Speaker Verification Using Fisher Vector

Yao Tian¹, Liang He¹, Zhi-yi Li¹, Wei-lan Wu²,³, Wei-Qiang Zhang¹, Jia Liu¹

¹National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University
²State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences
³University of Chinese Academy of Sciences

tianyao11@mails.tsinghua.edu.cn

Abstract

This paper introduces an approach based on Fisher vector feature representation for speaker verification. The Fisher vector is originated from Fisher Kernel and represents each utterance as a high-dimensional vector by encoding the derivatives of the log-likelihood of the UBM model with respect to its mean and variances. This representation captures the average first and second order differences between the utterance and each of the Gaussian centers of the UBM model. And the Fisher vector is further projected to a low-dimensional space using PPCA which is conducted in a similar way of factor analysis. We compare the proposed method with the state-of-art i-vector approach on the telephone-telephone condition of NIST SRE2010 female and male core task. The experimental results indicate that the proposed Fisher vector based method is competitive with i-vector. It can also provide complementary information to i-vector and the fusion of these two approach obtains a relative improvement of 11.8% and 14.7% in EER and 9.2% and 2.7% in minDCF for female and male than i-vector alone.

Index Terms: speaker verification, i-vector, Fisher vector

1. Introduction

The core task of speaker verification is to decide whether two utterances belong to the same person. The Gaussian Mixture Models-Universal Background Models(GMM-UBM)[1] lay the foundation of modeling speaker space and many approaches based on GMM-UBM framework has been proposed to improve the performance of speaker verification including Support Vector Machine(SVM)[2][3], Joint Factor Analysis(JFA)[4][5], total variability factor analysis(i-vector)[6] and etc.

The classical JFA try to model the speaker and session variability by defining two distinct subspaces: the speaker space and the channel space. In contrast, the i-vector which originates from JFA aims at modeling both the speaker and channel space by generating a low-dimensional space named total variability space since the channel space estimated using JFA also contains information about speakers[6]. Then each utterance is represented as a compact, fixed-length vector called i-vector in this low-dimensional space. Intersession compensation techniques, such as Linear Discriminant Analysis(LDA) and Within Class Covariance Normalization(WCCN) are further carried out in the total variability space[6] as opposed to the high-dimensional GMM supervector space[4] to improve the performance of speaker verification.

Recently, the Fisher vector representation[7] based on GMM has been successfully used in image classification[8][9] and face verification[10]. The local descriptors extracted from large number of images are used to train a Gaussian Mixture Model firstly which is similar to the process of building UBM in speaker verification. Then each image is represented as a high-dimensional vector called Fisher vector by encoding the derivatives of the log-likelihood of the GMM model with respect to its mean and variances using local descriptors[10]. This leads to the representation which captures the average first and second order differences between the features and each of the GMM centers.

In this study, we propose a novel feature representation method by using Fisher vector for speaker verification. And we further project it to a low-dimensional space using PPCA[11][12] which is conducted in a similar way as the factor analysis in the i-vector model approach. There are three motivations for our work: Firstly, Fisher vector representation takes data’s higher-order(second-order) moments into consideration directly while i-vector doesn’t. Secondly, Fisher vector models each utterance in a different perspective compared with i-vector approach which may provide complementary information. Thirdly, kernel methods had been wildly discussed in speaker verification area[2][3], but most of them remained at the supervector-level and it was not quite convenient to do verification as the amount of data grows dramatically nowadays.

The rest of this paper is organized as followed. in Section2, we give a review of the background of i-vector modeling, the approach of LDA, WCCN and the cosine distance scoring classifier. Section3 presents the process of extracting Fisher vector and dimension reduction method. Experimental setup and the results are showed in Section4. Finally, conclusions and future directions are discussed in Section5.

2. Theoretical Background

2.1. i-vector

Inspired by the classical Joint Factor Analysis(JFA)[4][5] model which is based on speaker and channel factors separately, Dehak et al.[6] proposed the i-vector based speaker verification technique which defines a compact low-dimensional space called total variability space. In total variability space modeling, no distinction is made between the speaker effects and channel effects in GMM supervector space since experiments showed that the channel space estimated using JFA also contains information about speakers[6].

Given an utterance, the speaker-and-channel dependent G-
MM supervector can be written as:

\[ M = m + T \omega \]  

(1)

where \( m \) is the speaker-and-channel independent component which can be taken to be the UBM supervector. \( T \) is a rectangular matrix of low rank representing the primary directions of the total variability space. \( \omega \) is a random vector having a standard distribution \( N(0, I) \) and the i-vector is obtained as the MAP point estimate of \( \omega \).

The process of training the the matrix \( T \) is the same as the training of eigenvoice in JFA except for that all the utterances from the same speaker is regarded as produced by different speakers. The i-vector modeling can be seen as a simple factor analysis, which allows us to represent an utterance by a compact, low-dimensional single factor.

2.2. Intersession Compensation

After the extraction of i-vector, the intersession compensation can be carried out in the total variability space. In our experiments, we use linear discriminant analysis (LDA) and within class covariance normalization (WCCN) for intersession compensation.

1) Linear discriminant analysis: linear discriminant analysis (LDA)[13] is a technique for dimensionality reduction that is widely used in the field of pattern recognition. It tries to seek new orthogonal axes that maximizing between-class variance and minimizing intra-class variance. All the utterances from the same speaker are labeled as an individual class. The LDA optimization problem can be defined by maximizing the following Fisher criteria:

\[ J(v) = \frac{v^T S_b v}{v^T S_w v} \]  

(2)

Where \( v \) is the projection direction. \( S_b \) and \( S_w \) are the between-class covariance and within-class covariance defined as:

\[ S_b = \sum_{s=1}^{S} (\bar{\omega}_s - \bar{\omega})(\bar{\omega}_s - \bar{\omega})^T \]  

(3)

\[ S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\omega^t_s - \bar{\omega}_s)(\omega^t_s - \bar{\omega}_s)^T \]  

(4)

\( S \) is the number of speakers, \( \bar{\omega}_s \) is the mean of i-vectors for speaker \( s \) can be computed as \( \bar{\omega}_s = (1/n_s) \sum_{i=1}^{n_s} \omega^t_s \). \( n_s \) is the number of i-vectors for speaker \( s \). \( \bar{\omega} \) is the speaker population mean.

The projection matrix \( V \) is obtained by stacking the top \( N \) eigenvectors of the general matrix \( S_w^{-1} S_b \).

2) Within class covariance normalization: Within class covariance normalization (WCCN) tries to use the within-class covariance to normalize the cosine kernel function in order to compensate for intersession variability[14]. All utterances from the same speaker are labeled as an individual class. The within-class covariance matrix is denoted as:

\[ W = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\omega^t_s - \bar{\omega}_s)(\omega^t_s - \bar{\omega}_s)^T \]  

(5)

\( S \) is the number of speakers, \( \bar{\omega}_s \) is the mean of i-vectors for speaker \( s \) and is computed as \( \bar{\omega}_s = (1/n_s) \sum_{i=1}^{n_s} \omega^t_s \). \( n_s \) is the number of i-vectors for speaker \( s \). The feature mapping matrix \( L \) can be obtained through a Cholesky decomposition of the matrix \( W^{-1} = LL^T \).

2.3. Classifier

Cosine distance scoring: In an i-vector based speaker verification system, cosine distance scoring[6] has been proven to be a fast and efficient way of measuring the similarity of a pair of i-vectors. The cosine distance between \( \omega_{\text{target}} \) and \( \omega_{\text{test}} \) is:

\[ \text{score}(\omega_{\text{target}}, \omega_{\text{test}}) = \frac{\langle \omega_{\text{target}}, \omega_{\text{test}} \rangle}{\|\omega_{\text{target}}\| \|\omega_{\text{test}}\|} \]  

(6)

3. Fisher Vector

Fisher vector is derived from the Fisher Kernel introduced by Jaakkola and Haussler[7] and applied successfully to image classification[8][9] and face verification[10] lately. Let \( X = \{x_t, t = 1...T\} \) be the set of frames of an utterance and the dimension of \( x_t \) is denoted as \( d \). They are modeled by UBM \( u_\lambda \) with parameters \( \lambda = \{\omega_k, \mu_k, \sigma_k\} \), where \( k \) is the mixture index, and \( \omega_k, \mu_k, \sigma_k \) are the mixture weights, means, and diagonal covariances of the UBM. \( X \) can be described by the gradient vector[7]:

\[ G^X_\lambda = \frac{1}{T} \nabla_{\lambda} \log u_\lambda(X) \]  

(7)

The gradient of the log-likelihood describes the direction in which the data parameters should be modified. A natural kernel between two utterances \( X \) and \( Y \) on these gradients is[7]

\[ \text{Kernel}(X, Y) = G^X_\lambda F_{\lambda}^{-1} G^Y_\lambda \]  

(8)

Here \( F_{\lambda} \) is the Fisher information matrix of \( u_\lambda \) and is denoted as:

\[ F_{\lambda} = E_{x \sim u_\lambda} \left[ \nabla_{\lambda} \log u_\lambda(x) \nabla_{\lambda} \log u_\lambda(x)^T \right] \]  

(9)

With Cholesky decomposition, \( F_{\lambda} \) can be denoted as \( F_{\lambda} = L_{\lambda} L_{\lambda}^T \). Thus Kernel\((X, Y)\) can be rewritten as a dot-product between normalized vectors \( \phi \):

\[ \phi^X_\lambda = L_{\lambda} G^X_\lambda \]  

(10)

Here \( \phi^X_\lambda \) is referred as Fisher vector which encodes the derivatives of the log-likelihood of the UBM model with respect with its parameters[7]. Similar to the supervector derived from KL divergence[3], the Fisher vector is also generated from a kernel but it can take data’s higher-order(second-order) moments into consideration. In this paper, we use Fisher vector as a novel feature representation method for speaker verification and only consider the derivatives with respect to the means and variances since the derivatives to weights doesn’t bring additional improvements in our experiments. They are computed as follows:

\[ \phi^X_{\mu,k} = \frac{1}{T \sqrt{\omega_k}} \sum_{t=1}^{T} \gamma_t(k) \left( \frac{x_t - \mu_k}{\sigma_k} \right) \]  

(11)

\[ \phi^X_{\sigma,k} = \frac{1}{2T \sqrt{\omega_k}} \sum_{t=1}^{T} \gamma_t(k) \left( \frac{(x_t - \mu_k)^2}{\sigma_k^2} - 1 \right) \]  

(12)

Here, \( \gamma_t(k) \) is the soft assignment weight of the \( t \)-th frame \( x_t \) to the \( k \)-th Gaussian component \( u_k \) and is computed as:

\[ \gamma_t(k) = \frac{\omega_k u_k(x_t)}{\sum_{i=1}^{K} \omega_i u_i(x_t)} \]  

(13)

Fisher vector \( \phi^X_\lambda \) is then obtained by concatenating the derivatives of each Gaussian as: \( \phi^X_\lambda = [\phi_{\mu,1}^X, \phi_{\sigma,1}^X, ..., \phi_{\mu,K}^X, \phi_{\sigma,K}^X] \).
After the extraction of Fisher vector, we normalized each Fisher vector to unit length by \( L2 \) normalization which can be seen as a gaussianization step to satisfy the gaussian assumption in PPCA model.

The dimension of Fisher vector is \( 2Kd \) which is usually very high(39936 for \( K=512 \) and \( d=39 \)). To further compare its performance with the i-vector approach, we use the Probabilistic Principle Component Analysis(PPCA) to project the Fisher vector to a low-dimensional space denoted \( T' \). Actually, the i-vector modeling can be considered as a classical PPCA model[12]. In this study, the PPCA projection matrix is conducted with EM algorithm in a similar way as the i-vector approach. After the projection, the intersession compensation methods mentioned above are used as well. For convenience, the Fisher vector in the Experiments Section denotes the projected Fisher vector.

4. Experiments

4.1. Experimental setup

The equal error rate(EER) and the minimum decision cost function(minDCF) were selected as metrics for evaluation[15]. \( 12 \) Mel Frequency Cepstral Coefficients(MFCC) together with log energy were extracted using a 20ms Hamming window and 10ms frame shift. This 13 dimensional feature vector was subjected to feature warping, using a 3s sliding window. Then Delta and delta-delta coefficients were calculated to produce final 39-dimensional feature vectors.

The training data included Swb(Switchboard) II, Swb cellular, NIST SRE 2004, 2005, 2006 and 2008 corpora. The gender-dependent UBM was trained with diagonal covariance and 512 mixture components. The dimension of i-vector and the Fisher vector were set to 400. The dimension of LDA was set to 200. The cosine distance scoring(CDS) was adopted as classifier to train the UBM, i-vector matrix, Fisher vector matrix, LDA, WCCN were listed in Table 1.

The experiments were carried out on the telephone-telephone condition of the NIST SRE2010 female and male core task. The training data included Swb(Switchboard) II, Swb cellular, NIST SRE 2004, 2005, 2006 and 2008 corpora. The gender-dependent UBM was trained with diagonal covariance and 512 mixture components. The dimension of i-vector and the Fisher vector were set to 400. The dimension of LDA was set to 200. The cosine distance scoring(CDS) was adopted as classifier to train the UBM, i-vector matrix, Fisher vector matrix, LDA, WCCN were listed in Table 1.

### Table 1: Corpora used to estimate the UBM, i-vector matrix \( T \), Fisher vector matrix \( T' \), LDA, WCCN.

<table>
<thead>
<tr>
<th></th>
<th>UBM</th>
<th>T</th>
<th>T'</th>
<th>LDA</th>
<th>WCCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swb II</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swb cellular</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST04</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST05</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST06</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST08</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.2. Experimental results

In Table 2, we give the performance of the state-of-art i-vector based system and our proposed Fisher vector based system without any intersession-compensation process on NIST SRE2010 female and male tel-tel core condition(condition 5). The score fusion result is also presented. From the result we can see that the EER(%) of Fisher vector approach outperforms i-vector. The fusion of these two approach show consistent improvement in both EER and minDCF where the relative improvement is 12.9% and 15.3 % in EER and 8.4% and 2.5% in minDCF compared with i-vector for female and male. The DET curves of i-vector system, Fisher vector system and fusion result are presented as Figure 1.

In Table 3, the system performance of i-vector and Fisher vector using LDA and WCCN on female and male are compared. The fusion result is presented as well. It can be seen from Table 3 when LDA and WCCN is used, the i-vector approach outperforms the Fisher vector method. But the fusion of these two methods still show improvements over both EER and minDCF. The relative improvement of EER and minDCF is 11.8% and 9.2% for female, as well as 14.7% relative improvement in EER and 2.7% in minDCF for male. The DET curves of i-vector system, Fisher vector system and fusion result are presented as Figure 2.

From the results above we can see that the Fisher vector approach is competitive with i-vector. It can also provide complementary information to i-vector and the fusion of these two systems further improve the performance of speaker verification. Another interesting observation is that the Fisher vector approach is inferior to the i-vector approach while intersession compensation is applied. We thought that \( L2 \) normalization at the supervector-level helps gaussianizing Fisher vector while breaking some utterances'(short duration) structure. We'll investigate this situation in our later work.

### Table 2: EER(%) and minDCF of i-vector system, Fisher vector system and fusion result without intersession compensation on the NIST SRE2010 female and male tel-tel core condition.

<table>
<thead>
<tr>
<th>system</th>
<th>Data Set</th>
<th>EER(%)</th>
<th>minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>i-vector</td>
<td>female</td>
<td>7.12</td>
<td>0.569</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>7.08</td>
<td>0.562</td>
</tr>
<tr>
<td>Fisher vector</td>
<td>female</td>
<td>6.76</td>
<td>0.561</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>6.82</td>
<td>0.570</td>
</tr>
<tr>
<td>score fusion</td>
<td>female</td>
<td>6.20</td>
<td>0.521</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>6.00</td>
<td>0.548</td>
</tr>
</tbody>
</table>

### Table 3: EER(%) and minDCF of i-vector system, Fisher vector system and fusion result with intersession compensation(LDA, WCCN) on the NIST SRE2010 female and male tel-tel core condition.

<table>
<thead>
<tr>
<th>system</th>
<th>Data Set</th>
<th>EER(%)</th>
<th>minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>i-vector</td>
<td>female</td>
<td>4.48</td>
<td>0.468</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>3.41</td>
<td>0.446</td>
</tr>
<tr>
<td>Fisher vector</td>
<td>female</td>
<td>4.79</td>
<td>0.465</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>3.69</td>
<td>0.443</td>
</tr>
<tr>
<td>score fusion</td>
<td>female</td>
<td>3.95</td>
<td>0.425</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>2.91</td>
<td>0.434</td>
</tr>
</tbody>
</table>

5. Conclusions

In this paper we propose a new approach based on Fisher vector feature representation to speaker verification. This feature represents each utterance in a different perspective compared with i-vector by using the derivatives of the log-likelihood of the UBM’s means and variances which takes both utterances’ first-order and second-order moments into consideration. Experimental results on NIST SRE 2010 female and male telephone-telephone core condition demonstrate the effectiveness of the
proposed method. The fusion result improve the state-of-art i-vector approach with a relative 11.8% and 14.7% in EER and 9.2% and 2.7% in minDCF for female and male respectively. The fusion result outperforms the state-of-art i-vector approach. In future work, we would like to test the Fisher vector method on interview conditions.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 61370034.

7. References


