IEICE TRANS. INF. & SYST., VOL.E92–D, NO.9 SEPTEMBER 2009

LETTER

Approximate Decision Function and Optimization for GMM-UBM Based Speaker Verification*

Xiang XIAO[†a)], **Xiang ZHANG**[†], **Haipeng WANG**[†], ***Nonmembers***, **Hongbin SUO**[†], ***Student Member***, **Qingwei ZHAO**[†], ***Member***, ***and*** **Yonghong YAN**[†], ***Nonmember***

SUMMARY The GMM-UBM framework has been proved to be one of the most effective approaches to the automatic speaker verification (ASV) task in recent years. In this letter, we first propose an approximate decision function of traditional GMM-UBM, from which it is shown that the contribution to classification of each Gaussian component is equally important. However, research in speaker perception shows that a different speech sound unit defined by Gaussian component makes a different contribution to speaker verification. This motivates us to emphasize some sound units which have discriminability between speakers while de-emphasize the speech sound units which contain little information for speaker verification. Experiments on 2006 NIST SRE core task show that the proposed approach outperforms traditional GMM-UBM approach in classification accuracy.

key words: *automatic speaker verification, contribution weight re-estimation, optimization*

1. Introduction

The GMM-UBM framework has proved to be a common technique in an automatic speaker verification (ASV) task [1]. A GMM-UBM system consists of two probabilistic models, one is the target model which is estimated via Maximum a Posterior (MAP) criterion using target utterance, and the other one is the universal background model (UBM) which is used to denote the non-target model. During the testing procedure, the decision function, defined as the likelihood ratio of the testing utterance between the target speaker model and the UBM is computed and then compared to a universal threshold, to make the decision whether the trial should be accepted (the given test utterance is spoken by the claimed target speaker) or rejected.

It is proposed in [2] that the GMM can be used to represent the underlying process that generates the multilingual data. The individual Gaussian components are trained to represent the underlying set of speech sound units (vowel, nasal, fricative etc.) in a self-organized manner.

In this letter, we first proposed an approximate estimation of traditional GMM-UBM based decision function. It is shown that the decision function of GMM-UBM can be treated approximately as an equally weighted linear combination of each Gaussian component's occupation probabilities, which means that in traditional GMM-UBM approach, the contribution of each Gaussian component is assumed to be equally weighted. However, the assumption is not suitable for classification. It is considered that different speech sound unit defined by Gaussian component may have different contribution weight to classification. Voiced sound units are more significant to speaker verification since they catch most of the vocal tract characteristics, and different types of vowels have different contribution weights to speaker verification as discussed in [3]. Thus, some speech sound units which are discriminative between speakers should be emphasized while some other speech sound units with little discriminability among speakers should be given less weight.

The outline of this letter is as follows. In Sect. 2, the decision function of traditional GMM-UBM approach is introduced. In Sect. 3, we proposed the approximate estimation of GMM-UBM based decision function. In Sect. 4, the re-estimation of the contribution weight of each Gaussian component's occupation probabilities is shown. The experiment results on NIST speaker recognition evaluation 2006 task data are listed in Sect. 5 and we present our conclusion in Sect. 6.

2. GMM-UBM Decision Function

The speaker verification can be viewed as a hypothesis test between

H_0: X is from the hypothesized speaker

and

H_1: X is not from the hypothesized speaker.

To make a decision between the two hypothesis, we define the likelihood ratio decision function as follows [1]

$$\Lambda(X) = \frac{P(X|H_0)}{P(X|H_1)} = \begin{cases} \geq threshold & \text{accept} \\ < threshold & \text{reject} \end{cases} \quad (1)$$

where $P(X|H_i)$ is the probability density function of hypothesis H_i given observed sequences $X=\{x_1, x_2, \ldots, x_T\}$. In the GMM-UBM approach, H_0 is represented by a GMM model λ_S, which characterizes the hypothesized target speaker, and the alternative hypothesis H_1 is represented by the universal background model (UBM) λ_U which is also a GMM model. If we take the logarithm of the likelihood ratio decision function, we get

$$\Lambda(X) = logP(X|\lambda_S) - logP(X|\lambda_U) \quad (2)$$

Manuscript received February 23, 2009.
Manuscript revised May 6, 2009.

[†]The authors are with the ThinkIT Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, China.

*This work is partially supported by MOST (973 program2004CB318106), National Natural Science Foundation of China (10574140, 60535030), The National High Technology Research and Development Program of China (863 program, 2006AA0101022006AA01Z195).

a) E-mail: xiaoxiang@hccl.ioa.ac.cn

DOI: 10.1587/transinf.E92.D.1798

The feature vectors of X are assumed to be independent, and we usually take the average log-likelihood value and the decision function will be

$$\Lambda(X) = \frac{1}{T}\sum_{t=1}^{T}(logP(x_t|\lambda_S) - logP(x_t|\lambda_U)) \tag{3}$$

Eq. (3) is the commonly used decision function of GMM-UBM approach [1].

3. Approximate Decision Function of GMM Based Speaker Verification

If we use $P(x_t)$ to denote the total probability of both λ_S and λ_U given observation x_t, where

$$P(x_t) = P(x_t|\lambda_S) + P(x_t|\lambda_U) \tag{4}$$

Eq. (3) can be rewritten as follows

$$\Lambda(X) = \frac{1}{T}\sum_{t=1}^{T}\left(log\frac{P(x_t|\lambda_S)}{P(x_t)} - log\frac{P(x_t|\lambda_U)}{P(x_t)}\right) \tag{5}$$

Two terms of the Taylor series, $log(x) \approx x - 1$ are used to obtain the approximation of Eq. (5) and we discard the -1 since the change will not affect the classification decision [4]

$$\tilde{\Lambda}(X) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{P(x_t|\lambda_S)}{P(x_t)} - \frac{P(x_t|\lambda_U)}{P(x_t)}\right) \tag{6}$$

Here we use $\tilde{\Lambda}(X)$ to denote the approximate estimation of $\Lambda(X)$.

Notice that the Taylor series, $log(x) \approx x - 1$ should satisfy that $x \approx 1$, however, we still have several reasons for using the Taylors series approximation. One reason is that the ratio $P(x_t|\lambda_i)/P(x_t)$ vary over a small dynamic range. Second the approximation preserves score order so it will not affect the classification results significantly.

While in a Gaussian mixture model (GMM) λ, the probability density $P(x_t|\lambda)$ given observed vector x_t is

$$P(x_t|\lambda) = \sum_{m=1}^{M} w_m p_m(x_t) = \sum_{m=1}^{M} g_m(x_t) \tag{7}$$

where M is the Gaussian component number, w_m is the weight of the m-th Gaussian component $p_m(*)$. Here we adopt $g_m(*)$ to denote the m-th weighted Gaussian component for simplification. The Eq. (6) can be written in a GMM form, that is

$$\tilde{\Lambda}(X) = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{m=1}^{M}\frac{g_m^{Spk}(x_t) - g_m^{Ubm}(x_t)}{P(x_t)}\right) \tag{8}$$

where $g_m^{Spk}(x_t)$ denotes the m-th Gaussian component of target speaker GMM and $g_m^{Ubm}(x_t)$ denotes the m-th Gaussian component of UBM.

Note that the since $P(x_t)$ is the total probability of both target model λ_S and UBM λ_U, we can combine the target GMM and UBM together, to form a combined GMM λ_C which has $2M$ Gaussian components

$$P(x_t|\lambda_C) = \sum_{m=1}^{2M}\hat{g}_m(x_t) \tag{9}$$

where

$$\hat{g}_m(x_t) = \begin{cases} \frac{1}{2}g_m^{Spk}(x_t) & \text{if } m \le M \\ \frac{1}{2}g_{m-M}^{Ubm}(x_t) & \text{if } m > M \end{cases}$$

Here we have each Gaussian component's weight half to satisfy that the sum of Gaussian component's weight of λ_C is 1. The new GMM λ_C can be thought to define an acoustic space which consists of sound units both from target speaker and non target speakers.

We can define

$$\gamma_m(t) = P(m|x_t, \lambda_C) = \frac{\hat{g}_m(x_t)}{P(x_t|\lambda_C)} \tag{10}$$

as the occupation probability $\gamma_m(t)$ of the m-th Gaussian mixture component given observation x_t and

$$\bar{\gamma}_m(X) = \frac{1}{T}\sum_{t=1}^{T}\gamma_m(t) \tag{11}$$

as the average occupation probability of the m-th Gaussian mixture component for observation sequence X. We can rewrite Eq. (8) as follows

$$\tilde{\Lambda}(X) = \langle\Phi(X), \Theta\rangle \tag{12}$$

where $\langle\cdot,\cdot\rangle$ denotes inner product, and

$$\Phi(X) = [\bar{\gamma}_1(X), \bar{\gamma}_2(X), \ldots, \bar{\gamma}_{2M}(X)]^t \tag{13}$$

is the $2M$ dimensional occupation vector. Θ is a $2M$ dimensional vector, which consists of M ones followed by M negative ones. Θ is named as the contribution weight vector since it denotes the contribution weight of each Gaussian component's occupation to decision function.

$$\Theta = [1, 1, \ldots, 1, -1, -1, \ldots, -1]^t \tag{14}$$

Here, the average occupation probability $\bar{\gamma}_m(X)$ can be thought to represent the frequency of occurrence of sound unit m in the whole observation sequences X and the value of $\bar{\gamma}_m(X)$ represents the contribution to decision function of Gaussian component m. It is shown in Eq. (14) that the contribution of each Gaussian component to decision function is assumed to be equally weighted. However, as we discussed in Sect. 1, different speech sound unit defined by Gaussian component may have different contribution weight to classification, and the problem of re-estimation of the contribution weight vector Θ can be converted into the problem of how to obtain an optimum solution of the linear classifier motivated by the inner product form of Eq. (12).

It is noticed that since we obtain an approximate estimation of the decision function, upon which the following

optimization is based, the effect on the classification after introducing the two terms of Taylor series should be examined first. In Sect. 5, before we give the experiment results of GMM-UBM and proposed approach, we show that the classification results of GMM-UBM are almost not affected after we introduce the Taylor series.

4. Re-Estimation of The Contribution Weight Vector

4.1 MSE Criterion

In this section, we will show how to obtain an optimum solution of the contribution weight vector Θ based on the minimizing the sum-of-squares error function (MSE) criterion [5].

Suppose we have a training set consist of n trials, each of which can be mapped into a vector Φ_i, we labeled $y_i = 1$ for true trial vectors(meaning two utterances of the trial are from the same speaker) and $y_i = -1$ for the false trial vectors. For convenience, we define $\Psi_i = y_i\Phi_i$ and we aim to find a Θ^* that for each Ψ_i, we have $\langle \Psi_i, \Theta^* \rangle > 0$, and the sum-of-squares error function $\mathbf{J}_\Theta$ is defined as

$$\mathbf{J}_\Theta = \sum_{i=1}^{n} \left(\Psi_i^t \Theta - b_i \right)^2 \tag{15}$$

where b_i is an arbitrary positive constant, if we define matrix $\mathbf{Y}$ as follows

$$\mathbf{Y} = [\Psi_1, \Psi_2, \ldots, \Psi_n]^t$$

we have

$$\mathbf{J}_\Theta = \|\mathbf{Y}\Theta - \mathbf{b}\|^2 \tag{16}$$

Here $\mathbf{b}$ is a vector consisting of arbitrary positive constants. To solve the minimizing problem, we can find

$$\nabla \mathbf{J}_\Theta = 2\mathbf{Y}^t(\mathbf{Y}\Theta - \mathbf{b}) \tag{17}$$

Let $\nabla \mathbf{J}_\Theta = 0$ and we can get the optimum Θ as follows

$$\Theta^* = \left(\mathbf{Y}^t\mathbf{Y} \right)^{-1} \mathbf{Y}^t\mathbf{b} \tag{18}$$

However, the MSE solution depends on the vector $\mathbf{b}$, it is obvious that different $\mathbf{b}$ will give the solution of different properties. Next part, we will show how to get a largest margin solution based on support vector machines (SVMs) [4].

4.2 MSE Mapping for SVMs

For a testing trial vector Ψ_x, the decision function is

$$\tilde{\Lambda}(X) = \Psi_x^t\Theta^* = \Psi_x^t \left(\mathbf{Y}^t\mathbf{Y} \right)^{-1} \mathbf{Y}^t\mathbf{b} \tag{19}$$

Here we set $\mathbf{R} = \mathbf{Y}^t\mathbf{Y}$ and $\hat{\Psi}_y = \mathbf{Y}^t\mathbf{b} = \sum_{i=1}^{n} b_i\Psi_i$. Eq. (19) will be

$$K_{\mathtt{MSE}} = \Psi_x^t\mathbf{R}^{-1}\hat{\Psi}_y \tag{20}$$

Here are several comments for the MSE kernel defined by Eq. (20).

First, the value of b_i has no effect on the SVM solution. A typical SVM solution is

$$\Theta^* = \sum_{i=0}^{n} \left(\alpha_i\mathbf{R}^{-1}x_i \right) + d \tag{21}$$

Given the kernel $\mathbf{R}$, and all of the training vectors $x_i, i \in \{1, 2, \cdots, n\}$, SVMs training procedure will find the corresponding α_i for each x_i and an universal d. In this letter, the solution defined by Eq. (21) is re-written as follows

$$\begin{aligned}\Theta^* &= \sum_{i=0}^{n} \left(\alpha_i\mathbf{R}^{-1}b_i\Psi_i \right) + d \\ &= \sum_{i=0}^{n} \left(\alpha_i b_i\mathbf{R}^{-1}\Psi_i \right) + d \\ &= \sum_{i=0}^{n} \left(\alpha_i^*\mathbf{R}^{-1}\Psi_i \right) + d\end{aligned}$$

That is, when the kernel $\mathbf{R}$ and the training vectors Ψ_i are given, no matter what the value of b_i is, the solution Θ^* obtained by SVMs is unique.

Second, please note that $\mathbf{R}$ defined by Eq. (20) is not a semi-positive matrix, so it is not a kernel yet, and an approximation is necessary for the MSE kernel application. An useful assumption is that the trial vectors Ψ_i of the training set are independent, and thus, we only take $\hat{\mathbf{R}}$ as the diagonal matrix of $\mathbf{R}$, this approximation significantly reduces the computation and makes $\mathbf{R}$ semi-positive. Eq. (20) can be re-written as

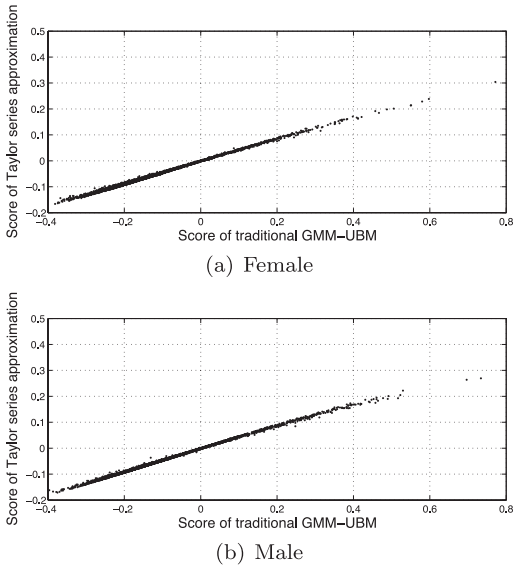
$$K_{\mathtt{MSE}} = \left(\hat{\mathbf{R}}^{-\frac{1}{2}}\Psi_x \right)^t \left(\hat{\mathbf{R}}^{-\frac{1}{2}}\hat{\Psi}_y \right) \tag{22}$$

Here $\hat{\mathbf{R}}^{-\frac{1}{2}}\Psi_x$ is defined as the MSE mapping of Ψ_x.

Eq. (22) has shown the way of re-estimating Θ^*. First, each trial vector of the training set is used to compute the diagonal matrix $\hat{\mathbf{R}}$. Second the MSE mapping of each trial vector, both for training and testing is calculated, and a linear kernel based SVMs is adopted to find the solution.

5. Experiment

As we discussed in Sect. 3, the experiments are conducted on 2006 NIST speaker recognition evaluation (SRE) core task [6] gender dependently (22490 male trials and 29934 female trials) to examine the effect on classification after introducing the Taylor series. Score of each trial is obtained respectively from traditional GMM-UBM approach (defined in Eq. (3)) and Taylor series approximation (defined in Eq. (12)). The relationship of scores are shown in Fig. 1. It is obvious that the relationship between scores obtained from traditional GMM-UBM and Taylor series approximation are almost linear which means that for the purpose of classification, the introducing of Taylor series makes almost no difference on the results.



(a) Female

(b) Male

Fig. 1 The effect on classification of approximation by taylor series.

Table 1 Trial numbers of 2004 and 2006 NIST-SRE core task.

Year-Gender	Trial Number
2004-female	14069
2004-male	11228
2006-female	29934
2006-male	22490

We compare the performance of proposed approach to traditional GMM-UBM approaches on the 2006 SRE core task gender dependently, without score normalization and with speaker adaptive test score normalization (AT-Norm) respectively [7].

All the systems are based on the same UBM, which has 1024 Gaussian components, gender dependently, and during the enrollment procedure, only the means are adapted for the target model with a relevance factor of 12. The relevance factor is used to compute the adaptation coefficient and further details can be found in [1]. We do not apply any feature compensation techniques in this letter. The speech is first processed by a feature extraction module that computes a 18-dimension frame of MFCC every 20 milliseconds with 10 milliseconds overlap (100 frames per second). Delta-cepstral coefficients are then computed and appended to the cepstral vector, producing a 36-dimensional feature vector. For the AT-Norm, the pool of cohort model consists of 300 speaker models and top 60 models are being selected.

For the proposed system, the contribution weight vector is estimated gender dependently with trials only from the 2004 NIST-SRE core task. We do not apply any trials from 2005 NIST-SRE task since we found that there is overlap between the data set of 2005 NIST-SRE task and 2006 NIST-SRE task.

The data sets of 2004 and 2006 NIST-SRE are listed in Table 1. The linear kernel based SVMs training procedure is conducted with the package SVM-light which is released by University of Dortmund [8].

Table 2 Experiment result of 2006 NIST-SRE core task.

System/minDCF*100	female	male	all
(1)Traditional GMM-UBM	4.47	3.92	4.25
(2)GMM-UBM Approximation Without Re-Estimation	4.47	3.92	4.25
(3)GMM-UBM Approximation With Re-Estimation	4.01	3.61	3.84
Relative Improvement	10.3%	7.9%	9.6%
(2)+AT-Norm	3.99	3.68	3.87
(3)+AT-Norm	3.68	3.47	3.59
Relative Improvement	7.8%	5.7%	7.2%

We used the minimum decision cost function (MinDCF) [6] for evaluation, which is

$$DCF = C_{Miss}P_{Target}P_{(Miss/Target)} + C_{FA}P_{NonTarget}P_{(FA/NonTarget)} \quad (23)$$

where $C_{Miss} = 10$, $C_{FA} = 1$ and $P_{Target} = 0.01$, $P_{NonTarget} = 0.99$.

Experiment results are listed in Table 2. Comparison between (1) and (2) shows that the approximation of Taylor series makes no difference on the classification results. Comparing (2) to (3), we can observe that the re-estimation of contribution weights significantly improves the classification accuracy both without and with AT-Norm.

6. Conclusion

In this letter, we introduced and evaluated the contribution weight re-estimation method for GMM-UBM based ASV task. The motivation of re-estimating the contribution weight was based on the concept that different speech sound unit defined by Gaussian component has different contribution weight to speaker verification. Experiments on NIST-SRE 2006 data set show that the re-estimation of contribution weights adopted by SVMs improves the performance of speaker verification significantly.

References

[1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," Digit. Signal Process., vol.10, no.1-3, pp.19–41, 2000.

[2] K. Lee, C. You, and H. Li, "Spoken language recognition using support vector machines with generative front-end," Proc. IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP 2008, pp.4153–4156, 2008.

[3] T. Kitamura and P. Mokhtari, "Effects of vowel types on perception of speaker characteristics of unknown speakers," International Workshop on Nonlinear Circuits and Signal Processing, NCSP 06, 2006.

[4] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," Comput. Speech Lang., vol.20, no.2-3, pp.210–229, 2006.

[5] R. Duda, P. Hart, and D. Stork, Pattern classification, Wiley, New York, 2001.

[6] M. Przybocki, A. Martin, and A. Le, "NIST speaker recognition evaluation chronicles-part 2," Speaker and Language Recognition Workshop, IEEE Odyssey 2006, pp.1–6, 2006.

[7] D. Sturim and D. Reynolds, "Speaker adaptive cohort selection for

tnorm in text-independent speaker verification," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP'05), 2005.

[8] T. Joachims, "SVMLight: Support vector machine," SVM-Light Support Vector Machine, http://svmlight.joachims.org/, University of Dortmund, 1999.