Speaker Verification Using Gaussian Posteriorgrams on Fixed Phrase Short Utterances

Sarfaraz Jelil[1], *Rohan Kumar Das*[2], *R. Sinha*[2] *and S. R. Mahadeva Prasanna*[2]

[1]Department of Information Technology
North-Eastern Hill University, Shillong-793022,
[2]Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati-781039
sarfarazjelil@ymail.com; {rohankd,rsinha,prasanna}@iitg.ernet.in

Abstract

This work explores the speaker verification using fixed phrase short utterances. A novel speaker verification system using Gaussian posteriorgrams is proposed in which the posteriorgram vectors are computed from speaker specific Gaussian mixture model (GMM). The enrollment utterances for each of the target speakers are labeled with GMM trained on the corresponding speaker's data. The test trials are then labeled with the claimed speaker's GMM model. Dynamic time warping (DTW) is used to find a match score between the posteriorgrams of the claimed speaker and that of test trial. The proposed approach is evaluated on the fixed pass phrase subset of the recent RSR2015 database. For contrast purpose, we have also developed state-of-the-art i-vector system including probabilistic linear discriminant analysis (PLDA) classifier. The proposed framework is found to result in highly improved performance when compared with the i-vector based contrast system. We hypothesize that the cause of this large improvement lies in the use of speaker specific variances information in generation of the posteriorgram representations. On evaluating the proposed framework with non-speaker specific variances, it resulted in significant performance degradation which confirmed our hypothesis.

Index Terms: speaker verification, fixed phrase, short utterances, Gaussian posteriorgrams, i-vector, PLDA, DTW

1. Introduction

Speaker verification (SV) refers to the task of authenticating the claim of a speaker based on the information present in his/ her speech signal [1–3]. It can be classified into two types depending upon the text spoken during the enrollment and the test sessions. If the text spoken is the same during both the sessions, the modality is termed as text-dependent SV [4, 5]. The other modality, text-independent SV, puts no such constraints on the text to be spoken [6]. This work focuses on the former mode of SV that requires minimal amount of speech data typically containing a fixed phrase of around 3-4 seconds duration.

Fixed phrase based SV classically uses dynamic time warping (DTW) to take advantage of the temporal information present in the speech utterances [7]. Another dominant approach in this field for the past few years has been the hidden Markov model (HMM), that exploits the temporal as well as the feature distribution information [8, 9]. Motivated by the success of i-vector based speaker representation in text-independent SV, recently it has been also explored for fixed phrase based SV [10, 11]. The addressing of the session/ channel variability is an integral part of all SV systems. Typically it is addressed using joint factor analysis (JFA), within class covariance normalization (WCCN), probabilistic linear discriminant analysis (PLDA) etc [12–14]. We hypothesize that in addition to explicit channel/ session compensation techniques, smoothing of the features with the help of Gaussian posteriorgrams should enhance the performance further. However, to the best of our knowledge Gaussian posteriorgrams have never been used explicitly for SV.

Gaussian posteriorgrams have been previously used successfully for audio keyword spotting, where each speech frame of the utterances are represented by corresponding Gaussian posteriorgrams [15]. The authors then use segmental dynamic time warping (SDTW) algorithm to find a similarity between all of the trained keywords and new test data. In [16], a technique for multi-speaker unsupervised speech pattern discovery is presented using Gaussian posteriorgrams, where a Gaussian mixture model (GMM) is trained in an unsupervised way and then speech frames are mapped to their corresponding Gaussian posteriorgrams. SDTW is used to find similar acoustic segments and then these segments are grouped using a graph clustering method. More recently, Gaussian posteriorgrams have also been used for accent recognition [17].

This paper explores the feasibility of using Gaussian posteriorgram based speaker representation for fixed phrase SV. A GMM is trained for every speaker with the training utterances and then each speech frame of a particular utterance is mapped with a vector, where the j^{th} component of the vector shows the posterior probability of being generated by the j^{th} component of the GMM. To find a match between the train and the test utterances, the DTW algorithm is used [18]. The experiments are conducted on the RSR2015 database prepared for fixed phrase short utterance studies [19]. To compare the performance of the proposed system, a contrast SV system is also developed using the i-vector based paradigm with PLDA for channel/ session compensation [13, 14]. The primary contribution of this work is the calculation of posteriorgrams from speaker specific GMMs, which boosts the respective speaker information.

The rest of the paper is organized as follows: Section 2 presents a brief overview of Gaussian posteriorgrams and its use in SV. Section 3 describes the proposed speaker verification system using Gaussian posteriorgrams. In Section 4 a contrast system using i-vector based framework is described, which is developed for comparing to the proposed system. Section 5 shows the experimental results and discussions. The summary and conclusions are given in Section 6.

2. Gaussian Posteriorgrams for Speaker Modeling

In this section, we provide a basic introduction to Gaussian posteriorgrams for speaker modeling. Given a few training utterances for a speaker, first a GMM of size N is trained using those utterances. It is followed by the computation of Gaussian posteriorgrams for the training utterances. Consider one utterance (in our case one sentence) speech data $\mathbf{S}$ which contains p number of frames. Symbolically, it can be represented as $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_p]$. Then, the posterior vector with respect to the speaker specific GMM corresponding to i^{th} feature frame is derived as,

$$\mathbf{x}_i = [P(\lambda_1|\mathbf{s}_i), P(\lambda_2|\mathbf{s}_i), \ldots, P(\lambda_N|\mathbf{s}_i)]^T \qquad (1)$$

where, λ_j denotes the j^{th} Gaussian component of the speaker specific GMM and N is the total number of Gaussian components.

On concatenating the posterior vectors in each of the frames in an utterance one derives the corresponding Gaussian posteriorgram which is given as,

$$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p] \qquad (2)$$

The Gaussian posteriorgrams for all the training utterances for a speaker are used as templates for modeling that speaker. The test utterances are also similarly represented by Gaussian posteriorgrams with reference to the claimed speaker specific GMM. The detailed procedure for SV using Gaussian posteriorgrams is discussed in the next section.

3. Fixed Phrase Gaussian Posteriorgrams Based Proposed Speaker Verification System

This section explains the process of development of the proposed SV system using Gaussian posteriorgrams. The process involves training a speaker specific GMM using the trained data and then calculating the posterior probabilities of train and the test utterances, which is described in details in the ensuing subsections.

3.1. Database

The studies in this work are carried out on the recently made available RSR2015 database designed specifically for text-dependent SV. It contains data from 300 speakers that comprises 143 female and 157 male speakers [19]. The database consists of three parts based on the variants of duration and text type, out of which *Part I* is taken to conduct the proposed work. In this subset there are 30 fixed pass phrase sentences of 3-4 seconds duration from each of the speaker for 9 different sessions. Out of this, 3 sessions are taken from each speaker to consider as enrollment session and the remaining as testing session for each of the fixed pass phrases. The 300 speakers are further divided into three groups, namely background, development and evaluation sets. The background set is used to develop the PLDA model for the i-vector based paradigm as discussed later, while the development and evaluation sets are used to test the performances of both the proposed as well as the contrast system. The development set consists of 47 female and 50 male speakers, while the evaluation set contains 49 female and 57 male speakers. The trials are evaluated under the condition of impostors speaking the correct pass phrases as mentioned in [19].

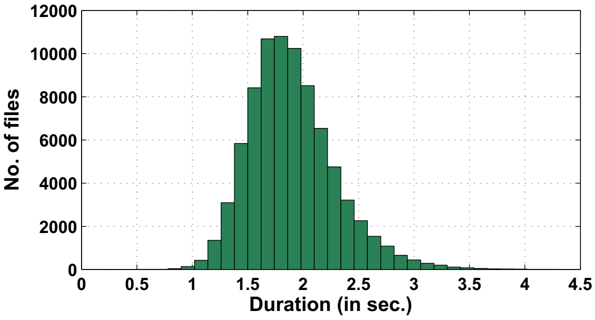

Figure 1: Histogram depicting the duration of utterances after end-point detection

3.2. Feature extraction

The train and the test utterances are processed in blocks of 20 ms window size with a shift of 10 ms. For each Hamming windowed frame, 39 dimensional MFCCs (13-base + 13-velocity + 13-acceleration coefficients) are extracted using 22 logarithmically spaced filters. The begin and end points of the utterances are detected using an energy based approach, where the frame energy is compared to the average energy of the utterance. MFCC features between the begin and end point of the utterance are then considered for further processing and saved as feature vectors after performing cepstral mean and variance normalization (CMVN) [20].

3.3. Development of proposed system using Gaussian Posteriorgrams

Gaussian posteriorgrams are probability based representation of feature vectors that indicate the probability of these feature vectors being generated by the different components of a GMM. Thus the dimension of the posterior vectors is equal to the number of components of the GMM. In order to obtain the Gaussian posteriorgrams, first the 39-dimensional normalized MFCCs of the training utterances of a speaker for one particular fixed pass phrase are taken to build a speaker specific GMM having 8 mixtures. This is done for all the 30 different fixed phrases separately, resulting in 30 GMMs for each speaker. The consideration of 8 mixtures for the GMM is justified by the fact that after end-point detection only about 2 seconds of speech data on an average is available in an utterance and three such sessions of a speaker are used to train the speaker specific GMM. This observation is depicted in Figure 1.

Once the GMMs are built, the next step is to map the train and test MFCCs into Gaussian posteriorgrams. This is done by using the feature vectors of the MFCCs and the respective speaker GMM and calculating the posterior probability vectors with the help of the Equation (1). The posterior vectors of each utterance are concatenated to form the Gaussian posteriorgram representation for that utterance as shown is Equation (2). The DTW algorithm is then applied between the train and the test posteriorgrams to measure the distance score which is used for the verification of a trial [18]. The DTW scoring mechanism utilizes the temporal sequence information present in the train and the test posteriorgrams with minimal amount of complexity. This process of development of proposed SV approach is illustrated in Figure 2. The performance of this proposed system is evaluated in terms of equal error rate (EER) and decision cost function (DCF) as per the evaluation procedure mentioned in [19].

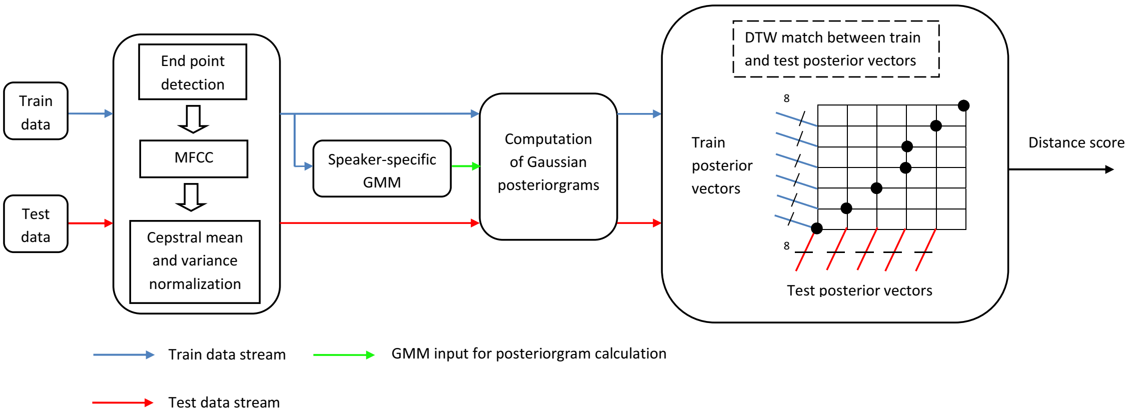
Train data
Test data
End point detection
MFCC
Cepstral mean and variance normalization
Speaker-specific GMM
Computation of Gaussian posteriorgrams
DTW match between train and test posterior vectors
Train posterior vectors
Test posterior vectors
8
8
Distance score
Train data stream
GMM input for posteriorgram calculation
Test data stream

Figure 2: Block diagram of the proposed Gaussian posteriorgram using DTW based fixed phrase SV system

4. Fixed Phrase i-vector Framework Based Contrast Speaker Verification System

The i-vector based paradigm has shown decent performance on evaluating it on the fixed pass phrase based SV as found in the literature [10, 11]. This section highlights the development of a contrast system using the i-vector framework for comparison with the proposed system.

4.1. i-vector framework for SV

The i-vector based speaker modeling technique provides a compact representation of GMM supervector of an utterance into a low dimensional subspace that posses the dominant information. The GMM mean vectors of an utterance are concatenated to form a GMM mean supervector, which is then projected using the total variability transformation matrix (T-matrix) to form the low rank compact vector called i-vector. For an utterance having the mean supervector $\boldsymbol{M_s}$, the corresponding i-vector $\boldsymbol{w_s}$ can be obtained as,

$$\boldsymbol{M_s} = \boldsymbol{m_u} + \boldsymbol{T}\boldsymbol{w_s} \tag{3}$$

where, $\boldsymbol{m_u}$ and $\boldsymbol{T}$ represent the universal background model (UBM) mean and the T-matrix respectively [13].

The train and the test utterances undergo similar transformation for generation of respective i-vectors. The generated i-vector requires channel/ session compensation for explicitly capturing the speaker information. PLDA based modeling can decompose the i-vector into speaker component and channel component, which is used for the i-vector system developed in this work [14].

4.2. Experimental setup

The SV studies over i-vector based framework are conducted on the same RSR2015 database for fixed phrase short utterances using 39 dimensional MFCCs as basic features over which CMVN is applied. The gender independent UBM having 1024 components is created using equal amount of male and female data, about 20 hours in total from the Switchboard Corpus II database for this setup. The zeroth and the first order statistics (sufficient statistics) of the train and the test utterances are computed using the UBM and then respective i-vectors are estimated by using a T-matrix of rank 400, that is learned using the sufficient statistics of the Switchboard Corpus II database. Then a gender independent PLDA model of 400 dimension is trained from the background set of speakers of the RSR2015 database consisting 50 male and 47 female speakers. This PLDA model is used as classifier for the development and evaluation set of the stated dataset to measure the contrast system performance.

5. Experimental Results and Discussions

The performances of the proposed system using Gaussian posteriorgrams and the contrast system using i-vector framework over the fixed pass phrase subset (*Part I*) of the RSR2015 database for both the development and evaluation sets are shown in Table 1. The performance of the contrast system based on the i-vector paradigm is found to be comparable to that reported in the recent literature over the same database [19]. The proposed system is noted to perform 4 folds better than the contrast system on an average. Figure 3 shows the detection error tradeoff (DET) curves for the proposed and the contrast SV system. To explain the noted improvements with our proposed approach we provide the following reasonings:

1. Use of speaker specific GMM, that mostly reflects in terms of speaker specific variances
2. In addition, creation of GMMs using text-dependent data helps in focusing on the required acoustic space
3. Further, DTW exploiting the temporal sequence information due to the text-dependency

The first reasoning is based on the ground that computation of Gaussian posteriorgrams involves the use of speaker specific variances. However, the i-vector based framework involves the global UBM variance, which is quite large as compared to the speaker specific variances obtained from a particular speaker specific GMM. This large variance of the UBM deemphasizes the speaker information under low data conditions of fixed pass phrases. But the use of speaker specific variances in Gaussian posteriorgrams based approach enhances the speaker characteristics and thus provides greater separability of true and false trial scores. To justify the same a controlled experimental setup is designed for the proposed system in which the variance parameters of all the mixtures in the GMM are deactivated. To do so a global variance of the feature vectors for all the speakers for each fixed phrase is computed. The variances of all the Gaussian mixtures in a speaker specific GMM for that particular

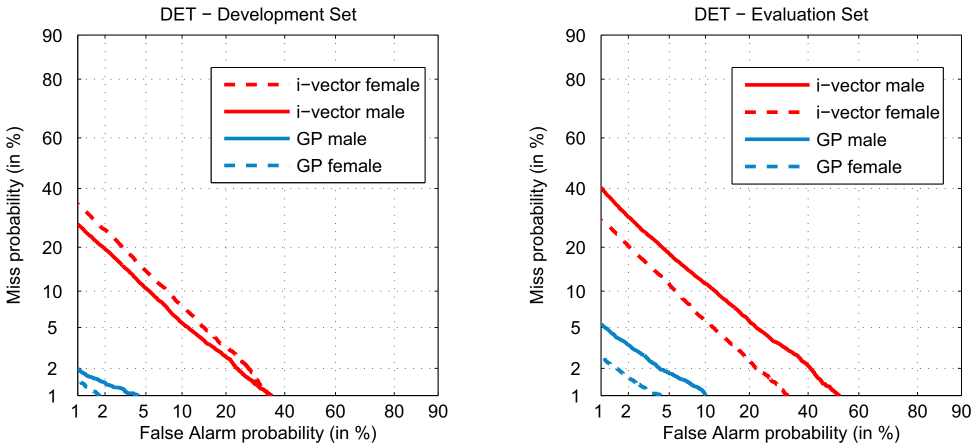

Figure 3: DET curves of i-vector and Gaussian posteriorgram (GP) based SV approaches

Table 1: *Performance comparison of the proposed SV system using Gaussian posteriorgrams (GP) against i-vector based contrast SV system on RSR2015 database*

System	Development Set				Evaluation Set			
	Female		Male		Female		Male	
	EER (%)	DCF×100	EER (%)	DCF×100	EER (%)	DCF×100	EER (%)	DCF×100
Proposed: GP	**1.26**	**0.59**	**1.54**	**0.73**	**1.75**	**0.81**	**2.77**	**1.26**
Contrast: i-vector	8.68	4.35	7.51	3.66	7.67	3.82	10.82	4.85

Table 2: *Performances of controlled experimentation done to verify the role of variance parameters of the speaker specific GMM over male speakers of RSR2015*

Development Set		Evaluation Set	
EER (%)	DCF×100	EER (%)	DCF×100
3.13	1.49	4.97	2.14

fixed phrase is replaced with the 10% scaled version of the computed global variance vector. In this way the role of the variance parameters gets deactivated and the computed Gaussian posteriorgram values are simply decided by the speaker specific mean parameters only. The experiments are conducted for only male speakers of the development and evaluation sets and the results are given in Table 2. It can be noticed that the performance degrades under this condition for the proposed Gaussian posteriorgram based approach, which supports the reasoning.

The second reasoning claims that a part of the gain has come due to the text-dependent data used for GMM training, that models only the acoustic space of interest. However, in the i-vector computation the UBM is trained using the text-independent data which expands the acoustic space and thus making the adapted GMM less specific for the fixed pass phrases. These less specific adapted GMMs are expected to result in a poorer posterior probability for a given test data in comparison to speaker and pass phrase specific GMM used in this work.

The third reasoning refers to explicit exploitation of temporal ordering of the acoustic events in DTW which is expected to enhance the performance of fixed pass phrase SV. In i-vector paradigm, the temporal information gets lost due to the involved projections and as a result the i-vector can not take explicit advantage of the same. From these observations, it seems that by creating the i-vectors with UBM conditioned on the fixed phrases present in the task can be beneficial in achieving improved performance in case of text-dependent SV.

6. Conclusion

This work proposes a novel way to use Gaussian posteriorgrams for fixed phrase short utterance based SV. The studies are conducted on the RSR2015 database. The proposed approach outshines the performance obtained from the contrast system developed over i-vector system by 4 folds on an average. The improvements obtained in Gaussian posteriorgram based approach is due to the fact that speaker specific weighting is used while calculating the posteriorgrams. This makes Gaussian posteriorgrams more speaker dependent and thus helps in discriminating among different speakers. This fact is further strengthened by the controlled experiment conducted, where non-speaker specific variances are used in the computation of Gaussian posteriorgrams. It is also argued that the text-dependent nature of the task is better exploited by the proposed method. The use of fixed pass phrases for GMM computation limits the acoustic space and thus results in higher posterior probabilities. Also the exploitation of the temporal sequence information by DTW enhances the performance of the proposed system as compared to that obtained from the contrast system. In the future, efforts will be directed towards making the i-vector framework utilize the text dependency of the task along with creating pass phrase specific UBMs for better speaker modeling.

7. Acknowledgement

This work is in part supported by a project grant 12(6)/2012-ESD for the project entitled “Development of Speech-Based Multi-level Person Authentication System” funded by the Department of Electronics and Information Technology (DeitY), Govt. of India.

8. References

[1] B.S. Atal, “Automatic recognition of speakers from their voices,” *Proceedings of the IEEE*, vol. 64, no 4, pp. 460–475, April, 1976.

[2] G. Doddington, “Speaker recognition identifying people by their voices,” *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1651–1664, November, 1985.

[3] J. Naik, “Speaker verification: A tutorial,” *IEEE Communications Magazine*, vol. 28, Issue 1, pp. 42–48, January, 1990.

[4] M. Hèbert, “Text-dependent speaker recognition,” *Springer-Verlag Heidelberg*, pp. 743–762, 2008.

[5] B. Yegnanarayana, S.R.M. Prasanna, J. Zachariah, and C. Gupta, “Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, July, 2005.

[6] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, vol. 52, pp. 12 – 40, 2010.

[7] T. Dutta, “Dynamic time warping based approach to text-dependent speaker identification using spectrograms,” in *Congress on Image and Signal Processing 2008, CISP ’08*, vol. 2, May 2008, pp. 354–360.

[8] D. Munteanu and S. Toma, “Automatic speaker verification experiments using HMM,” in *8th International Conference on Communications (COMM) 2010*, June 2010, pp. 107–110.

[9] A. Rosenberg, C. Lee, and S. Gokcen, “Connected word talker verification using whole word hidden Markov models,” in *International Conference on Acoustics, Speech, and Signal Processing 1991 (ICASSP-91)*, Apr 1991, pp. 381–384 vol.1.

[10] A. Larcher, P. Bousquet, K. A. Lee, D. Matrouf, H. Li, and J.-F. Bonastre, “i-vectors in the context of phonetically-constrained short utterances for speaker verification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2012*, March 2012, pp. 4773–4776.

[11] A. Larcher, K. A. Lee, B. Ma, and H. Li, “Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013*, May 2013, pp. 7673–7677.

[12] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[14] D. Garcia-Romero and C. Y. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *In Proc. Interspeech*, Aug 2011, pp. 249–252.

[15] Y. Zhang and J. Glass, “Unsupervised spoken keyword spotting via segmental dtw on Gaussian posteriorgrams,” *ASRU*, pp. 398–403, 2009.

[16] Y. Zhang and J. Glass, “Towards multi-speaker unsupervised speech pattern discovery,” in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 2010*, March, 2010.

[17] M. Bahari, R. Saeidi, H. Van hamme, and D. Van Leeuwen, “Accent recognition using i-vector, Gaussian mean supervector and Gaussian posterior probability supervector for spontaneous telephone speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013*, May, 2013.

[18] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1993.

[19] A. Larcher, K. A. Lee, B. Ma, and H. Li, “Text-dependent speaker verification: Classifiers, databases and RSR2015,” *Speech Communication*, vol. 60, pp. 56 – 77, 2014.

[20] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr 1981.