CHANNEL ADAPTATION OF PLDA FOR TEXT-INDEPENDENT SPEAKER VERIFICATION

Liping Chen [1, 2], *Kong Aik Lee*[2], *Bin Ma*[2], *Wu Guo*[1], *Haizhou Li*[2], and *Li Rong Dai*[1]

[1]National Engineering Laboratory for Speech and Language Information Processing, USTC, China

[2]Institute for Infocomm Research, A*STAR, Singapore

clp2011@mail.ustc.edu.cn, kalee@i2r.a-star.edu.sg

ABSTRACT

Probabilistic linear discriminant analysis (PLDA) has shown to be effective for modeling channel variability in the i-vector space for text-independent speaker verification. Speaker verification is a binary hypothesis testing. Given a test segment, the verification score could be computed as the log-likelihood ratio between a *speaker-adapted* PLDA and the *universal* PLDA model. This work proposes to infer the channel factor specific to each test segment and to include the channel estimate in the PLDA models, which essentially shifts the scoring function to better match that of the test channel. We also explore the influence of covariance adaptation in both speaker and channel adaptations. Experimental results on NIST SRE'08 and SRE'10 dataset confirm that the proposed channel adaptation can be effective when the covariance is kept un-adapted, while the covariance adaptation is necessary in the speaker adaptation.

Index Terms— speaker verification, PLDA scoring, speaker adaptation, channel adaptation

1. INTRODUCTION

Over the past few years, many approaches have been proposed to improve the channel robustness of text-independent speaker verification system [1, 2, 3, 4]. Among others, subspace model, like eigenchannel and joint factor analysis (JFA) [3, 4, 5 6], has shown to be extremely effective. Following the same framework as the JFA, the i-vector was proposed in [7] and soon became the mainstream front-end for speaker verification. An i-vector is a fixed-length representation of a speech utterance, which is typically of variable length. Furthermore, it has a much lower dimensionality compared to the mean supervector of a Gaussian mixture model (GMM) [5]. This allows channel compensation techniques, for instance, within-class covariance normalization [8], linear discriminant analysis (LDA) [9], and notably, the probabilistic LDA (PLDA) [10] to be applied effectively. This paper focuses on the use of PLDA in combination with i-vector for text-independent speaker verification.

Speaker verification is a binary hypothesis testing problem, where a decision has to be made between two hypotheses – whether the enrollment and test utterances are from the same or different speakers [11]. With PLDA, the hypotheses test leads to a symmetric scoring function where the roles of the enrollment and test utterances are interchangeable [10, 11, 12]. In [13], it was shown that an equivalent form of scoring function could be derived by translating the verification problem to a likelihood-ratio test between a *speaker-adapted* and the *universal* PLDA models, much similar to the speaker adaptation in the classical GMM-UBM paradigm. In [14], we used similar idea to handle multisession enrollment and showed that a better performance could be achieved using speaker prior derived based on the minimum divergence criterion. In a similar spirit to speaker adaptation, we show how channel adaptation could be performed and the benefit of doing so in this paper.

In i-vector space, we use PLDA to model the speaker and channel variability. The channel variability component plays a key role in providing the general distribution of the distortion due to the channel variation. At testing time, channel compensation is accomplished by factoring out the contribution of channel subspace in the scoring function [12]. All test segments are therefore scored under a common channel assumption, which is suboptimal. We propose to account for the specific channel condition pertaining to each test segment by inserting the channel estimate to the PLDA model, resulting in a form of channel adaptation. The objective here is to shift the PLDA model to match the test channel. This idea of channel adaptation has been the key element of the eigenchannel approach [5]. In this paper, we derive the channel adaptation formula for PLDA. We also investigate the impact of performing channel adaptation either with a full posterior distribution, or in part with only the posterior mean estimate, on the performance.

The rest of this paper is organized as follows. Section 2 gives a brief review of i-vector and PLDA. In Section 3, we illustrate the concept of speaker adaptation with the help of graphical model. In Section 4 presents the theory of channel adaptation for PLDA. Experiment results are then presented in Section 5. Finally, Section 6 concludes the paper.

2. THE I-VECTOR PLDA PARADIGM

The central idea of i-vector extraction is to find a fixed length, and usually reduced dimension, representations of variable-length speech utterances [7]. The fundamental assumption is that the feature vector sequence $\mathcal{O}$, extracted from an utterance, is generated by a session-specific GMM. The mean supervector $\mathbf{m}$ of the GMM is constrained to reside in the subspace $\mathbf{T}$ with origin $\mathbf{m}_o$, as follows

$$\mathbf{m} = \mathbf{m}_o + \mathbf{T}\mathbf{w}. \tag{1}$$

An i-vector is taken as the maximum *a posteriori* estimate of the latent variable $\mathbf{w}$:

$$\phi = \arg\max_{\mathbf{w}} p\left(\mathcal{O} \mid \mathbf{m}_o + \mathbf{T}\mathbf{w}\right)\mathcal{N}\left(\mathbf{w} \mid 0, \mathbf{I}\right). \tag{2}$$

The low-rank matrix $\mathbf{T}$ captures the *total* variability, which is usually understood to reflect both speaker and channel variabilities. An i-vector therefore represents the speaker and channel information both being conflated to a low-dimensional space.

Channel compensation is applied on the i-vectors to suppress the channel effects. With PLDA, this is achieved by introducing two separate subspaces to segregate the channel variation from that of the speaker. In particular, PLDA assumes an i-vector extracted from the *r*-th session of speaker *s* is generated as

$$p\left(\boldsymbol{\phi}_{s,r} \mid \mathbf{h}_s, \mathbf{x}_{s,r}\right) = \mathcal{N}\left(\boldsymbol{\phi}_{s,r} \mid \boldsymbol{\mu} + \mathbf{F}\mathbf{h}_s + \mathbf{G}\mathbf{x}_{s,r}, \boldsymbol{\Sigma}\right). \quad (3)$$

The vector $\boldsymbol{\mu}$ denotes the global mean of all i-vectors. The latent variable $\mathbf{h}_s$ accounts for the identity of a speaker while $\mathbf{x}_{s,r}$ represents the channel effects. The modelling capability of PLDA relies on the speaker and channel loading matrices denoted as $\mathbf{F}$ and $\mathbf{G}$, respectively. We refer to the set $\{\boldsymbol{\mu}, \mathbf{F}, \mathbf{G}, \boldsymbol{\Sigma}\}$ as the parameters of the PLDA model, which could be determined by fitting the model onto a given set of training data using the expectation maximization (EM) algorithm [9]. Details of training procedure used in this paper can be found in [12, 15].

3. IDENTITY INFERENCE WITH PLDA

Let the prior over the latent variables $\mathbf{h}$ and $\mathbf{x}$ be a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. By integrating out the the latent variables, the marginal density can be obtained as follows

$$\begin{aligned} p(\boldsymbol{\phi}) &= \int \mathcal{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu} + \mathbf{F}\mathbf{h} + \mathbf{G}\mathbf{x}, \boldsymbol{\Sigma}\right) \mathcal{N}\left(\mathbf{h} \mid 0, \mathbf{I}\right) \mathcal{N}\left(\mathbf{x} \mid 0, \mathbf{I}\right) d\mathbf{h}\, d\mathbf{x} \\ &= \mathcal{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \mathbf{F}\mathbf{F}^{\mathrm{T}} + \mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right). \end{aligned} \quad (4)$$

From the above, it can be seen that a PLDA model is essentially a Gaussian distribution with a structured covariance model comprising of a speaker and a channel component. In particular, the term $\mathbf{F}\mathbf{F}^{\mathrm{T}}$ corresponds to the speaker variability and the term $\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}$ represents the channel variability.

3.1 Hypothesis test

In addition to channel compensation, PLDA serves as a means to derive the scoring function for the speaker verification task. Given a pair of i-vectors, one from the enrollment and the other from the test segment, the task is to determine whether they are from the same speaker or not. This question gives rise to the following hypotheses:

$$\begin{aligned} &\mathcal{H}_0 : \boldsymbol{\phi}_{\mathrm{t}} \text{ and } \boldsymbol{\phi}_{\mathrm{s}} \text{ are from the same speaker} \\ &\mathcal{H}_1 : \boldsymbol{\phi}_{\mathrm{t}} \text{ and } \boldsymbol{\phi}_{\mathrm{s}} \text{ are from different speakers} \end{aligned} \quad (5)$$

where $\boldsymbol{\phi}_{\mathrm{t}}$ and $\boldsymbol{\phi}_{\mathrm{s}}$ are the i-vectors estimated from the test and enrolment segment, respectively. The log-likelihood ratio for the hypothesis test is

$$l\left(\boldsymbol{\phi}_{\mathrm{t}}, \boldsymbol{\phi}_{\mathrm{s}}\right) \equiv \log \frac{p\left(\boldsymbol{\phi}_{\mathrm{t}}, \boldsymbol{\phi}_{\mathrm{s}} \mid \mathcal{H}_0\right)}{p\left(\boldsymbol{\phi}_{\mathrm{t}}, \boldsymbol{\phi}_{\mathrm{s}} \mid \mathcal{H}_1\right)} = \frac{p\left(\boldsymbol{\phi}_{\mathrm{t}}, \boldsymbol{\phi}_{\mathrm{s}}\right)}{p\left(\boldsymbol{\phi}_{\mathrm{t}}\right) p\left(\boldsymbol{\phi}_{\mathrm{s}}\right)}, \quad (6)$$

where the likelihood terms are evaluated using the PLDA model in (4). Refer to [11, 12, 15] for the details of evaluating the log-likelihood function.

3.2 Speaker-adapted PLDA

Using the chain rule, we replace $p\left(\boldsymbol{\phi}_{\mathrm{t}}, \boldsymbol{\phi}_{\mathrm{s}}\right)$ with $p\left(\boldsymbol{\phi}_{\mathrm{t}} \mid \boldsymbol{\phi}_{\mathrm{s}}\right) \times p\left(\boldsymbol{\phi}_{\mathrm{s}}\right)$ in (4). Cancelling out common term, we arrive at

$$l\left(\boldsymbol{\phi}_{\mathrm{t}}, \boldsymbol{\phi}_{\mathrm{s}}\right) = \log \frac{p\left(\boldsymbol{\phi}_{\mathrm{t}} \mid \boldsymbol{\phi}_{\mathrm{s}}\right)}{p\left(\boldsymbol{\phi}_{\mathrm{t}}\right)}. \quad (7)$$

The numerator in (7) could be further decomposed as

$$p\left(\boldsymbol{\phi} \mid \boldsymbol{\phi}_{\mathrm{s}}\right) = \int p\left(\boldsymbol{\phi} \mid \mathbf{h}\right) p\left(\mathbf{h} \mid \boldsymbol{\phi}_{\mathrm{s}}\right) d\mathbf{h}, \quad (8)$$

where $p\left(\mathbf{h} \mid \boldsymbol{\phi}_{\mathrm{s}}\right)$ is the posterior distribution of the latent variable $\mathbf{h}$ given the enrollment i-vector $\boldsymbol{\phi}_{\mathrm{s}}$. It could be shown (see [12] for details) that the posterior $\mathbf{h} \mid \boldsymbol{\phi}_{\mathrm{s}} \sim \mathcal{N}\left(\mathbf{m}_{\mathrm{s}}, \mathbf{L}_{\mathrm{s}}^{-1}\right)$ is also Gaussian with mean

$$\mathbf{m}_{\mathrm{s}} = \mathbf{L}_{\mathrm{s}}^{-1} \cdot \mathbf{F}^{\mathrm{T}} \left(\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1} \left(\boldsymbol{\phi}_{\mathrm{s}} - \boldsymbol{\mu}\right), \quad (9)$$

and covariance

$$\mathbf{L}_{\mathrm{s}}^{-1} = \left[\mathbf{F}^{\mathrm{T}} \left(\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1} \mathbf{F} + \mathbf{I}\right]^{-1}. \quad (10)$$

Using (9) and (10) in $p\left(\mathbf{h} \mid \boldsymbol{\phi}_{\mathrm{s}}\right)$ and integrating out the latent variable, (8) is now given by

$$p\left(\boldsymbol{\phi} \mid \boldsymbol{\phi}_{\mathrm{s}}\right) = \mathcal{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu} + \mathbf{F}\mathbf{m}_{\mathrm{s}}, \mathbf{F}\mathbf{L}_{\mathrm{s}}^{-1}\mathbf{F}^{\mathrm{T}} + \mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right). \quad (11)$$

Comparing (11) to (4), $\boldsymbol{\mu} + \mathbf{F}\mathbf{m}_s$ and $\mathbf{F}\mathbf{L}_s^{-1}\mathbf{F}^{\mathrm{T}} + \mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}$ are the mean and covariance which have been adapted to the target speaker. Using (11) and (4) in (7), the verification score

$$l\left(\boldsymbol{\phi}_{\mathrm{t}}, \boldsymbol{\phi}_{\mathrm{s}}\right) = \frac{\mathcal{N}\left(\boldsymbol{\phi}_{\mathrm{t}} \mid \boldsymbol{\mu} + \mathbf{F}\mathbf{m}_{\mathrm{s}}, \mathbf{F}\mathbf{L}_{\mathrm{s}}^{-1}\mathbf{F}^{\mathrm{T}} + \mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)}{\mathcal{N}\left(\boldsymbol{\phi}_{\mathrm{t}} \mid \boldsymbol{\mu}, \mathbf{F}\mathbf{F}^{\mathrm{T}} + \mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)} \quad (12)$$

can therefore be interpreted as the log-likelihood ratio between the *speaker-adapted* PLDA and the *universal* PLDA model, in a way much similar to the idea of the *universal background model* (UBM) [2]. The major difference is that the PLDA model is adapted through the latent variable $\mathbf{h}$ in the current case. Figure 1 illustrates this idea in the form of graphical model. This interpretation was first conceived in [13] and developed further in [14] to deal with the issue of multisession enrollment.

4. CHANNEL-ADAPTED PLDA

The columns of the channel matrix $\mathbf{G}$ represent the subspace where the unwanted channel variation correlates the most. In (9) and (10), channel compensation is imposed during the estimation of speaker parameters $\left\{\mathbf{m}_{\mathrm{s}}, \mathbf{L}_{\mathrm{s}}^{-1}\right\}$ through the term $\mathbf{G}\mathbf{G}^{\mathrm{T}}$. In the scoring function (12), $\mathbf{G}$ appears as a part of the covariance which models the speaker, channel and residual variability. As shown in Figure 1, test segments are scored under the same channel condition assuming a non-informative prior on the channel variable, i.e., $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$. This is suboptimal as the channel characteristics of the test segments are generally different, though we assume that the specific differences reside in the subspace span by the columns of the channel matrix $\mathbf{G}$.

4.2 Channel adaptation

We propose to adapt the scoring function with respect to the specific channel characteristic of the test segment. This is ac-

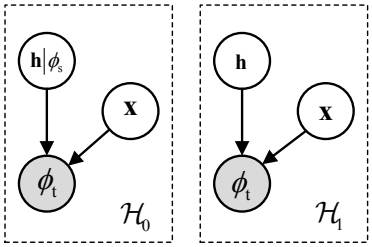

Figure 1: PLDA scoring using the speaker-adapted PLDA (left) and universal PLDA model (right).

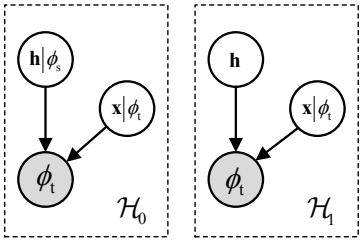

Figure 2: PLDA scoring with the latent variable **x** adapted to the channel characteristic of the test i-vector ϕ_t.

complished through the use of channel posterior $\mathbf{x}|\phi_t \sim \mathcal{N}\left(\mathbf{m}_t, \mathbf{L}_t^{-1}\right)$ estimated for each test i-vector ϕ_t. More specifically, the posterior mean $\mathbf{m}_t$ and covariance $\mathbf{L}_t^{-1}$ are computed as follows:

$$\mathbf{m}_t = \mathbf{L}_t^{-1} \cdot \mathbf{G}^T \boldsymbol{\Sigma}^{-1} \left(\phi_t - \boldsymbol{\mu} - \mathbf{F}\mathbf{m}_s\right) \tag{13}$$

$$\mathbf{L}_t^{-1} = \left[\mathbf{G}^T \boldsymbol{\Sigma}^{-1} \mathbf{G} + \mathbf{I}\right]^{-1}. \tag{14}$$

In (13), $\mathbf{m}_s$ is obtained using (9) and (10). Following the same step as in Section 3, the *channel and speaker* adapted PLDA model could be obtained, as follows

$$\begin{aligned} p\left(\phi|\phi_s,\phi_t\right) &= \int \mathcal{N}\left(\phi \mid \boldsymbol{\mu} + \mathbf{F}\mathbf{h} + \mathbf{G}\mathbf{x}, \boldsymbol{\Sigma}\right) p\left(\mathbf{h}|\phi_s\right) p\left(\mathbf{x}|\phi_t\right) d\mathbf{h}\, d\mathbf{x} \\ &= \mathcal{N}\left(\phi|\boldsymbol{\mu} + \mathbf{F}\mathbf{m}_s + \mathbf{G}\mathbf{m}_t, \mathbf{F}\mathbf{L}_s^{-1}\mathbf{F}^T + \mathbf{G}\mathbf{L}_t^{-1}\mathbf{G}^T + \boldsymbol{\Sigma}\right) \end{aligned} \tag{15}$$

while the *channel*-adapted PLDA model is

$$p\left(\phi \mid \phi_t\right) = \mathcal{N}\left(\phi \mid \boldsymbol{\mu} + \mathbf{G}\mathbf{m}_t, \mathbf{F}\mathbf{F}^T + \mathbf{G}\mathbf{L}_t^{-1}\mathbf{G}^T + \boldsymbol{\Sigma}\right). \tag{16}$$

Using (15) and (16) in (7), we obtain the scoring function being adapted to the channel characteristic of the test segment, as follows:

$$l\left(\phi_t,\phi_s\right) = \frac{\mathcal{N}\left(\phi_t|\boldsymbol{\mu} + \mathbf{F}\mathbf{m}_s + \mathbf{G}\mathbf{m}_t, \mathbf{F}\mathbf{L}_s^{-1}\mathbf{F}^T + \mathbf{G}\mathbf{L}_t^{-1}\mathbf{G}^T + \boldsymbol{\Sigma}\right)}{\mathcal{N}\left(\phi_t \mid \boldsymbol{\mu} + \mathbf{G}\mathbf{m}_t, \mathbf{F}\mathbf{F}^T + \mathbf{G}\mathbf{L}_t^{-1}\mathbf{G}^T + \boldsymbol{\Sigma}\right)}. \tag{17}$$

Similar to that of $\mathbf{h}|\phi_s \sim \mathcal{N}\left(\mathbf{m}_s, \mathbf{L}_s^{-1}\right)$, the channel posterior $\mathbf{x}|\phi_t \sim \mathcal{N}\left(\mathbf{m}_t, \mathbf{L}_t^{-1}\right)$ resides in a subspace within the i-vector space. The channel adaptation is done through the latent variable $\mathbf{x}|\phi_t$. Figure 2 illustrates the idea in the form of graphical model. Notice that we use the subscript 's' to indicates speaker adaptation and the subscript 't' to indicates channel adaptation specific to each to test segment.

4.2 Necessity of covariance adaptation

In the above, the adaptation of PLDA is executed in two steps. At enrollment time, the universal PLDA model is adapted through the speaker factor **h** given an enrolment i-vector. This leads to the speaker PLDA model in (11). At test time, both speaker and universal PLDA models are first adapted to the test channel through the channel factor **x**. A verification score is then computed using (17). Notice that the mean adaptation appears in the form of shifting, while the covariance adaptation appears in the form of rotation subspaces.

Previous study on GMM-UBM [2] and JFA [4] has shown that speaker characteristic is mainly reflected by the Gaussian mean vectors while the covariance matrices are not as necessary. Here, the speaker and channel adaptation could also be implemented without the covariance, i.e., using only the posterior mean estimate of the speaker and channel factors. It turns out that covariance adaptation is necessary for speaker adaptation while redundant for channel adaptation, as we shall show further in Section 5.

5. EXPERIMENTS

5.1 Dataset and system setup

Experiments were conducted on the telephone-only trials of NIST SRE'08 and SRE'10. For SRE'08, we used common condition (CC) 6 and 7 of the *short2-short3* task. For SRE'10 we used CC 5, 6 and 8 of the *core-core* task. In all of these tasks, the enrollment and test utterances were recorded over telephone line (mobile or landline) using myriad types of handsets thereby posing a difficult challenge for channel compensation. Furthermore, the language variability in CC 6 of SRE'08 and the vocal effort variability in CC 5, 6, and 8 of SRE'10 make the speaker verification task become even more difficult, though we do not address the influence of these factors in the current paper. The performance was evaluated based on the equal-error-rate (EER) and the detection cost function (DCF)

$$C_{\text{DET}} = P_{\text{tar}} P_{\text{miss}}\left(\theta\right) + \left(1 - P_{\text{tar}}\right) P_{\text{fa}}\left(\theta\right).$$

We consider the minimum DCF at three different operation points, namely, DCF08, DCF10 and DCF12 [16, 17, 18].

We used gender-dependent setup. Two UBMs consisting of 512 Gaussians (with full covariance matrices) were trained with SRE'04 dataset. The acoustic features were 57-dimensional *mel frequency cepstral coefficents* (MFCC) with first and second derivatives appended. The total variability space, with 400 dimensions, was trained with the telephone data from SRE'04, 05 and 06. The same set of data was used to train the speaker and channel subspaces, **F** and **G**, of the PLDA. The rank of the channel loading matrix **G** was set to 50, while the rank of speaker loading matrix **F** was 200. Length normalization and whitening [19] were used. We found that it is beneficial to include microphone data (drawn from SRE'05 and 06) to the residual as

$$\boldsymbol{\Sigma} \leftarrow \mathbf{G}_{\text{mic}} \mathbf{G}_{\text{mic}}^T + \boldsymbol{\Sigma}. \tag{18}$$

The additional channel matrix $\mathbf{G}_{\text{mic}}$ was trained in a decoupled manner on top of $\boldsymbol{\Sigma}$ and $\mathbf{G}_{\text{tel}}$ following the method as proposed in [20]. Note that for the posterior estimation of the channel

Table I: Results on CC6 and CC7 of SRE'08 and CC5, CC6 and CC8 of SRE'10 with full posterior used for speaker adaptation.

	EER (%)	DCF08	DCF10	DCF12
Male				
CC6 (08)	4.6496	0.2462	0.7541	0.6300
CC7 (08)	2.3246	0.1336	0.6903	0.4888
CC5 (10)	3.0974	0.1548	0.4808	0.4031
CC6 (10)	5.2442	0.3553	0.7360	0.6766
CC8 (10)	0.8182	0.0699	0.2589	0.2156
Female				
CC6 (08)	6.9085	0.3451	0.9906	0.8680
CC7 (08)	3.3350	0.1543	0.9911	0.7018
CC5 (10)	4.6418	0.2001	0.4930	0.4429
CC6 (10)	9.1476	0.4441	0.8634	0.7962
CC8 (10)	2.4027	0.0916	0.4264	0.3147

Table II: Results on CC6 and CC7 of SRE'08 and CC4, CC6, and CC8 of SRE'10 with posterior mean estimate used for speaker adaptation (i.e., without covariance adaptation)

	EER (%)	DCF08	DCF10	DCF12
Male				
CC6(08)	5.2303	0.2732	0.7655	0.6601
CC7(08)	3.9787	0.1636	0.7973	0.5743
CC5(10)	3.6514	0.1905	0.5913	0.4761
CC6(10)	7.6213	0.3633	0.7809	0.7594
CC8(10)	1.5697	0.0888	0.3513	0.2621
Female				
CC6 (08)	8.1773	0.4061	0.9884	0.8777
CC7 (08)	4.0699	0.2051	0.9899	0.7248
CC5 (10)	4.8792	0.2378	0.6147	0.5342
CC6 (10)	9.8644	0.4389	0.9180	0.8230
CC8 (10)	2.2861	0.1058	0.5400	0.3882

Table III: Results on CC6 and CC7 of SRE'08 and CC5, CC6 and CC8 of SRE'10 with full posterior used for both speaker and channel adaptations.

	EER (%)	DCF08	DCF10	DCF12
Male				
CC6 (08)	4.6637	0.2436	0.7506	0.6263
CC7 (08)	2.3319	0.1314	0.6811	0.4820
CC5 (10)	3.0140	0.1498	0.4723	0.3981
CC 6 (10)	4.9446	0.3482	0.7247	0.6653
CC 8 (10)	0.7849	0.0663	0.2589	0.2108
Female				
CC6 (08)	6.8662	0.3409	0.9906	0.8636
CC7 (08)	3.2758	0.1502	0.9924	0.7036
CC5 (10)	4.5800	0.1982	0.4958	0.4364
CC6 (10)	9.0161	0.4432	0.8579	0.7958
CC8 (10)	2.3270	0.0908	0.4097	0.2979

Table IV: Results on CC6 and CC7 of SRE'08 and CC4, CC6, and CC8 of SRE'10 with full posterior used for speaker adaptation and posterior mean estimate used for channel adaptation (i.e., without covariance adaptation).

	EER (%)	DCF08	DCF10	DCF12
Male				
CC6 (08)	4.6248	0.2390	0.7484	0.6195
CC7 (08)	2.2562	0.1273	0.6629	0.4740
CC5 (10)	2.8501	0.1504	0.4638	0.3898
CC6 (10)	4.9482	0.3454	0.7247	0.6618
CC8 (10)	0.7630	0.0624	0.2589	0.2108
Female				
CC6 (08)	6.8009	0.3379	0.9911	0.8620
CC7 (08)	3.2480	0.1471	0.9924	0.6999
CC5 (10)	4.4000	0.1972	0.4823	0.4280
CC6 (10)	8.9356	0.4347	0.8525	0.7834
CC8 (10)	2.1974	0.0897	0.3762	0.2809

variable, $\mathbf{G}_{\text{mic}}$ is not used. The scoring with only speaker adaptation as described in (12) is used as the baseline system.

5.2 Results

In the first experiment, we investigate the role of the posterior covariance in speaker adaptation of PLDA. Table I shows the results for the case where the full posterior was used for speaker adaptation (these serve as our baseline). This corresponds to the scoring function in (12). Table II shows the results when the posterior mean was used for speaker adaptation. More specifically, we set the posterior covariance $\mathbf{L}_s^{-1}$ to an identity matrix in (12) which essential reduces (11) to a mean only adaptation. Comparing Table I and II, we can see that the posterior covariance plays a significant role in speaker adaptation of PLDA model. The performance deteriorates significantly when only the mean estimate was used in place of the full posterior. We used full posterior in speaker adaptation for subsequent experiments.

Tables III and IV show the performance of channel adaptation using full posterior and mean only estimate of the channel factor, respectively. In particular, we used the scoring function in (17) for the results in Table III. For the case of mean-only adaptation in Table IV, we set $\mathbf{L}_t^{-1} = \mathbf{I}$ in the numerator and denominator of (17), which essentially switch off the covariance adaptation. Comparing Table III and IV, it is clear that the use of posterior covariance should be avoided for the case of channel adaptation. Now, comparing the results in Table IV to the baseline in Table I, it is clear that channel mean adaptation of PLDA model improve the performance consistently across the five common conditions for both SRE'08 and SRE'10. This is also partially true even when the channel adaption is done using full posterior of the channel factor (comparing Table III and I).

6. CONCLUSION

We proposed the idea of channel adaptation for PLDA and showed that it could be accomplished through the use of the posterior distribution of the channel factor. Conventional PLDA scoring is suboptimal as the channel characteristics of test segments are generally different. By means of channel adaptation, we shift the PLDA model to match each test channel, which has shown to be critical in previous study in the context of eigenchannel GMM. We also showed that channel adaptation could be done in full, or in part using only the posterior mean estimate. Experimental results on SRE'08 and SRE'10 confirm that posterior covariance plays a significant role in speaker adaptation but not for channel adaptation. The results suggest that channel adaptation of PLDA model should be performed using only the posterior mean estimate of the channel factor with the covariance kept unchanged.

7. ACKNOWLEDGEMENTS

The work of Liping Chen was funded by the National Nature Science Foundation of China (Grant No. 61273264) and the electronic information industry development fund of China (Grant No. 2013-472).

8. 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, Jan. 2010.

[2] 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.

[3] P. Kenny, G. Boulianne and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435-1447, 2007.

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

[5] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP estimators for speaker recognition," in *Proc. the 8th European Conference on Speech Communication and Technology*, 2003, pp. 2691-2964.

[6] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, "Speaker adaptation using an eigenphone basis." *IEEE Trans. Audio, Speech, and Language Processing*, vol. 12, no. 6, pp. 579-589, 2004.

[7] 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.

[8] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, September 2006.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

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

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

[12] Y. Jiang, K. A. Lee, L. Wang, "PLDA in the i-supervector space for text-independent speaker verification," *EURASIP Journal on Audio, Speech, and Music Processing*, 2014, 2014:29.

[13] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P.Dumouchel, "PLDA for speaker verification with utterance of arbitrary duration," in *Proc. IEEE ICASSP*, 2013, pp. 7649 - 7653.

[14] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, L. R. Dai, "Minimum estimation of speaker prior in multi-session PLDA scoring," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, May, 2014, pp.4035-4039.

[15] K. A. Lee, A. Larcher, C. H. You, B. Ma, and H. Li, "Multi-session PLDA scoring of i-vector for partially open-set speaker detection," in *Proc. INTERSPEECH*, 2013, pp. 3651-3655.

[16] National Institute of Standards and Technology, *The NIST 2008 SRE Evaluation Plan*, 2008.

[17] National Institute of Standards and Technology, *The NIST 2010 SRE Evaluation Plan*, 2010.

[18] R. Saeidi, K. A. Lee, et al, "I4U submission to NIST SRE 2012: a large-scale collaborative effort for noise-robust speaker verification," in *Proc. Interspeech*, 2013, pp. 1986-1990.

[19] P. M. Bousquet, A. Larcher, D. Matrouf , et al, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2012, pp. 157-164.

[20] M. Senoussaoui, P. Kenny, N. Dehak, P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2010, pp. 28- 3.