PLDA Modeling in I-Vector and Supervector Space for Speaker Verification

Ye Jiang, Kong Aik Lee, Zhenmin Tang, Bin Ma, Anthony Larcher, Haizhou Li

PLDA Modeling in I-Vector and Supervector Space for Speaker Verification

Ye Jiang[1, 2], *Kong Aik Lee*[2], *Zhenmin Tang*[1], *Bin Ma*[2], *Anthony Larcher*[2], *and Haizhou Li*[2,3]

[1]School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China
[2]Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore
[3]School of EE&T, University of New South Wales, Australia
208060141@njust.edu.cn, kalee@i2r.a-star.edu.sg

Abstract

In this paper, we advocate the use of uncompressed form of i-vector. We employ the *probabilistic linear discriminant analysis* (PLDA) to handle speaker and session variability for speaker verification task. An i-vector is a low-dimensional vector containing both speaker and channel information acquired from a speech segment. When PLDA is used on i-vector, dimension reduction is performed twice – first in the i-vector extraction process and second in the PLDA model. Keeping the full dimensionality of i-vector in the supervector space for PLDA modeling and scoring would avoid unnecessary loss of information. The drawback of using PLDA on uncompressed i-vector is the inversion of large matrices, which we show can be solved rather efficiently by portioning large matrix into smaller blocks. We also introduce the Gaussianized rank-norm, as an alternative to whitening, for feature normalization prior to PLDA modeling.

Index Terms: speaker verification, i-vector, probabilistic LDA

1. Introduction

Over recent years, many approaches based on the use of Gaussian mixture models (GMM) in a GMM-UBM framework [1] have been proposed to improve the performance of speaker verification system. Inspired by the *joint factor analysis* (JFA) approach [2, 3], it was shown in [4] that the problem of session variability can be coped with by confining the variability within a low-dimensional subspace, referred to as the *total variability space*, in the parameter space of GMM.

The idea of defining a total variability space is to extract low dimensional *identity vectors* (the so-called i-vectors), by which speech segments of *variable-length* can be represented as *fixed-length* vectors [4]. Such a representation greatly simplifies the modeling and scoring processes in speaker verification. For instance, we could assume that the i-vectors are generated from a Gaussian density instead of the mixture of Gaussian densities usually considered for the case of acoustic features [1]. In this regard, linear discriminant analysis (LDA) [4], probabilistic LDA (PLDA) [5], and the heavy-tailed PLDA [6] have shown to be effective for such fixed-length data. In this paper, we shall focus on PLDA with Gaussian prior instead of heavy-tailed prior as it was recently shown in [7] that the advantage of the heavy-tailed assumption diminishes with a simple length-normalization on the i-vector preceding PLDA modeling.

Since the total variability space is always spanned by a low-rank rectangular matrix, a dimension reduction process is also imposed by the i-vector extractor. In this paper, we advocate the use of uncompressed form of i-vector. Similar to that in [4], our extractor converts speech sequence into fixed-length vector, but retains its dimensionality in the full supervector space. Modeling of speaker and session variability is then carried out with PLDA, which has shown to be effective in handling high-dimensional data [5]. By doing so, we avoid reducing the dimensionality of the i-vector twice – first in the extraction process and second in the PLDA model. Any dimension reduction procedure will unavoidably discard information. Our intention is therefore to keep the full dimensionality till the scoring stage with PLDA and to investigate the performance of PLDA in the supervector space.

The downside of using uncompressed form of i-vector (we call this *i-supervector* hereafter to avoid confusion) with PLDA is that we have to deal with large matrices. The size of the matrices becomes enormous when more sessions are available for each speaker in the development data[1]. One option is to estimate the subspaces in a decoupled manner, which might lead to suboptimal solution [2, 3]. In this paper, we show how the subspaces can be jointly estimated by partitioning large matrices into sub-matrices, thereby making the matrix inversion and the joint estimation feasible. A significant advantage of PLDA approach for speaker verification is the use of Bayes factor [8] in computing the verification score. In this regard, we show how to manipulate large matrices efficiently in computing the PLDA score. In addition, we also look into various normalization methods and introduce the use of Gaussianized rank-norm for PLDA.

The paper is organized as follows. In Section 2, we look at the scenario where dimension reduction is performed twice when PLDA is used on i-vector. Section 3 shows that inversion of large matrices encountered in PLDA can be solved by exploiting some inherent structure of the matrices. Section 4 deals with PLDA scoring and introduces the Gaussianized rank norm. We present some experimental results in Section 5 and conclude the paper in Section 6.

2. PLDA for i-vector and i-supervector

2.1. From i-vector to i-supervector

An i-vector represents a variable-length speech utterance as a low-dimensional vector (low as compared to the dimensionality of the mean supervector). The generative equation is given by

$$\mathbf{m} = \mathcal{M} + \mathbf{T}\mathbf{x}, \tag{1}$$

where $\mathbf{m}$ and $\mathcal{M}$ are the mean supervectors of the speaker (and session) dependent GMM and the UBM, respectively. The subspace spanned by the columns of the matrix $\mathbf{T}$ captures the speaker and session variability (hence the name *total variability*).

[1] The number of sessions is usually limited in face recognition for which PLDA was originally proposed in [5].

The latent variable $\mathbf{x}$ is taken to be a low-dimensional random vector with a standard normal distribution. For each observation sequence $\mathcal{X}$, the i-vector is given by the posterior mean ϕ of the latent variable $\mathbf{x}$, i.e., $E\{\mathbf{x}|\mathcal{X}\}=\phi$. Since $\mathbf{T}$ is always a low-rank rectangular matrix, the dimensionality D of the i-vector is much smaller compared to that of the supervector, i.e., $D \ll C\times F$. Here, F is the dimensionality of the acoustic feature and C is the number of mixture in the GMM.

Consider the case where we allow the latent variable to grow into the full supervector space, in which $D=C\times F$. The generative equation is now given by

$$\mathbf{m}=\mathcal{M}+\mathbf{D}\mathbf{z}\,. \tag{2}$$

Similar to that in (1), the i-supervector is taken as the posterior mean of the latent variable: $E\{\mathbf{z}|\mathcal{X}\}=\phi$. The difference here is that $\mathbf{D}$ is a *CF*-by-*CF* diagonal matrix so that the i-supervector has the same dimensionality as the mean supervector $\mathbf{m}$. The i-supervector extractor can easily be implemented by adopting the diagonal model in JFA [2, 3] with a slight modification. The matrix $\mathbf{D}$ is trained per utterance instead of per speaker basis in order to capture both speaker and session variability.

2.2. PLDA

The advantage of i-vector and i-supervector representations is that they represent a speech segment as a fixed-length vector instead of a variable-length sequence of vectors. Let ϕ_{ij} be a fixed-length vector representing the *j*-th session of the *i*-th speaker with the assumption that each speaker has multiple sessions in the development set. Taking ϕ_{ij} as input, PLDA assumes that it is generated from a Gaussian density, as follows

$$p(\phi_{ij})=\mathcal{N}(\phi_{ij}|\boldsymbol{\mu},\mathbf{F}\mathbf{F}^{\mathrm{T}}+\mathbf{G}\mathbf{G}^{\mathrm{T}}+\boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{\mu}$ denotes the mean vector and $\boldsymbol{\Gamma}=\mathbf{F}\mathbf{F}^{\mathrm{T}}+\mathbf{G}\mathbf{G}^{\mathrm{T}}+\boldsymbol{\Sigma}$ is the covariance matrix.

The modeling capability of PLDA is determined by the structured modeling of the covariance $\boldsymbol{\Gamma}$. To understand this, we rewrite (3) in the form of marginal density, as follows

$$p(\phi_{ij})=\iint p(\phi_{ij}|\mathbf{h}_i,\mathbf{w}_{ij})p(\mathbf{h}_i)p(\mathbf{w}_{ij})d\mathbf{h}_i d\mathbf{w}_{ij}\,, \tag{4}$$

where the conditional and prior densities are given by

$$p(\phi_{ij}|\mathbf{h}_i,\mathbf{w}_{ij})=\mathcal{N}(\phi|\boldsymbol{\mu}+\mathbf{F}\mathbf{h}_i+\mathbf{G}\mathbf{w}_{ij},\boldsymbol{\Sigma}), \tag{5}$$

$$p(\mathbf{h}_i)=\mathcal{N}(0,\mathbf{I}), \tag{6}$$

$$p(\mathbf{w}_{ij})=\mathcal{N}(0,\mathbf{I}). \tag{7}$$

In the above equations, $\mathbf{h}_i$ is the speaker-specific latent variable pertaining to the *i*-th speaker, while $\mathbf{w}_{ij}$ is the session-specific latent variable corresponds to the *j*-th session of the *i*-th speaker. The low-rank matrices $\mathbf{F}$ and $\mathbf{G}$ are used to model the subspaces pertaining to speaker and session variability, while the diagonal matrix $\boldsymbol{\Sigma}$ mops up the remaining variability. Looking at (5), the mean of the conditional distribution is given by

$$E\{\phi_{ij}|\mathbf{h}_i,\mathbf{w}_{ij}\}=\boldsymbol{\mu}+[\mathbf{F}\ \ \mathbf{G}]\left[\mathbf{h}_i^{\mathrm{T}}\ \ \mathbf{w}_{ij}^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{8}$$

Comparing (1) and (8), we see that both i-vector extractor and PLDA model involve dimension reduction via similar form of subspace modeling. This observation motivates us to explore the use of PLDA on i-supervector in the original supervector space. The extraction process serves as the front-end which converts a variable-length sequence $\mathcal{X}$ to a fixed-length vector without reducing the dimension.

3. Joint-estimation of posterior means

We estimate the parameters $\{\boldsymbol{\mu}, \mathbf{F}, \mathbf{G}, \boldsymbol{\Sigma}\}$ of the PLDA model using the expectation maximization (EM) algorithm. To this end, we assume that our development set consists of speech samples from *N* speakers each having *J* number of sessions. Notice that the number of sessions *J* could be different for each speaker. In the E-step, the inference of the posterior means involves the inversion of the precision matrix $\mathbf{L}$, as shown below:

$$\mathbf{L}^{-1}=\left[\begin{array}{c|cccc} J\mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{I} & \mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{G} & \mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{G} & \cdots & \mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{G} \\ \hline \mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F} & \mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{G}+\mathbf{I} & 0 & \cdots & 0 \\ \mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F} & 0 & \mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{G}+\mathbf{I} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F} & 0 & 0 & \cdots & \mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{G}+\mathbf{I} \end{array}\right]^{-1}. \tag{9}$$

The matrix is large as we consider the joint inference of latent variables representing the speaker, $\mathbf{h}_i$, and for all sessions $\{\mathbf{w}_{i,1}, \mathbf{w}_{i,2}, \ldots, \mathbf{w}_{i,J}\}$ from the same speaker. The size of the matrix increases with the number of sessions *J*, though more sessions is always desirable for more robust estimation of parameter.

The matrix $\mathbf{L}$ possesses a unique structure since all sessions from the same speakers are tied to one speaker-specific latent variable. As shown in (9), the matrix $\mathbf{L}$ can be partitioned into four smaller blocks. Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\boldsymbol{\Phi}$ denotes the four sub-matrices, the inverse is given by

$$\mathbf{L}^{-1}=\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \boldsymbol{\Phi}\end{bmatrix}^{-1}=\begin{bmatrix}\mathbf{M} & -\mathbf{M}\mathbf{B}\boldsymbol{\Phi}^{-1}\\ -\boldsymbol{\Phi}^{-1}\mathbf{C}\mathbf{M} & \boldsymbol{\Phi}^{-1}+\boldsymbol{\Phi}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\boldsymbol{\Phi}^{-1}\end{bmatrix}. \tag{10}$$

Inversion of $\boldsymbol{\Phi}$, in the left hand side of (10), is simple as it is block diagonal. Let $\mathbf{Q}=\left(\mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{G}+\mathbf{I}\right)^{-1}$ and $\boldsymbol{\Lambda}=\mathbf{Q}\mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F}$, the sub-matrix $\mathbf{M}$ is given by

$$\mathbf{M}=\left[J\,\mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{F}-\mathbf{G}\boldsymbol{\Lambda})+\mathbf{I}\right]^{-1}. \tag{11}$$

Using above results, it can be shown that the posterior means of the speaker and session specific latent variables are given by

$$E[\mathbf{h}_i]=\mathbf{M}\left[\sum_{j=1}^{J}\mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\phi'_{ij}\right]-\mathbf{M}\boldsymbol{\Lambda}^{\mathrm{T}}\left[\sum_{j=1}^{J}\mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\phi'_{ij}\right], \tag{12}$$

$$E[\mathbf{w}_{ij}]=\mathbf{Q}\left(\mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\phi'_{ij}\right)-\boldsymbol{\Lambda}E[\mathbf{h}_i], \tag{13}$$

where $\phi'_{ij}=\phi_{ij}-\boldsymbol{\mu}$ denotes the i-vectors or i-supervectors centralized with the global mean $\boldsymbol{\mu}=(1/NJ)\sum_{i,j}\phi_{ij}$.

Similarly, the M-step can also be formulated in terms sub-matrices. Let $\tilde{\mathbf{w}}_{ij}^{\mathrm{T}}=\left[\mathbf{h}_i^{\mathrm{T}}\ \mathbf{w}_{ij}^{\mathrm{T}}\right]$ by appending $\mathbf{h}_i$ to each session $\mathbf{w}_{ij}$ belonging to the same speaker. We update the PLDA model, as follows:

$$[\mathbf{F}\ \mathbf{G}]=\left\{\sum_{i,j}\phi'_{ij}E\left[\tilde{\mathbf{w}}_{ij}\right]^{\mathrm{T}}\right\}\left\{\sum_{i,j}E\left[\tilde{\mathbf{w}}_{ij}\tilde{\mathbf{w}}_{ij}^{\mathrm{T}}\right]\right\}^{-1}, \tag{14}$$

$$\boldsymbol{\Sigma}=\frac{1}{NJ}\sum_{i,j}diag\left[\phi'_{ij}\phi'^{\mathrm{T}}_{ij}-[\mathbf{F}\ \mathbf{G}]E\left[\tilde{\mathbf{w}}_{ij}\right]\phi'^{\mathrm{T}}_{ij}\right]. \tag{15}$$

The second-order moment in (14) is obtained for each individual session of all speakers, as follows

$$E\left[\tilde{\mathbf{w}}_{ij}\tilde{\mathbf{w}}_{ij}^{\mathrm{T}}\right]=\begin{bmatrix}\mathbf{M} & -\mathbf{M}\boldsymbol{\Lambda}^{\mathrm{T}}\\ -\boldsymbol{\Lambda}\mathbf{M} & \mathbf{Q}+\boldsymbol{\Lambda}\mathbf{M}\boldsymbol{\Lambda}^{\mathrm{T}}\end{bmatrix}+E\left[\tilde{\mathbf{w}}_{ij}\right]E\left[\tilde{\mathbf{w}}_{ij}\right]^{\mathrm{T}}. \tag{16}$$

Notice that in (14), the subspaces $\mathbf{F}$ and $\mathbf{G}$ are estimated jointly, whereby the correlation between speaker and session variability are considered.

4. PLDA verification score

Given two i-supervectors (or i-vector) $\{\phi_1, \phi_2\}$ that correspond to the train and test segments, respectively, the verification score is calculated as the log-likelihood ratio between two hypotheses $\{H_0, H_1\}$:

$$s = \log p(\phi_1, \phi_2 | H_0) - \log p(\phi_1, \phi_2 | H_1) . \quad (17)$$

Here, H_0 and H_1 correspond to the models as shown in Fig. 1. The model H_0 hypothesizes that $\{\phi_1, \phi_2\}$ belong to the same speaker and hence share the same speaker-specific latent variable $\mathbf{h}_{1,2}$. On the other hand, the model H_1 hypothesizes that they belong to different speakers and hence have separate latent variables $\mathbf{h}_1$ and $\mathbf{h}_2$. The solution for (17) can be given by

$$s = \frac{1}{2}\left[\sum_{l=1}^{2}\phi_l''\right]^{\mathrm{T}}(2\mathbf{K}+I)^{-1}\left[\sum_{l=1}^{2}\phi_l''\right] - \frac{1}{2}\sum_{l=1}^{2}(\phi_l'')^{\mathrm{T}}(\mathbf{K}+I)^{-1}(\phi_l'') + K \quad (18)$$

where K is a constant consisting of the determinant of matrices, which diminishes when score normalization is applied. The other two variables are defined as follows:

$$\mathbf{K} = \mathbf{F}^{\mathrm{T}}\left(\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{F} , \quad (19)$$

$$\phi_l'' = \mathbf{F}^{\mathrm{T}}\left(\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}(\phi_l - \boldsymbol{\mu}) . \quad (20)$$

The matrix inversion $\left(\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}$ can be solved using the matrix inversion lemma, as follows

$$\left(\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{G}\left(\mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{G} + \mathbf{I}\right)^{-1}\mathbf{G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} . \quad (21)$$

Notice that in (20), the centralized vector is projected onto the subspace **F** where speaker information co-vary the most (i.e., dimension reduction), while de-emphasizing the subspace pertaining to channel variability. Notice that the transformation matrix $\mathbf{F}^{\mathrm{T}}\left(\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}$, in (19) into (20), should be computed by multiplying $\mathbf{F}^{\mathrm{T}}$ to the right-hand-side of (21).

Another prerequisite for good performance with PLDA is that the i-supervectors have to follow a normal distribution, as in (3). It has been shown in [7], for the case of i-vector, that whitening followed by length normalization help toward this goal. However, whitening is not feasible for i-supervector due to data scarcity. To this end, we propose in this paper a Gaussianized version of rank norm [9]. The i-supervector is processed element-wise with warping functions mapping each dimension to a standard Gaussian distribution (instead of uniform distribution as in rank norm). To put it mathematically, let $\phi_l^{(m)}$, $m = 1, 2, \ldots, CF$, denotes the elements of the i-supervector ϕ_l. We first get the normalized rank of $\phi_l^{(m)}$ with respect to a background set $B^{(m)}$, as follows

$$r^{(m)} = \frac{\left|\left\{b \in B^{(m)} : b < \phi_l^{(m)}\right\}\right|}{\left|B^{(m)}\right|} , \quad (22)$$

where $|\cdot|$ denotes the cardinality of a set. The Gaussianized value is then obtained by using the inverse CDF of a standard Gaussian distribution (i.e., the probit function), as follows

$$\phi_l^{(m)} \leftarrow \sqrt{2}\ \mathrm{erf}^{-1}\left(2r^{(m)} - 1\right) , \quad (23)$$

where $\mathrm{erf}^{-1}(\cdot)$ denotes the inverse error function. This can then be followed by length normalization prior to PLDA modeling.

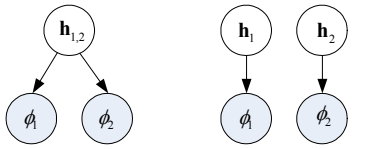

Figure 1: PLDA for verification task. The null hypothesis H_0 states that the observations $\{\phi_1, \phi_2\}$ are from the same speakers. The alternative hypothesis states that they are from different speakers.

5. Experiment

Experiments were carried out on the core task (short2-short3) of NIST SRE08. We use two well-known metrics in evaluating the performance, namely, *equal error rate* (EER) and *minimum detection cost* (MinDCF). Two gender-dependent UBMs consisting of 512 Gaussians were trained using data drawn from the SRE04. Speech parameters were represented by a 54-dimensional vector of *mel frequency cepstral coefficients* (MFCC) with first and second derivatives appended.

The total variability matrix **T** in (1) and **D** in (2) were both trained with similar set of data drawn from Switchboard, SRE04 and SRE05. We use 500 factors for **T**, while **D** is a diagonal matrix by definition. The dimensionality of i-vector is therefore 500, while i-supervector is of dimensionality *CF* = 27648. The rank of the matrices **F** and **G** in the PLDA model is set to 300 and 200, respectively, for the case of i-supervector. For i-vector, best result is found with the rank of **F** set to 300 and using a full matrix for $\boldsymbol{\Sigma}$, in which **G** is no longer required. This observation is consistent with that reported in [6].

5.1. Feature and score normalization

The first set of experiments is to investigate the effectiveness of different normalization methods on i-vector and i-supervector prior to PLDA modeling (i.e., length normalization, whitening and Gaussianized rank norm) and on the score (we used s-norm as reported in [6]). For simplicity, we used only telephone data and report the results on `det6` (i.e., `tel-tel` subtask) in TABLE I. Length normalization (`len`) always outperforms `raw` for both i-vector and i-supervector. Whitening followed by length normalization (`white+len`) further improves the performance for i-vector. Similarly in the case of i-supervector, we used Gaussianized rank-norm followed by length normalization (`grank+len`) to cope with the high dimensionality. Finally, we also notice that s-norm gives consistent improvement for both i-vector and i-supervector.

Without any normalization (`raw`) i-supervector performs better than i-vector. One possible reason is that the Gaussian assumption in (3) can be better fulfilled in the supervector space with higher dimensionality compared to that of the i-vector. However, after applying a full normalization (`white+len+snorm`, `grank+len+snorm`), i-vector outperforms i-supervector. This is consistent for both MALE and FEMALE sets in terms of EER and MinDCF. Notice that i-vector gain huge improvement from length normalization. For the MALE case, we observed 20.0% and 6.5% of relative improvement in EER when length normalization was applied on i-vector and i-supervector, respectively. One avenue to explore for i-supervector is a better normalization method beside length normalization and Gaussianized rank-norm.

5.2. Performance comparison

We compared the performance of i-supervector and i-vector using JFA as baseline and under different `train-test` channel conditions, namely, `det1` (`int-int`), `det4` (`int-tel`), `det5` (`tel-mic`) and `det6` (`tel-tel`) as defined in NIST SRE08 short2-short3 core task. The PLDA models used for i-vector and i-supervector were the same as described in Section 5.1. In addition, we included microphone data (drawn from SRE05 and SRE06) for the whitening transform, Gaussianized rank-norm and s-norm to handle the interview (`int`) and microphone (`mic`) channel conditions. The JFA was trained as follows. The eigenvoice loading matrix (with 300 factors) was train on telephone data drawn from Switchboard. The eigenchannel loading matrix (with 150 factors) was trained using both telephone and microphone data drawn from SRE04, SRE05 and SRE06. Finally, the diagonal model was trained using telephone data drawn from SRE04.

TABLE II shows the results when full normalization (i.e., `white+len+snorm` for i-vector, `grank+len+snorm` for i-supervector, and `zt-norm` for JFA) was applied. Here, we consider the EER and MinDCF by pooling together the male and female scores. Similar to the observation in Section 5.1, i-vector gives better performance than i-supervetor for the case with full normalization, except in `det5` where i-supervector gives a much lower EER though the MinDCF is slightly worse. This again shows that current normalization strategy (Gaussianized rank-norm followed by length normalization) thought effective, has to be further improved. Compared to the JFA baseline, PLDA methods (both i-vector and i-supervector) give competitive performance with slightly lower EER and MinDCF in `det4`, `det5` and `det6`.

6. Conclusions

We have introduced the use of uncompressed form of *i-vector* (i.e., the *i-supervector*) for PLDA-based speaker verification. Similar to i-vector, an i-supervector represents a variable-length speech utterance as a fixed-length vector. But different from i-vector, we keep the total variability space having the same dimensionality as the original supervector space. To this end, we showed how manipulation of large matrices can be done efficiently in training and scoring with the PLDA model. We also introduced the use of Gaussianized rank-norm for feature normalization prior to PLDA modeling.

Compared to i-vector, we found that i-supervector performs better when no normalization (on both feature and score) was applied. This suggests that Gaussian assumption imposed by PLDA becomes less stringent and easier to fulfill in the higher dimensional i-supervector space. However, the performance improvement given by the high dimensionality diminishes when full normalization is applied. As such, current normalization strategy, though effective, has to be improved for better performance. This is a point for future work. Finally, it is also evident that PLDA methods based on i-vector or i-supervector give competitive performance compared to the state-of-the-art JFA.

7. Acknowledgements

The work of Ye Jiang is partially supported by Jiangsu Provincial Natural Science Foundation of China under Grant No. CX2211_0261.

TABLE I: Performance comparison of various normalization methods on i-vector and i-supervector evaluated on NIST SRE08, `det6` subtask of short2-short3.

I-VECTOR	MALE		FEMALE	
	EER	MinDCF	EER	MinDCF
`raw`	6.1785	3.1206	8.1486	3.7028
`len`	4.9411	2.6286	6.4409	3.0581
`white+len`	4.5458	2.4546	6.3193	3.0065
`white+len+snorm`	4.3478	2.2155	6.1530	3.0034
I-SUPERVECTOR	**MALE**		**FEMALE**	
	EER	MinDCF	EER	MinDCF
`raw`	5.2632	2.6605	6.7976	3.3368
`len`	4.9199	2.6271	6.3667	3.3624
`grank+len`	4.8982	2.6676	6.0976	3.2588
`grank+len+snorm`	4.5888	2.3737	6.2639	3.1132

TABLE II: Performance comparison under various train-test channel conditions: `det1`, `det4`, `det5` and `det6` subtasks of short2-short3, NIST SRE08.

	`det1 (int-int)`		`det4 (int-tel)`	
	EER	MinDCF	EER	MinDCF
i-vector	7.2964	3.5189	5.7919	2.7576
i-supervector	7.8769	3.6724	5.9421	3.0541
JFA	7.1404	3.7991	7.6815	3.0350

	`det5 (tel-mic)`		`det6 (tel-tel)`	
	EER	MinDCF	EER	MinDCF
i-vector	6.0462	2.0975	5.5602	2.7556
i-supervector	4.7554	2.2475	5.7740	2.8949
JFA	7.2690	2.6126	6.8126	3.0960

8. References

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

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-Based speaker verification," *IEEE Trans. Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1448-1460, May 2007.

[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio Speech and Language Processing*, vol. 16, no. 5, pp. 980-988, July 2008.

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

[5] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. International Conference on Computer Vision*, 2007.

[6] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey: Speaker and Language Recognition Workshop*, Jun. 2010.

[7] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, pp. 249-252, Aug. 2011.

[8] N. Brümmer, "A farewell to SVM: Bayes factor speaker detection in supervector space," Available: http://sites.google.com/site/nikobrummer, 2006.

[9] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification," in *Proc. ICASSP,* pp. 1577-1580, 2008.