COMBINED INTER-FRAME AND INTRA-FRAME FAST SCORING METHODS FOR EFFICIENT IMPLEMENTATION OF GMM-BASED SPEAKER VERIFICATION SYSTEMS

H. R. Sadegh Mohammadi, R. Saeidi**, M. R. Rohani**, and R. D. Rodman****

* Iranian Research Institute for Electrical Engineering, No. 166, Heidarkhani Ave., Narmak, Tehran, I. R. Iran
** Research Center of Intelligent Signal Processing, Shariati Ave., Tehran, I. R. Iran
*** Department of Computer Science, North Carolina State University, Raleigh, NC 27695-8206, US

ABSTRACT

In this paper a new inter-frame fast scoring scheme is proposed for Gaussian mixture model universal background model (GMM-UBM) speaker verification systems. It is combined with a recently introduced intra-frame efficient scoring method called the sorted Gaussian mixture model (SGMM) classifier which itself uses a sorted UBM known as the sorted background model (SBM). To enhance the performance of the system a GMM identifier is applied as a post-processing block. Experimental results show that the performance of this combined method compares favorably with the baseline GMM-UBM system, while the computational load of the proposed system is greatly less than that of the baseline system.

Index Terms— Speaker verification, GMM-UBM, fast scoring, speed-up, decimation

1. INTRODUCTION

Speaker verification research has been motivated in the past decade by the potential applications in several areas including e-business. A popular method for speaker verification is to model speakers with the Gaussian mixture model (GMM). Currently the Gaussian mixture model universal background model (GMM-UBM) method for speaker verification is considered to be the dominant approach in text-independent speaker verification [1]. In many speaker verification applications, accuracy and computational complexity are two major criteria for the selection of a proper system. In the GMM-UBM speaker verification method, the major part of the computational load is related to the likelihood calculation for all mixtures of the UBM, which select the highest scoring mixtures (top-C mixtures), and to the likelihood calculations for the associated mixtures in the claimed speaker model [1]. Such a system tends to use the majority of the processing time for scoring the Gaussian densities.

Several techniques have been investigated to increase the computational efficiency in a GMM-UBM speaker verification system while achieving an acceptable tradeoff between accuracy and the complexity, these include both inter-frame fast scoring schemes, such as the method introduced in [2], and intra-frame fast scoring techniques, e.g., the method proposed in [3]. While the former techniques normally use some sort of frame decimation in the feature domain, the latter ones exploit the benefits of structured GMM models for intelligent search and scoring tasks. It is well-known that the use of such fast scoring techniques normally results in degraded performance compared to the baseline GMM-UBM system. Therefore, a post-processing stage is usually applied to compensate for such degradation. Often it may even outperform the baseline system. Such a post-processing block may use a neural-network [4], a supported vector machine (SVM) [5], or even another GMM [6] classifier for this purpose.

In this paper we propose the use of a speaker verification system that exploits the benefits of a new decimation method for the inter-frame speed up combined with the optimized sorted version GMM-UBM [6] as the core classifiers, where a very low complexity GMM identifier is applied as a post-processor. Such a system provides both good performance and low computational costs which suits most speaker verification applications. The remainder of the paper is organized as follows. In Section 2, a brief description of the new decimation method is presented. Sections 3 reviews the principles of the optimized sorted version GMM-UBM and the GMM identifier method, and also describes their training scheme. The computer simulation and experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

2. INTER-FRAME FAST SCORING METHOD

In a GMM-UBM speaker verification system, where the order of UBM is M, the likelihood computational load for each frame of speech is of order $(M+C)$. If the test utterance following any silence or unvoiced speech contains N frames, then the total likelihood calculation complexity for the verification test is related to $N(M+C)$. By applying a frame-layer decimation as defined by Chan *et al.* [7], the complexity of the system can be reduced. The use

of the inter-frame decimation method as a pre-processing stage for the reduction of computational cost was proposed by McLaughlin *et al.* [2]. They used three different decimation techniques: fixed frame rate (FFR) decimation, variable frame rate (VFR) decimation, and adaptive frame rate (AFR) decimation. They reported that FFR and AFR decimation schemes outperformed the VFR method. Considering an FFR method, say by selection of one feature vector of each segment of D vectors, the computational cost is reduced to $N(M+C)/D$. It is expected that such a decimation method, even if D is chosen optimally, would slightly degrade system performance.

The reason for so little degradation despite several times computational load reduction, is that generally the feature vectors from adjacent frames change very little and therefore end up with nearly the same likelihood scores. The proposed method used for inter-frame decimation in this work, described in [8], uses an intelligent VFR decimation scheme. In this method, letting d_n be the L1-norm of the delta MFCC coefficients of frame n, their sum over l adjacent frames starting from frame n is defined as follows

$$y_{n,l} = \sum_{i=n}^{n+l-1} d_i \qquad (1)$$

where y is taken to be the main variable for the segmentation of speech. Starting from the n-th frame the length of the segment of feature vectors is l if

$$y_{n,l} < \tau \quad \text{and} \quad y_{n,l+1} \geq \tau, \qquad \tau \text{ a preset threshold} \qquad (2)$$

or if a silent or unvoiced speech frame is reached at frame $n+l-1$. In this work only the feature vector located in the middle of each segment is considered for likelihood computation, and the other frames of the segment are discarded. In this variable frame rate scheme, the likelihood computational cost for the same test speech utterance has an order of $N(M+C)/\widetilde{D}$, where $\widetilde{D}$ is the average decimation rate which is equal to the average segment length. Apparently, by increasing the threshold value, τ, the average decimation rate is also increased. This normally provides higher degradation of the verification performance.

3. INTRA-FRAME FAST SCORING METHOD AND POST-PROCESSING STAGE

For the intra-frame fast scoring stage, the sorted GMM method reported in [9] is applied which is briefly outlined: Given an L-dimensional feature vector $\mathbf{x}_t = [x_{1t}, x_{2t}, ..., x_{Lt}]^T$ related to the speech frame at the time interval t, and an M order GMM, a *sorting parameter* is defined as $s_t = f(x_{1t}, x_{2t}, ..., x_{Lt})$, where $f(\cdot)$ is a suitable function known called a sorting function. It is chosen in such a way that neighboring target feature vectors provide neighboring values near s_t. In this study $f(\cdot)$ is considered simply as the summation of the elements of the feature vector. The mixtures of the GMM are sorted in ascending order of the associated sorting parameter, according to the vector $\mathbf{S} = [s_1, s_2, ..., s_M]^T$ with $s_1 \leq s_2 \leq ... \leq s_M$.

To compute the likelihood of each input feature vector, the first step is to scalar quantize s_t by $\mathbf{S}$. Suppose s_i is the result of the scalar quantization, with $1 \leq i \leq M$. The index of s_i (i.e., i) is called the central index. In the next step, the input feature vector's likelihood is evaluated using the ordinary method by an extensive local search in the neighborhood of the central index, which includes an M_s mixtures subset taken from the entire mixtures, $M_s < M$. For example, only the mixtures with indices within the range of $i-k+1$ to $i+k$ may be searched, where k is an offset value ($k = M_s/2$).

To achieve a better performance for the sorted GMM, always $2k$ mixtures are searched, i.e., for the case of $i \leq k$, the first $2k$ mixtures in the GMM are considered for the local search, and for $i \geq M-k$ the last $2k$ mixtures are evaluated for the likelihood calculation. Generally, the computational complexity of this method grows linearly with M_s, which normally is set to be less than M. Therefore, for a test speech utterance with N voiced speech frames, which results in an N feature vectors likelihood evaluation, the sorted GMM computational load for $M_s > C$ and $M_s \leq C$ is equal to $N(M_s + C)$ and $2NM_s$, respectively, compared to the $N(M+C)$ for the ordinary GMM-UBM. For instance, for the case $M = 64, M_s = 16$, and $C = 5$, the likelihood computational costs of sorted GMM and GMM-UBM are $69N$ and $21N$, respectively. It is reported that the overall performance of sorted GMM speaker verification can be enhanced by the proper selection of the sorting parameter, by applying a suitable GMM optimization algorithm for the background model and finally adapting the speakers' GMMs using the ordinary method [9].

For the post-processing stage, use is made of a GMM identifier which has been reported to provide good performance enhancements [6]. That is, if the overall scores of UBM-GMM target speakers in a test speech utterance are considered to be a simple two dimensional vector, the performance enhancement comes when classifying the vector either as the target speaker or imposter using a dedicated GMM classifier with proper order. It is noteworthy that the computational load of such a GMM identifier is negligible since only two likelihood computations are performed for each test trial.

4. PERFORMANCE ASSESSMENT

To evaluate the performance of the proposed combined fast scoring method several experiments were performed and the results evaluated. This section explains different aspects of these trials.

4.1. Database

The speaker verification experiments were conducted using a set of TV speech database recorded by the authors. The database is a collection of conversational speech in Farsi, recorded from different different television channels using a Winfast® TV card installed on a PC. Recordings were taken when the speakers talked in noise free studios without crosstalk or musical background. The speech signals were recorded using single channgel PCM with a 11025 Hz sampling rate and 16 bit quantization. We used 123 minutes of speech from 130 male speakers for the training of a UBM with 64 Gaussian mixtures. Each speaker had 13 to 70 seconds of speech samples in the UBM training data set.

About three to four minutes of speech from a set of 110 separate male speakers was also recorded to create the set of target speakers for the test stage. The target speakers' speech utterances and models were separated into two different subsets. The first subset included the speech and models of 80 speakers for the main test set, and the other one included the speech and models for the remaining 30 speakers for the auxiliary test set, from which verification scores were drawn to train the GMM identifier employed in the post-processing stage. No speaker overlap exists between the UBM training data and the main and auxiliary test subsets.

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 the detection cost function (DCF) defined in [10]

$$DCF = C_{miss} E_{miss} P_{target} + C_{fa} E_{fa} (1 - P_{target}) \qquad (3)$$

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

4.3. Experimental Setup

At first, an ordinary GMM-UBM of order 64 was trained using the training dataset in two stages. The feature extraction used in the experiments was similar to the one reported in [6]. Later, speakers' GMMs were adapted using 30 seconds of speech samples according to the Bayesian or maximum *a posteriori* (MAP) adaptation method using the corresponding speaker's speech data [1]. Hereafter, this will be addressed as the baseline GMM-UBM. Then, the optimization algorithm reported in [9] was used to create the background model for a sorted GMM with $M_s = 16$ (M_s is the number of mixtures for which the likelihood is evaluated in each frame), and the optimized model was applied in creating the speakers' model adaptation. These optimized background model and adapted speakers' GMM models were applied in all tests except for the baseline GMM-UBM test. We carried out 50534 verification trials of three seconds each using speech from the main test set in the test stage, i.e., 4594 target speaker trials and 45940 trials for impostors. The NIST guidelines were applied in the evaluations [10]. No normalization scheme was employed in this work.

4.3. Experiments Results

In the first step, an experiment was conducted to evaluate the effects of the inert-frame fast scoring method described in Section 2. In this experiment the decimation scheme was applied combined with the ordinary GMM-UBM method. Fig. 1 shows the equal error rate (EER) and speed-up factor versus the change of the threshold value for this experiment. Results for such a system with threshold value τ equal to 1.9 are also given in Table 1 (System Type I).

In the second experiment, only the sorted GMM method was utilized for speaker verification. In this study, M_s was changed from 1 to 64 to evaluate the performance of the system as a result of such variations which is also related to the system computational load. The results are depicted in Fig. 2. System Type II in Table 1 show the results of this system for $M_s = 16$.

The third experiment used both the decimation pre-processing stage and the sorted GMM method. In this test based on the results shown in Figs. 1 and 2, M_s =16 and τ =1.9 were selected. This value of τ corresponds to a speed-up factor of 9.08 (System Type III in Table 1).

An experiment similar to the previous one was applied for the auxiliary test subset with 18942 trials, and its background and speakers' scores were used to train a GMM identifier in the score domain. Again the ratio of target to impostor trials is 1:10. The order of this GMM identifier was changed from 2 to 128 and the best performance was observed when the order 32. This trained GMM-UBM was applied in the final system.

In the last experiment, the proposed speaker verification system that uses both inter-frame and intra-frame fast scoring methods was combined with the GMM identifier post-processing block, and was evaluated using the main test set. The equal error rates and the minimum DCF values for this system are presented as System Type IV in Table 1 in comparison with the baseline GMM-UBM system (Baseline System) performance tested on the same test set. As observed in this table, the proposed speaker verification system presents very good results, while providing a considerable (29.82) computational load reduction, in comparison with the baseline system. Fig. 3 shows the DET curves [10] for the aforementioned experiments. The considerable improvement in performance

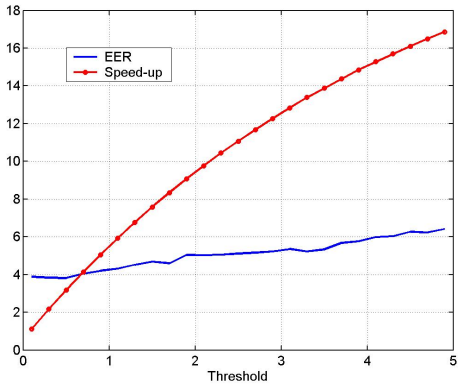

Fig. 1. Equal error rate and computation speed–up factor versus the change of threshold τ , for a speaker verification system with inter-frame fast scoring stage.

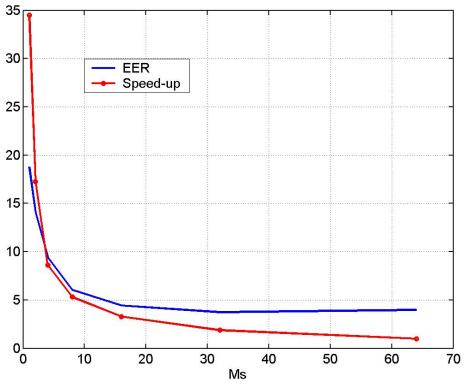

Fig. 2. Equal error rate and computation speed–up factor versus the change of M_s , for a speaker verification system with intra-frame fast scoring stage.

of the proposed system is quite noticeable in this figure.

5. CONCLUSIONS

In this paper an efficient speaker verification system which uses an inter-frame and intra-frame fast scoring algorithm combined with a GMM based post-processor is proposed. The experimental results show that this method performs well compared to the GMM-UBM baseline system, while its computational cost for the likelihood calculation is greatly less than the baseline system. Such computational saving provides enough space to incorporate more complicated algorithms with less error rates on DSP chips.

6. REFERENCES

[1] 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.

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

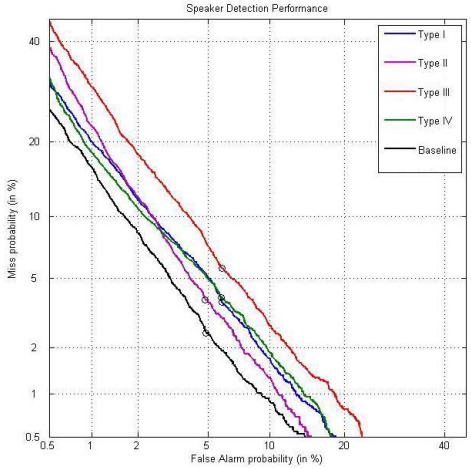

Fig. 3. Comparison of DET curves of five different GMM based speaker verification systems which use 30 and 3 seconds of speech segments for the training and test, respectively.

Table 1. Comparison of different experimental results

Speaker Verification System	Speed-up factor	EER%	Minimum DCF
Baseline	1	3.83	0.037
Type I	9.08	5.05	0.048
Type II	3.29	4.44	0.043
Type III	29.82	5.81	0.058
Type IV	29.82	4.98	0.049

[3] 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.

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

[5] Sh. Fine, J. Navratil, and R. A. Ganesh, "A hybrid GMM/SVM approach to speaker identification," in *Proc. ICASSP'01,* vol. 1, pp. 417-420*,* Salt Lake City, Utah, May 2001.

[6] R. Saeidi, H. R. Sadegh Mohammadi, and M. Khalaj Amir-Hosseini, "An efficient GMM classification post-processing method for structural Gaussian mixture model based speaker verification," in *Proc.*ICASSP'06, vol. 1, pp. 909-912, , Toulouse, France, May 2006.

[7] A. Chan, R. Mosur, A. Rudnicky, and J. Sherwani, "Four-layer categorization scheme of fast GMM computation techniques in large vocabulary continuous speech recognition systems," in *Proc. INTERSPEECH'2004*, pp. 689-692, 2004.

[8] R. Saeidi, H. R. Sadegh Mohammadi, R. D. Rodman, and T. Kinnunen, "A new segmentation algorithm combined with transient frames power for text independent speaker verification," submitted to ICASSP'07.

[9] H. R. Sadegh Mohammadi and R. Saeidi, "Efficient implementation of GMM based speaker verification using sorted Gaussian mixture model," in *Proc. EUSIPCO'06*, Florence, Italy, Sept. 4-8, 2006.

[10] *The 2000 NIST Speaker Recognition Evaluation*, http://www.nist.gov/speech/tests/spk/2000/index.htm