FACTOR ANALYSIS OF LAPLACIAN APPROACH FOR SPEAKER RECOGNITION

Jinchao Yang, Chunyan Liang, Lin Yang, Hongbin Suo, Junjie Wang, Yonghong Yan

Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences, Beijing
{yangjinchao, liangchunyan, yanglin, hsuo, wangjunjie, yonghong.yan}@hccl.ioa.ac.cn

ABSTRACT

In this study, we introduce a new factor analysis of Laplacian approach to speaker recognition under the support vector machine (SVM) framework. The Laplacian-projected supervector from our proposed Laplacian approach, which finds an embedding that preserves local information by locality preserving projections (LPP), is believed to contain speaker dependent information. The proposed method was compared with the state-of-the-art total variability approach on 2010 National Institute of Standards and Technology(NIST) Speaker Recognition Evaluation (SRE) corpus. According to the compared results, our proposed method is effective.

Index Terms— speaker recognition, factor analysis, Laplacian, locality preserving projections, support vector machine

1. INTRODUCTION

The Gaussian mixture model (GMM)[1] is the most widely used approach for text-independent speaker recognition. Support vector machine (SVM)[2][3] has been proved to be an effective method for speaker recognition task. In the GMM-SVM system, we combine the SVM method with the GMM supervector [4]. Recently, factor analysis approach has been successfully used for speaker recognition system to compensate the variability from the change in channel, gender and environment. Some approaches combining factor analysis and SVM have been successfully used in speaker recognition. For example, the supervector obtained from joint factor analysis (JFA)[5] was applied as the input feature to SVM classifier in speaker recognition.

Recently, total variability approach has been proposed in speaker recognition [5][6], which uses the factor analysis to define a new low-dimensional space named total variability space. In contrast to classical joint factor analysis (JFA), the speaker and the channel variability are contained simultaneously in this new space. The intersession compensation can be carried out in low-dimensional space.

Actually, we can consider total variability approach as a classical application of the probabilistic principal component analysis (PPCA)[7]. The factor analysis of the total variability approach can obtain useful information by reducing the dimension of the space of GMM supervectors. All utterances could in fact be well represented in a low-dimensional space. Recent work[8] shows that a different structure in terms of nonlinear manifolds exist within the high-dimensional space if the evaluation data is available a-priori. In this study, we think over whether additional nonlinear structure could be used if the evaluation data is not available a-priori.

A number of researches show that the face images have reside on a nonlinear submanifold, and we are interested in the laplacianface which is proposed in [9][10]. The face images are mapped into a face subspace for analysis by locality preserving projections (LPP) in the laplacianface approach which finds an embedding that preserves local information. In this paper, we introduce Laplacian approach as a new factor analysis approach to speaker recognition under the support vector machine(SVM) framework.

This paper is organized as follows: In section 2, we give a simple review of total variability, support vector machine, Gaussian mixture model supervector and laplacianface. In section 3, Laplacian approach for speaker recognition is presented in detail. Section 4 gives experimental setup and experimental results. Finally, we conclude in section 5.

2. BACKGROUND

2.1. Total Variability

In speaker recognition, unlike in classical joint factor analysis (JFA), the total variability approach defines a new low-dimensional space that is named total variability space, which contains the speaker and the channel variability simultaneously. The total variability approach in speaker recognition releases the independent assumption between speaker and channel variability spaces in JFA speaker recognition [11].

For a given utterance, the speaker and channel variability dependent GMM supervector is denoted in equation (1).

$$M = m_{ubm} + Tw \tag{1}$$

where m_{ubm} is the UBM supervector, T is total variability space, and the member of the vector w is total factor.

2.2. Support Vector Machine

SVM [12] is used as a classifier for our proposed laplacian-projected supervector. An SVM is a two-class classifier constructed from sums of a kernel function $K(\cdot, \cdot)$:

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

where N is the number of support vectors, t_i is the ideal output, α_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 upon whether the corresponding support vector belongs to class 0 or class 1. For classification, a class decision is based upon whether the value, $f(x)$, is above or below a threshold.

2.3. GMM Supervector

Suppose we have a Gaussian mixture model universal background model (UBM),

$$g(\eta) = \sum_{i=1}^{N} w_i p(\eta; m_i; \Sigma_i) \quad (3)$$

where w_i, $i = 1, ..., N$, are the mixture weights, N is the number of mixtures , $p(\cdot)$ is a Gaussian, and m_i and Σ_i are the mean and covariance of the Gaussians respectively. We assume Σ are diagonal covariances.

For an utterance, GMM-UBM training is implemented by MAP adaptation [13] of the mean. The mean vectors of all mixture components are concatenated to form one GMM supervector for each utterance. In this study, our proposed approach is based on GMM supervector.

3. LAPLACIAN APPROACH

Factor analysis approach in speaker recognition was similar to researches on face recognition about eigenfaces. In recent research, laplacianface is proposed[9][10]. The face images are mapped into a face subspace for analysis by locality preserving projections in the laplacianface approach. In face recognition, compared to eigenfaces which effectively find the Euclidean structure of face space, LPP, which finds an embedding that preserves local information, is used in face recognition and obtains good performance successfully. We introduce Laplacian approach with locality preserving projection(LPP) as a new factor analysis approach to speaker recognition under the support vector machine(SVM) framework.

3.1. Laplacian-projected supervector

In laplacianface, LPP projection is carried on after principal component analysis (PCA) projection, which is the most popular method to process and compress data. In our proposed system, the PCA projection is as follows to each GMM supervector x

$$x \longrightarrow w = A_{PCA} x \quad (4)$$

We consider w as a low-dimensional representation of GMM supervector x. In our proposed Laplacian approach for speaker recognition, the probabilistic principal component analysis (PPCA) [7][14] with EM algorithm is used in ours process instead of PCA. Actually, the total variability in recent research can be considered as a classical PPCA model[15]. That is, in our proposed approach, PCA projection is carried out similar to total variability approach.

Locality preserving projection (LPP) [9][10][14][16] is different from PCA and Linear Discriminant Analysis (LDA) which effectively preserve global structure and linear manifold. LPP considers the manifold structure which is modeled by a nearest-neighbor graph. LPP can gain an embedding that preserves local information. In this way, the variability resulting from changes in channel, gender and environment may be eliminated or reduced.

$$w^{'} = A_{LPP} w \quad (5)$$

By LPP transformation matrix A_{LPP} in equation (5), the supervector w after PCA projection is projected to $w^{'}$ to preserve local information.

Firstly, for training LPP transformation matrix, we construct the nearest-neighbor graph. Let G denote a graph with m nodes. The ith node corresponds to the supervector w_i. We put an edge between nodes i and j while i is among k nearest neighbors of j, or j is among k nearest neighbors of i. In this paper, k is set to be 3. If nodes i and j are connected, let

$$E_{ij} = e^{-\frac{(w_i - w_j)^2}{t}} \quad (6)$$

The justification for this choice of weights can be traced back to [17].

Then, we compute the eigenvectors and eigenvalues for generalized eigenvector problem:

$$WLW^T a = \theta W D W^T a \quad (7)$$

where D is a diagonal matrix whose entries are column sums of E, $D_{ij} = \sum_j E_{ji}$. $L = D - E$ is the Laplacian matrix. The ith row of matrix W is w_i. Let $a_0, a_1, ..., a_{\tau-1}$ be the solution to (7), sorted by their eigenvalues, $0 \leq \theta_0 \leq \theta_1 \leq ... \leq \theta_{\tau-1}$. Thus, the LPP transformation matrix is as follows:

$$A_{LPP} = (a_0, a_1, ..., a_{\tau-1}) \quad (8)$$

Thus the embedding is as follows to each GMM supervector x:

$$x \longrightarrow y = Ax \quad (9)$$

$$A = A_{LPP} A_{PCA} \quad (10)$$

where A denote the Laplacian transformation matrix, And We call y (or $w^{'}$) laplacian-projected supervector.

3.2. Intersession Compensation

After the new feature extractor, the intersession compensation can be carried out in low-dimensional space. In our experiment, we use the linear discriminant analysis (LDA) approach [4] [18] and within class covariance normalization (WCCN) [19] [5][6] approach for intersession compensation.

3.2.1. Linear Discriminant Analysis

All of the total factor vectors of the same speaker are recorded as the same class in linear discriminant analysis.

$$w^* = A_{LDA} w^{'} \quad (11)$$

By LDA transformation in equation (11), the total factor vector w is projected to new axes that maximize the variance between speakers and minimize the intra-class variance. The matrix A_{LDA} is contained of the eigenvectors of equation (12).

$$S_b \nu = \lambda S_w \nu \quad (12)$$

where λ is the diagonal matrix of eigenvalues. The matrix S_b is the between class covariance matrix and S_w is the within class covariance matrix.

3.2.2. Within Class Covariance Normalization

WCCN is presented in detail in [19] and is successfully applied in speaker recognition [5][6]. All utterances of a given speaker are considered to belong to one class.

$$W = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_i^{'s} - \overline{w_s})(w_i^{'s} - \overline{w_s})^t \quad (13)$$

where $\overline{w_s} = \frac{1}{n_s}\sum_{i=1}^{n_s} w_i^{'s}$ is the mean of laplacian-projected supervectors of each speaker, S is the number of speakers and n_s is the number of utterances of speaker s.

$$w^{\tau} = Bw^{'} \tag{14}$$

$$W^{-1} = BB^{t} \tag{15}$$

where B is the Cholesky decomposition of W^{-1}. w^{τ} is used as feature for SVM classifier.

4. EXPERIMENTS

4.1. Experimental Setup

We performed experiments on the 2010 NIST SRE corpus (condition5 ,condition6 and condition8). We focus on speaker detection in the context of conversational speech. The task is to determine whether a specified speaker is speaking during a given segment of conversational speech. We use equal error rate (EER) and the minimum decision cost value (minDCF) as metrics for evaluation [20].

For cepstral feature extraction, a 20 ms Hamming window with 10 ms shift is used. Each utterance is converted into a sequence of 36-dimensional feature vectors, each consisting of 12 MFCC coefficients and their first and second derivatives. An energy-based speech detector is applied to discard vectors from low-energy frames. Feature warping, cepstral mean subtraction and variance normalization are applied to the features to mitigate channel effects.

Gender-dependent UBMs with 1024 mixture gauss number were trained using EM with the data from the corpora: NIST01, NIST02, NIST04, NIST05. The full background training dataset consisted 4668 female and 3414 male conversation, and these data are chosen from NIST04, NIST05, SwbC. We used all of the training data for estimating the total variability space. The NIST SRE 2004, 2005 and 2006 datasets were used for training WCCN and the LDA matrix. The SVMLight toolkit [3] was used for SVM modeling.

4.2. Experimental Result

In Table 1, we give the performances of the state-of-the-art total variability and our proposed Laplacian speaker recognition systems on NIST 2010 SRE task in female and male. Then we compare the results of the state-of-the-art total variability system to Laplacian system. It is observed that our proposed Laplacian system produces better performance than total variability system. It leads to a relative improvement of 12.0% in EER and 11.3% in minDCF in female, and 28.9% in EER and 9.8% in minDCF in male.

Table 1. EER(%) And minDCF*100 of 2010 NIST-SRE task in female and male.

	female		male	
System	EER	minDCF*100	EER	minDCF*100
total variability	9.84	4.16	8.42	3.36
Laplacian	8.66	3.69	5.99	3.03

Table 2 shows the results of the state-of-the-art total variability and our proposed Laplacian speaker recognition systems with the intersession compensation techniques of LDA and WCCN in female. EER and minDCF are observed. In our experiments, it is observed that the two intersession compensation techniques of LDA and WCCN is effective for the state-of-the-art total variability and our proposed Laplacian speaker recognition systems. With the performance comparison, total variability approach with LDA obtains better performance than Laplacian approach with LDA, and Laplacian approach with WCCN performs better than total variability approach with WCCN. The experimental results in male are shown in table 3. It is observed that the trends of experimental results are similar to the experimental results in female.

Table 2. EER(%) And minDCF*100 of 2010 NIST-SRE task in female with LDA or WCCN.

	female	
System	EER	minDCF*100
total variability + LDA	7.40	2.86
Laplacian + LDA	7.65	3.11
total variability + WCCN	9.32	3.76
Laplacian + WCCN	6.22	2.70

Table 3. EER(%) And minDCF*100 of 2010 NIST-SRE task in male with LDA or WCCN.

	male	
System	EER	minDCF*100
total variability + LDA	5.07	2.22
Laplacian + LDA	5.55	2.46
total variability + WCCN	6.77	2.66
Laplacian + WCCN	5.08	2.28

Lastly, experimental results with female and male combined are given in table 4. Comparing to total variability approach, our proposed Laplacian approach obtains relative improvement of 17.8% in EER and 10.5% in minDCF without intersession compensation techniques, which demonstrates that our proposed Laplacian approach is feasible. When intersession compensation techniques are used, our proposed Laplacian approach with WCCN may achieves best performance though total variability approach with LDA performs better than Laplacian approach with LDA. We notice that WCCN is much more suitable for our proposed Laplacian approach than LDA. Therefore we suggest using WCCN as intersession compensation technique for our proposed Laplacian approach. Figure 1 shows that the relative improvement of speaker recognition performance is observable with our proposed Laplacian approach.

Table 4. EER(%) And minDCF*100 of 2010 NIST-SRE task without intersession compensation, and with LDA or WCCN.

System	EER	minDCF*100
total variability (a)	9.27	3.81
Laplacian (b)	7.62	3.41
total variability + LDA(c)	6.23	2.57
Laplacian + LDA (d)	6.60	2.85
total variability + WCCN(e)	8.54	3.24
Laplacian + WCCN (f)	5.70	2.52

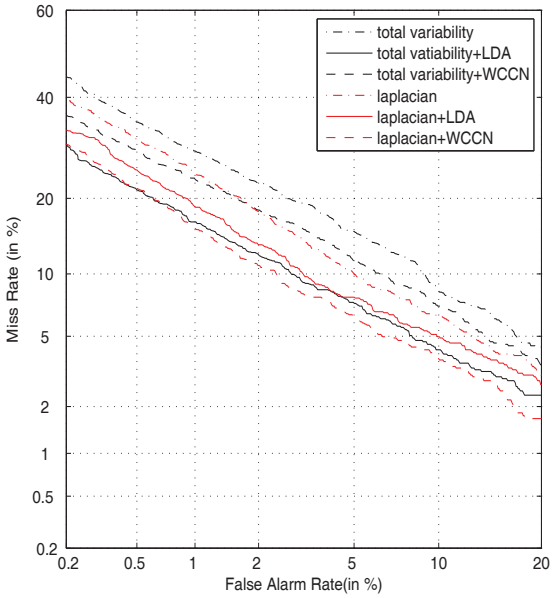

Fig. 1. DET curves for each system

5. CONCLUSIONS

In this paper, we propose a new factor analysis of Laplacian approach to speaker recognition by introducing laplacianfaces to speaker recognition. Our experiments show that projecting GMM supervectors into Laplacian space still contains speaker dependent information. Comparing to supervector from the state-of-the-art total variability approach, our proposed Laplacian approach can achieve better performance on some condition. Future work will examine whether the new factor analysis of Laplacian approach is useful for other pattern recognition problem to obtain much more desired useful information.

Acknowledgments

This work is partially supported by The National Science and Technology Pillar Program (2008BAI50B00), National Natural Science Foundation of China (No. 10925419, 90920302,10874203, 60875014).

6. REFERENCES

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," 2001.

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

[4] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and PA Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[5] N. Dehak, R. Dehak, 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," in *Proc. International Conferences on Spoken Language Processing (ICSLP)*, 2009, pp. 1559–1562.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *submitted to IEEE Transaction on Audio, Speech and Language Processing*.

[7] M.E. Tipping and C.M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[8] Zahi N.Karam and William M.Campbell, "Gragh-Embedding for Speaker Recognition," in *Proc. interspeech*, 2010.

[9] X. He and P. Niyogi, *Locality preserving projections*, Citeseer, 2005.

[10] X. He, S. Yan, Y. Hu, P. Niyogi, and H.J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 328–340, 2005.

[11] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[12] N. Cristianini and J. Shawe-Taylor, "Support Vector Machines," *Cambridge University Press, Cambridge, UK*, 2000.

[13] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE transactions on speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.

[14] J. Yang, X. Zhang, B Hong, L. Lu, J. Zhang, and Y. Yan, "Low-dimensional Representation of Gaussian Mixture Model Supervector For Language Recognition," *Accepted by EURASIP Journal on Advances in Signal Processing*, 2011.

[15] John H.L.Hansen Yun Lei, "Speaker Recognition using Supervised Probabilistic Principal Component Analysis," in *Proc. interspeech*, 2010.

[16] J. Yang, X. Zhang, L. Lu, J. Zhang, and Y. Yan, "Language recognition with locality preserving projection," in *The Sixth International Conference on Digital Telecommunications (ICDT 2011,)*, 2011, pp. 46–50.

[17] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in neural information processing systems*, vol. 1, pp. 585–592, 2002.

[18] E. Singer, PA Torres-Carrasquillo, TP Gleason, WM Campbell, and D.A. Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language identification," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[19] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proc. Interspeech*. Citeseer, 2006, vol. 4.

[20] "2008 NIST Speaker Recognition Evaluation Plan," *http://www.itl.nist.gov/iad/mig//tests/sre/2008/index.html*, 2008.