INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Speaker Verification via Estimating Total Variability Space Using Probabilistic Partial Least Squares

Chen Chen, Jiqing Han, Yilin Pan

School of Computer Science and Technology, Harbin Institute of Technology, China
chenc53@126.com, jqhan@hit.edu.cn, yilinpan.hit@gmail.com

Abstract

The i-vector framework is one of the most popular methods in speaker verification, and estimating a total variability space (TVS) is a key part in the i-vector framework. Current estimation methods pay less attention on the discrimination of TVS, but the discrimination is so important that it will influence the improvement of performance. So we focus on the discrimination of TVS to achieve a better performance. In this paper, a discriminative estimating method of TVS based on probabilistic partial least squares (PPLS) is proposed. In this method, the discrimination is improved by using the priori information (labels) of speaker, so both the correlation of intra-class and the discrimination of interclass are fully utilized. Meanwhile, it also introduces a probabilistic view of the partial least squares (PLS) method to overcome the disadvantage of high computational complexity and the inability of channel compensation. And also this proposed method can achieve a better performance than the traditional TVS estimation method as well as the PLS-based method.

Index Terms: speaker verification, i-vector, total variability space, probabilistic partial least squares

1. Introduction

Speaker verification refers to verifying speakers from their voices. One important issue in speaker verification is how to represent utterances. As a fixed-dimensional representation of an utterance, mean supervector is effective but high-dimensional [1]. Thus, the problem has been converted into the task of learning the discrimination of the supervector space and then reducing dimensionality of mean supervectors [2]. To solve these problems, the joint factor analysis (JFA) [3] provides an idea of discrimination exploration and dimensionality reduction as well as channel compensation.

As an improvement of JFA, the combination of i-vector [4] and probabilistic linear discriminant analysis (PLDA) [5] has become a typical baseline system in speaker verification [6–8]. In this system, the mean supervectors are mapped onto low-dimensional space named total variability space (TVS) by using the method of factor analysis (FA). To estimate the TVS, a number of approaches have been proposed [9–11]. However, all of these works only explore the internal structure of mean supervectors, but ignore external constraints (such as speaker category). Particularly, the external constraints can force different categories of data to disperse in the TVS to improve the discrimination.

A new breakthrough point of improving the discrimination was proposed in work [12] by using the partial least squares (PLS) [13–15]. It introduces category labels to train the speaker models. However, the label of this work is a two-classified label, to make a discrimination between any two different speakers, it has to train a model for each speaker. Once a new target speaker is enrolled, it should train a speaker model for this new speaker again. Furthermore, it is mostly used in situations where only a small number of leading eigenvectors are required, but that cannot avoid having to evaluate the supervector covariance matrix as an intermediate step. Thus, the computational complexity of the PLS-based method is high when the speaker number is large and the dimensionality is high. And also this work is not based on the i-vector/ PLDA framework, thus it is unable to process the channel variability, although it can directly provide the matching score for speaker verification.

Considering the disadvantages of the current PLS-based method, and no effective use of category information of the training data for modeling the TVS, we propose a new TVS estimating method by using the probabilistic partial least squares (PPLS) [16–18]. In this method, the category information (speaker labels) is effectively used for modeling the TVS. Different from using the approach of the label for PLS-based work, the PPLS-based method introduces multi-classified labels to estimate the TVS, thus it can solve the multi-classified problem only by using the estimated TVS. Furthermore, we derive an algorithm based on PPLS to avoid evaluating the supervector covariance matrix. As a consequence, the PPLS-based TVS contains more speaker variability, and it saves more space and time than the method of training a model for each speaker. And also the PPLS-based method allows the use of channel compensation technology.

2. Related work

2.1. TVS estimated based on factor analysis

The method of FA estimates a TVS containing both speaker and channel variability. Given an utterance, the mean supervector $\boldsymbol{M}$ is rewritten as follows [19]:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{1}$$

where $\boldsymbol{m}$ is the UBM mean supervector, $\boldsymbol{T}$ is the low rank total variability matrix whose columns are vectors spanning the TVS, and $\boldsymbol{w}$ is the i-vector. Assume $c(c = 1, \ldots, C)$ to be the index of Gaussian components, the zero-order Baum-Welch statistics to be N_c, and the centralized first-order statistics for a given utterance u to be $\boldsymbol{F}_c(u)$. The posterior distribution of $\boldsymbol{w}(u)$ is Gaussian with covariance matrix $\mathrm{cov}[\boldsymbol{w}(u), \boldsymbol{w}(u)]$ and mean $E[\boldsymbol{w}(u)]$:

$$\begin{cases} \mathrm{cov}[\boldsymbol{w}(u), \boldsymbol{w}(u)] = \boldsymbol{L} = (\boldsymbol{I} + \boldsymbol{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{N}(u)\boldsymbol{T})^{-1} \\ E[\boldsymbol{w}(u)] = \boldsymbol{L}^{-1}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F}(u) \end{cases} \tag{2}$$

where $\boldsymbol{N}(u)$ is a diagonal matrix of dimension $CF \times CF$ whose diagonal blocks are $N_c\boldsymbol{I}$ (F is the dimension of feature vectors). $\boldsymbol{F}(u)$ is a supervector of dimension $CF \times 1$ obtained by concatenating all first-order Baum-Welch statistics $\boldsymbol{F}_c(u)$ for the

http://dx.doi.org/10.21437/Interspeech.2017-633

given utterance u. $\boldsymbol{\Sigma}$ is a diagonal covariance matrix of dimension $CF \times CF$ whose diagonal blocks are $\boldsymbol{\Sigma}_c\boldsymbol{I}$. The i-vector $\boldsymbol{w}$ for a given utterance can be obtained as $E[\boldsymbol{w}(u)]$.

2.2. PLS framework for speaker verification

The PLS method aims at modeling the relationship between the mean supervector $\boldsymbol{M}$ and the corresponding speaker label y (1 for speaker and -1 for impostor) using projection into latent spaces. Given the variable pairs $\{\boldsymbol{M}_n, y_n\}, n = 1, ..., N$, the set of mean supervectors $\widetilde{\boldsymbol{M}} = (\boldsymbol{M}_1^{\mathrm{T}}, \ldots, \boldsymbol{M}_n^{\mathrm{T}}, \ldots, \boldsymbol{M}_N^{\mathrm{T}})^{\mathrm{T}}$ is assumed to be the matrix made up of N mean supervectors, and $\widetilde{\boldsymbol{Y}} = (y_1, \ldots, y_n, \ldots, y_N)^{\mathrm{T}}$ to be the corresponding label vector. The PLS decomposes $\widetilde{\boldsymbol{M}}$ and $\widetilde{\boldsymbol{Y}}$ as:

$$\begin{cases} \widetilde{\boldsymbol{M}} = \boldsymbol{W}\boldsymbol{T}^{\mathrm{T}} + \boldsymbol{E} \\ \widetilde{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{Q}^{\mathrm{T}} + \boldsymbol{F} \end{cases} \tag{3}$$

where $\boldsymbol{T}$ and $\boldsymbol{Q}$ are loading vectors, $\boldsymbol{W}$ and $\boldsymbol{U}$ are latent vectors, $\boldsymbol{E}$ and $\boldsymbol{F}$ are residual matrices. The detailed analysis of PLS framework for speaker verification can be found in [12]. This optimization problem can be converted to compute the eigenvalue of $\widetilde{\boldsymbol{M}}^{\mathrm{T}}\widetilde{\boldsymbol{Y}}\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\widetilde{\boldsymbol{M}}$. As a result, the PLS-based method trains a regression coefficient $\boldsymbol{B}$ which transforms supervector $\boldsymbol{M}$ to the predicted label $\widehat{y}$ for each target speaker:

$$\widehat{y} = \boldsymbol{M}\boldsymbol{B}. \tag{4}$$

Then the predicted label $\widehat{y}$ is directly used as the matching score for speaker verification.

3. TVS estimated based on probabilistic partial least squares

It can be seen from the analysis of the above methods that the FA-based TVS is lack of discrimination, and the PLS-based method uses label to improve the discrimination, but still needs some improvements: (1) Since y is a two-classified label, the PLS-based method has to train a model for each speaker. It is a waste of time and space. (2) The complexity of PLS does not perform well for large sample sizes and large number of features [12]. (3) The predicted label $\widehat{y}$ is directly used as the matching score for speaker verification. That is to say, this work is unable to process the channel variability.

In order to overcome these problems, we propose a PPLS-based method to estimate the TVS: (1) The label $\boldsymbol{Y}$ is assigned to be a multi-class label vector to introduce all the speaker variability. (2) A probabilistic view of PLS via EM algorithm is derived to estimate the TVS. (3) The PPLS-based method extracts the i-vector through the discriminative TVS and allows the use of channel compensation technology channel compensation technology to remove the channel variability.

3.1. Posterior calculation for PPLS model

The mean supervector and category label are represented as the variable $\boldsymbol{M}$ and $\boldsymbol{Y}$. Suppose that there are K speaker classes, different with the label of work [12], $\boldsymbol{Y}$ is a K dimensional binary vector with only one non-zero element with 1, e.g. $\boldsymbol{Y}_1 = (1, 0, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^K$ expresses the category label of mean supervector $\boldsymbol{M}_1$. Since i-vector $\boldsymbol{w}$ is a low dimensional representation of supervector $\boldsymbol{M}$, in order to make the i-vector more discriminative, we introduce an external constraint to create a mapping relationship from label vector $\boldsymbol{Y}$ to i-vector $\boldsymbol{w}$. Therefore, $\boldsymbol{M}$ and $\boldsymbol{Y}$ are related by $\boldsymbol{w}$ as:

$$\begin{cases} \boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} + \boldsymbol{\varepsilon} \\ \boldsymbol{Y} = \boldsymbol{\mu_Y} + \boldsymbol{Q}\boldsymbol{w} + \boldsymbol{\zeta} \end{cases} \tag{5}$$

where $\boldsymbol{m}$ is the UBM mean supervector; $\boldsymbol{T}$ is the total variability matrix, and the columns of $\boldsymbol{T}$ span a linear subspace within the data space (supervector space) which corresponds to the principal subspace; $\boldsymbol{w} \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{I})$ is the i-vector; $\boldsymbol{\varepsilon} \sim \mathbb{N}(\boldsymbol{0}, \sigma_{\boldsymbol{M}}^2\boldsymbol{I})$ is the residual variability not captured by the total variability matrix $\boldsymbol{T}$; $\boldsymbol{\mu_Y}$ is the mean of all $\boldsymbol{Y}$; $\boldsymbol{Q}$ is the conversion matrix; $\boldsymbol{\zeta} \sim \mathbb{N}(\boldsymbol{0}, \sigma_{\boldsymbol{Y}}^2\boldsymbol{I})$ is the residual variability. The conditional distribution of $\boldsymbol{M}$ and $\boldsymbol{Y}$, conditioned on the value of the latent variable $\boldsymbol{w}$, are Gaussian of the form:

$$\begin{cases} p(\boldsymbol{M}|\boldsymbol{w}) = \mathbb{N}(\boldsymbol{M}|\boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}, \sigma_{\boldsymbol{M}}^2\boldsymbol{I}) \\ p(\boldsymbol{Y}|\boldsymbol{w}) = \mathbb{N}(\boldsymbol{Y}|\boldsymbol{\mu_Y} + \boldsymbol{Q}\boldsymbol{w}, \sigma_{\boldsymbol{Y}}^2\boldsymbol{I}). \end{cases} \tag{6}$$

Thus the joint distribution of $\boldsymbol{M}$ and $\boldsymbol{Y}$ has the Gaussian distribution:

$$p(\boldsymbol{MY}|\boldsymbol{w}; \boldsymbol{\Theta}) = \mathbb{N}(\boldsymbol{\mu}_{\boldsymbol{MY}|\boldsymbol{w}}, \boldsymbol{\Sigma}_{\boldsymbol{MY}|\boldsymbol{w}}) \tag{7}$$

where

$$\begin{cases} \boldsymbol{\Sigma}_{\boldsymbol{MY}|\boldsymbol{w}} = \begin{pmatrix} \sigma_{\boldsymbol{M}}^2\boldsymbol{I} & 0 \\ 0 & \sigma_{\boldsymbol{Y}}^2\boldsymbol{I} \end{pmatrix} \\ \boldsymbol{\mu}_{\boldsymbol{MY}|\boldsymbol{w}} = \begin{pmatrix} \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \\ \boldsymbol{\mu_Y} + \boldsymbol{Q}\boldsymbol{w} \end{pmatrix} \\ \boldsymbol{\Theta} = \{\boldsymbol{T}, \boldsymbol{Q}, \boldsymbol{\mu_{MY}}, \sigma_{\boldsymbol{M}}^2, \sigma_{\boldsymbol{Y}}^2\}. \end{cases} \tag{8}$$

Suppose that $\boldsymbol{\Lambda} = (\boldsymbol{T}^{\mathrm{T}}, \boldsymbol{Q}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{Z} = (\boldsymbol{M}^{\mathrm{T}}, \boldsymbol{Y}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\mu_{MY}} = (\boldsymbol{m}^{\mathrm{T}}, \boldsymbol{\mu_Y}^{\mathrm{T}})^{\mathrm{T}}$. Similar to Eq.(2), the posterior conditioned distribution of $\boldsymbol{w}$ can be written as:

$$p(\boldsymbol{w}|\boldsymbol{MY}; \boldsymbol{\Theta}) = \mathbb{N}(\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{MY}}, \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{MY}}) \tag{9}$$

where

$$\begin{cases} \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{MY}} = (\boldsymbol{I} + \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{MY}|\boldsymbol{w}}^{-1}\boldsymbol{\Lambda})^{-1} \\ \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{MY}} = \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{MY}}\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{MY}|\boldsymbol{w}}^{-1}(\boldsymbol{Z} - \boldsymbol{\mu_{MY}}). \end{cases} \tag{10}$$

3.2. Likelihood calculation for PPLS model

As mentioned above, the PPLS model can be expressed in terms of a marginalization over a continuous latent space $\boldsymbol{w}$ for each $\boldsymbol{Z}$ ($\boldsymbol{M}$ and $\boldsymbol{Y}$). Meanwhile, given training pairs of mean supervector and category label vector $\{\boldsymbol{M}_n, \boldsymbol{Y}_n; n = 1, \ldots, N\}$, the logarithm likelihood of the parameters can be written as:

$$\ln p(\boldsymbol{MY}; \boldsymbol{\Theta}) = \sum_{n=1}^{N} \ln p(\boldsymbol{M}_n, \boldsymbol{Y}_n; \boldsymbol{\Theta}). \tag{11}$$

Therefore the EM algorithm can be used to get the maximum likelihood estimates of the model parameters $\boldsymbol{\Theta}$. The expectation of the complete-data with respect to the posterior distribution of the latent distribution evaluated using 'old' parameter values. maximization of this expected completed-data logarithm likelihood then yields the 'new' parameter values. Using Eq.(9) and Eq.(10), $Q(\boldsymbol{w}_n) = p(\boldsymbol{w}_n|\boldsymbol{M}_n, \boldsymbol{Y}_n)$ can be denoted as an auxiliary function. And the logarithm likelihood function can be written as:

$$\begin{aligned}
&\sum_{n=1}^{N} \ln p(\boldsymbol{M}_n, \boldsymbol{Y}_n; \boldsymbol{\Theta}) \\
&= \sum_{n=1}^{N} \ln \int p(\boldsymbol{M}_n, \boldsymbol{Y}_n, \boldsymbol{w}_n; \boldsymbol{\Theta}) d\boldsymbol{w}_n \qquad (12) \\
&= \sum_{n=1}^{N} \ln \int Q(\boldsymbol{w}_n) \frac{p(\boldsymbol{M}_n, \boldsymbol{Y}_n, \boldsymbol{w}_n; \boldsymbol{\Theta})}{Q(\boldsymbol{w}_n)} d\boldsymbol{w}_n.
\end{aligned}$$

Since logarithm fuction $f(x) = \ln(x)$ is a concave function, based on the Jensen's inequality rule, $E[f(x)] \leq f(E[x])$ can be obtained. So the Eq.(12) can be transformed into:

$$\begin{aligned}
&\sum_{n=1}^{N} \ln \int Q(\boldsymbol{w}_n) \frac{p(\boldsymbol{M}_n, \boldsymbol{Y}_n, \boldsymbol{w}_n; \boldsymbol{\Theta})}{Q(\boldsymbol{w}_n)} d\boldsymbol{w}_n \\
&\leq \int Q(\boldsymbol{w}_n) \ln \frac{p(\boldsymbol{M}_n, \boldsymbol{Y}_n, \boldsymbol{w}_n; \boldsymbol{\Theta})}{Q(\boldsymbol{w}_n)} d\boldsymbol{w}_n \\
&= \sum_{n=1}^{N} E_{\boldsymbol{w}_n \sim Q}[\ln p(\boldsymbol{M}_n, \boldsymbol{Y}_n | \boldsymbol{w}_n; \boldsymbol{\Theta}) + \ln p(\boldsymbol{w}_n) - \ln Q(\boldsymbol{w}_n)]
\end{aligned} \tag{13}$$

where $E_{\boldsymbol{w}_n \sim Q}$ indicates that the expectations with respect to $\boldsymbol{w}_n$ are drawn from distribution $Q(\boldsymbol{w}_n) = p(\boldsymbol{w}_n | \boldsymbol{M}_n, \boldsymbol{Y}_n; \Theta)$. It has been proved that EM always monotonically improves the logarithm likelihood by maximizing the lower bound corresponding to parameters $\boldsymbol{\Theta}$, which equals to:

$$\max \sum_{n=1}^{N} E_{\boldsymbol{w}_n \sim Q}[\ln p(\boldsymbol{M}_n, \boldsymbol{Y}_n | \boldsymbol{w}_n; \boldsymbol{\Theta})]. \tag{14}$$

Similar to the probabilistic principal component analysis (PPCA) [20, 21], taking the expectation with respect to the posterior distribution over the latent variables, Eq.(14) can be obtained as:

$$\begin{aligned}
&\sum_{n=1}^{N} E_{\boldsymbol{w}_n \sim Q}[\ln p(\boldsymbol{M}_n, \boldsymbol{Y}_n | \boldsymbol{w}_n; \boldsymbol{\Theta})] = \\
&- \sum_{n=1}^{N} \{ \frac{D_{\boldsymbol{M}}}{2} \ln(2\pi\sigma_{\boldsymbol{M}}^2) - \frac{1}{\sigma_{\boldsymbol{M}}^2} E[\boldsymbol{w}_n]^{\mathrm{T}} \boldsymbol{T}^{\mathrm{T}} (\boldsymbol{M}_n - \boldsymbol{m}) \\
&+ \frac{1}{2\sigma_{\boldsymbol{M}}^2} \|\boldsymbol{M}_n - \boldsymbol{m}\|^2 + \frac{1}{2\sigma_{\boldsymbol{M}}^2} \mathrm{Tr}(E[\boldsymbol{w}_n \boldsymbol{w}_n^{\mathrm{T}}] \boldsymbol{T}^{\mathrm{T}} \boldsymbol{T}) \\
&+ \frac{D_{\boldsymbol{Y}}}{2} \ln(2\pi\sigma_{\boldsymbol{Y}}^2) - \frac{1}{\sigma_{\boldsymbol{Y}}^2} E[\boldsymbol{w}_n]^{\mathrm{T}} \boldsymbol{Q}^{\mathrm{T}} (\boldsymbol{Y}_n - \mu_Y) \\
&+ \frac{1}{2\sigma_{\boldsymbol{Y}}^2} \|\boldsymbol{Y}_n - \boldsymbol{m}\|^2 + \frac{1}{2\sigma_{\boldsymbol{Y}}^2} \mathrm{Tr}(E[\boldsymbol{w}_n \boldsymbol{w}_n^{\mathrm{T}}] \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{Q}) \}.
\end{aligned} \tag{15}$$

Eq.(15) is taken derivative to all parameters, and the parameters can be computed as: Eq.(16) and Eq.(17):

$$\begin{cases}
\boldsymbol{T} = [\sum_{n=1}^{N} (\boldsymbol{M}_n - \boldsymbol{m}) E[\boldsymbol{w}_n]^{\mathrm{T}}][\sum_{n=1}^{N} E[\boldsymbol{w}_n \boldsymbol{w}_n^{\mathrm{T}}]]^{-1} \\
\boldsymbol{Q} = [\sum_{n=1}^{N} (\boldsymbol{Y}_n - \boldsymbol{\mu}_{\boldsymbol{Y}}) E[\boldsymbol{w}_n]^{\mathrm{T}}][\sum_{n=1}^{N} E[\boldsymbol{w}_n \boldsymbol{w}_n^{\mathrm{T}}]]^{-1} \\
\sigma_{\boldsymbol{M}}^2 = \frac{1}{D_{\boldsymbol{M}} N} \sum_{n=1}^{N} \{ \mathrm{Tr}(E[\boldsymbol{w}_n \boldsymbol{w}_n^{\mathrm{T}}] \boldsymbol{T}^{\mathrm{T}} \boldsymbol{T}) \\
\qquad - 2E[\boldsymbol{w}_n]^{\mathrm{T}} \boldsymbol{T} (\boldsymbol{M}_n - \boldsymbol{m}) + \|\boldsymbol{M}_n - \boldsymbol{m}\|^2 \}
\end{cases} \tag{16}$$

and

$$\begin{aligned}
\sigma_{\boldsymbol{Y}}^2 = \frac{1}{D_{\boldsymbol{Y}} N} \sum_{n=1}^{N} \{ &\mathrm{Tr}(E[\boldsymbol{w}_n \boldsymbol{w}_n^{\mathrm{T}}] \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{Q}) \\
&- 2E[\boldsymbol{w}_n]^{\mathrm{T}} \boldsymbol{Q} (\boldsymbol{Y}_n - \boldsymbol{\mu}_{\boldsymbol{Y}}) + \|\boldsymbol{Y}_n - \boldsymbol{\mu}_{\boldsymbol{Y}}\|^2 \}
\end{aligned} \tag{17}$$

where

$$\begin{cases}
E[\boldsymbol{w}_n] = \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{MY}}(n) \\
E[\boldsymbol{w}_n \boldsymbol{w}_n^{\mathrm{T}}] = \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{MY}}(n) + E[\boldsymbol{w}_n] E[\boldsymbol{w}_n]^{\mathrm{T}}
\end{cases} \tag{18}$$

where $\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{MY}}(n)$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{MY}}(n)$ are computed in Eq.(10). And the parameters need to be renewed iteratively until converged. The algorithm for TVS modeling based on PPLS is shown follows:

```
Algorithm 1: Algorithm for TVS modeling based on PPLS
Input:
  $\boldsymbol{M}$: speaker mean supervector; $\boldsymbol{Y}$: category label;
  $\boldsymbol{m}$: speaker UBM mean supervector; $R$: size of i-vector;
Initialize parameters:
  Random initialize $\boldsymbol{T}$ and $\boldsymbol{Q}$;
  $\boldsymbol{\mu}_{\boldsymbol{Y}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{Y}_n$;
  $\sigma_{\boldsymbol{M}}^2 = 1; \sigma_{\boldsymbol{Y}}^2 = 1$;
Parameter Estimation:
  1: Compute the posterior expectation $E[\boldsymbol{w}_n]$ and posterior
     variance $E[\boldsymbol{w}_n \boldsymbol{w}_n^{\mathrm{T}}]$ using Eq.(18);
  2: Compute total variability matrix $\boldsymbol{T}$, mapping matrix $\boldsymbol{Q}$,
     variance of the conditional distribution $\sigma_{\boldsymbol{M}}^2$ and $\sigma_{\boldsymbol{Y}}^2$ using
     Eq.(16) and Eq.(17);
  3: Go to step 1 until convergence;
Return:
  The parameters of the PPLS model: $\boldsymbol{\Theta} = \{\boldsymbol{T}, \boldsymbol{Q}, \sigma_{\boldsymbol{M}}^2, \sigma_{\boldsymbol{Y}}^2\}$.
```

3.3. I-vector extraction

After estimating the parameters of the PPLS model, the relation between the mean supervector $\boldsymbol{M}$ and the label $\boldsymbol{Y}$ can be determined by the marginal distribution along $\boldsymbol{w}$:

$$p(\boldsymbol{Y}|\boldsymbol{M}) = \int p(\boldsymbol{Y}|\boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{M}) d\boldsymbol{w}. \tag{19}$$

And the posterior distribution of $\boldsymbol{Y}$ can be determined as $\mathbb{N}(\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{M}}, \boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{M}})$ with:

$$\begin{cases}
\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{M}} = \boldsymbol{Q}\boldsymbol{C}^{-1} \boldsymbol{T}^{\mathrm{T}} (\boldsymbol{M} - \boldsymbol{m}) + \boldsymbol{\mu}_{\boldsymbol{Y}} & \text{(20a)} \\
\boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{M}} = \boldsymbol{Q} \sigma_{\boldsymbol{M}}^2 \boldsymbol{Q}^{\mathrm{T}} + \sigma_{\boldsymbol{Y}}^2 \boldsymbol{I} & \text{(20b)} \\
\boldsymbol{C} = \boldsymbol{T}^{\mathrm{T}} \boldsymbol{T} + \sigma_{\boldsymbol{M}}^2 \boldsymbol{I}. & \text{(20c)}
\end{cases}$$

Therefore, the predicted label $\widehat{\boldsymbol{Y}}$ will be estimate as the mean value of its conditional distribution on $\boldsymbol{M}$ using Eq.(20a).

Since the i-vector is calculated by evaluating the posterior expectation of the latent variables $\boldsymbol{w}$, the i-vector $\boldsymbol{w}(u)$ for a given utterance u can be obtained by using the following equation:

$$E[\boldsymbol{w}(u)] = (\boldsymbol{I} + \boldsymbol{\Lambda}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{MY}|\boldsymbol{w}}^{-1} \boldsymbol{\Lambda})^{-1} \boldsymbol{\Lambda}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{MY}|\boldsymbol{w}}^{-1} (\widehat{\boldsymbol{Z}}(u) - \boldsymbol{\mu}_{\boldsymbol{MY}}) \tag{21}$$

where $\widehat{\boldsymbol{Z}}(u) = (\boldsymbol{M}(u)^{\mathrm{T}}, \widehat{\boldsymbol{Y}}(u)^{\mathrm{T}})^{\mathrm{T}}$, and $\widehat{\boldsymbol{Y}}(u)$ is the predicted label for $\boldsymbol{M}(u)$. After the PPLS-based i-vector extraction, the classification methods steps are the same as the traditional i-vector modeling.

4. Experiments and discussion

4.1. Database

The King-ASR-009 database was used for experiments. It is a Chinese mandarin speech recognition database which contains the utterances spoken by 200 different native speakers (87 males, 113 females). Each speaker read 120 short message sentences which are specially designed for both training and testing. We carried out experiments by creating our own list. The list contains 320000 trials, in which 1600 trials are target trials, and 318400 trials are non-target trials. And the Equal error rate (EER) and the Minimum detection cost function (Min DCF) are used as metrics.

4.2. Experimental setup

The experiments operated on cepstral features which was extracted using a 32-ms Hamming window. Every 10 ms, 13 Mel frequency cepstral coefficients (MFCCs) were calculated. Delta and delta-delta coefficients were then calculated to produce 39-dimensional feature vectors. We used a gender-independent UBM containing 1024 Gaussians. This UBM was trained with 80% of recordings from the King-ASR-009 database which were not used for evaluation. The recordings mentioned above also were used for estimating the TVS. And the dimension of the TVS was set from 200 to 600. The Linear discriminant analysis (LDA) was used for channel compensation, and the PLDA was used for verification.

4.3. Results and discussions

The performance is compared between different dimensions (from 200 to 600) of FA- and PPLS-based i-vector. The results are shown in Figure 1.

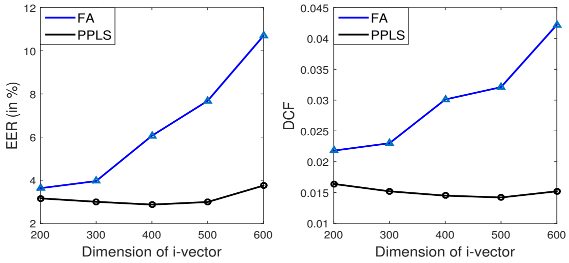

Figure 1: *Relationship between dimension of i-vector and performance*

As shown in Figure 1, it can be seen that: (1) The curve of PPLS with different dimensions is below the curve of FA in an overall view. Obviously, the PPLS-based i-vector has lower dimension and better performance. And the results show that introducing category information into TVS estimation has the effect of increasing the correlation of intra-class and the discrimination of inter-class, thus the PPLS-based i-vector is more discriminative. (2) When the dimensions of FA- and PPLS-based i-vectors are 200 and 400 respectively, the speaker verification system achieves the best performance.

According to the best results of FA- and PPLS-based methods, the dimension of FA- and PPLS-based i-vectors are set to be 200 and 400 separately, the dimension of LDA is chosen from 300 to 50 (the dimension of FA-based method is from 200 to 30), and the best performance was obtained with the dimension of 50 (FA-based) and 150 (PPLS-based). The performance of PPLS-based method is also compared with the PLS-based method [12]. The results of the experiments using the King-ASR-009 database are shown in Table 1 and the DET curves for the experiments are shown in Figure 2.

Table 1: *Performance of PPLS-based i-vector model against PLS-based method and FA-based i-vector model*

method	EER(%)	DCF
PLS	4.75	0.018
FA	3.63	0.022
FA+LDA(dim=50)	3.13	0.018
PPLS	2.87	0.014
PPLS+LDA(dim=150)	**2.31**	**0.012**

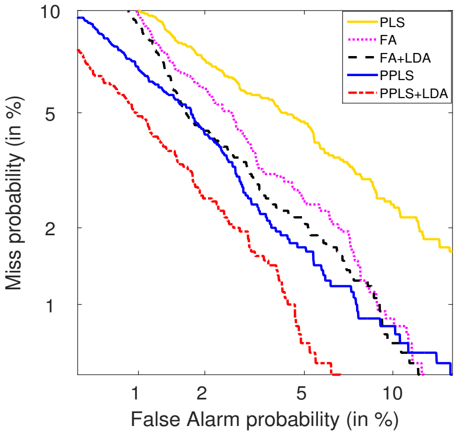

Figure 2: *DET curves of PPLS-based i-vector model against PLS-based method and FA-based i-vector model*

From the results, it can be seen that: (1) PPLS with LDA gives the best performance (EER is 2.31% and DCF is 0.012). These results prove that the application of PPLS gives a better performance than FA applied to the TVS. The EER is decreased from 3.13% (FA+LDA) to 2.31% (PPLS+LDA). The relative EER is decreased by 26.2%. (2) Without LDA, PPLS gives a better performance than PLS-based method. It can be seen that, the probabilistic view of PLS can achieve a better performance than the simple linear transformation of PLS. (3) The channel compensation technique can be used to remove the nuisance direction from the PPLS-based i-vector. And PPLS with LDA can also give a better performance than the PLS-based method.

5. Conclusions

In this paper, a new PPLS-based TVS estimation method is proposed. This method gives a probabilistic view of PLS via using both speaker variability and the relationship between speaker features and their category labels to make the TVS more discriminative. The experiment has proved that the proposed PPLS-based method is able to achieve better performances than the FA- and PLS-based methods.

6. Acknowledgements

This research is partly supported by the National Natural Science Foundation of China under grant No.61471145 and 91120303.

7. References

[1] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

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

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *Digital Signal Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proceedings of IEEE International Conference on Computer Vision 2007*, 2007, pp. 1–8.

[6] D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, C. S. Greenberg, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, “Summary and initial results of the 2013-2014 speaker recognition i-vector machine learning challenge,” in *INTERSPEECH 2014*, 2014, pp. 368–372.

[7] J. H. Hansen and T. Hasan, “Speaker recognition by machines and humans: a tutorial review,” *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[8] M. A. Nematollahi and S. A. R. Al-Haddad, “Distant speaker recognition: An overview,” *International Journal of Humanoid Robotics*, vol. 13, no. 2, pp. 1–45, 2016.

[9] Z. Lei and Y. Yang, “Maximum likelihood i-vector space using PCA for speaker verification,” in *INTERSPEECH 2011*, 2011, pp. 2725–2728.

[10] V. Hautamäki, Y. Cheng, P. Rajan, and C. Lee, “Minimax i-vector extractor for short duration speaker verification,” in *INTERSPEECH 2013*, 2013, pp. 3708–3712.

[11] L. Chen, K. Lee, B. Ma, W. Guo, H. Li, and L. Dai, “Local variability modeling for text-independent speaker verification,” in *Proceedings of Odyssey 2014: Speaker and Language Recognition Workshop*, 2014, pp. 54–59.

[12] B. V. Srinivasan, D. N. Zotkin, and R. Duraiswami, “A partial least squares framework for speaker recognition,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5276–5279.

[13] P. Geladi and B. R. Kowalski, “Partial least-squares regression: a tutorial,” *Analytica Chimica Acta*, vol. 185, no. 86, pp. 1–17, 1986.

[14] Q. Zhao, L. Zhang, and A. Cichocki, “Multilinear and nonlinear generalizations of partial least squares: an overview of recent advances,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 2, pp. 104–115, 2014.

[15] T. Mehmood and B. Ahmed, “The diversity in the applications of partial least squares: an overview,” *Journal of Chemometrics*, vol. 30, no. 1, pp. 4–17, 2015.

[16] S. Li, J. Gao, and J. O. Nyagilo, “Probabilistic partial least square regression: A robust model for quantitative analysis of Raman spectroscopy data,” in *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, 2011, pp. 526–531.

[17] S. Li, J. Gao, J. O. Nyagilo, D. P. Dave, B. Zhang, and X. Wu, “A unified probabilistic PLSR model for quantitative analysis of surface-enhanced Raman spectrum (SERS),” in *Proceeding of the Second International Conference on Communications, Signal Processing, and Systems*, 2014, pp. 1095–1103.

[18] S. Li, J. O. Nyagilo, and D. P. Dave, “Probabilistic partial least squares regression for quantitative analysis of Raman spectra,” *International journal of data mining and bioinformatics*, vol. 11, no. 2, pp. 223 – 243, 2015.

[19] P. Kenny and G. Boulianne, “Eigenvoice modeling with sparse training data,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[20] M. E. Tipping and C. M. Bishop, “Probabilistic principal component analysis,” *Journal of the Royal Statistical Society*, vol. 61, no. 3, pp. 611–622, 1999.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.