AN EFFICIENT GMM CLASSIFICATION POST-PROCESSING METHOD FOR STRUCTURAL GAUSSIAN MIXTURE MODEL BASED SPEAKER VERIFICATION

R. Saeidi, H. R. Sadegh Mohammadi**, M. Khalaj Amirhosseini**

* Electrical Engineering Department, Iran University of Science and Technology, Narmak, Tehran, I. R. Iran
** Iranian Research Institute for Electrical Engineering, No. 166, Heidarkhani Ave., Narmak, Tehran, I. R. Iran

ABSTRACT

In this paper a Gaussian mixture model (GMM) classifier, called GMM identifier, is proposed as an efficient post-processing method to enhance the performance of a GMM-based speaker verification system; such as Gaussian mixture model universal background model (GMM-UBM) and structural Gaussian mixture models with structural background model (SGMM-SBM) speaker verification schemes. The proposed classifier shows good performance while its computational load is almost negligible compared to the main GMM system. Experimental results show the superior performance of this post-processing method in comparison with a neural-network post-processor for such applications.

1. INTRODUCTION

Speaker verification has been an attractive research area in the past decades and statistical methods are dominant approach in this area since they provide superior performance compared to the other methods. A popular method for speaker verification is to model the speakers with the Gaussian mixture model (GMM) based on the maximum-likelihood (ML) criterion, which has been shown to outperform several other existing techniques [1]. The Gaussian mixture model universal background model (GMM-UBM) method for speaker verification has also demonstrated high performance in several NIST evaluations and has become the dominant approach in text-independent speaker verification [2]. In many speaker verification applications, accuracy and computational complexity are two major criteria for the selection of a proper system. In GMM-UBM speaker verification method, the major computation loads are the likelihood calculation for all mixtures of the UBM to select the highest scoring mixtures (top-C mixtures) and the likelihood calculation for the claimed speaker model [2]. Such a system with no optimization tends to use the majority of the processing time for scoring the Gaussian densities.

Several straightforward techniques have been investigated to increase the computational efficiency in a GMM-UBM speaker verification system while achieving an acceptable tradeoff between accuracy and complexity [3], [4]. In [4], a structural adaptation scheme is proposed which assumes a hierarchical structure of model common to all speakers and a multi-resolution GMM is used whose mean vectors are organized in a tree structure, with coarse-to-fine resolution when going down the tree. Xiang and Berger suggested the use of a neural network as a post-processor for the combination of such multi-resolution GMM which improves the performance of the system [5].

In this paper a GMM classifiers, called GMM identifier is proposed as a post-processor block with low complexity to enhance the performance of GMM based speaker verification systems. The remainder of the paper is organized as follows. In Section 2, a brief description of SGMM-SBM speaker verification is provided along with a short review of tree construction method for such application. Sections 3 explains the principles of the GMM identifier method and its training scheme. The computer simulation and experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

2. SGMM-SBM SPEAKER VERIFICATION METHOD

In GMM-UBM speaker verification, speakers are modeled with GMMs. A speaker-independent UBM is first trained using a large speech corpus which contains speech utterances from a rather large number of speakers. Then each speaker model's is derived from the UBM via Bayesian or maximum *a posterior* (MAP) adaptation method using the corresponding speaker's speech data [2]. Since the UBM and speakers' models are associated to each others, a fast scoring technique can be used as follows. For each input feature vector, all the UBM mixtures are scored to determine the top C highest scoring mixtures, and the speaker model likelihood is calculated using only the C speaker mixtures corresponding to the top C from the UBM, where C is much smaller than the order of speakers' GMM model (usually C is equal to 4 or 5).

Since in the GMM-UBM method, all mixtures of a UBM are used to calculate likelihood for each input vector, a heavy computational load is implies for the system. To efficiently find the top C mixtures, UBM's Gaussian

mixtures can be clustered hierarchically to organize a tree structure, which is called structural background model (SBM) [5]. In this way, the top mixtures for a given vector can be found easily by searching the tree. Then the speakers' models entitled structural Gaussian mixture models (SGMM) are adapted from SBM model. The evaluation of the claimed speaker's model is performed in similar manner to the GMM-UBM method. In the final stage of speaker verification the scores from the SBM and claimed speaker's SGMM models can be compared to accept/reject the claimed identities. While the simple subtraction can be used for such comparison, it has been shown that the use of a post-processor block such as a neural-network can improve the system performance [5].

3. GMM IDENTIFIER

Inspired by the use of neural networks as a post-processor block in [5], in this paper we propose the application of a second GMM classifier, which hereafter called GMM identifier, as a post-processor block instead of a neural network. It is noteworthy that the GMM identifier operates on the SGMM and SBM scores; therefore, its computational load is negligible compared to the main SGMM and SBM stage which operates on the feature vectors. Similar to the neural network post-processor, the GMM identifier can combine the scores from different layers of the tree-structured model. In the GMM identifier method, two separate GMM models are trained, one for the target speakers' trial scores and the other for the imposter speakers' trial scores.

Moreover, in the training stage of the GMM identifier one may train the target speakers' scores and the imposter speakers' scores model separately or train one of them first and then use the adaptation scheme to train the other one.

The number of mixtures for the GMM identifier affects both the performance of the overall systems and its computational complexity; therefore, it should be chosen properly to achieve the best performance.

4. PERFORMACE ASSESSMENT EXPERIMENTS

To evaluate the performance of the proposed post-processing method several experiments were performed and the results are compared with the competitive schemes. This section explains different aspect of these trials.

4.1. Database

The speaker verification experiments were conducted using a set of TV recorded speech database that recorded by the authors [6]. The database is a collection of conversational speech in Farsi, recorded from different channels of Iranian Broadcasting TV using a Winfast® TV card installed on a PC. Recordings were done when the speakers talked in noise free studios and there were no crosstalks or any musical background. The speech signals were recorded with PCM 11025 Hz, 16 bit and mono format. Ninety minutes of speech from 100 male speakers used for the UBM training. About three minutes speech for a set of separate 90 male speakers were also recorded to form the target speakers in the test stage. Two minutes of target speakers' speeches used for speaker model adaptation from speakers' models and the last one minutes of speeches were applied in the test procedure.

4.2. Evaluation Measure

The evaluation of the speaker verification system is based on detection error tradeoff (DET) curves, which show the tradeoff between false alarm (FA) and false rejection (FR) errors. We also used detection cost function (DCF) defined as [7]

$$DCF = C_{miss}.E_{miss}.P_{t\arg et} + C_{fa}.E_{fa}(1 - P_{t\arg et}) \qquad (1)$$

where P_{target} is the a priori probability of target tests with $P_{t\arg et} = 0.01$ and the specific cost factors $C_{miss} = 10$ and $C_{fa} = 1$.

4.3. Experimental Setup

At first, an SBM-SGMM system was trained using aforementioned database. 100 speakers used for UBM training and 90 other speakers used for the training of speakers models. Among them 30 speakers scores are held out for the training of the post-processing blocks. 33000 verification trials from the other 60 speakers are used in the test stage. No speaker overlap exists between the UBM and post-processing block training data and the test data. The duration of train segments in two sets of experiments was 15 and 45 seconds that were tested with test segments of 3 and 7 seconds, respectively. The ratios between target and impostor trials in both evaluations are 1:10. We used NIST guidelines in our evaluations. More details about the NIST evaluation guidelines can be found in [7].

The post-processing blocks use the scores of SBM and SGMM models for further processing to achieve higher performance for the entire speaker verification system. The overall diagram of such system is shown in Fig. 1, which is inspired by the block diagram used in [5]. Two post-processing blocks are considered for the results comparison. In the first system, this block is a multi-layered perception with one hidden layer similar to that presented in [5]. For the 15 seconds and 45 seconds speech training cases, 30 and 20 nodes in hidden layer were used, respectively; which provide good performance for the system under test [8]. In the second system, the post-processing block is comprised of a GMM identifier as explained in the previous section of

this paper. The zero normalization (Znorm) technique is also applied in the scores domain for all systems in our experiments [9]. In practice, a speaker model is tested against a set of speech signals produced by some impostors during the training stage, resulting in a speaker dependent impostor similarity score distribution. Speaker-dependent mean and variance normalization parameters are estimated from this distribution and applied on scores yielded by the speaker verification system in the test phase. One of the advantages of incorporating the Znorm in the system is that the estimation of the normalization parameters can be performed offline during the speaker model training stage.

In all tests variants GMMs of order 64 were employed during the experiments reported in this paper according to the findings reported in [6]. Moreover, the applied SBM-SGMM systems have 1-4-64 nodes tree structure.

4.3. Experiments Results

In the first stage, an experiment was conducted to find the best Gaussian model order for the GMM identifier for the system under investigation. Figs. 2 and 3 show the minimum DCF for the tests which use 15 and 45 seconds of training speeches of the 30 speakers subset of the speech database for different model order of the GMM identifier. In these experiments the duration of test utterances was 3 and 7 seconds, respectively. It can be seen that a GMM identifier with the model order of 16 provides the best performance in both cases. Also, it is observed from min DCF point of view, the SBM-SGMM systems with GMM identifier post-processor outperforms the SBM-SGMM systems with MLP post-processor. Moreover, its superiority is more intense for the short training and test utterances case.

In the second stage, the 60 speakers subset of the speech database was applied to evaluate the performances of three different SBM-SGMM speaker verification systems. The first system used no post-processing block and the final score is computed simply by the subtraction of SGMM score from SBM score. The second one is a SBM-SGMM speaker verification system with an MLP neural network with one hidden layer contains 30 (20) nodes for the 15 (45) seconds training speeches and 3 (7) seconds test speeches, respectively. These numbers of nodes confirmed to be the best choices for similar experiments with changing the number of hidden layer nodes within the range of 10 to 90 [8]. The third system is similar to the latter one except that a GMM identifier post-processing block of order 16 is applied in substitution for the neural network. Finally, the last system is the baseline UBM-GMM speaker verification system which apparently has higher computational complexity compared to the SBM-SGMM systems and just used for the comparison of systems' performances.

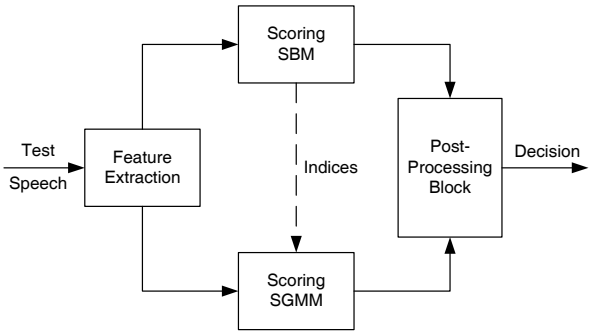

Fig. 1. The block diagram of the SBM-SGMM speaker verification system equipped with post-processing block.

Figs. 4 and 5 show the DET curves for aforementioned speaker verification systems of 15 (3) and 45 (7) seconds training (test) cases, respectively. The DET scores were computed following the guidelines presented in [7], i.e., 3000 and 30000 evaluations on true speakers and imposters, respectively. These results also confirm that the GMM identifier post-processor provides better performance from its MLP counterpart in this application.

5. CONCLUSIONS

In this paper a novel post-processor block called GMM identifier is proposed to enhance the performance of fast scoring GMM based speaker verification systems, such as SBM-SGMM system. The experiment results proves that the suggested method presents a desirable performance and it outperforms the already known neural network post-processing block which itself provides an acceptable performance for fast scoring GMMs.

ACKNOWLEDGEMENTS

The authors would like to acknowledge partial supports of the Iranian Research Institute for Electrical Engineering (IRIEE) and Iranian Telecommunication Research Center (ITRC) in the course of this research.

6. REFERENCES

[1] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech Audio Processing*, vol. 3, no. 1, pp. 72-83, Jan. 1995.

[2] D.A. Reynolds, T F. Quatieri, and R B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, Jan. 2000.

[3] J. Mclaughlin, D. A. Reynolds, and T. Gleason, "A study computation speed-ups of the GMM-UBM speaker recognition system," in *Proc. Eurospeech'99*, pp. 1215-1218, 1999.

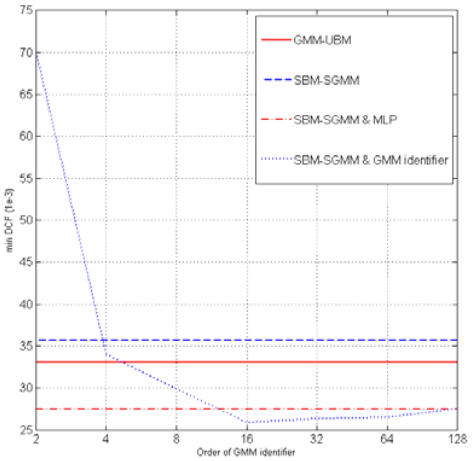

Fig. 2. Performance of SBM-SGMM system combined with GMM identifier post-processor in terms of minimum DCF with respect to the model order of GMM identifier using 15 and 3 seconds of speech segments for the training and test, respectively; in comparison with other systems.

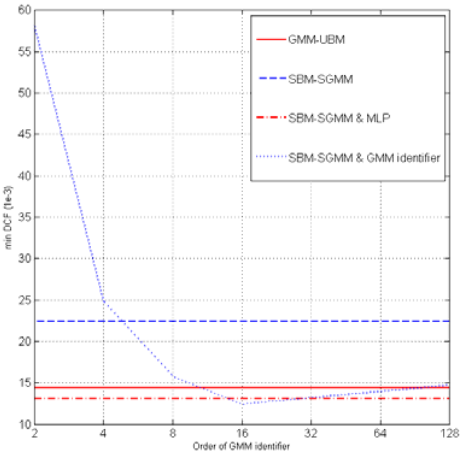

Fig. 3. Performance of SBM-SGMM system combined with GMM identifier post-processor in terms of minimum DCF with respect to the model order of GMM identifier using 45 and 7 seconds of speech segments for the training and test, respectively; in comparison with other systems.

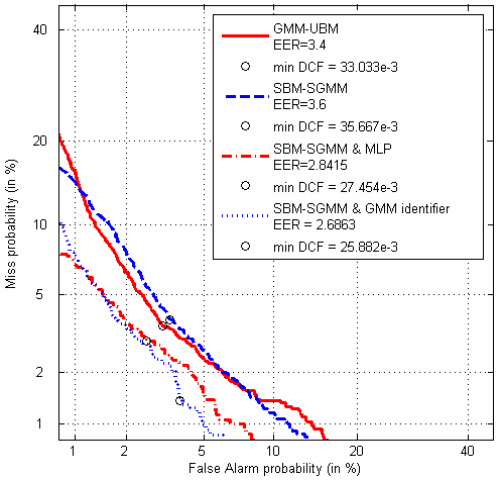

Fig. 4. Comparison of DET curves of four different GMM based speaker verification systems which use 15 and 3 seconds of speech segments for the training and test, respectively.

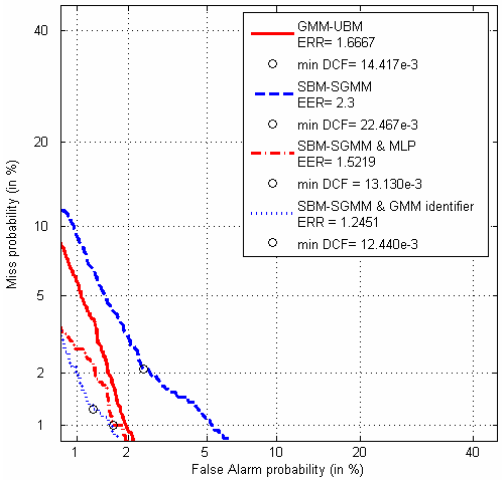

Fig. 5. Comparison of DET curves of four different GMM based speaker verification systems which use 45 and 7 seconds of speech segments for the training and test, respectively.

[4] K. Shinoda and C. H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 3, pp. 276-287, May 2001.

[5] B. Xiang and T. Berger, "Efficient text-independent speaker verification with structural Gaussian mixture models and neural network," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 447-456, Sept. 2003.

[6] R. Saeidi, H.R. Sadegh Mohammadi, and M. Khalaj Amirhosseini "Study of model parameters effects in adapted Gaussian mixture models based text independent speaker verification", in *Proc. International Symp. of Telecommunications, IST2005*, vol. 1, pp. 387-392, Shiraz, Iran, Sept. 2005.

[7] *The NIST Year 2000 Speaker Recognition Evaluation*, http://www.nist.gov/speech/tests/

[8] R. Saeidi, H.R. Sadegh Mohammadi, and M. Khalaj Amirhosseini "Efficient GMM-UBM system in text independent speaker verification using structural Gaussian mixture models", in *Proc. International Symp. of Telecommunications, IST2005*, vol. 1, pp. 39-44, Shiraz, Iran, Sept. 2005.

[9] G. Gravier, J. Kharroubi, and G. Chollete "On the use of prior knowledge in normalization schemes for speaker verification", *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 213-225, Jan. 2000.