IMPROVED GMM-UBM/SVM FOR SPEAKER VERIFICATION

Minghui Liu[1,2], *Beiqian Dai*[2], *Yanlu Xie*[1,2], *Zhiqiang Yao*[1,2]

[1]MOE-Microsoft Key Laboratory of Multimedia Computing and Communication, University of Science and Technology of China

[2]Department of Electronic Science and Technology, University of Science and Technology of China

(liumh@ustc.edu, bqdai@ustc.edu.cn, xieyl@mail.ustc.edu.cn, zqyao@ustc.edu)

ABSTRACT

This paper combines Gaussian Mixture Model-Universal Background Model (GMM-UBM) and Support Vector Machine (SVM) through post processing the GMM-UBM scores of different dimension feature parameter with SVM in speaker verification. Because different dimension feature makes different contribution to recognition performance and SVM has good discriminability, this combining approach yields significant performance improvements on decision-making. Experiments on text-independent speaker verification in NIST'05 8conv4w-1conv4w data showed that the actual detection cost function (DCF) of the test system was reduced to 0.0290 from 0.0343.

1. INTRODUCTION

Current state-of-the-art approaches for text-independent speaker verification are based on Gaussian Mixture Model (GMM), which has good scalability and excellent ability in handling variable size sequences. In more recent years, Gaussian Mixture Model-Universal Background Model (GMM-UBM) has become the basis of the top performing systems in the NIST SREs for better performance and better robustness [1,2].

For better decision-making in the GMM-UBM speaker verification system, some decision-making methods have been proposed recently [2,3,4]. Traditionally, the decision-making is based on a likelihood ratio of an utterance to the hypothesized speaker GMM and UBM. Because of its discriminative properties, Support Vector Machine (SVM) performs better performance in static classification [5,6,7] and can construct flexible decision boundaries [8]. Thus, SVM can be combined with GMM-UBM by post processing scores obtained from GMM-UBM. In [4], Bengio processed the scores of GMM and UBM with an SVM instead of the traditional log-likelihood ratio and made a better decision.

In this paper, we propose an improved GMM-UBM/SVM to incorporate different dimension features' GMM-UBM scores using Support Vector Machines in text-independent speaker verification, because different dimension feature gives different contribution to recognition performance.

The paper is organized as follows. In section 2, we recall the classical speaker verification based on the traditional GMM-UBM. In section 3, we first give a brief introduction to SVM and the GMM-UBM/SVM, then our new method is described in detail. The experimental results are shown in Section 4 while some conclusions are given in Section 5.

2. BASELINE GMM-UBM

We have used Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker verification system as a baseline system [1]. The UBM is a large GMM trained to represent the speaker-independent distribution of features, which is constructed from a set of background speakers. The speaker model is derived from the UBM using MAP adaptation with the corresponding training data. All the models are diagonal covariance GMMs. For better performance, only the mean vectors are adapted.

Speaker verification can be described in terms of a two-hypothesis problem in which the verifier must decide whether the speech presented was from the hypothesized speaker H_0 or from an imposter H_1. Given an input sequence of T short-time speech feature vectors, $O = \{o_1, o_2, ..., o_T\}$, the hypothesis can be tested using the likelihood ratio,

$$\Lambda(O) = \frac{p(O \mid H_0)}{p(O \mid H_1)} = \frac{p(O \mid \lambda_{Tar})}{p(O \mid \lambda_{UBM})} \quad (1)$$

Where λ_{Tar} and λ_{UBM} represent models for the hypothesized speaker and imposter respectively. Furthermore, the log-likelihood ratio can be expressed as,

$$\log\Lambda(O) = \log p(O \mid \lambda_{Tar}) - \log p(O \mid \lambda_{UBM}) \quad (2)$$

During processing the log-likelihood ratio is compared with a threshold, θ, in order to decide hypothesis H_0 or H_1. The observations are assumed statistically independent, therefore the log-likelihoods of the observation sequence to

the hypothesized speaker model and the imposter model are given by,

$$\log p(\mathrm{O} \mid \lambda_{\mathrm{Tar}}) = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \log p(\mathrm{o}_{\mathrm{t}} \mid \lambda_{\mathrm{Tar}}) \quad (3)$$

$$\log p(\mathrm{O} \mid \lambda_{\mathrm{UBM}}) = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \log p(\mathrm{o}_{\mathrm{t}} \mid \lambda_{\mathrm{UBM}}) \quad (4)$$

3. POST PROCESSING GMM-UBM SCORES WITH SVM

3.1. Support Vector Machine (SVM)

Post processing GMM-UBM scores with SVM can be treated as a pattern classification problem if the given scores are considered as input patterns to be labeled as accepted/rejected. Under this point of view, any learning machine approach can be applied. In the last years, Support Vector Machine (SVM) has become an extremely successful discriminative approach to pattern classification.

The principle of SVM relies on a linear separation in a high dimension feature space where the data have been previously mapped, in order to take into account the eventual non-linearities of the problem. An SVM classifier has the general form:

$$f(x) = \sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b \quad (5)$$

Where $x_i \in R^n, i = 1,2,...,l$ are the training data. Each point of x_i belongs to one of the two classes identified by the label $y_i \in \{-1,1\}$. The coefficients α_i and b are the solutions of a quadratic programming problem [9]. α_i is non-zero for support vectors (SV) and is zero otherwise. *K* is the kernel function. Classification of a test data point **x** is performed by computing the right-hand side of equation (5).

Typical choices for kernel function K are:

Linear Kernel: $K(x_i, x_j) = < x_i, x_j >$ (6)

Radial Basis Function (RBF) kernel:

$$K(x_i, x_j) = \exp(-\| x_i - x_j \|^2 / 2\sigma^2) \quad (7)$$

3.2. GMM-UBM/SVM

In [4], Bengio incorporated the two scores obtained from GMM and UBM with an SVM. By Bayes decision rule, equation (2) is optimal so long as the hypothesized speaker and impostors are well modelled. Bengio proposed that the probability estimates are not perfect and that a better version would be,

$$a \log p(\mathrm{O} \mid \lambda_{\mathrm{Tar}}) - b \log p(\mathrm{O} \mid \lambda_{\mathrm{UBM}}) + c \quad (8)$$

Where a, b and c are adjustable parameters. Given a set of training data and labels, these parameters may be estimated using any learning algorithm. Bengio learned these parameters by post processing the scores with an SVM. The input to the SVM is the two dimensional vector made up of the log-likelihoods of the hypothesized speaker GMM and the UBM.

3.3. Improved GMM-UBM/SVM

The log-likelihood ratio for a test sequence of feature vectors O can be computed as

$$\log \Lambda(\mathrm{O}) = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} [\log p(\mathrm{o}_{\mathrm{t}} \mid \lambda_{\mathrm{Tar}}) - \log p(\mathrm{o}_{\mathrm{t}} \mid \lambda_{\mathrm{UBM}})] \quad (9)$$

The hypothesized speaker model was adapted from the UBM and the components of the adapted GMM retain a correspondence with the mixtures of the UBM, we can use a faster scoring method instead of merely evaluating the two GMMs [1].For each feature vector, select the top C scoring mixtures in the UBM and compute UBM likelihood using only these top C mixtures. Next, score the vector against only the corresponding C components in the adapted speaker model to evaluate the speaker's likelihood. Usually, a value of C=5 is chosen. Through experiments, as shown in Figure 2, we found that, when C=1 was chosen, similar performance was achieved. So in this paper, we determined C=1 in all GMM-UBM scoring experiments. Thus, GMM score can be replaced by a Gaussian mixture, equation (9) can be written as

$$\begin{aligned}
&\log \Lambda(\mathrm{O}) \\
&= \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} (\log[\omega_i \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2} \sum_{d=1}^{D} \frac{(x_{td} - \mu_{id})^2}{\sigma_{id}^2}\right\}] \\
&- \log[\omega_i \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2} \sum_{d=1}^{D} \frac{(x_{td} - \nu_{id})^2}{\sigma_{id}^2}\right\}]) \\
&= \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} (\frac{1}{2} \sum_{d=1}^{D} [\frac{(x_{td} - \nu_{id})^2}{\sigma_{id}^2} - \frac{(x_{td} - \mu_{id})^2}{\sigma_{id}^2}]) \\
&= \sum_{d=1}^{D} \frac{1}{2\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} [\frac{(x_{td} - \nu_{id})^2}{\sigma_{id}^2} - \frac{(x_{td} - \mu_{id})^2}{\sigma_{id}^2}]
\end{aligned} \quad (10)$$

Where ω_i and Σ_i are the mutual weight and diagonal covariance of UBM and the GMM, μ_{id} and ν_{id} are their dissimilar means of the *d*th dimension feature.

From equation (10), we can see that the score of system can be written as the sum of different dimension features' GMM-UBM scores. If the score of different feature was regarded as the final system score respectively, the performance of the systems vary remarkably, as shown in Figure 1. Thus, we can easily get that the scores of different dimension features make different contribution to recognition performance.

Because the scores of different dimension features make different contribution to recognition performance, if we incorporate them using a SVM just like the GMM-UBM/SVM system, the processed score will be more discriminable and a better decision can be made. For the feature is 32-dimensional in this paper, the input to the SVM is the 32-dimensional vector.

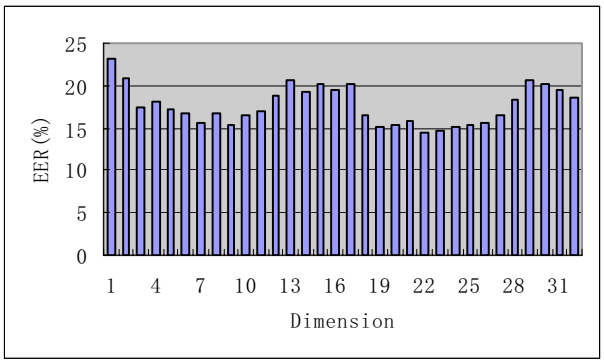

Figure 1. The EERs of the systems based on different dimension feature.

4. EXPERIMENTS

4.1. Database

To compare different approaches, we used the subset of the Switchboard telephone corpus used for NIST speaker recognition evaluations in 2005. Here, we used the 8conv4w-1conv4w data. About 2 hour of speech from NIST'04 Dev training data was used to build the UBM with 2048 Gaussians. Target models are derived by MAP estimation of the UBM parameters using the EVAL designated training data, only the mean vectors are adapted. Each of the 295 target speakers has about 20 minutes of speech for training. Among them 145 speakers are held out for training the SVM and deciding the threshold. 6882 verification trials from the other 150 speakers are used for the test. So there is no speaker overlap between the SVM training data and test data. The duration of each test segment is about 2 minutes. The ratios between target and impostor trials in both evaluations are roughly 1:10. More details about the NIST evaluation can be found at [10].

The frame rate is set to 10 ms. 16-dimensional Mel-frequency cepstral coefficients (MFCC) are extracted from silence-removed and bandlimited data first. The 16 delta coefficients are calculated based on the MFCCs and appended to form a 32-dimensional feature vector which is used in all the following experiments.

4.2. Detection cost Function (DCF)

For better evaluation of the speaker verification system, we used a detection cost function (DCF) defined for the NIST evaluation:

$$C_{Det} = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1 - P_{Target}) \quad (11)$$

The parameters of this cost function are the relative costs of detection errors, $C_{Miss} = 10$ and $C_{FalseAlarm} = 1$, and the a priori probability of the specified target speaker, $P_{Target} = 0.01$.

4.3. Experimental result

4.3.1. GMM-UBM top C selection

Three GMM-UBM verification results were compared when the top C mixtures were selected as 1, 5 and 2048 in NIST'05 8conv4w-1conv4w data. The experimental result was shown as Figure 2. We can see that the performance of the three systems was similar. So, we decided C=1 in the later experiments.

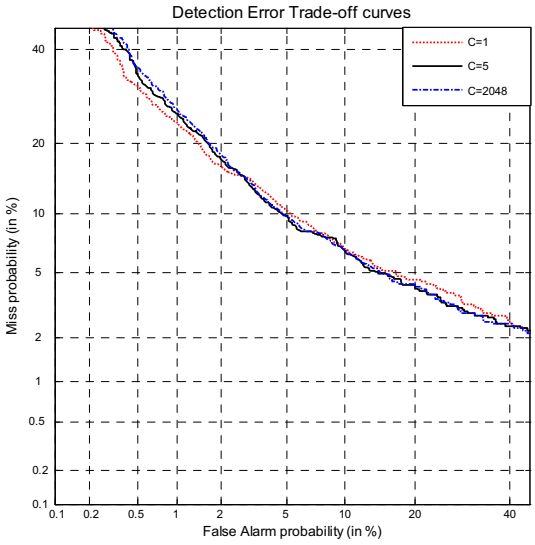

Figure 2. The DET curve of GMM-UBM when C=1, C=5 and C=2048.

4.3.2. Improved GMM-UBM/SVM

To compare the performance of different systems' decision-making, we trained the SVM and recorded the threshold where the DCF is minimal using the training data. With the SVM and the estimated threshold, we got the test system's actual DCF using the test data.

Table 1 gives the results of these experiments. The GMM-UBM/SVM system used a Linear SVM. In the improved GMM-UBM/SVM system, we compared a Linear

SVM as well as an SVM with an RBF kernel. These two systems were compared to the classical GMM-UBM system. We can see that GMM-UBM system yielded the worst results, while the improved GMM-UBM/SVM system with Linear kernel gave the best results. The relative improvement between the improved GMM-UBM/SVM and GMM-UBM is equal to 15.5%, which is particularly significant.

	Train (min DCF)	Test (actual DCF)
GMM-UBM	0.0324	0.0343
GMM-UBM/SVM	0.0315	0.0338
Improved GMM-UBM/SVM (RBF)	0.0293	0.0333
Improved GMM-UBM/SVM (Linear)	0.0248	0.0290

Table 1. The minimal DCFs on the training data and the actual DCFs on the test data with the estimated threshold from the training data.

From Table 1, we can also see that, in the improved GMM-UBM/SVM, Linear kernel performed much better than RBF kernel because of better robustness.

Figure 3 shows the performance comparison of GMM-UBM, GMM-UBM/SVM and the improved GMM-UBM/SVM. It's clear from the result that the improved GMM-UBM/SVM significantly outperforms both the traditional GMM-UBM and GMM-UBM/SVM. Obviously, substantial reduction was achieved in EER from 7.34% to 6.03%.

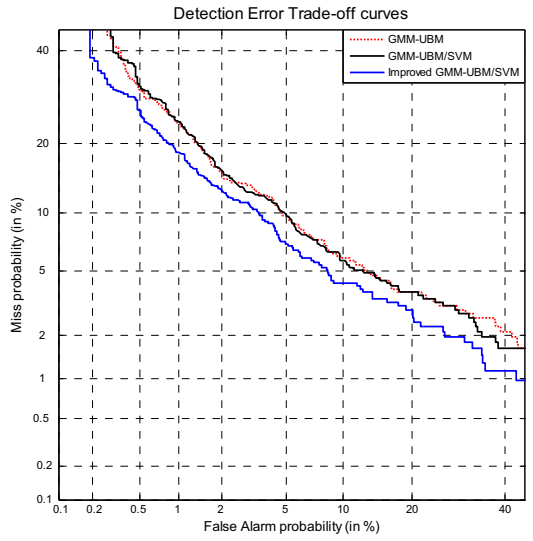

Figure 3. The DET curve of GMM-UBM, GMM-UBM/SVM and improved GMM-UBM/SVM.

5. CONCLUSION

In this paper, we have presented a new approach to replace the traditional GMM-UBM log-likelihood ratio with an incorporated score using SVM in text-independent speaker verification. In the improved GMM-UBM/SVM, we post processed different dimension features' GMM-UBM scores using SVM. This approach yields significant performance improvement on both decision-making and DET curve. Experiments in NIST'05 8conv4w-1conv4w data proved it.

6. ACKNOWLEDGMENTS

This research has been funded by the National Science Foundation of China (Project number 60272039). This research was also supported by the Science Research Fund of MOE-Microsoft Key Laboratory of Multimedia Computing and Communication (Grant No.05071810). The experiments using SVM have been realized using LIBSVM [11].

7. REFERENCES

[1] Douglas A. Reynolds, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp.19-41, 2000.

[2] G.R. Doddington, M.A. Przybocki, A.F. Martin, and D.A. Reynolds, "The NIST speaker recognition evaluation—overview, methodology, systems, results, perspective," *Speech Communication,* vol. 31, pp. 225-254, 2000.

[3] Ke Chen, "Towards better making a decision in speaker verification," *Pattern Recognition*, vol. 36, pp. 329-346, 2003.

[4] S. Bengio, "Learning the decision function for speaker verification," *Proc. IEEE ICASSP*, 2001.

[5] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 1–47, 1998.

[6] S. Fine, J. Navratil, and R. Gopinath, "A hybrid GMM/SVM approach to speaker identification," in *Proc. IEEE ICASSP*, 2001.

[7] V. Wan and S. Renals, "SVMSVM: Support Vector Machine speaker verification methodology," in *Proc. IEEE ICASSP*, 2003.

[8] D. Garcia-Romero, J. Fierrez-Aguilar, "Support vector machine fusion of idiolectal and acoustic speaker information in Spanish conversational speech", in *Proc. IEEE ICASSP*, 2003.

[9] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

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

[11] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "A Practical Guide to Support Vector Classification", available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.