GMM Kernel by Taylor Series for Speaker Verification

Minqiang Xu¹², Xi Zhou², Beiqian Dai¹, Thomas S. Huang²

¹Department of Electronic Science and Technology, USTC, Hefei Anhui, China
²Department of Electrical and Computer Engineering, UIUC, USA
xumq@uiuc.edu, xizhou2@uiuc.edu, bqda@ustc.edu.cn, huang@ifp.uiuc.edu

Abstract

Currently, approach of Gaussian Mixture Model combined with Support Vector Machine to text-independent speaker verification task has produced the state-of-the-art performance. Many kernels have been reported for combining GMM and SVM.

In this paper, we propose a novel kernel to represent the GMM distribution by Taylor expansion theorem and it’s regarded as the input of SVM. The utterance-specific GMM is represented as a combination of orders of Taylor series expanding at the the means of the Gaussian components. Here we extract the distribution information around the means of the Gaussian components in the GMM as we can naturally assume that each mean position indicates a feature cluster in the feature space. And then the kernel computes the ensemble distance between orders of Taylor series.

Results of our new kernel on NIST speaker recognition evaluation (SRE) 2006 core task have been shown relative improvements of up to 7.1% and 11.7% in EER for male and female compared to K-L divergence based SVM system.

Index Terms: speaker verification, Taylor Series, GMM Kernel, Support Vector Machine

1. Introduction

Gaussian Mixture Models (GMMs) have become the dominant model to describe the distribution of the acoustic features in text-independent speaker verification since last decade. Later, many efforts have been taken to combine GMM with Support Vector Machine (SVM), which can further increase the discriminative capability of the verification system. In the early attempts, the combination are mainly performed on the decision level [1]. That is, learning GMM and SVM systems separately, and fusing the output scores of the two systems. Recently, some suggested using GMM and SVM in a sorted order. In the sorted combinations, GMMs are commonly used as an intermediate step to summarize the information of acoustic features and form the kernels as the input of SVMs and has achieved significant performance gain comparing with the traditional GMM based speaker verification system.

One interesting point in the framework is how to design the kernel from GMMs. In [2], Fisher kernel based on generative model was first proposed, and later, S. Fine [3] and Wan[4] used it for speaker verification task, respectively. In [5][6], an augmented Gaussian mixture model, which was expanded from Fisher kernel, was designed for speech and speaker recognition. In [7], SVM was used as a post-processing model for GMM scores. One of the most popular kernels in text-independent speaker verification is Kullback-Leibler divergence (KLD) based supervector kernel [8]. In the method, utterance-specific GMMs are first estimated for each speech utterance, and then the distance between a utterance pair can be naturally represented by the similarity kernel between their corresponding GMM pairs by KLD. The kernel can be good representation when the utterance-specific GMM is accurate estimated in the whole feature space, which is a strong assumption by just using a single utterance. As the limited feature samples are quite sparse in the high-dimensional feature space, it is better to emphasis the regions with more feature samples when calculating the density similarities instead of comparing the densities on the whole feature space.

In this paper, we propose a novel kernel to represent GMM by Taylor expansion theorem. The utterance-specific GMM is represented as a combination of orders of Taylor series expanding at the the means of the Gaussian components and then the kernel computes the ensemble distance between orders of Taylor series. Here we extract the distribution information around the means of the Gaussian components in the GMM as we can naturally assume that each mean position indicates a feature cluster in the feature space. We demonstrate our new kernel on NIST speaker recognition evaluation (SRE) 2006 core task, our Taylor series based method achieves relative improvements of up to 7.1% and 11.7% in equal-error rate for male and female compared to K-L divergence based SVM system.

This paper is organized as follows. The Section 2, we give a brief introduction to the mentioned system principle and diagram. Detail in representing GMM by employing Taylor series is outlined in the section 3. And Section 4 demonstrates the process for constructing SVM from the orders of Taylor series. Experimental results are shown in Section 5 while some conclusions are given in Section 6.

2. The Principle of GMM kernel by Taylor series for SVM speaker verification

Figure 1 shows the diagram of the proposed method for speaker identification. Firstly, for a given utterance, the GMM probability density function(p.d.f.) can be obtained from Unii-

![Diagram](image_url)
versal Background Model(UBM) with mean-only MAP adaptation. Here, UBM is trained from many utterances, and GMM-UBM is used as the basic system[9]. And then, Taylor expansion theorem is employed to represent the p.d.f. as a combination of orders of Taylor series at the means of the target speaker’s Gaussian mixture. Thirdly, orders of Taylor series are used for constructing SVMs. Finally, a linear fusion in conjunction with those SVMs is implemented to produce the output score.

3. GMM kernel by Taylor series

3.1. Gaussian Mixture Model(GMM)

Gaussian Mixture Model is a linear combination of Gaussian mixtures. Given an input sequence \( O = (o_1, o_2, \ldots, o_T) \) the probability density function(p.d.f.) is as follows:

\[
p(O|\omega, \mu, \Sigma) = \left[ \prod_{t=1}^{T} \sum_{m=1}^{M} \omega_m N(o_t; \mu_m, \Sigma_m) \right]^{\frac{1}{2}}
\]  

(1)

where \( N(o_t; \mu_m, \Sigma_m) \) is a Gaussian, and \( \sum_{m=1}^{M} \omega_m = 1, 0 < \omega_m < 1, m = 1, \ldots, M, o_t \in \mathbb{R}^{d}, t = 1, \ldots, T. \)

The GMM for an utterance is obtained from UBM with mean-only MAP adaptation in this study, and the \( \Sigma \) is diagonal[9].

3.2. Taylor series for GMM

In this paper, we propose the concept of representing the speaker’s GMM by using Taylor series. The p.d.f. for each utterance is represented at some special points by employing Taylor series, and those points are selected by intuition which can most classify the target and imposter, such as the means of Gaussian mixtures. And then, each order of Taylor series can be applied to construct the SVM for the claimed speaker.

Given an utterance \( O = (o_1, o_2, \ldots, o_T) \), the corresponding p.d.f., after taking a log, is \( \log p(O|\omega, \mu, \Sigma) \). And the Taylor expansion at \( \mu_j \) is now:

\[
\log p(O|\omega, \mu, \Sigma)|_{\mu = \mu_j} = \log p(O|\omega, \mu_j, \Sigma_j) + \left( \sum_{i=1}^{\infty} \frac{\gamma_i}{i!} \left( \log p(O|\omega, \mu_j, \Sigma_j) \right)^{(i)} (\mu - \mu_j)^{i} \right)_{i=0}^{\infty}
\]  

(2)

where we define \( 0! = 1 \). It is obvious that Eq.(2) is always identical if \( |\mu - \mu_j| < 1 \) is satisfied[12]:

\[
\log p(O|\omega, \mu, \Sigma)|_{\mu = \mu_j} = \left( \sum_{i=1}^{\infty} C_i \gamma_i^{(i)} \left( \log p(O|\omega, \mu_j, \Sigma_j) \right) \right)_{i=0}^{\infty}
\]  

(3)

where, \( C_i \) is \( (\mu - \mu_j)^{i} \), and \( C_i \) is identical for all p.d.f. while \( \mu_j \) is given. So every GMM can be depicted by Eq.(3), and we focus on \( \gamma_i^{(i)} \left( \log p(O|\omega, \mu_j, \Sigma_j) \right) \) for speaker classification.

Suppose the component number of GMM is \( M \), and for each component \( \mu_j \in \mathbb{R}^{d} (j = 0, 1, \ldots, M) \). So by expansion at each \( \mu_j \), the \( \gamma_i^{(i)} \left( \log p(O|\omega, \mu_j, \Sigma_j) \right) \), where \( \mu_0 = (\mu_1, \ldots, \mu_M) \), can be expressed as:

\[
\gamma_i^{(i)} \left( \log p(O|\omega, \mu, \Sigma) \right) = \begin{bmatrix}
\gamma_i^{(1)} \left( \log p(O|\omega, \mu_1, \Sigma_1) \right) \\
\gamma_i^{(2)} \left( \log p(O|\omega, \mu_2, \Sigma_2) \right) \\
\vdots \\
\gamma_i^{(M)} \left( \log p(O|\omega, \mu_M, \Sigma_M) \right)
\end{bmatrix}
\]  

(4)

If \( \log p(O|\omega, \mu, \Sigma) \) is a GMM’s p.d.f., and \( (\omega_m, \mu_m, \Sigma_m) (m = 1, \ldots, M) \) is trained from utterance \( O \), the covariance \( \Sigma_m (m = 1, \ldots, M) \) is set to be diagonal, the orders of Taylor series are now:

\[
\gamma_i^{(i)} (\log p(O|\omega, \mu_j, \Sigma_j)) = \log \left( \sum_{m=1}^{M} \omega_m N(\mu_j; \mu_m, \Sigma_m) \right)
\]  

\[
\gamma_i^{(1)} (\log p(O|\omega, \mu_j, \Sigma_j)) = \frac{1}{T} \sum_{t=1}^{T} \gamma_i^{(1)} \left( \Sigma_j^{-1} (\mu_t - \mu_j) \right)
\]  

\[
\gamma_i^{(2)} (\log p(O|\omega, \mu_j, \Sigma_j)) = \frac{1}{T} \sum_{t=1}^{T} \gamma_i^{(2)} \left( \Sigma_j^{-1} - (1 - \gamma_i^{(1)} (\mu_t - \mu_j)) \right)
\]  

\[
\gamma_i^{(3)} (\log p(O|\omega, \mu_j, \Sigma_j)) = \gamma_i^{(1)} \left( \left( -\frac{1}{T} \sum_{t=1}^{T} \gamma_i^{(1)} \left( \Sigma_j^{-1} - (1 - \gamma_i^{(1)} (\mu_t - \mu_j)) \right) \right) \right)
\]  

\[
\gamma_i^{(i)} (\omega, \mu, \Sigma) = \frac{\omega_j N(\mu_j; \mu_m, \Sigma_m)}{\sum_{m=1}^{M} \omega_j N(\mu_j; \mu_m, \Sigma_m)}
\]  

(5)

For a given utterance \( O_0 \), the utterance-specific GMM \( \lambda_0 = (\omega_0, \mu_0, \Sigma_0) \) is obtained from UBM with mean-only MAP adaptation, then \( \gamma_i^{(i)} (\log p(O_0|\omega_0, \mu_0, \Sigma_0)) \) is the derivative of the log p.d.f. at itself means \( \mu_0 \). While for the impostor utterance \( O \), the corresponding GMM p.d.f. is \( p(O|\omega, \mu, \Sigma) \). So \( \gamma_i^{(i)} (\log p(O|\omega, \mu, \Sigma)) \) represents the derivative of \( p(O|\omega, \mu, \Sigma) \) at the means of target speaker. Thus the differences of \( \gamma_i^{(i)} (\log p(O|\omega, \mu, \Sigma)) \) can be used for pattern classification.

4. GMM kernel for SVM speaker verification

4.1. Support Vector Machine

Support Vector Machine(SVM)[10] is a powerful algorithm for classification tasks, and for two classification problem, it’s defined as:

\[
f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b
\]  

(6)

where, \( \alpha_i \geq 0 \) and \( y_i \) are the two classes identified by label \( y_i \in \{-1, +1\} \), \( b \) is a constant, \( \sum_{i=1}^{n} y_i \alpha_i = 0 \). The vectors \( x_i \) are support vectors while \( \alpha_i \) and \( b \) are obtained from the training set by solving the quadratic programming problem.

The kernel \( K(x_i, x_j) \) is very important and constrained to have certain properties (the Mercer condition), so that \( K(x_i, x_j) \) can be expressed as:

\[
K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)
\]  

(7)
4.2. GMM kernel function

According to Eq.(4) and Eq.(7), we define the normalized kernel:

\[ K(O_i, O_j) = \Phi_{\mu_i}(O_i) G^{-1} \Phi_{\mu_j}(O_j) \]
\[ G = \frac{1}{N} \sum_{i=1}^{N} (\Phi_{\mu_i}(O_i) - \mu) (\Phi_{\mu_i}(O_i) - \mu)^T \]
\[ \mu = \frac{1}{N} \sum_{i=1}^{N} \Phi_{\mu_i}(O_i) \]  \( (8) \)

where \( O_i \) is the speech feature vector of speaker \( i \), \( \Phi_{\mu_i}(O_i) \) is the kernel mapping for speaker \( i \), and \( G \) is diagonal only.

In this paper, we use different kernels for each subsystem. For the \( i^{th} \) SVM, the \( \Phi(O) \) is defined as follows

\[ \Phi(O) = \sum_{i=1}^{M} (\log p(O|\omega, \mu, \Sigma)) \]  \( (9) \)

For the \( 0^{th} \) order of Taylor series, \( \Phi(O) \) is the log likelihood at \( \mu_0 \) \((i = 1, \ldots, M)\), here \( \mu_0 \) is the mean of \( i^{th} \) Gaussian component of the target speaker. For the \( 1^{st} \) order of Taylor series, is a super-vector with dimensions of \( Md+1 \). And for orders of \( 0^{th} (i = \geq 2) \), only diagonal elements are considered.

4.3. Linear fusion for SVMs

After constructing the SVM for each order of Taylor series, we use linear weight method for fusion those SVMs:

\[ \text{Score} = \sum_{i=0}^{K} \beta_i \text{Score}(i) \quad (\sum_{i=0}^{K} \beta_i = 1) \]  \( (10) \)

Here, \( \text{Score}(i) \) is the output of the \( i^{th} \) SVM.

4.4. Performance evaluation

Performance is primarily evaluated using the Equal Error Rate (EER), DCF and the curve of DET[13].

5.2. Experimental results

Four methods are considered in this work, the Taylor series’ based speaker verification system (Taylor_SVM, for short), traditional GMM_UBM system[9], K-L divergence based SVM system(KL_SVM)[8], and Fisher kernel based SVM system(Fisher_SVM)[4].

Fig. 2 depicts the DET curves of our proposed method and the other three. The dash red curve is the result of the traditional GMM_UBM system, the dash-dot blue one is from the dimensions of \( \sum_{\mu=0}^{1} (\log p(O|\omega, \mu, \Sigma)) \) is 256 * 1 for \( i = 0 \) and 8192 * 1 for \( i \geq 1 \). All of the orders are normalized according Eq.(8). For an utterance of target speaker, these orders are labeled as positive class, while for an impostor, it’s labeled as negative class. In this paper, there is only one utterance named as +1 class, and the number of -1 class data is 09 for male task and 09 for female task, which of them are random selected from NIST05 SRE core task.

Performance is primarily evaluated using the Equal Error Rate(EER), DCF and the curve of DET[13].

5.2. Experimental results

Four methods are considered in this work, the Taylor series’ based speaker verification system(Taylor_SVM, for short), traditional GMM_UBM system[9], K-L divergence based SVM system(KL_SVM)[8] and Fisher kernel based SVM system(Fisher_SVM)[4].

Fig. 2 depicts the DET curves of our proposed method and the other three. The dash red curve is the result of the traditional GMM_UBM system, the dash-dot blue one is from the dimensions of \( \sum_{\mu=0}^{1} (\log p(O|\omega, \mu, \Sigma)) \) is 256 * 1 for \( i = 0 \) and 8192 * 1 for \( i \geq 1 \). All of the orders are normalized according Eq.(8). For an utterance of target speaker, these orders are labeled as positive class, while for an impostor, it’s labeled as negative class. In this paper, there is only one utterance named as +1 class, and the number of -1 class data is 09 for male task and 09 for female task, which of them are random selected from NIST05 SRE core task.

Performance is primarily evaluated using the Equal Error Rate(EER), DCF and the curve of DET[13].

5.2. Experimental results

Four methods are considered in this work, the Taylor series’ based speaker verification system(Taylor_SVM, for short), traditional GMM_UBM system[9], K-L divergence based SVM system(KL_SVM)[8] and Fisher kernel based SVM system(Fisher_SVM)[4].

Fig. 2 depicts the DET curves of our proposed method and the other three. The dash red curve is the result of the traditional GMM_UBM system, the dash-dot blue one is from the dimensions of \( \sum_{\mu=0}^{1} (\log p(O|\omega, \mu, \Sigma)) \) is 256 * 1 for \( i = 0 \) and 8192 * 1 for \( i \geq 1 \). All of the orders are normalized according Eq.(8). For an utterance of target speaker, these orders are labeled as positive class, while for an impostor, it’s labeled as negative class. In this paper, there is only one utterance named as +1 class, and the number of -1 class data is 09 for male task and 09 for female task, which of them are random selected from NIST05 SRE core task.

Performance is primarily evaluated using the Equal Error Rate(EER), DCF and the curve of DET[13].

5.2. Experimental results

Four methods are considered in this work, the Taylor series’ based speaker verification system(Taylor_SVM, for short), traditional GMM_UBM system[9], K-L divergence based SVM system(KL_SVM)[8] and Fisher kernel based SVM system(Fisher_SVM)[4].

Fig. 2 depicts the DET curves of our proposed method and the other three. The dash red curve is the result of the traditional GMM_UBM system, the dash-dot blue one is from the dimensions of \( \sum_{\mu=0}^{1} (\log p(O|\omega, \mu, \Sigma)) \) is 256 * 1 for \( i = 0 \) and 8192 * 1 for \( i \geq 1 \). All of the orders are normalized according Eq.(8). For an utterance of target speaker, these orders are labeled as positive class, while for an impostor, it’s labeled as negative class. In this paper, there is only one utterance named as +1 class, and the number of -1 class data is 09 for male task and 09 for female task, which of them are random selected from NIST05 SRE core task.

Performance is primarily evaluated using the Equal Error Rate(EER), DCF and the curve of DET[13].
Fisher-SVM and KL-SVM. Compared with the KL-SVM, the gain of EER and MinDCF are 7.0% and 5.1% for male, and 11.7% and 10.9% for female.

Table 1: Comparison between Taylor_SVM and other systems

<table>
<thead>
<tr>
<th>System</th>
<th>EER(%)</th>
<th>MinDCF</th>
<th>EER(%)</th>
<th>MinDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM_UBM</td>
<td>11.82</td>
<td>0.0484</td>
<td>12.98</td>
<td>0.0559</td>
</tr>
<tr>
<td>KL_SVM</td>
<td>9.93</td>
<td>0.0473</td>
<td>11.29</td>
<td>0.0509</td>
</tr>
<tr>
<td>Fisher_SVM</td>
<td>10.32</td>
<td>0.0490</td>
<td>10.96</td>
<td>0.0492</td>
</tr>
<tr>
<td>Taylor_SVM</td>
<td>9.28</td>
<td>0.0450</td>
<td>10.11</td>
<td>0.0459</td>
</tr>
</tbody>
</table>

5.3. Further Analysis about Taylor Series System

5.3.1. Taylor Series’ each order for speaker verification

Results of each order of Taylor series \(T_i\), \(i = 0, 1, \ldots\), (for short) are shown in Table 2. Besides the Taylor series’ system, traditional GMM_UBM system and KL divergence based SVM system (KL_SVM) are presented, in order to show the effectiveness of each order of Taylor series system.

Table 2: Each order for speaker verification.

<table>
<thead>
<tr>
<th>System</th>
<th>EER(%)</th>
<th>MinDCF</th>
<th>EER(%)</th>
<th>MinDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM_UBM</td>
<td>11.82</td>
<td>0.0484</td>
<td>12.98</td>
<td>0.0559</td>
</tr>
<tr>
<td>KL_SVM</td>
<td>9.93</td>
<td>0.0473</td>
<td>11.29</td>
<td>0.0509</td>
</tr>
<tr>
<td>T0</td>
<td>15.11</td>
<td>0.0578</td>
<td>16.13</td>
<td>0.0634</td>
</tr>
<tr>
<td>T1</td>
<td>10.32</td>
<td>0.0490</td>
<td>10.96</td>
<td>0.0492</td>
</tr>
<tr>
<td>T2</td>
<td>13.73</td>
<td>0.0601</td>
<td>17.01</td>
<td>0.0751</td>
</tr>
<tr>
<td>T3</td>
<td>11.29</td>
<td>0.0521</td>
<td>11.74</td>
<td>0.0545</td>
</tr>
</tbody>
</table>

The results show the competitive performance of each order of Taylor series. The 1st order \(T0\) and the 3rd order \(T3\) of Taylor series systems are even better than the traditional GMM_UBM system, and only a little worse than the KL divergence based SVM system, which can achieve the state of the art. The 1st order of Taylor series gives the best result compared with all the other orders.

5.3.2. Combination of each order for speaker verification

According to Eq. (3), each order of Taylor series is complementary with each other. Linear weighted method is used for combining each order of Taylor series. For orders of Taylor series, \(T0 : T1 : T2 : T3 = 15 : 85 : 15 : 20\) is chosen, as our experiences on NIST SRE 01 task showed it worked well.

Table 3: Combination of each order for speaker verification.

<table>
<thead>
<tr>
<th>System</th>
<th>Male EER(%)</th>
<th>Male MinDCF</th>
<th>Female EER(%)</th>
<th>Female MinDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>10.32</td>
<td>0.0490</td>
<td>10.96</td>
<td>0.0492</td>
</tr>
<tr>
<td>T0+T1</td>
<td>9.78</td>
<td>0.0457</td>
<td>10.25</td>
<td>0.0462</td>
</tr>
<tr>
<td>T0+T1+T2</td>
<td>9.28</td>
<td>0.0450</td>
<td>10.11</td>
<td>0.0459</td>
</tr>
<tr>
<td>T0+T1+T2+T3</td>
<td>9.20</td>
<td>0.0453</td>
<td>10.10</td>
<td>0.0463</td>
</tr>
</tbody>
</table>

Results are listed in Table 3. For none combined system, \(T1\) has the best performance. And all of the combinations are better than \(T1\). The best EER and MinDCF are achieved while assembling \(T0, T1, T2\). The relative improvement for male, compared with \(T1\) solo system, is 11.2% and 8.9%, while for female, the gain is 8.4% and 7.2%, respectively.

Although the performance of the three-combining system does not put the four-assembling one in the shade, even the later is a bit better than the first one, we still choose the three-combining system as our final system. That’s because it’s much hard to set the appropriate weights for the system while the number of subsystems grows from 3 to 4.

6. Conclusion

In this paper, we explore the technology how to combine SVM and GMM for speaker verification. A novel kernel to represent the GMM distribution by Taylor expansion theorem is proposed. The utterance-specific GMM is represented as a combination of orders of Taylor series expanding at the the means of the Gaussian components. And then those orders are set as the input of the SVMs for calculating the output. Results has been shown the effective of the orders of Taylor series, and better result is achieved by linear fusion those orders for the complementary of orders. Compared with the KL_SVM, the gain of EER is 7.0% for male, and 11.7% for female.

7. References