# ADAPTATION OF PLDA FOR MULTI-SOURCE TEXT-INDEPENDENT SPEAKER VERIFICATION

*Liping Chen* [1], *Kong Aik Lee*[2], *Bin Ma*[2], *Long Ma*[1], *Haizhou Li*[2,3], *Li-Rong Dai*[4]

[1] Microsoft Search Technology Center Asia, Beijing, China
[2] Institute for Infocomm Research, A*STAR, Singapore
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapre
[4] National Engineering Laboratory for Speech and Language Information Processing, USTC, China

lipch@microsoft.com kalee@i2r.a-star.edu.sg

## ABSTRACT

Probabilistic linear discriminant analysis (PLDA) is widely described as an effective model for text-independent speaker verification in the i-vector space. The PLDA scoring function is typically formulated as the likelihood ratio between the speaker-adapted and the universal PLDAs. In this case, the adaptation of PLDA was performed through the speaker factors. In this paper, we show that the channel factors of the PLDA could be equivalently exploited to deal with the multi-source conditions. In speaker verification, with the proposed method, a PLDAmodel trained on conversational telephone speech could be adequately adapted for interview-style microphone recordings. Experimental results on NIST SRE'08 and SRE'10 datasets confirm that the proposed method is effective, especially for the case whereby enrollment and test utterances were captured from different sources.

***Index Terms—*** multi-source speaker verificaiton, channel adaptation, channel prior estimation, probabilistic linear discriminant analysis

## 1. INTRODUCTION

In recent years, factor analysis techniques have been successfully applied to deal with the session variability in text-independent speaker verification. Among others, eigenchannel was first proposed to model the channel variability with a low-rank subspace [1, 2]. Joint factor analysis (JFA) was then proposed to model both the speaker and channel variabilities with independent linear subspaces [3, 4]. Thereafter, the i-vector was proposed in [5]. In the i-vector framework, a vector of low- and fixed-dimension was used to represent the session variability contained in a speech utterance, including both the speaker and channel variabilities. When i-vector is used for speaker recognition, session compensation is performed using the *probabilistic linear discriminant analysis* (PLDA) [6]. In these methods, the channel variability is compensated with the channel variation to be a part of the covariance and all the test and enrollment speech utterances are compared on the global channel condition.

For speaker recognition with heterogeneous datasets, where the speech utterances were acquired under different recording scenarios (e.g. telephone conversation versus interview style, broadband microphone versus narrow-band telephone speech), channel subspace needs to cover all the possible sources. A commonly used method is concatenating the channel subspaces that are independently trained for individual sources [7]. However, since the rank of the final channel subspace is decided by the number of the recording sources, the subspace matrix will become over-complete with the number of columns exceeding the dimensionality of the i-vectors. Thus a trade-off between the rank of the channel subspace and the modeling capability on all the sources needs to be considered. Another method proposed in [8] is through the use of informative prior during i-vector extraction, which became complicated when only i-vectors are available. In this paper, we propose an alternative solution to handle the multi-source condition by adapting the PLDA model using source-specific prior.

In [9], we proposed to adapt the PLDA model to the channel condition depending on the test i-vector in every single trial during scoring. In this manner, the training and test utterances can be compared upon the same channel, providing a recipe to the multi-source speaker verification. But it didn't bring significant performance improvement. According to our analysis, two possible explanations can be responsible for such a phenomenon. One is that the posterior estimation of the channel variable given a single i-vector may not be accurate enough to represent a kind of channel condition. The other is that the PLDA scoring models vary among the testing trials, bringing vibration to the scores and causing uncertainty in setting the decision threshold.

In this paper, for multiple sources, we propose to adapt the PLDA model according to the recording sources. Given the type of recording sources of the enrollment and test utterances, the corresponding prior distribution can be chosen for PLDA scoring. The adaptation is performed through the prior distribution of the channel variable on the specific source type. Experiments were carried out on NIST SRE'08 and SRE'10 where three recording sources, i.e., telephone, microphone and interview were present [10, 11]. Experimental results confirm the effectiveness of such an adaptation in multi-source speaker verification tasks.

## 2. THE I-VECTOR PLDA PARADIGM

The crux of i-vector extraction lies in finding a fixed-length and usually dimension-reduced representation of a given utterance which is always of variable length. The fundamental assumption is that for a set of speech utterances, the variability among them lies in a linear and low-rank subspace, denoted with $\mathbf{T}$. Given a feature sequence $\mathcal{O}_r$, extracted from an utterance, the mean supervector of the utterance-specific GMM $\mathbf{m}_r$ can be modeled as:

$$\mathbf{m}_r = \mathbf{m}_0 + \mathbf{T}\mathbf{w}_r \tag{1}$$

In (1), $\mathbf{m}_0$ is the mean supervector of the universal background model (UBM), denoting the common variability shared by all the utterances, $\mathbf{w}_r$ is the session-specific variable corresponding to $\mathcal{O}_r$, containing the session variability in $\mathcal{O}_r$ that is modeled by $\mathbf{T}$ [5].

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

$$\phi_r = \arg\max_{\mathbf{w_r}} p\left(\mathcal{O}_r \left| \mathbf{m}_0 + \mathbf{T}\mathbf{w}_r \right.\right) \mathcal{N}\left(\mathbf{w}_r \left| \mathbf{0}, \mathbf{I} \right.\right) \tag{2}$$

where the prior distribution of $\mathbf{w}_r$ is standard normal as $\mathbf{w}_r \sim \mathcal{N}\left(\mathbf{w}_r \left| \mathbf{0}, \mathbf{I} \right.\right)$.

The low-rank subspace $\mathbf{T}$ captures not only the speaker variability, but also channel variability. For better speaker comparison, the impact of channel needs to be compensated, for which PLDA is among the mainstream techniques. From the perspective of its generative property, PLDA assumes an i-vector extracted from the $r$-th session of speaker $s$ to be generated by:

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

The vector $\boldsymbol{\mu}$ denotes the global mean of all i-vectors. The latent variable $\mathbf{h}_s$ accounts for the speaker identity while $\mathbf{x}_{s,r}$ represents the channel effects. The modeling capability of PLDA relies on the speaker and channel loading matrices, denoted as $\mathbf{F}$ and $\mathbf{G}$, respectively. $\boldsymbol{\Sigma}$ models the residual variation that cannot be accounted for by $\mathbf{F}$ and $\mathbf{G}$. We refer to the parameters $\theta = \{\boldsymbol{\mu}, \mathbf{F}, \mathbf{G}, \boldsymbol{\Sigma}\}$ as the parameter set of the PLDA model, which can be estimated by fitting the model onto a given set of training data using the expectation maximization (EM) algorithm [12, 13, 14].

## 3. PLDA SCORING MODEL

In PLDA, the prior over the latent variables $\mathbf{h}$ and $\mathbf{x}$ are standard normal distribution as $\mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$. Integrating out the latent variables, the marginal density of an i-vector $\phi$ can be obtained as follows:

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

From the above equation, it can be seen that a PLDA model is essentially a Gaussian distribution with its covariance matrix composed of both the speaker and channel variabilities, $\mathbf{F}\mathbf{F}^{\mathsf{T}}$ and $\mathbf{G}\mathbf{G}^{\mathsf{T}}$ respectively. As a kind of probabilistic model, PLDA has its inherent likelihood computation according to (4) which is used in the two-hypothesis scoring model [6].

### 3.1. Scoring models on speaker and channel adaptation

Given a pair of i-vectors, $\phi_{\mathrm{e}}$ for enrollment and $\phi_{\mathrm{t}}$ for testing, the speaker verification task is to determine whether the two i-vectors come from the same speaker or not. Besides the conventional two-hypothesis scoring model, in [15], we proposed an equivalent scoring model from the perspective of model adaptation which can be described mathematically as follows:

$$\begin{aligned} l\left(\phi_{\mathrm{e}}, \phi_{\mathrm{t}}\right) &= \log \frac{p\left(\phi_{\mathrm{t}} \left| \phi_{\mathrm{e}} \right.\right)}{p\left(\phi_{\mathrm{t}}\right)} \\ &= \log \frac{\mathcal{N}\left(\phi_{\mathrm{t}} \left| \boldsymbol{\mu} + \mathbf{F}\mathbf{m}_s, \mathbf{F}\mathbf{L}_s^{-1}\mathbf{F}^{\mathsf{T}} + \mathbf{G}\mathbf{G}^{\mathsf{T}} + \boldsymbol{\Sigma} \right.\right)}{\mathcal{N}\left(\phi_{\mathrm{t}} \left| \boldsymbol{\mu}, \mathbf{F}\mathbf{F}^{\mathsf{T}} + \mathbf{G}\mathbf{G}^{\mathsf{T}} + \boldsymbol{\Sigma} \right.\right)} \end{aligned} \tag{5}$$

where $\mathbf{m}_s$ and $\mathbf{L}_s^{-1}$ are the posterior mean and covariance estimates of the speaker variable $\mathbf{h}_s$ given the i-vector $\phi_{s,r}$. The denominator represents the likelihood on the universal PLDA model and the numerator denotes the likelihood of $\phi_{\mathrm{t}}$ given the PLDA model that is adapted to speaker $s$ with the posterior distribution parameters of the speaker variable $\mathbf{h}$.

Furthermore, in [9], we proposed to adapt the PLDA model to the channel of the test utterances with the scoring model to be:

$$\begin{aligned} &l\left(\phi_{\mathrm{e}}, \phi_{\mathrm{t}}\right) \\ &= \log \frac{\mathcal{N}\left(\phi_t \left| \boldsymbol{\mu} + \mathbf{F}\mathbf{m}_s + \mathbf{G}\mathbf{m_t}, \mathbf{F}\mathbf{L}_s^{-1}\mathbf{F}^{\mathsf{T}} + \mathbf{G}\mathbf{L}_{\mathrm{t}}^{-1}\mathbf{G}^{\mathsf{T}} + \boldsymbol{\Sigma} \right.\right)}{\mathcal{N}\left(\phi_t \left| \mu + \mathbf{G}\mathbf{m_t}, \mathbf{F}\mathbf{F}^{\mathsf{T}} + \mathbf{G}\mathbf{L}_{\mathrm{t}}^{-1}\mathbf{G}^{\mathsf{T}} + \boldsymbol{\Sigma} \right.\right)} \end{aligned} \tag{6}$$

where $\mathbf{m}_{\mathrm{t}}$ and $\mathbf{L}_{\mathrm{t}}^{-1}$ are the posterior mean and covariance of the channel variable $\mathbf{x}$ estimated on the test i-vector $\phi_{\mathrm{t}}$. However, the performance comparison given in [9] shows that the channel adaptation is not that effective. Two possible reasons can be responsible. One is the inaccuracy of the posterior distribution parameters estimated with only one i-vector in representing its specific channel. The other is the uncertainty in the decision threshold on the scores computed on the models that are adapted to the different testing i-vectors.

## 4. MULTI-SOURCE ADAPTATION WITH INFORMATIVE PRIOR

In the conventional PLDA model, the prior distribution of the channel variable is assumed to be standard normal, which is non-informative about the specialty of the source of the recordings. The model performs well when the test utterances are from the same source with the training dataset. However, when the model is tested on data from different sources, the performance is always poor. In this section, we propose to adapt the PLDA model to specific sources through prior distribution imposed on the channel variable.

Given a training dataset of the target source composed of $S$ speakers with the feature vectors denoted as $\mathcal{O} = \bigcup_{s=1}^{S} \mathcal{O}_s$, where $\mathcal{O}_s$ is the feature vector of the utterances from the $s$-th speaker. $\mathcal{O}_s$ can be specified with respect to the utterances as $\mathcal{O}_s = \{\mathcal{O}_{s,1}, ..., \mathcal{O}_{s,R_s}\}$, with $R_s$ to be the number of utterances of speaker $s$ for $s = 1, ..., S$. For each of the speech utterances, an i-vector can be estimated, collectively denoted as $\bigcup_{s=1}^{S} \bigcup_{r=1}^{R_s} \phi_{s,r}$. In the following, we describe the process of adapting a well-trained PLDA model, $\theta = \{\boldsymbol{\mu}, \mathbf{F}, \mathbf{G}, \boldsymbol{\Sigma}\}$, to fit the target source which results in the target model with parameter set $\theta = \{\boldsymbol{\mu}_{\mathrm{t}}, \mathbf{F}, \mathbf{G}_{\mathrm{t}}, \boldsymbol{\Sigma}\}$. To achieve this goal, we use the informative prior distribution parameters for the channel variable on the training set of the target source estimated by minimizing its divergence with the posterior distribution [16].

### 4.1. Channel posterior estimation

Given the i-vector $\phi_{s,r}$ from the i-vector set of speaker $s$ as $\bigcup_{r=1}^{R_s} \phi_{s,r}$, [13] presented the posterior estimates of its channel variable $\mathbf{x}_{s,r}$. To be specific here, the posterior mean and covariance of the speaker variable $\mathbf{h}_s$ needs to be estimated first as follows:

$$\mathbf{L}_{h_s}^{-1} = \left[\mathbf{I} + R_s \mathbf{F}^{\mathsf{T}} \left(\mathbf{G}\mathbf{G}^{\mathsf{T}} + \boldsymbol{\Sigma}\right)^{-1} \mathbf{F}\right]^{-1} \tag{7}$$

$$\mathrm{E}\left[\mathbf{h}_s\right] = \mathbf{L}_{h_s^{-1}} \sum_{r=1}^{R_s} \mathbf{F}^{\mathsf{T}} \left(\mathbf{G}\mathbf{G}^{\mathsf{T}} + \boldsymbol{\Sigma}\right)^{-1} \left(\phi_{s,r} - \boldsymbol{\mu}\right) \tag{8}$$

with $\mathbf{L}_{h_s}^{-1}$ and $\mathrm{E}[\mathbf{h}_s]$ to be the posterior covariance and mean estimates of $\mathbf{h}_s$ respectively, given $\bigcup_{r=1}^{R_s}\phi_{s,r}$.

In [13], the posterior mean estimation of $\mathbf{x}_{s,r}$ is given by:

$$\mathbf{m}_{s,r} = \mathrm{E}[\mathbf{x}_{s,r}] = \mathbf{Q}\mathbf{G}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\{\phi_{s,r} - \boldsymbol{\mu} - \mathbf{F}\mathrm{E}[\mathbf{h}_s]\} \quad (9)$$

And the posterior covariance estimation of $\mathbf{x}_{s,r}$ is given by

$$\mathbf{L}_{x_{s,r}}^{-1} = \mathbf{Q}\left(\mathbf{I} + \mathbf{G}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F}\mathbf{L}_{h_s}^{-1}\mathbf{F}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{Q}\right) \quad (10)$$

where $\mathbf{Q} = \left(\mathbf{I} + \mathbf{G}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{G}\right)^{-1}$.

From (9) and (10), we can see that the posterior estimations of the mean and covariance of $\mathbf{x}_{s,r}$ for every single i-vector $\phi_{s,r}$ is influenced by the other $(R_s - 1)$ i-vectors from that speaker $\{\phi_{s,1}, ..., \phi_{s,r-1}, \phi_{s,r+1}, ..., \phi_{s,R_s}\}$. This is reasonable for the mean estimation since it is a determinant that represents the variability contained in an i-vector not modeled as the speaker variability. But this is not the case for the posterior covariance. In fact, the covariance models the uncertainty of approximating the value of $\mathbf{x}_{s,r}$ with its posterior mean estimate and should be estimated within the utterance itself. In this paper, the posterior covariance of the channel variable $\mathbf{x}_{s,r}$ given $\phi_{s,r}$ is estimated within the i-vector itself as:

$$\mathbf{C}_{x_{s,r}} = \left[\mathbf{I} + \mathbf{G}^{\mathsf{T}}\left(\mathbf{F}\mathbf{F}^{\mathsf{T}} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{G}\right]^{-1} \quad (11)$$

Note that when $R_s = 1$, the estimation given in (10) reduces to (11). And for any given i-vector, (11) only depends on the model parameters and is free from the specific speech utterances.

### 4.2. Channel prior estimation

For each i-vector in the dataset for a target source $\bigcup_{s=1}^{S}\bigcup_{r=1}^{R_s}\phi_{s,r}$, the posterior distribution of the latent channel variable $\mathbf{x}_{s,r}$ can be computed according to (9) and (11), i.e., $p(\mathbf{x}_{s,r}|\phi_{s,r}) = \mathcal{N}(\mathbf{m}_{s,r}, \mathbf{C})$. For brevity, the subscript for indexing the utterance in $\mathbf{C}$ as used in (11) is ignored. Assume the priori of the channel variable for the target source condition is normally distributed as $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\omega}, \mathbf{P})$ with $\boldsymbol{\omega}$ and $\mathbf{P}$ to be the mean and covariance respectively. The parameters can be obtained by minimizing the Kullback-Leibler (KL) divergence of $\mathcal{N}(\boldsymbol{\omega}, \mathbf{P})$ from the set of posteriors $p(\mathbf{x}_{s,r}|\phi_{s,r})$ where $s = 1, ..., S$ and $r = 1, ..., R_s$. The objective function is defined as:

$$D(\theta_{\mathrm{MD}}) = \sum_{s=1}^{S}\sum_{r=1}^{R_s}\mathrm{E}\left\{\log\frac{\mathcal{N}(\mathbf{x}_{s,r}|\mathbf{m}_{s,r}, \mathbf{C})}{\mathcal{N}(\mathbf{x}_{s,r}|\boldsymbol{\omega}, \mathbf{P})}\right\} \quad (12)$$

The expectation is taken with respect to the posterior probability. The solution to the prior distribution parameters are obtained by setting the derivatives of $D(\theta_{\mathrm{MD}})$ with respect to $\boldsymbol{\omega}$ and $\mathbf{P}$ to 0. To be specific, the minimum divergence estimates can be expressed in closed form as follows:

$$\boldsymbol{\omega} = \frac{1}{\sum_{s=1}^{S} R_s}\sum_{s=1}^{S}\sum_{r=1}^{R_s}\mathbf{m}_{s,r} \quad (13)$$

$$\mathbf{P} = \mathbf{C} + \mathbf{S} \quad (14)$$

where $\mathbf{S}$ is the covariance matrix on the posterior mean vectors, computed as:

$$\mathbf{S} = \frac{1}{\sum_{s=1}^{S} R_s}\sum_{s=1}^{S}\sum_{r=1}^{R_s}(\mathbf{m}_{s,r} - \boldsymbol{\omega})(\mathbf{m}_{s,r} - \boldsymbol{\omega})^{\mathsf{T}} \quad (15)$$

### 4.3. Source adaptation with channel prior estimation

Assume that the PLDA model trained on the original source is described mathematically as:

$$p(\phi_{s,r}|\mathbf{h}_s, \mathbf{x}_{s,r}) = \mathcal{N}(\phi_{s,r}|\boldsymbol{\mu} + \mathbf{F}\mathbf{h}_s + \mathbf{G}\mathbf{x}_{s,r}, \boldsymbol{\Sigma}) \quad (16)$$

where the prior distribution of $\mathbf{h}_s$ and $\mathbf{x}_{s,r}$ are assumed to be standard normal, i.e., $\mathbf{h}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{x}_{s,r} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

For the target source with training i-vectors $\bigcup_{s=1}^{S}\bigcup_{r=1}^{R_s}\phi_{s,r}$, the prior distribution parameters for $\mathbf{x}$ can be estimated using (13) and (14), denoted as $\mathbf{x}_{\mathrm{t}} \sim (\boldsymbol{\omega}, \mathbf{P})$, where the subscript $_{\mathrm{t}}$ specifies the target channel. By absorbing $\boldsymbol{\omega}$ and $\mathbf{P}$ into the model parameters, the target PLDA model can be described as:

$$p(\phi_{s,r}|\mathbf{h}_s, \mathbf{x}_{s,r}) = \mathcal{N}(\phi_{s,r}|\boldsymbol{\mu}_{\mathrm{t}} + \mathbf{F}\mathbf{h}_s + \mathbf{G}_{\mathrm{t}}\mathbf{x}_{s,r}, \boldsymbol{\Sigma}) \quad (17)$$

where $\boldsymbol{\mu}_{\mathrm{t}}$ and $\mathbf{G}_{\mathrm{t}}$ are the global mean vector and channel subspace adapted to the target source as follows:

$$\boldsymbol{\mu}_{\mathrm{t}} = \boldsymbol{\mu} + \mathbf{G}\boldsymbol{\omega} \quad (18)$$

$$\mathbf{G}_{\mathrm{t}} = \mathbf{G}\mathbf{L} \quad (19)$$

with $\mathbf{L}$ to be the lower triangular matrix of the cholesky decomposition of $\mathbf{P}$, i.e., $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{P}$. Note that by absorbing the prior information to the model parameters, in the source-adapted model (17), the prior of $\mathbf{x}_{s,r}$ falls back to standard normal distribution as $\mathbf{x}_{s,r} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Note that during the source adaptation, the speaker subspace $\mathbf{F}$ and the residual covariance $\boldsymbol{\Sigma}$ are kept unadapted.

## 5. MULTI-SOURCE PLDA SCORING

We adopt the speaker-adaptation PLDA scoring model proposed in [15] for scoring on the given two i-vectors in a trial, denoted as $\phi_{\mathrm{e}}$ and $\phi_{\mathrm{t}}$ for enrollment and testing respectively. In the enrollment phase, denote the model parameter set of the source type of $\phi_{\mathrm{e}}$ as $\theta_{\mathrm{e}}$. The posterior mean and covariance of the speaker variable $\mathbf{h}_s$ can be estimated given $\phi_{\mathrm{e}}$ as $p(\mathbf{h}_s|\phi_{\mathrm{e}}; \theta_{\mathrm{e}}) = \mathcal{N}\left(\mathbf{m}_{\mathrm{e}}, \mathbf{L}_{\mathrm{e}}^{-1}\right)$ for enrolling speaker adaptation.

In testing, the PLDA model parameter set is chosen according to the source condition of $\phi_{\mathrm{t}}$ for testing, denoted as $\theta_{\mathrm{tst}} = \{\boldsymbol{\mu}_{\mathrm{tst}}, \mathbf{F}, \mathbf{G}_{\mathrm{tst}}, \boldsymbol{\Sigma}\}$. The score of the trial can be computed as the log-likelihood ratio between the speaker adapted and the universal PLDA models on the test source. Mathematically, it is computed as follows:

$$\begin{aligned} &l(\phi_{\mathrm{e}}, \phi_{\mathrm{t}}; \theta_{\mathrm{e}}, \theta_{\mathrm{tst}}) \\ &= log\frac{\mathcal{N}\left(\boldsymbol{\mu}_{\mathrm{tst}} + \mathbf{F}\mathbf{m}_{\mathrm{e}}, \mathbf{F}\mathbf{L}_{\mathrm{e}}^{-1}\mathbf{F}^{\mathsf{T}} + \mathbf{G}_{\mathrm{tst}}\mathbf{G}_{\mathrm{tst}}^{\mathsf{T}} + \boldsymbol{\Sigma}\right)}{\mathcal{N}\left(\boldsymbol{\mu}_{\mathrm{tst}}, \mathbf{F}\mathbf{F}^{\mathsf{T}} + \mathbf{G}_{\mathrm{tst}}\mathbf{G}_{\mathrm{tst}}^{\mathsf{T}} + \boldsymbol{\Sigma}\right)} \end{aligned} \quad (20)$$

## 6. EXPERIMENTS

We carried out our experiments on all the common conditions in *short2-short3* of NIST SRE'08 and *coreext-coreext* of SRE'10. The equal error rate (EER) and the detection cost function (DCF) are used to evaluate the performance. We consider the minimum DCF at the operation points of DCF08 and DCF10 [10, 11].

For the i-vector system, we use the i-vector proposed in [17] and [18] whose Baum-welch statistics are computed on the posteriors given by a deep neural network (DNN) trained as the acoustic model for speech recognition. The DNN is trained on Fisher and Switchboard datasets using Kaldi [19] with 4112 senones modeled. The network feature is 40 filterbank features appended with its first

**Table 1**. Performance comparison between the baseline and the source adaptation on all the *common conditions* (CC) of *short2-short3* of NIST SRE'08 (telephone / source-adapted PLDAs)

| | EER(%) | DCF08 | DCF10 |
|---|---|---|---|
| **Male** | | | |
| CC1(itv-itv) | 3.868/**2.947** | 0.147/**0.106** | 0.257/**0.196** |
| CC2(itv-itv) | 0.164/**0.000** | 0.008/**0.000** | 0.008/**0.000** |
| CC3(itv-itv) | 4.051/**3.087** | 0.154/**0.111** | 0.271/**0.205** |
| CC4(itv-tel) | 2.635/**2.456** | 0.127/**0.112** | 0.248/**0.205** |
| CC5(tel-mic) | 2.161/**1.712** | 0.097/**0.075** | 0.415/**0.337** |
| CC6(tel-tel) | 4.984/**4.939** | 0.257/**0.256** | 0.667/**0.657** |
| CC7(tel-tel) | 1.410/**1.398** | 0.071/**0.069** | 0.376/0.387 |
| CC8(tel-tel) | **0.397**/0.429 | **0.021**/0.022 | **0.193**/0.219 |
| **Female** | | | |
| CC1(itv-itv) | 6.300/**3.637** | 0.271/**0.150** | 0.526/**0.321** |
| CC2(itv-itv) | 0.000/0.000 | 0.000/0.000 | 0.000/0.000 |
| CC3(itv-itv) | 6.605/**3.817** | 0.285/**0.158** | 0.552/**0.337** |
| CC4(itv-tel) | 4.916/**4.142** | 0.209/**0.186** | 0.698/**0.676** |
| CC5(tel-mic) | 3.698/**2.608** | 0.116/**0.095** | 0.159/**0.155** |
| CC6(tel-tel) | 6.782/**6.662** | 0.375/**0.372** | **0.989**/0.990 |
| CC7(tel-tel) | 1.974/**1.950** | 0.097/**0.095** | 0.863/**0.846** |
| CC8(tel-tel) | 2.123/**2.114** | 0.082/**0.080** | **0.976**/0.984 |

and second order derivatives. For network input, the feature vector of each frame is concatenated with 5 frames on its left and right sides. The network structure is $1320 - 5 \times 2048 - 4112$. Among the 4112 senones, 20 were removed, including *laughter*, *silence*, *noise* and *OOV*, leaving 4092 senones for total variability modeling.

For session variability modeling, we used a gender-dependent setup. The UBM was trained with the NIST SRE'04 dataset. The acoustic features are 13-dimensional MFCC with the first and second order derivatives appended, leading to 39-dimensional feature vectors. In the test set, three different sources are included, telephone, microphone and interview. In our training datasets, we had telephone data from NIST SRE'04, SRE'05, SRE'06 and the Switchboard; microphone and interview datasets were from NIST SRE'05, SRE'06 and Mixer. Compared with the telephone dataset, the microphone and interview datasets are significantly smaller in scale, so the two sources are modeled as one channel with their training datasets pooled together.

As the primary data in our training datasets, the telephone datasets are used to train the total variability model of rank 400. Besides, it has been shown in researches such as [1] that in multi-source tasks, the speaker model trained on the primary dataset gives the best performance. So in our experiments, considering the modeling for both speaker and channel, we trained a PLDA model on the telephone datasets with the ranks of the speaker and channel subspaces to be 200 and 400 respectively with a diagonal covariance. The telephone PLDA was set as the baseline, on which we adapted for microphone (and interview).

Given the original PLDA model well trained on the telephone channel, we estimated the channel prior on the telephone and microphone (and interview) datasets respectively using their training datasets. Then the original PLDA model was adjusted on the telephone and adapted to microphone (and interview) using the estimated priors specific to the sources. In testing, the adapted PLDA models were chosen based on the recording sources of the enrollment and test utterances for speaker enrollment and scoring respectively.

**Table 2**. Performance comparison between the baseline and the source adaptation on all the *common conditions* (CC) of *coreext-coreext* of NIST SRE'10(telephone / source-adapted PLDAs)

| | EER(%) | DCF08 | DCF10 |
|---|---|---|---|
| **Male** | | | |
| CC1(itv-itv) | 2.183/**2.146** | 0.074/**0.071** | 0.373/**0.349** |
| CC2(itv-itv) | 3.952/**3.576** | 0.150/**0.126** | 0.543/**0.489** |
| CC3(itv-tel) | 3.164/**3.080** | 0.121/**0.118** | 0.342/**0.319** |
| CC4(itv-itv) | 2.409/**2.147** | 0.081/**0.070** | 0.243/**0.216** |
| CC5(tel-tel) | 1.121/**1.110** | 0.055/**0.054** | 0.199/**0.197** |
| CC6(tel-tel) | 1.908/**1.823** | 0.118/**0.115** | 0.471/**0.460** |
| CC7(mic-mic) | 2.355/**1.669** | 0.105/**0.084** | 0.413/**0.365** |
| CC8(tel-tel) | 0.754/**0.747** | 0.037/**0.035** | 0.178/**0.162** |
| CC9(mic-mic) | 0.850/**0.815** | 0.038/**0.030** | 0.102/**0.068** |
| **Female** | | | |
| CC1(itv-itv) | 3.265/**2.893** | 0.133/**0.115** | 0.519/**0.444** |
| CC2(itv-itv) | 6.802/**5.691** | 0.305/**0.235** | 0.820/**0.667** |
| CC3(itv-tel) | 3.140/**2.932** | 0.145/**0.129** | 0.494/**0.453** |
| CC4(itv-itv) | 2.493/**2.086** | 0.124/**0.096** | 0.440/**0.325** |
| CC5(tel-tel) | 1.833/**1.794** | 0.094/**0.091** | 0.290/**0.291** |
| CC6(tel-tel) | **3.168**/3.185 | 0.166/**0.165** | 0.575/**0.566** |
| CC7(mic-mic) | 4.606/**2.657** | 0.237/**0.177** | 0.626/**0.519** |
| CC8(tel-tel) | **1.216**/1.223 | 0.071/**0.070** | 0.311/**0.312** |
| CC9(mic-mic) | 1.116/**0.596** | 0.045/**0.040** | 0.129/**0.088** |

In our experiments, we compared the performances of the telephone PLDA with the source-adapted PLDA model. Tables 1 and 2 present the performances of the two models. In the tables, *tel*, *mic* and *itv* are abbreviations for telephone, microphone and interview respectively. From the comparison, we observe that with the channel adapted to the enrollment and test i-vectors respectively according to their recording sources, better performances can be achieved when microphone and interview channels are included for testing, i.e. where the source mismatch between testing and training exists. This confirms that the source mismatch between the data for PLDA training and testing can be solved with the source adaptation on the PLDA model. For the cross-source tasks, such as CC4, CC5 in SRE'08 and CC3 in SRE'10, the models for the enrollment and testing were chosen according to their recording sources respectively, and the results show its effectiveness in coping with the channel mismatch problem. However, for the telephone-telephone trials, the performance of adaptation is not consistently better than the baseline. This is reasonable since there is no source mismatch problem on the speech utterances between model training and testing.

## 7. CONCLUSION

We propose to adapt a generic PLDA model to a target source with the prior distribution of the channel variable to be informative of the target source. We resort to the criterion of minimum divergence between the posterior and prior distributions of the channel variable for prior estimation. For speaker verification task where the speech utterances from multiple sources are present, by choosing an appropriate prior according to the sources of the enrollment and test utterances, the source mismatch problem between the training, enrollment and test utterances is resolved.

## 8. REFERENCES


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

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

[3] 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.

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

[5] 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.

[6] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th International Conference on Computer Vision (ICCV)*, Oct 2007, pp. 1–8.

[7] J. Gonzalez-Dominguez, B. Baker, R. Vogt, J. Gonzalez-Rodriguez, and S. Sridharan, "On the use of factor analysis with restricted target data in speaker verification.," in *Proc. Odyssey: Speaker and Language Recognition Workshop*, 2010, pp. 103–108.

[8] S. E. Shepstone, K. A. Lee, H. Li, Z.-H. Tan, and S. H. Jensen, "Total variability modeling using source-specific priors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 501–517, 2016.

[9] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. R. Dai, "Channel adaptation of PLDA for text-independent speaker verification," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5251–5255.

[10] "The NIST year 2008 speaker recognition evaluation plan," 2008, `http://www.itl.nist.gov/iad/mig/tests/sre/2008/`.

[11] "The NIST year 2010 speaker recognition evaluation plan," 2010, `http://www.itl.nist.gov/iad/mig/tests/sre/2010/`.

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

[13] Y. Jiang, K. A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, "PLDA modeling in i-vector and supervector space for speaker verification," in *Proc. INTERSPEECH*, 2012.

[14] Y. Jiang, K. A. Lee, and L. Wang, "PLDA in the i-supervector space for text-independent speaker verification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–13, 2014.

[15] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. R. Dai, "Minimum divergence estimation of speaker prior in multi-session PLDA scoring," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4007–4011.

[16] N. Brümmer, "EM for probabilistic LDA," *Agnitio Research, Cape Town, Tech. Rep*, 2010.

[17] Y Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1695–1699.

[18] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," in *Proc. Odyssey: Speaker and Language Recognition Workshop*, 2014, pp. 293–298.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The Kaldi speech recognition toolkit," in *Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011, number EPFL-CONF-192584.