LETTER

Factor Analysis of Neighborhood-Preserving Embedding for Speaker Verification*

Chunyan LIANG†(a), Lin YANG†, Nonmembers, Qingwei ZHAO‡, Member, and Yonghong YAN†, Nonmember

SUMMARY In this letter, we adopt a new factor analysis of neighborhood-preserving embedding (NPE) for speaker verification. 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.

key words: speaker verification, neighborhood-preserving embedding, total variability, support vector machine, cosine distance scoring

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] has been adopted as a powerful discriminative classifier. Factor analysis has led to the development of an effective method of compensating for intersession variability in speaker verification. 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 spaces: the speaker space and the channel space. With the application of SVM, the speaker GMM supervector obtained with JFA or the speaker factor from JFA [7] is used as the input for SVM. In contrast, in the total variability factor analysis, which is also referred to as i-vector method, 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. Channel compensation techniques, such as Linear Discriminant Analysis (LDA) and within-class covariance normalization (WCCN), are carried out in the total variability factor space. A new cosine distance scoring method based directly on the cosine kernel has also been successfully applied on the total factors [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. 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.

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

Copyright © 2012 The Institute of Electronics, Information and Communication Engineers
2.2 Cosine Distance Scoring

Cosine distance scoring [6] directly uses the value of the cosine kernel between the target speaker vector \( w_{\text{target}} \) and the test vector \( w_{\text{test}} \) as the decision score:

\[
\text{score}(w_{\text{target}}, w_{\text{test}}) = \frac{< w_{\text{target}}, w_{\text{test}} >}{\| w_{\text{target}} \| \| w_{\text{test}} \|} \leq \theta
\]  

(2)

The value of this kernel is then compared to the threshold \( \theta \) in order to take the final decision.

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 by MAP adaptation of the mean for an utterance. The mean values of all mixture components are concatenated to form one GMM supervector and our proposed NPE approach is based on GMM supervector.

3. Neighborhood-Preserving Embedding Approach

3.1 The NPE Algorithm

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

For the first step, we construct an adjacency graph \( G \) with nodes. Given \( n \) 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,

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

(3)

with constrains

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

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

\[
WNW^T a = \lambda W W^T a
\]  

(4)

where

\[
W = (w_1, w_2, \ldots, w_n)
\]

\[
N = (I - E)^T (I - E)
\]

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

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

\[
ANPE = (a_1, a_2, \ldots, a_K)^T
\]  

(6)

Above is the supervised mode of NPE in which the speaker label information is used. Without the label information, NPE can also be performed in the unsupervised mode where the adjacency graph \( G \) in the NPE algorithm can be constructed by \( K \) nearest neighbors.

3.2 Speaker Verification Based NPE

In our proposed system, the NPE projection is carried combined with the traditional principal component analysis (PCA) projection.

Firstly, the PCA projection is performed to each GMM supervector \( x \) as follows

\[
x \rightarrow w = A_{PCA} x
\]  

(7)

where \( A_{PCA} \) is the PCA projection matrix and \( w \) is the projected low-dimensional representation of GMM supervector \( x \). The probabilistic principal component analysis (PPCA) with EM algorithm is used in our implementation instead of PCA. Actually, the total variability can be considered as a classical PPCA model [9]. That is, in this study, the PCA projection in (7) is performed in a way similar to the total variability approach.

Then, the NPE projection is implemented after the PCA projection as

\[
w \rightarrow w' = ANPE w
\]  

(8)

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 \( ANPE \) 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 \( x \):

\[
x \rightarrow w' = Ax
\]  

(9)

\[
A = ANPE A_{PCA}
\]  

(10)

where \( A \) denotes the final NPE transformation matrix. The new space, we refer to as the NPE space, can be defined by the transformation matrix \( A \). We name \( w' \) as the NPE-projected vector. Since the NPE algorithm is performed in a supervised mode, the speaker class information can be effectively utilized.
4. Intersession Compensation

After the new feature extractor as described in Sect. 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 recorded as the same class in linear discriminant analysis. With the LDA transformation matrix $A_{LDA}$, the NPE-projected vector $w'$ can be transformed by the following form

$$w' = A_{LDA}^T w$$

(11)

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. A feature mapping function can be defined as follows:

$$\phi(w') = A_{WCCN} w'$$

(12)

where $A_{WCCN}$ is the transformation matrix for WCCN.

The detail information of $A_{LDA}$ and $A_{WCCN}$ can be found in [6].

5. Experiments

5.1 Experimental Setup

The experiments for different systems based on the 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 input 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 to a Gaussian distribution, CMN and CVN were performed to the feature vectors.

The gender dependent UBM models with 1024 mixture components were trained using the NIST SRE 2004 1side training corpus. The background data for SVM system were selected from the data of NIST SRE 2004 and NIST SRE 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 datasets were used for training the NPE, WCCN and LDA matrix. In order to find the optimal dimension of the projection matrices for PCA (total variability), NPE and LDA, we compared the performance by varying the number of their spatial dimensions from 100 to 400. The best results were obtained with the dimension of 400 for PCA (total variability), 300 for NPE and 250 for LDA. The results below are based on the optimal dimension of the projection matrices.

SVM with linear kernel was used as the classifier in both the conversional total variability approach and the proposed NPE method. The SVMLight toolkit [17] was used for SVM modeling. We also tested the cosine scoring method.

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

5.2 SVM

In this subsection, the results are based on the SVM classifier.

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 SRE 2008 task respectively across all the male and female speakers. NPE was tested in both unsupervised and supervised modes. It is observed that our proposed NPE systems (supervised and unsupervised) produce better performances than the total variability system. The supervised NPE system 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. We can also see that NPE in supervised mode outperforms that in unsupervised mode but the difference is not significant, which indicates that the speaker label information is useful but not the dominant factor. Thus, we infer that the local information may play the key role. To test the effectiveness of preserving the local information, we also carried out the experiments where NPE was performed alone, not along with

| system                  | Male | Female |
|-------------------------|------|--------|
| EER         | 6.49 | 8.54   |
| minDCF      | 0.306| 0.385  |
| NPE (supervised) | 5.26 | 6.48   |
| EER         | 0.248| 0.321  |
| minDCF      | 0.260| 0.338  |
| NPE (unsupervised) | 5.35 | 6.62   |
| EER         | 0.293| 0.373  |
| minDCF      | 0.246| 0.351  |
| NPE (without PCA) | 6.27 | 8.40   |
| EER         | 0.293| 0.373  |
| minDCF      | 0.246| 0.351  |
Table 2  EER (%) and minDCF of different factor analysis methods based on SVM with LDA on the NIST SRE 2008 tel-tel condition.

| system                  | Male EER | Male minDCF | Female EER | Female minDCF |
|-------------------------|----------|-------------|------------|---------------|
| total variability + LDA | 5.28     | 0.251       | 7.01       | 0.320         |
| NPE + LDA               | 4.20     | 0.212       | 5.60       | 0.273         |

Table 3  Performance of different factor analysis methods based on SVM with WCCN on the NIST SRE 2008 tel-tel condition.

| system                        | Male EER | Male minDCF | Female EER | Female minDCF |
|-------------------------------|----------|-------------|------------|---------------|
| total variability + WCCN      | 5.42     | 0.234       | 7.16       | 0.329         |
| NPE + WCCN                    | 4.57     | 0.221       | 5.99       | 0.292         |

Table 4  Comparison of different factor analysis methods based on SVM with both LDA and WCCN on the NIST SRE 2008 tel-tel condition.

| system                        | Male EER | Male minDCF | Female EER | Female minDCF |
|-------------------------------|----------|-------------|------------|---------------|
| total variability+LDA+WCCN    | 4.31     | 0.214       | 6.08       | 0.276         |
| NPE+LDA+WCCN                  | 4.11     | 0.205       | 5.52       | 0.252         |

PCA (total variability). The results of NPE without PCA (total variability) is just slightly better than that of PCA (total variability), which indicates that the reason for the good performance of the proposed method (along with PCA) is in the fact that it preserves both the global and local information through projection but not only the local information.

The following results of NPE are all based on the supervised mode and along with PCA (total variability).

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 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.

5.3 Cosine Distance Scoring

The cosine distance scoring is based on the same i-vectors and NPE projected vectors as the previous SVM system.

Table 5 compares the performance of two different factor analysis methods (NPE and total variability) with the intersession compensation of both LDA and WCCN using cosine distance scoring. We can see that the results obtained with cosine distance scoring slightly outperform those obtained with SVM classifier and our NPE method still gives better performance than the total variability method with the cosine distance scoring.

6. Conclusion

In this letter, we propose a new factor analysis method of neighborhood-preserving embedding (NPE) to speaker verification. The NPE method can preserve local 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 telephone speech.

References

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digit. Signal Process., vol.10, pp.19–41, 2000.
[2] N. Cristianini and J. Shawe-Taylor, Support vector machines, Cambridge University Press, Cambridge, UK, 2000.
[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” IEEE Trans. Audio Speech Language Process., vol.15, no.4, pp.1453–1447, 2007.
[4] P. Kenny, P. Ouellet, N. Dekah, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” IEEE Trans. Audio Speech Language Process., vol.16, no.5, pp.980–988, 2008.
[5] N. Dekah, R. Dekah, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification,” Proc. International Conferences on Spoken Language Processing (ICSLP), pp.1559–1562, 2009.
[6] N. Dekah, P. Kenny, R. Dekah, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Trans. Audio Speech Language Process., vol.19, no.4, pp.788–798, May 2011.
[7] N. Dekah, P. Kenny, R. Dekah, O. Glembek, P. Dumouchel, L. Burget, and V. Hubeika, “Support vector machines and joint factor analysis for speaker verification,” ICASSP2009, pp.4237–4240, Taipei, Taiwan, 2009.
[8] M.E. Tipping and C.M. Bishop, “Probabilistic principal component analysis,” J. Royal Statistical Society: Series B (Statistical Methodology), vol.61, no.3, pp.611–622, 1999.
[9] J.H.L. Hansen and Y. Lei, “Speaker recognition using supervised probabilistic principal component analysis,” Proc. Interspeech, pp.382–385, 2010.
[10] X. He, D. Cai, S. Yan, and H.-J. Zhang, “Neighborhood preserving embedding,” Tenth IEEE International Conference on Computer
[11] I. Jolliffe, Principal component analysis, Springer New York, 2002.

[12] N. Dehak and G. Chollet, “Support vector GMMs for speaker verification,” Proc. IEEE Odyssey: The Speaker and Language Recognition Workshop (Odyssey 2006), pp.1–4, San Juan, Puerto Rico, June 2006.

[13] S. Roweis and L.K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” Science, vol.290, 22 Dec. 2000.

[14] J.P. Hespanha, P.N. Belhumeur, and D.J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” IEEE Trans. Pattern Anal. Mach. Intell., vol.19, no.17, pp.711–730, 1997.

[15] A. Hatch, S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for SVM-based speaker recognition,” International Conference on Spoken Language Processing, pp.1471–1474, Pittsburgh, PA, USA, Sept. 2006.

[16] “The NIST year 2008 speaker recognition evaluation plan,” http://www.nist.gov/speech/tests/spk/2008/index.html, 2008.

[17] T. Joachims, “Svmlight: Support vector machine,” SVM-Light Support Vector Machine, http://svmlight.joachims.org/, University of Dortmund, vol.19, 1999.