On the Use of Gaussian Component Information in the Generative Likelihood Ratio Estimation for Speaker Verification

Rong Zheng¹, Bo Xu¹ ²

¹ Digital Content Technology Research Center, Institute of Automation
² National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing 100190, China
{rzheng,xubo}@hitic.ia.ac.cn

Abstract

This paper presents an experimental study of exploiting Gaussian component information for speaker verification. The motivation of the proposed algorithm is to examine detailed component information by using individual Gaussian component’s contribution to the final output score. Analysis of component-specific score is important to understand in-depth Gaussian mixture’s impact on performance. We present a new method, called Gaussian component information based likelihood ratio (GCILR), to introduce and weight component-dependent information based on adapted Gaussian mixture models. Performance evaluations comparing our system to the well-known technique, generative likelihood ratio estimation, are provided. The paper discusses how the performance is influenced by different significance in the informative component-specific scores. Comparison experiments conducted on the NIST 2006 SRE dataset show the effectiveness of the proposed method.

Index Terms: Speaker verification, generative likelihood ratio, GMM-UBM, Gaussian component information

1. Introduction

For text-independent speaker verification, the Gaussian mixture model (GMM) and support vector machine (SVM) are the most popularly used methods for many years. The Gaussian Mixture Model-Universal Background Model (GMM-UBM) framework has been proven to be highly successful due to good performance in several NIST evaluations [1, 2].

The decision-making in the GMM-UBM recognition system is commonly based on a likelihood ratio of an utterance between the hypothesized speaker GMM and UBM model [3]. Some methods towards making a better decision have been presented to improve the conventional log-likelihood-ratio. Chen proposed a method to use the more reliable statistics for threshold-setting and eliminate abnormal data for better estimation of underlying statistics [4]. In [5], instead of the traditional log-likelihood-ratio, Bengio processed the scores of GMM and UBM with SVM and provided good performance. Liu et al. presented an improved GMM-UBM/SVM to incorporate different dimension features’ GMM-UBM scores using SVM [6]. However, all the techniques mentioned above can be considered as score-level post-processing techniques.

Moreover, some recent studies deal with the conception of component likelihood ratios. Li et al. developed AdaBoost-GMM method to boost GMM based speaker verification by combining individual Gaussian mixture models using AdaBoost learning algorithm [7]. In [8], Gaussian information bottleneck method assumed that both the source and target variables were high dimensional multivariate Gaussian variables, and then GIB was applied to the super vector dimension reduction for speaker recognition.

In this paper we explore an in-depth study of better score computation method using Gaussian component-level information for improved speaker verification without much computational load. We make the assumption that the generative likelihood ratio estimation in GMM-UBM is not enough to retain Gaussian component-specific relevant information and relative importance. More detailed component information needs to be examined in order to improve the speaker verification performance.

Firstly, the average log-likelihood-ratio of test speech is approximated and decomposed into Gaussian component-dependent likelihood calculation. Then different impact on recognition performance is found when each largest scoring component’s score as singleton classifier for verification. Component-specific score combination with weighting strategy is finally applied and experimental evaluations are compared with generative likelihood ratio estimation.

The remainder of this paper is organized as follows. In section 2, a brief introduction to the conventional GMM-UBM speaker verification is provided. Section 3 describes the proposed Gaussian component-dependent likelihood estimation method. The implementation issues are also detailed. The experimental results are shown in section 4. Finally, some conclusions are given in section 5.

2. GMM-UBM based speaker verification

In GMM-UBM based speaker verification, a UBM is trained using speech utterances from a large group of speakers to represent the characteristics of all different speakers. Each speaker model is derived from the UBM by employing maximum a posteriori (MAP) adaptation using speaker-specific training speech [3]. The UBM and speaker models are modeled by GMM that is a weighted sum of multivariate Gaussian probability distributions. The classification for speaker verification may be performed based on the generative log-likelihood-ratio (LLR). Given a sequence of input feature vectors which are assumed statistically

\* This work was supported by a grant from the National Natural Science Foundation of China (No. 90820303)
independent, \( X = \{ x_1, x_2, ..., x_n, x_r \} \), the verifier can be tested using the following generative LLR,

\[
\text{LLR} = \frac{1}{T} \sum_{t=1}^{T} \left[ \log p(x_t | \lambda_{\text{tar}}) - \log p(x_t | \lambda_{\text{UBM}}) \right]
\]

where \( \lambda_{\text{tar}} \) and \( \lambda_{\text{UBM}} \) represent the target speaker model and UBM model. The LLR is compared with a preset threshold to provide the verification result.

\[3. \text{ Exploring Gaussian Component Information in Generative Likelihood Ratio Estimation}\]

\[3.1. \text{Gaussian component information}\]

The generative likelihood ratio estimation calculates one likelihood score for each input vector. Only the sum of Gaussian component likelihood values is considered, which corresponds to the total contribution for this speech frame. In this section, individual Gaussian component’s importance is included as its relevant information for the vector sequence. We can denote this component-dependent likelihood calculation as the detailed component information. Then, all the component-dependent scores are discriminatively weighted and combined to produce the likelihood ratio. Finally, averaging can be accomplished by dividing the number of speech frames, which can be regarded as rough duration compensation.

A GMM with \( M \) Gaussian components is parameterized mathematically by the notation \( \lambda = [\mu, \Sigma, w] \), \( \mu \in \{ \mu_1, \Sigma_1 \} \), \( \Sigma \) is the mixture weight, mean, and diagonal covariance of the \( m \)-th Gaussian component. Thus, generative LLR can be expressed as a Gaussian mixture.

\[
\text{LLR} = \frac{1}{T} \sum_{t=1}^{T} \log \frac{\sum_{m=1}^{M} w_m p(x_t | g_m^w)}{\sum_{n=1}^{M} w_n p(x_t | g_n^w)}
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \log \left( \frac{\sum_{m=1}^{M} w_m p(x_t | g_m^w) - w_m p(x_t | g_m^\text{UBM})}{\sum_{m=1}^{M} w_m p(x_t | g_m^\text{UBM})} \right)
\]

Because the hypothesized speaker model is adapted from the UBM and the components of the adapted GMM retain a correspondence with the mixtures of the UBM. For each feature vector, only a few of components contribute significantly to the likelihood score. So the faster scoring method can be used in generative likelihood ratio estimation [3], and then we obtain a subset of largest scoring components, \( S_{\text{tar}}(t) \), for each frame \( X_t \), that is, top-N scoring components (N=10 is used in this paper). The difference of component likelihood value out of this subset between the hypothesized speaker model and UBM is trivial and can be ignored. As indicated in equation (3), LLR can be decomposed into dominant and residual score elements if it belongs to the largest scoring components or not,
\[ CSS_{(i,j)} = \log\left[1 + \frac{u_i^m \cdot p(x_i / g_i^m) - u_i^{\text{w}m} \cdot p(x_i / g_i^{\text{w}m})}{\sum_{j=1}^{H} u_j^m \cdot p(x_i / g_j^{\text{w}m})}\right] \]

**CSS** 
- Component-Specific Score (CSS) of Gaussian component \( s, s \in S(t) \), in the largest scoring components for the feature vector \( x_I \);
- Weighting factor: weighting factor for component \( s, s \in S(t) \);
- Sum of weights: sum of weighting factors for feature vector \( x_I \);
- Count: number of the largest scoring components for \( x_I \), i.e., \( N \), so that GCILR can be comparable with equation (6).

Two weighting strategies are implemented using UBM and assumed to be fixed for given utterance in this paper. The first set of weighting factors is calculated using the average occurrence frequency of top-N scoring components in the UBM, that is, for each test speech, the more appearances of Gaussian components are in the top-N components of UBM, the larger weighting factors are obtained for these mixture components. We define this method as Frequency-based weighting strategy.

The second weighting technique we adopt is mutual information (MI). Define the value of MI between a Gaussian component \( m \) and all frames of the test utterance, \( X \), as follows:

\[ MI(m; X) = \sum_{i=1}^{N} p(m, x_i) \log \frac{p(m, x_i)}{p(m)p(x_i)} \]

\[ = \sum_{i=1}^{N} p(m)p(x_i / m) \log \frac{p(x_i / m)}{\sum_{i=1}^{N} p(x_i / m)p(m)} \]  

where \( p(m, x_i) \) is the probability that the speech frame \( x_i \) of test utterance appears in the Gaussian component \( m \). The Gaussian component with larger value of mutual information is more informative in the final output score. In this implementation, we first calculate the probabilistic alignment of the feature vectors against the UBM mixture components, that is, for each Gaussian component \( m \) in the UBM, the posterior probability of frame at time instant \( t \) is computed. \( p(m) \) is the prior probability that assigns equal probability. Thus the weighting factors including MI can be provided by the outcome of equation (8), which is defined as MI-based weighting strategy.

4. Experiments and Results

4.1. Database

Experiments have been conducted on two test scenarios of the NIST SRE2006 dataset, that is, the 1conv4w-1conv4w core test (tel-tel) and 1conv4w-1convmic auxiliary microphone (tel-mic) task. The core test of the NIST SRE2006 database involves 3612 true trials, 47836 false trials, and no cross-gender trials. In the tel-mic task, the training speech for a speaker is collected from a single telephone number, but the test speech comes from one of eight different microphones, which includes 2662 true trials and 22308 false trials, no cross-gender trials. A detailed description of the evaluation corpus can be found in [9].

4.2. System description and performance evaluation

For cepstral feature extraction, speech utterances are divided into 25-ms window progressing at a 10-ms frame rate. 16 MFCCs and 16 delta coefficients are calculated to produce 32 dimensional feature vectors. The feature streams are processed through energy-based speech activity detection to eliminate low-energy frames. Feature warping is applied to mitigate the channel effects using a 3s sliding window. Feature-domain intersession compensation (FIDC) is then performed to reduce session variability during the training and test procedures [10].

The gender-independent UBM model with 512 Gaussian mixtures is trained using NIST SRE04 and SRE05 1side training data. For FIDC procedure, a telephone channel space with 50 telephone-channel factors is trained based on the telephone data from NIST SRE2004 and NIST SRE2005. Microphone channel space (50 microphone-channel factors) is trained based on the microphone data from NIST SRE2005. The full channel space (100 channel factors) is derived by appending the above two sub-spaces. All speaker models are created by MAP adaptation of the UBM with the mixture weight and covariance matrix unchanged, where the relevance factor is set to 16. It should be notable that the same set of hypothesized speaker models is used for tel-tel and tel-mic tasks. The baseline system is the GMM-UBM using generative likelihood ratio estimation as described in section 2. During the faster scoring procedure, top-10 largest scoring mixture components are obtained from UBM and the corresponding ones are used for hypothesized speaker model. We focus on comparing and analyzing the Gaussian component information in the GMM-UBM performance, so no score normalization is performed.

The performance assessment of speaker verification system is evaluated using Detection Error Trade-off (DET) curve [11] and Detection Cost Function (DCF) [9]. The DCF is defined to weight two types of errors, i.e., miss detections and false alarms.

4.3. Experimental results

To provide some perspective on Gaussian component information, as indicated by the first term in the right-hand part of equation (6), experiments are conducted on the score of each largest scoring Gaussian component as singleton classifier separately. The performances of singleton systems vary remarkably, as shown in Figure 1. We may arrive at a conclusion that the impact and information contained in each largest Gaussian component is different. In other words, it means that the largest scoring components carry much information (e.g. singleton classifier of top-1 gives 7.75% EER, comparing with 7.20% for generative LLR on core test is reported next), while the last top scoring mixtures are not so important as the first components for performance.

In order to validate our GCILR system, we compare its performance to generative LLR in conventional GMM-UBM baseline system. Two groups of experiments performed on tel-tel and tel-mic datasets of the NIST 2006 corpus are shown in Fig. 2 and Fig. 3 respectively. The EERs and minimum DCFs for these comparison experiments are summarized in Table 1.

For the core test (tel-tel) condition, the commonly used generative LLR of the conventional GMM-UBM system can achieve the performance with EER 7.20% and min.DCF 0.0389. The GCILR with frequency-based weight strategy can achieve a comparable performance with EER 6.92% and DCF 0.0371.
Our proposed verification approach with mutual information based weighting strategy can lead to the improvement of system performance to 6.78% and DCF to 0.0350, which brings 5.83% relative reduction of EER and 10.0% relative reduction of the minimal DCF compared to the generative LLR. Under the tel-mic test scenario, the system performance can be further enhanced by GCILR using MI-based weighting strategy with EER improvement from 8.90% to 7.63% and DCF falling from 0.0400 to 0.0354, which achieves 14.2% relative improvement of EER and 11.5% of the minimum DCF for tel-mic data. The results show that our GCILR method outperforms the generative likelihood ratio estimation in GMM-UBM system. Considering all the additional information is only extracted from the GMM-UBM framework, evaluations show that GCILR technique satisfactorily improves the performance. An interesting thing in the experiments is that GCILR method performs better in the cross-channel tel-mic condition than in the tel-tel condition. Due to the same configuration, including the same target speaker models and FDIC channel space, is used for evaluation, this may be explained by the language variability of test data characteristics. For example, the tel-tel condition involves many different languages, while only English trials are included in the tel-mic test scenario.

5. Conclusions

In this paper we investigate individual Gaussian component’s impact on speaker verification performance over NIST-SRE2006 dataset. We have proposed a new method of including detailed component-dependent score information and mutual information based weighting strategy, and then compare it to the generative likelihood ratio estimation in conventional GMM-UBM system. The resulting system consistently leads to considerable performance gains. Comparison experiments show that our GCILR algorithm performs better and it is efficient to enhance the adapted GMM system in a NIST speaker verification task.

6. References