TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.6, June 2014, pp. 4299 ~ 4305
DOI: 10.11591/telkomnika.v12i6.3994

Speaker Recognition Based on i-vector and Improved Local Preserving Projection

Di Wu*, Jie Cao, Jinhua Wang
College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, 730050, China
*Corresponding author, e-mail: wudi6152007@hotmail.com

Abstract

In this paper,a improved local preserve projection algorithm is proposed in order to enhance the recognition performance of the i-vector speaker recognition system under unpredicted noise environment. First , the non zero eigenvalue is rejected when we solve the optimal objective function and only the value greater than zero are used. A mapping matrix is obtained by solving a generalized eigenvalue problem, so can settle the singular value problem always occurred in traditional local preserve projection algorithm. The experiment results shown that the recognition performance of the method proposed in this paper is improved under several kinds of noise environments.

Keywords: *computer application, i-vector, local preserving projection, manifold learning, speaker recognition*

1. Introduction

With the past decades, speaker recognition has become a very popular area of research in pattern recognition, computer vision and machine learning [1]. Due to the mismatch between training and testing conditions caused by some inevitable reason such as channel distortion, different microphones, transmitting channels or encoder. One of the main causes of the performance degradation is the additive noise that may appear in many practical applications. There are a large number of different solutions to alleviate this problem. We can identify three main class of techniques for noise-robust ASR, namely feature enhancement method [2], model adaptation method [3] and score normalization method [4]. The feature enhancement method attempts to normalization the distorted feature, or estimate undistorted feature form the distorted speech, and do not require any explicit knowledge about the noise. Some examples are the cepstral mean normalization (CMN), cepstral mean and variance normalization (CMVN), relative spectra (RASTA) and feature mapping. In contrast, the model adaptation methods work in the backend to compensate by modifying the acoustic models and carried out by using some type of knowledge about the noise. Some typical examples are maximum likelihood linear regression (MLLR), maximum a posterior (MAP), factor analyse (FA) and vector Taylor series (VTS) etc. The score normalization method try to normalizing the output score using various normalization methods, such as HNorm, TNorm and ZNorm etc.

In the last years, the Gaussian Mixture Models based on Universal Background Model (GMM-UBM) [5] has become the most popular modeling approach in speaker recognition, some generative models such as Eigenvoices, Eigenchannels and the most powerful one, the Joint Factor Analysis, have built on the success of the GMM-UBM approach. Recently, a new method which inspired from the joint factor analysis and consists in finding a low dimensional subspace of the GMM supervector space, named the total variability space that represents both speaker and channel variability, the vectors in the low dimensional space are called i-vectors [6]. The i-vector method are become the main stream in the speaker recognition system at home and aboard for the reason of its leading role in the NIST test.

Locality Preserving Projections (LPP) [7] is a manifold learning method widely used in pattern recognition and computer vision, LPP is also well known as a linear graph embedding method. But the traditional LPP method was unsupervised and was proposed for only vector samples, not being able to be directly applied to image samples, so there are been several

Received July 27, 2013; Revised December 23, 2013; Accepted January 22, 2014

types improvements to conventional LPP [8]. The first type of the improvement is supervised LPP, which try to exploiting the class label information of samples in the training phase. The second type changes LPP to a nonlinear transform method by using the kernel trick. The third type of the improvement to LPP mainly focuses on directly implementing LPP for two dimensional rather than one dimensional vectors and its have higher computational efficiency. And the last improvement seeks to obtain LPP solutions with different solution properties, such as orthogonal locality preserving method and uncorrelated LPP feature extraction method.

From the modeling process of the i-vector method, the manifold learning method has been achieved well performance in automatic speaker recognition system. But the LPP algorithm always suffers from the small sample size (SSS) problem. A new solution scheme for LPP is proposed in this paper which can be directly implemented no matter whether there exists the SSS problem or not. We only using the eigenvectors corresponding positive eigenvalue when solving the optimized objective function and removing the zero eigenvalue.

The remainder of the paper is organized as follows: in Section 2 we introduce the conventional LPP and present our new LPP solution. In Section 3 the original i-vector ASR system is given and our new i-vector ASR system based on our new LPP solution is also proposed. In Section 4 we describe the experiment results. Section 5 offers our Conclusion.

2. The improved LPP method

2.1. Description of LPP

LPP was proposed as a way to transform samples into an new space and to ensure that samples that were in close proximity in the original space remain so in the new space. Consider there have l training samples $X = \{x_i\}_{i=1}^{l}$, the goal of LPP is to minimize the following function [9-10]:

$$\min(\sum_{i,j}(W^T x_i - W^T x_i)^2 S_{ij}) \tag{1}$$

The S_{ij} is a symmetric matrix and the element of the S_{ij} is defined as follows:

$$S_{ij} = \begin{cases} \dfrac{\exp\left\|x_i - x_j\right\|^2}{t} & if\ x_j \quad is \quad one \quad of \quad K \quad neighbors \quad of \quad x_i \\ 0 & else \end{cases} \tag{2}$$

From the optimized function equation (1) we can see the local structure of the feature space can preserved like in the original high dimension space after dimension reduction, which means close samples in the original space will still close in the new space, so the projection matrix W can be written as :

$$\begin{aligned} W &= \arg\min W^T XLX^T W \\ &= \arg\min W^T X(D-S)X^T W \end{aligned} \tag{3}$$

In Equation (3), D is diagonal matrix, $D_{ii} = \sum_j S_{ij}$, $L = D - S$, the solution of Equation (3) can be obtained by finding the generalized eigenvalue of the following function:

$$XLX^T W = \lambda XDX^T W \tag{4}$$

2.2. New LPP Solution Scheme

For the conventional LPP method, even if the neighbour samples are from different classes, in the transform space obtained using the conventional LPP solution they might also statistically have the same representation, which is disadvantages for pattern recognition

problems. In other words, it is possible for the conventional LPP solution to produce the same representation for samples from different classes, especially for the samples located on the border of two classes, all unsupervised LPP methods might suffer from this same drawback.

In this section, we describe our new improvement scheme to the conventional LPP solution scheme. First, we demonstrate the effective solution of the conventional LPP solution should be from a subspace XDX^T, for simplicity, we define matrix D_1, L_1 and S_1:

$$\begin{cases} D_1 = XDX^T \\ L_1 = XLX^T \\ S_1 = XSX^T \end{cases} \tag{5}$$

Suppose that $\vec{\partial}_1, \vec{\partial}_2, ..., \vec{\partial}_n$ are the eigenvectors corresponding to the positive eigenvalues of D_1 while $\vec{\partial}_{n+1}, \vec{\partial}_{n+2}, ..., \vec{\partial}_N$ are the eigenvectors corresponding to the zero eigenvalues, in this paper, we regard eigenvalues that are less than 0.2×10^{-10} are zero eigenvalues. According to the nature of LPP, the ability of the preserving the neighbour relationship can be measured by $W^T L_1 W / W^T D_1 W$, that means the smaller $W^T L_1 W / W^T D_1 W$ value is , the better the local structure of samples is preserved, so the Equation (4) can be rewrite as:

$$L_1 W = \lambda D_1 W \tag{5}$$

Then we design a matrix, $R = [\vec{\partial}_1, \vec{\partial}_2, ..., \vec{\partial}_n]$, using R, we respectively transform D_1, L_1, S_1 into the following matrices:

$$\begin{cases} \overline{D} = R^T D_1 R \\ \overline{L} = R^T L_1 R \\ \overline{S} = R^T S_1 R \end{cases} \tag{6}$$

We then construct the following eigen-equation:

$$\overline{L}\overline{W} = \lambda \overline{D}\overline{W} \tag{7}$$

Then we can directly solve the equation (7) since $\overline{D}$ is of full rank. Let $\vec{\beta}_1, \vec{\beta}_2, ..., \vec{\beta}_n$ denote the eigenvectors corresponding to eigenvalues $\vec{\lambda}_1, \vec{\lambda}_2, ..., \vec{\lambda}_n$ in the increasing order of Equation (7). Using the matrix R, we produce $W = X^T R$, then we transform W into Y bu carrying out $Y = WG$, where $G = [\vec{\beta}_1, \vec{\beta}_2, ..., \vec{\beta}_n]$, that is:

$$Y = WG = (X^T R)G \tag{8}$$

In this new method, the small sample size problem are solved because its directly implemented no matter whether there exists the SSS problem or not, and only the eigenvector which its eigenvalue greater than zero are used so the drawback of the conventional LPP algorithm can avoid.

3. The Improved i-vector System

3.1. Baseline i-Vector System Description

The main idea in traditional JFA is to find two subspace which represent the speaker and channel variabilities respectively. The experiment show that JFA is only partially successful in separating speaker and channel variabilities. While in the i-vector method proposed a single space that models the two variabilities and named it the total variability space [11-12]:

$$M = m + T\omega \tag{9}$$

Where M is the mean supervector which contain speaker and channel information, m is UBM supervector, T is a low rank matrix named total variability matrix, which represents a basis of the reduced total variability space and ω is a standard normal distributed vector, the components of ω are the factors and they represent the coordinates of the speaker in the reduced total variability space, these feature vectors are referred to as identity vectors or named i-vector for short.

The crucial step to the i-vector method is to compute total variability matrix T. At first, we train UBM using EM algorithm, and extract the Baum-Welch variables according to the trained UBM:

$$N_m = \sum_t \gamma_{m,t} \tag{10}$$

$$F_m = \sum_t \gamma_{m,t}(\xi_t - \mu_m) \tag{11}$$

The N_m and F_m represent zero order and first order statistic variable respectively, t is the frame numbers, m represent the m-th hybrid vectors of UBM, $\gamma_{m,t}$ is the Gaussian sharing rate, that:

$$\gamma_{m,t} = \frac{N(\xi_t;\mu_m, \sum_m)}{\sum_{i=1}^{M} N(\xi_t;\mu_i, \sum_i)} \tag{12}$$

$N(;\mu_m, \sum_m)$ is the Gaussian component which the meanis μ_m and variance is $\sum_m$, ξ_t is the random vector of the t frame, M is the mixed number of UBM. After calculate Baum-Welch variables, we can training matrix T using EM method as follows:

$$L = I + T^T \sum^{-1} NT \tag{13}$$

$$E(x) = L^{-1} T^T \sum^{-1} F \tag{14}$$

F is the vector arrangement of F_m, N、$\sum$ is the diagonal matrix of N_m、$\sum_m$ respectively.

3.2. The Proposed i-Vector System

After obtained the initial i-vector features, we complete the improved LPP algorithm proposed in this paper to the i-vector system, the specific procedure are taken as follows:

(1) Performing the dimension reduction process to the i-vector by the improved LPP method proposed in this paper.

(2) Further dimension reduction processing using LDA scheme.

(3) Taking the equivalent dimension mapping to the reduced i-vector by the WCCN transform [13]:

$$W_{WCCN} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_r}\sum_{i=1}^{n_r}(v_i^r - \overline{v_r})(v_i^r - \overline{v_r})^T \quad (15)$$

In this equation, R is represent the total numbers of speaker in the training set, $\overline{v_r}$ is the mean rth training speaker samples, v_i^r is represent the ith sample of the rth speaker, and n_r represent the training numbers of the rth speaker.

(4) Recognize the test sample using cosine distance score:

$$Score(\omega_{tar}, \omega_{test}) = \frac{\langle \omega_{tar}, \omega_{test} \rangle}{\|\omega_{tar}\| \quad \|\omega_{test}\|} \gtrless \theta \quad (16)$$

4. Experiment

4.1. Experiment Design

To evaluate our improved i-vector system, experiment were conducted on the database from CLEAR evaluation. Which consist of 200 voice segments, each voice segment is corresponding a face figure proposed in the above, and the length of each segment is 1 minute. Those 100 segments are used for training GMM parameters, and the rest used for testing. The HTK tools were used for experiments. In the fronted, speech was Hamming windowed every 10 ms with a window width of 20ms, the feature used were 13-D MFCC coefficients appended by their first and second order derivatives. The mixture components of UBM is 512, the column numbers in the total variability matrix T is 400, the new dimension after dimension reduction using improved LPP method is 350 while after LDA dimension processing is 200.

4.2. Evaluation Criterion

In order to test the performance of the new method proposed in this paper, we utilizing the Equal Error Rate(EER) and Min Detection Cost Function(MinDCF) as the evaluation criterion, the computation of MinDCF is taken as follows [14]:

$$MinDCF = \min_{\theta}\{C_{FR} \cdot F_R(\theta) \cdot P_{Tar} + C_{FA} \cdot F_A(\theta) \cdot P_{\mathrm{Im}p}\} \quad (17)$$

While C_{FR} and C_{FA} are the cost of error refuse and error accept respectively,in the NIST match, the C_{FR} is set as 10 and the C_{FA} is set as 1. P_{Tar} and $P_{\mathrm{Im}p}$ are the prior probability of genuine speaker and imposter speaker in the test set, naturally, P_{Tar} is set as 0.01and $P_{\mathrm{Im}p}$ is set as 0.99. F_R is the false refuse rate, and F_A is the false accept rate.

4.3. Experiment Result and Analysis

The simulation experiments in this paper are consist of two part:

(1) In the clean background, we compare the performance between the conventional LPP method, the improved LPP method proposed in this paper utilizing in the i-vector system and GMM method, the result are shown in the Table 1.

(2) Under different noise environments, we explore the robustness of the new LPP method utilizing in the i-vector system, the result are shown in the Table 2.

Table 1. Experiment Result Compared between Initial LPP Algorithm and Improved LPP Algorithm which used for i-vector Speaker Recognition System

Method	EER(%)	MinDCF
LPP(i-vector)	4.72	0.19
Improved LPP(i-vector)	4.45	0.17
Conventional GMM	7.32	0.53

From the results shown in the Table 1, it is clear that the recognition performance of the i-vector system is better than the initial GMM recognition system whether under EER criterion or MinDCF criterion.we can see that the EER is reducing 3% and MinDCF is reducing 0.35% compared to the initial GMM system, so the experiment result confirm the superiority of the i-vector system powerfully. While further to see the result shown in the Table 1, the performance given by the improved LPP algorithm are better than the performance given by initial LPP algorithm, the EER is reducing 0.27% and MinDCF is reducing 0.02%. This improved method can enhancing the recognition performance of the i-vector system for the reason of it can further discriminate the in-class samples and the near distance extra-class samples.

Table 2. Experiment Result Based on Improved LPP Algorithm under Different Noise Environment which used for i-vector Speaker Recognition System

Voice Enironment	SNR	EER(%)	MinDCF
Clean Background	>40dB	4.45	0.17
White Noise Environment	0dB	7.04	0.335
	5dB	6.72	0.295
	10dB	5.91	0.276
	15dB	5.36	0.242
	20dB	4.93	0.204
Babble Noise Environment	0dB	6.89	0.314
	5dB	6.49	0.282
	10dB	5.71	0.255
	15dB	5.02	0.228
	20dB	4.76	0.189

Form the experiment results shown in the Table 2, the performance given by the i-vector system based on the improved LPP scheme are better than the initial GMM method. The EER is 4.45% and the MinDCF is 0.17 under clean background and its decreasing 2.87% and 0.36 respectively compared to the initial GMM method.

The performance of the method proposed in this paper can reducing the EER and MinDCF certain degree under different signal noise rate(SNR) under white noise and babble noise environment. While the SNR is 20, the EER is 4.93% and 0.17 under white noise environment and babble environment, and its decreasing 2.29% and 2.56% respectively compared to the initial GMM method.

5. Conclusion

In this paper, a new method of enhancing the speaker recognition performance under i-vector system which its the most cutting edge recognition system in our knowledge is proposed, the new method is based on conventional LPP method and the motivation was that the conventional LPP method is always suffer from the SSS problem, and in this new scheme, We only using the eigenvectors corresponding positive eigenvalue when solving the optimized objective function and removing the zero eigenvalue.

Further work will concentrate on following two areas:

(1) Solving the small sample size (SSS) problem of the LPP method utilizing other

mathematical method forever.

(2) The computational requirements for training the i-vector systems and estimating the i-vectors, however, are too high for certain types of applications. A simply method to the original i-vector extraction and training which would dramatically decrease their complexity while retaining the recognition performance is insistent demand.

Acknowledgements

This work was supported by in part the National Science-technology Support Plan Project of China under contract 1214ZGA008, in part by the Nature Science Foundation of China under contract 61263031, in part by the Science Foundation of Gansu Province of China under contract 1010RJZA046.

References

[1] Kinnunen T, Li HZ. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*. 2010; 52: 12-40.
[2] Hamid Reza Tohidypour, Seyyed Ali Seyyedsalehi, Hossein Behbood, Hossein Roshandel. A new representation for speech frame recognition based on redundant wavelet filter banks. *Speech Communication.* 2012; 54: 256-271.
[3] Tyler K Perrachione, Stephanie N Del Tufo, John DE Gabrieli. Human Voice Recognition Depends on Language Ability. *Science*. 2011; 333: 595.
[4] Parvin Zarei Eskikanda, Seyyed Ali Seyyedsalehia. Robust speaker recognition by extracting invariant features. Procedia - Social and Behavioral Sciences. 2012; 32(3): 230-237.
[5] Shao Yang, Jin Zhaozhuang, Wang Deliang. An auditory based feature for robust speaker recognition. *ICASSP.* Taibei, Tanwan. 2009: 4625-4628.
[6] Di Wu, Jie Cao, Jinhua Wang, Wei Li. Multi-feature fusion face recognition based on Kernel Discriminate Local Preserve Projection Algorithm under smart environment. *Journal of Computers*. 2012; 7(10): 2479-2487.
[7] Jun Du, Qiang Huo. A Feature Compensation Approach Using High-Order Vector Taylor Series Approximation of an Explicit Distortion Model for Noisy speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing.* 2011; 19(8): 2285-2293.
[8] Jeong Y. *Speaker adaptation based on the multilinear decomposition of training speaker models*. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas, USA: IEEE. 2010; 4870-4873.
[9] Yongjun He, Jiqing Han. Gaussian Specific Compensation for Channel Distortion in speaker recognition*. IEEE Signal Processing Letters.* 2011; 18(10): 599-602.
[10] Omid Dehzangi, Bin Mab, Eng Siong Chng, Haizhou Li. Discriminative feature extraction for speaker recognition using continuous output codes. *Pattern Recognition Letters.* 2012; 33: 1703-1709.
[11] GU Xiao hua, GONG Wei guo, YANG Li ping. Supervised graph-optimized locality preserving projections. *Optics and Precision Engineering*. 2011; 19(3): 672-680.
[12] N Dehak, P Kenny, R Dehak, P Dumouchel, P Ouel-let. Front-end factor analysis for speaker verification. Audio, Speech, and Language Processing. *IEEE Transactions on*. 2010; 99.
[13] Tomas Pfister, Peter Robinson. Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis. *IEEE Transactions on Affective Computing*. 2011; 2(2): 66-78.
[14] C Santhosh Kumar, VP Mohandas. Robust features for multilingual acoustic modeling*. Int J Speech Technol*. 2011; 14: 147-155.