate (FRR) and false acceptance rate (FAR). These parameters compute to give equal error rate (EER).

In the present work, all experiments were carried out with a constant set of 356 train and 2559 test speakers from NIST2003 database and time span for each speaker sample varies around 4sec, 5sec and 6secs in order to make a relative comparison in the performance of speaker verification using two modelling techniques. In GMM, the parameters (mean vector, covariance matrix, mixture weights) were estimated using expectation maximization (EM) algorithm. In GMM, the speakers were modelled for Gaussian mixture of 16, 32 and 64. In case of UBM, speaker specific models were created by adapting only the mean vectors using maximum a posterior (MAP) adaption algorithm. UBM is modelled for Gaussian mixtures of 16, 32 and 64. The performance of both the methods is tabulated in table 1 and 2.

Table I: GMM verification performance EER (%) for different amounts of Training / Testing data and Gaussian mixtures.

Train/Test data	Gaussian Mixtures		
	16	32	64
4sec	44.30	45.70	47.40
5sec	42.68	44.12	45.21
6sec	42.09	41.96	44.98

Table II: GMM-UBM verification performance EER (%) for dif- ferent amounts of Training / Testing data and Gaussian mixtures.

Train/Test data	Gaussian Mixtures		
	16	32	64
4sec	39.52	38.66	39.06
5sec	38.34	37.85	37.75
6sec	36.72	37.12	37.17

Fig. 2 and 3 show the individual performance of GMM and GMM-UBM models, respectively. Fig. 4, 5 and 6 indicates the comparison between both the models for 16, 32 and 64 Gaussian mixtures. The Experimental results in Table 1 and 2 show that the performance of GMM is poor compared to the performance of GMM-UBM. This is because, GMM needs sufficient data to model the speaker well [9] and GMM fails as there are two many parameters that need to be estimated given the limited amount of training and test data. Since the GMM-UBM uses the background models obtained from large set of speakers data, it gives good performance even through the speaker-dependent data is small. It can be observed in Fig. 4, 5 and 6 that the GMM-UBM gives lower EER compared to

GMM. An average percentage reduction in EER of 11.24%, 13.72% and 17.2% was obtained for Gaussian mixture of 16, 32 and 64, respectively.

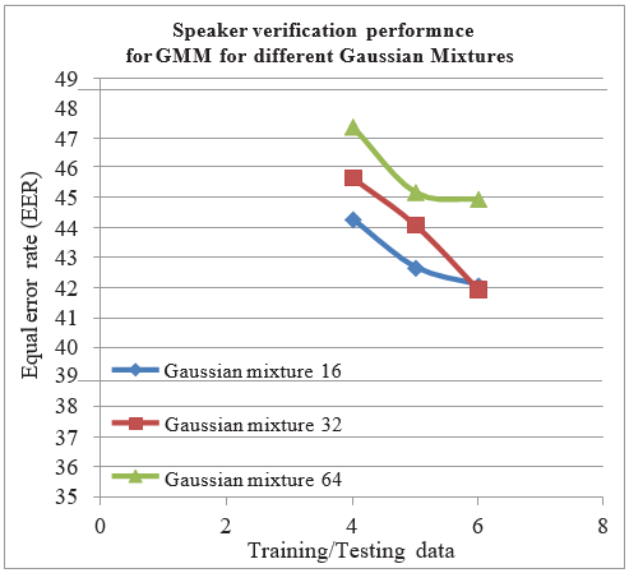

Fig. 2: Performance of speaker verification for GMM for 16,32 and 64 Gaussian mixtures

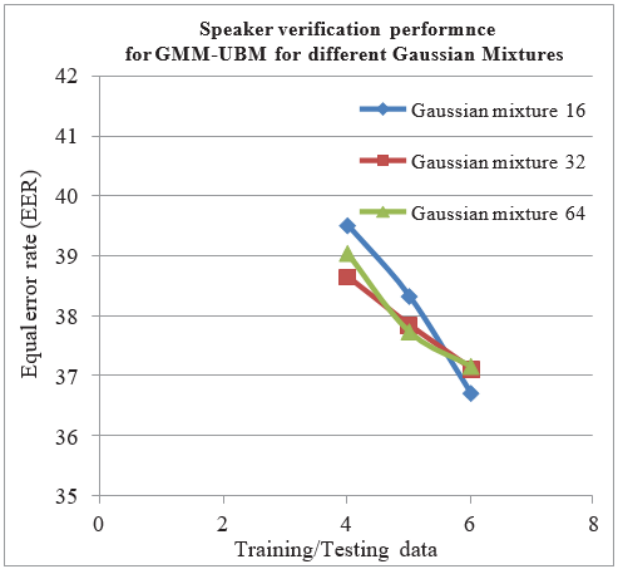

Fig. 3: Performance of speaker verification for GMM-UBM for 16,32 and 64 Gaussian mixtures

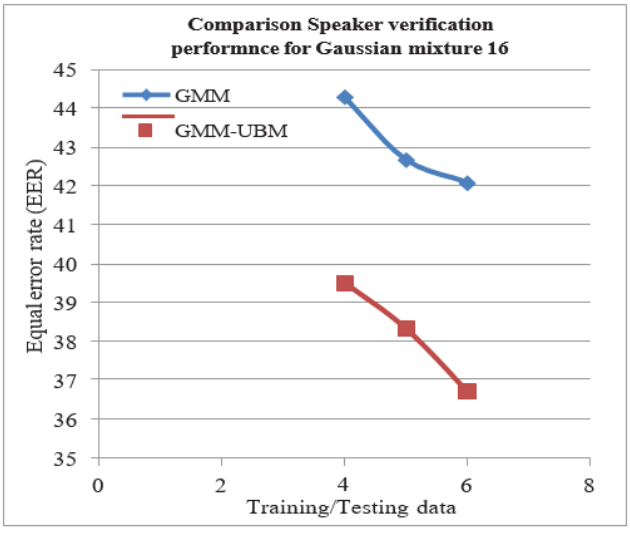

Fig. 4: Performance of speaker verification for Gaussian mixture 16

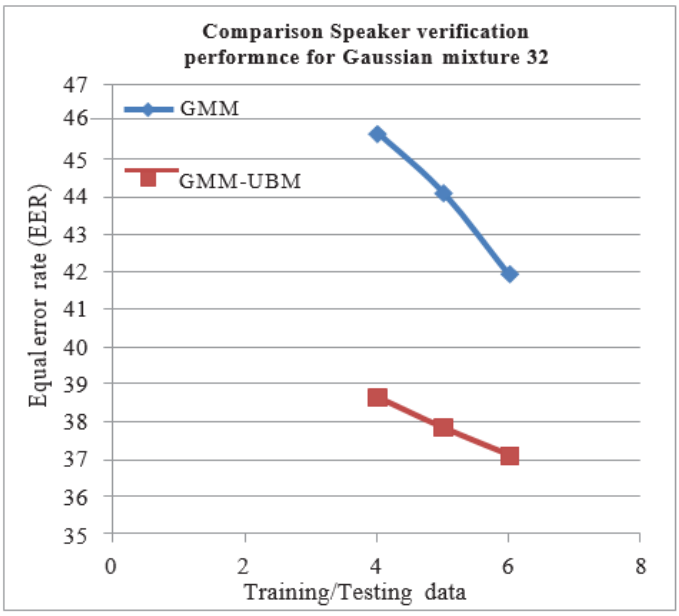

Fig. 5: Performance of speaker verification for Gaussian mixture 32

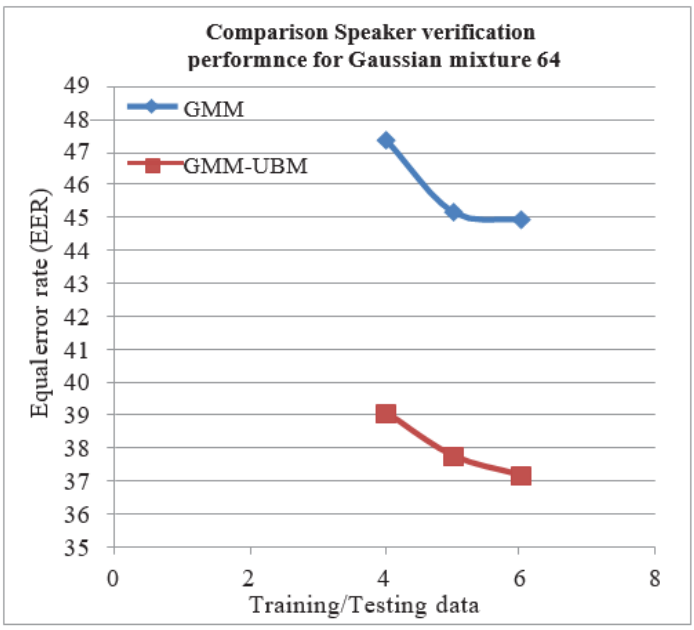

Fig. 6: Performance of speaker verification for Gaussian mixture 64

V. Conclusion

In this paper we have compared the performance of GMM and GMM-UBM modeling techniques using MFCC for speaker verification with the constraint of limited data. The results indicated that GMM-UBM gives lower EER compared to GMM in all the cases. Therefore, we suggest that the GMM- UBM can be used as modeling technique for speaker verifi- cation with the Constraint of limited data. The significance of different features need to be analyzed for speaker verification under limited data.

References

[1] B.S.Atal, "Automatic recognition of speakers from their voices," *proc IEEE*, vol. 64(4), pp. 460–475, Apr.1976.
[2] A.E.Rosenberg, "Automatic speaker verification a review," *Proc IEEE*, vol. 64(4), pp. 475–487, Apr.1976.
[3] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing*, vol. 88, pp. 18–32, Oct 2006.
[4] L. R. Rabiner and R. W. Schafer, "Digital processing of speech signals," *Prentice Hall,First edition*, 1978.
[5] Rajesh Ranjan, Sanjay Kumar Singh,Anupam Shukla and Ritu Tiwari, "Text-dependent multilingual speaker identification for indian languages using artificial neural network," *Proc IEEE*, pp. 632–635, 2010.
[6] H. S. Jayanna, "Limited data speaker recognition," *Ph.D disserta- tion,Indian Institute of Technology,Guwahati*, 2009.

[7] Hemant A Patil ,Sunayana Sitaram and Esha Sharma, "DA-IICT Crosslingual and Multilingual Corpora for Speaker Recognition," *Proc IEEE,Advances in Pattern Recognition,(Kolkata)*, p. 187190, 2009.

[8] B.G.Nagaraja and H.S Jayanna, "Mono and Cross lingual Speaker Identification with the Constraint of Limited Data," *Proc IEEE PRIME- 2012*, pp. 457–461, (March 2012).

[9] D.A.Reynods, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[10] A. Dempster , N. Laird and D.Rubin, "Maximum likelihood from incomplete data via the EM alogrithm," *Journal of Roval Statistical Society*, vol. 39, pp. 1–38, 1977.

[11] H.S Jayanna and S R Prasanna, "Analysis,Feature extraction,modeling and Testing techniques for Speaker Recognition," *IETE Technical Re- view*, vol. 26, pp. 181–190, May-june2009.

[12] D.A Reynolds, "Universal background models," *Encyclopedia of Bio- metric Recognition,Springer ,Journal Article*, Feb 2008.

[13] NIST2003, "http://www.itl.nist.gov/iad/mig//tests/sre/2003/2003-spkrec- evalplan-v2.2.pdf[online]."

[14] Ahmad Salman, Ejaz Muhammad and Khawar Khurshid, "Speaker Ver- ification Using Boosted Cepstral Features with Gaussian Distributions," *Proc IEEE*, 2007.

[15] J.W.Picone, "Signal modeling techniques in speech recognition," *Proc IEEE*, vol. 81(9), pp. 1215–1247, 1993.

[16] D.A.Reynolds and R.C.Rose, "Robust text-independent speaker identi- fication using Gaussian mixture speaker models," *IEEE Trans.Speech Audio Process*, vol. 3, pp. 72–83, 1995.

[17] Reynolds, T. Quateri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, Jan 2000