Abstract

In this paper, we adopt a new factor analysis of neighborhood preserving embedding (NPE) for speaker verification under the support vector machine (SVM) framework. NPE aims at preserving the local neighborhood structure on the data and defines a low-dimensional speaker space called neighborhood preserving embedding space. We compare the proposed method with the state-of-the-art total variability approach on the telephone-telephone core condition of the NIST 2008 Speaker Recognition Evaluation (SRE) dataset. The experimental results indicate that the proposed NPE method outperforms the total variability approach, providing up to 24% relative improvement.

1. Introduction

For the task of speaker verification, the Gaussian mixture model (GMM) [1] has always been the dominant method and support vector machine (SVM) [2] is a powerful discriminative classifier adopted. Factor analysis has led to the development of an effective method of compensating for intersession variability in speaker verification. The Joint Factor Analysis (JFA) [3][4] and total variability factor analysis [5][6] have been successfully applied in speaker verification.

The classical JFA is a model used to solve the problem of speaker and session variability in GMM’s and defines two distinct separate spaces: the speaker space and the channel space. In contrast, in the total variability factor analysis, the speaker and the channel variability are contained simultaneously in a new low-dimensional space named total variability space. The SVM can be trained using the extracted total variability factors. And channel compensation techniques, such as Linear Discriminant Aanlysis (LDA) and within class covariance normalization (WCCN), are carried out in the total variability factor space [6].

Actually, we can consider the total variability approach as an application of the probabilistic principal component analysis (PPCA) [8][9]. The factor analysis of the total variability approach can obtain useful information by reducing the dimension of the GMM supervectors so that the latent variables can be estimated well using limited data. As a type of PCA, the total variability model does not need speaker information. However, the speaker label is generally available for the training data. To incorporate the speaker label information into the dimension-reduction projection processing, neighborhood preserving embedding (NPE) [10], which is known as NPEface in face recognition, is introduced into speaker verification here. Besides, different from principal component analysis (PCA) [11], i.e. Eigenface, which aims at preserving the global structure, NPE aims at preserving the local manifold structure. In this study, we propose to use NPE as a novel factor analysis approach to speaker verification under the SVM framework.

The remainder of this paper is organized as follows. In Section 2, we give a review of the total variability, support vector machine and GMM supervector. Section 3 introduces the approach of NPE. The techniques of intersession compensation of LDA and WCCN are briefly introduced in Section 4. Experiments and results are presented in Section 5. Finally, we conclude the paper in Section 6.

2. Theoretical Background

2.1. Total Variability

Unlike the classical JFA modeling [3][4] which is based on speaker and channel factors separately, the total variability approach defines a total variability space, which contains simultaneously the speaker and channel variabilities [5][6]. In total variability factor analysis, no distinction is made between the effect of the speaker and that of the channel in the GMM supervector space.

Given an utterance, the speaker-and-channel-dependent GMM supervector is written as follows

\[ M = m + Tw \] (1)

where \( m \) is the speaker-and-channel-independent supervector (which can be taken to be the UBM supervector), the total variability space \( T \) is a rectangular matrix of low rank and the identity vector or i-vector \( w \) is a random vector having a standard normal distribution \( N(0, I) \). The components of the vector \( w \) are the total factors.
2.2. Support Vector Machine (SVM)

SVM [2] is used as a classifier for our proposed NPE-projected vector. An SVM is a supervised binary classifier which searches for the hyperplane that best discriminates two given classes of patterns according to a maximum separation margin criterion. It can be constructed from sums of a kernel function \( K(\cdot, \cdot) \)

\[
f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + d \tag{2}
\]

where \( N \) is the number of support vectors, \( t_i \) is the ideal output, \( \alpha_i \) is the weight for the support vector \( x_i \), \( \alpha_i > 0 \) and \( \sum_{i=1}^{N} \alpha_i t_i = 0 \); the ideal outputs are either 1 or -1, depending on whether the corresponding support vector belongs to class 0 or class 1. For classification, a class decision is made based on whether the value \( f(x) \) is above or below a threshold.

2.3. GMM Supervector

Since the universal background model (UBM) is included as a part in most speaker recognition systems, it provides a natural way to create supervectors [12]. This leads to hybrid classifier where the generative GMM-UBM model is used for creating "feature vector" for the discriminative SVM. In this study, GMM-UBM training is implemented in the form of all mixture components are concatenated by MAP adaptation of the mean for an utterance. The supervectors approach is based on GMM supervector.

The details about how to solve the above optimization can be found in [13].

Finally, the NPE projection can be obtained by solving the following generalized eigenvector problem

\[
WNW^T a = \lambda WW^T a \tag{4}
\]

where \( W = (w_1, w_2, \ldots, w_m) \)

\[
N = (I - E)^T (I - E) \tag{5}
\]

\[
I = \text{diag}(1, \ldots, 1)
\]

Let \( a_1, a_2, \ldots, a_K \) be the generalized eigenvectors for the solutions of equation (4) corresponding to the \( K \) largest eigenvalues. In this work, \( K \) is set to 300. Thus, the NPE transformation matrix is as follows

\[
A_{NPE} = (a_1, a_2, \ldots, a_K)^T \tag{6}
\]

3. Neighborhood Preserving Embedding Approach

3.1. The NPE Algorithm

In this section, the algorithmic training procedure of NPE is formally stated [10].

For the first step, we construct an adjacency graph \( G \) with nodes. Given \( m \) labeled training utterances, the \( i \)-th node corresponds to the supervector point \( w_i \) of the \( i \)-th utterance. We put a directed edge from node \( i \) to \( j \) if the supervectors \( w_i \) and \( w_j \) are from the same class, i.e. the same speaker.

Next, the weights on the edges are computed. Let \( E \) denote the weight matrix with \( E_{ij} \) having the weight of the edge from node \( i \) to node \( j \), and 0 if there is no such edge. The weights on the edges can be computed by minimizing the following objective function,

\[
\min \sum_i ||w_i - \sum_j E_{ij}w_j||^2 \tag{3}
\]

with constrains

\[
\sum_j E_{ij} = 1, j = 1, 2, \ldots, m.
\]

The new space, we refer to as the NPE space, can be defined by the transformation matrix \( A \). We name \( w' \) as the probabilistic principal component analysis (PPCA) projection.

Finally, the PCA projection in (7) is conducted in a way similar to the total variability approach.

Thus, in this study, the PCA projection in (7) is conducted in a way similar to the total variability approach.

Then, NPE projection is implemented after the PCA projection as

\[
w \rightarrow w' = A_{NPE}w
\]

Different from the PCA, NPE considers the manifold structure which is modeled by an adjacency graph and gains the embedding that preserves local information. Thus, after the NPE transformation matrix \( A_{NPE} \) in (8), the supervector \( w \) obtained through PCA projection according to (7) can be further projected to \( w' \), which is believed to preserve both global and local information.

Thus, the final embedding is as follows to each GMM supervector \( x \):

\[
x \rightarrow w' = Ax \tag{9}
\]

\[
A \rightarrow A_{NPE}A_{PCA} \tag{10}
\]
the NPE-projected vector. Since the NPE algorithm is performed in a supervised mode, the speaker class information can be effectively utilized.

4. Intersecion Compensation

After the new feature extractor as described in section 3, the intersession compensation can be carried out in a low-dimensional space where the NPE-projected vector \( w' \) lies. In our experiment, we use the linear discriminant analysis (LDA) approach and within class covariance normalization (WCCN) approach for intersession compensation [6].

4.1. Linear Discriminant Analysis

Linear discriminant analysis (LDA) [14] is a technique for dimensionality reduction that is widely used in the field of pattern recognition. The idea behind this approach is to seek new orthogonal axes to better discriminate between different classes. All of the NPE-projected vectors from the same speaker are regarded as the same class in linear discriminant analysis. The matrix LDA transformation matrix \( A_{LDA} \) is consisted of the eigenvectors of equation (11)

\[
S_{by} = \lambda S_{w}v
\]  

(11)

where \( \lambda \) is the diagonal matrix of eigenvalues. The matrix \( S_{by} \) is the between class covariance matrix and \( S_{w} \) is the within class covariance matrix. With the LDA transformation matrix \( A_{LDA} \), the NPE-projected vector \( w' \) can be transformed by the following form

\[
w^* = A_{LDA}^t w'
\]  

(12)

4.2. Within Class Covariance Normalization

The idea behind within class covariance normalization (WCCN) [15] is to minimize the expectation error rate of false alarms and false rejections during the SVM training step. WCCN is successfully applied in speaker recognition [6] [15]. All utterances of a given speaker are considered to belong to one class. The within class covariance matrix \( W \) can be obtained as follows:

\[
W = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_{s,i} - \bar{w}_s)(w_{s,i} - \bar{w}_s)^t
\]  

(13)

where \( \bar{w}_s = 1/n_s \sum_{i=1}^{n_s} w_{s,i} \) is the mean of NPE-projected vectors of each speaker, \( S \) is the number of speakers and \( n_s \) is the number of utterances of speaker \( s \). A feature mapping function can be defined as follows:

\[
\phi(w') = A_{WCCN}w'
\]  

(14)

where \( A_{WCCN} \) is obtained using a Cholesky decomposition of the matrix \( W^{-1} = A_{WCCN}A_{WCCN}^t \).

Table 1: EER(%) and minDCF of different factor analysis methods without any intersession compensation on the NIST SRE2008 tel-tel condition.

<table>
<thead>
<tr>
<th>System</th>
<th>Male</th>
<th>Female</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>total variability</td>
<td>6.49</td>
<td>0.306</td>
<td>8.54</td>
<td>0.385</td>
</tr>
<tr>
<td>NPE</td>
<td>5.26</td>
<td>0.248</td>
<td>6.48</td>
<td>0.321</td>
</tr>
</tbody>
</table>

5. Experiments

5.1. Experimental Setup

The experiments for different systems based on two kinds of factor analysis methods, including the total variability and the proposed NPE, were carried out on the NIST 2008 speaker recognition evaluation corpus. In this work, we focused on the telephone-telephone condition. Equal error rate (EER) and the minimum decision cost function (minDCF) were used as metrics for evaluation [16].

The speech utterance was first converted to a sequence of 36-dimensional feature vectors including 18 MFCC coefficients and their first order derivatives over 5 frames. To reduce channel effects, feature warping, CMN and CVN were performed to the feature vectors.

The gender dependent UBM models with 1024 mixture were trained using the NIST SRE 2004 side training corpus. The background data for SVM system were selected from the data of NIST SRE 2004 and 2005. We used the Switchboard II, Switchboard Cellular corpus as well as the telephone data from NIST SRE 2004, 2005 and 2006 corpus for estimating the total variability space. The NIST SRE 2004, 2005 and 2006 data sets were used for training the NPE, WCCN and LDA matrix.

The SVMLight toolkit [17] was used for SVM modeling.

The raw scores are speaker-normalized by means of gender-dependent ZTnorm with telephone utterances drawn from the NIST SRE 2006 corpus.

5.2. Experimental Results

In Table 1, we give the performance of the state-of-the-art total variability and our proposed NPE speaker recognition systems without any intersession-compensation on the NIST SRE2008 task respectively across all the male and female speakers. It is observed that our proposed NPE system produces better performance than the total variability system. It leads to a relative improvement of 18.9% in EER and 19.0% in minDCF for male, and improvement of 24.1% in EER and 16.6% in minDCF for female.

In Table 2, we compare the performance of NPE with total variability using the intersession compensation of LDA. It can be seen from Table 2 that when the LDA
Table 2: EER(%) and minDCF of different factor analysis methods with LDA on the NIST SRE2008 tel-tel condition

<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>total variability + LDA</td>
<td>5.28</td>
<td>0.251</td>
<td>7.01</td>
<td>0.320</td>
</tr>
<tr>
<td>NPE + LDA</td>
<td>4.20</td>
<td>0.212</td>
<td>5.60</td>
<td>0.273</td>
</tr>
</tbody>
</table>

Table 3: Performance of different factor analysis methods with WCCN on the NIST SRE2008 tel-tel condition

<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>total variability + WCCN</td>
<td>5.42</td>
<td>0.234</td>
<td>7.16</td>
<td>0.329</td>
</tr>
<tr>
<td>NPE + WCCN</td>
<td>4.57</td>
<td>0.221</td>
<td>5.99</td>
<td>0.292</td>
</tr>
</tbody>
</table>

compensation is used, the NPE system also outperforms the total variability method, yielding 20.5% relative improvement in EER and 15.5% in minDCF for male, as well as 20.1% relative improvement in EER and 14.7% in minDCF for female.

The performance of NPE and total variability with WCCN intersession compensation is listed in Table 3, which shows the same trend as the performances above. Compared with the total variability method, our proposed NPE method gives additional gains of 15.7% and 5.5% respectively in EER and minDCF for male, as well as 16.3% and 11.2% respectively in EER and minDCF for female.

Table 4 lists the performance of two different factor analysis methods (NPE and total variability) with the intersession compensation of both LDA and WCCN. We can see that our NPE method still gives better performance than the total variability method when both LDA and WCCN are used.

6. Conclusion

In this paper, we propose a new factor analysis method of neighborhood preserving embedding (NPE) to speaker verification. The NPE method can preserve global information as well as make good use of the labeled speaker information of the training data. The experimental results on NIST SRE 2008 telephone-telephone condition indicate that the proposed NPE method still contains speaker dependent information and is effective for the task of speaker verification. Our proposed method outperforms the state-of-the-art total variability factor analysis approach. In future work, we would like to test the NPE method on other conditions beyond the telephone speech.

7. Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319).

8. References