# MINIMUM DIVERGENCE ESTIMATION OF SPEAKER PRIOR IN MULTI-SESSION PLDA SCORING

*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, Agency for Science, Technology and 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 speaker and channel variability in the i-vector space for text-independent speaker verification. This paper shows that the PLDA scoring function could be formulated as model comparison between an adapted PLDA model and the *universal* PLDA. Based on this formulation, we show that a more robust adaptation could be attained by adapting the PLDA model through the use of minimum divergence estimate of speaker prior in the latent subspace. Experimental results on NIST SRE'10 and SRE'12 dataset confirm that the proposed method is effective in handling multi-session task. Notably, it is free from the covariance shrinkage problem typically found in the standard multi-session PLDA scoring.

***Index Terms***— multi-session speaker verification, PLDA scoring, speaker adaptation, minimum divergence

## 1. INTRODUCTION

Over the past few years, many approaches based on the Gaussian mixture model (GMM) in a GMM-UBM framework [1, 2] have been proposed to improve the performance of text-independent speaker verification system [3, 4]. Based upon the GMM supervector [5], the i-vector was proposed in [6] and soon became the mainstream front-end for speaker verification and spoken language recognition alike [7]. Similar to a GMM supervector, an i-vector is a fixed-length representation of a speech utterance, which is typically of variable length. Besides, an i-vector offers a much lower dimensionality than that of the GMM supervector. This allows channel compensation techniques, for instance, within-class covariance normalization [8], linear discriminant analysis (LDA) [9], and notably, probabilistic LDA (PLDA) [10] to be applied effectively with the low dimensional i-vectors.

With PLDA, a commonly used scoring method is based on the *likelihood-ratio test* between two hypotheses – whether the enrollment and test utterances are from the same or different speakers [11]. This leads to a symmetric scoring function whereby the roles of the enrollment and test utterances are interchangeable as far as the detection score is concerned. In this paper, we show that such PLDA scoring paradigm could be formulated in equivalent form as model comparison between an *adapted* and the *universal* PLDA models, much similar to the speaker adaptation in the classical GMM-UBM paradigm [1]. This new interpretation gives rise to the use of minimum divergence estimation for speaker adaptation proposed in this paper. For easier understanding, we illustrate the two scoring methods visually with probabilistic graphical model [9].

It is customary to assume that only one i-vector is available per speaker during enrolment. In this paper, we consider a more general setting, as in the recent NIST SRE'12 [12, 13], whereby multiple i-vectors are available for enrollment. Following the method as briefly described earlier (more details in Section 3), these i-vectors are used to adapt the universal PLDA to a speaker-specific PLDA model through a latent variable in the speaker space. One subtle problem with this procedure is the shrinkage of the posterior covariance when large numbers of enrollment i-vectors are available pre speaker. This is particularly problematic when the i-vectors are highly correlated as they might be extracted from simultaneous multi-channel recordings, shorter duration cuts or exact replicas of other utterances. In this paper, we propose the use of *minimum divergence* [14] to address this problem. We show that minimum divergence estimation leads to a simple procedure of taking the empirical mean and covariance in the speaker space. The covariance matrix estimated in this manner is always lower bounded by a fixed value determined by the loading matrices of the PLDA.

The rest of this paper is organized as follows. Section 2 provides a brief review of i-vector and PLDA. In Section 3, we show, for the general case of multi-session, that PLDA scoring could be interpreted as model comparison between an adapted and the universal PLDA models. We then look into the problem of covariance shrinkage and address this issue with the help of minimum divergence estimation. Section 5 presents some experiment results. Finally, Section 6 concludes the paper.

## 2. I-VECTOR AND PLDA

### 2.1. I-vector extraction

The central idea of i-vector extraction is to find a fixed length, and usually reduced dimension, representation of a variable-length speech utterance [6]. The fundamental assumption is that the feature vector sequence is generated by a session-specific GMM. Let $r$ be the session index, the mean supervector $\mathbf{m}_r$ of the GMM is constrained to lie in the subspace with origin $\mathbf{m}$, as follows

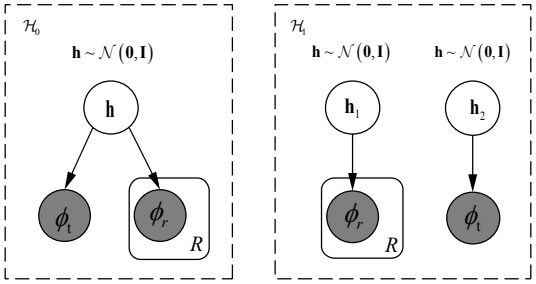


Figure1: Graphical model illustrating the two hypotheses of the likelihood-ratio test.

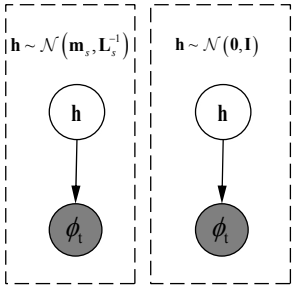


Figure 2: Graphical model illustrating the model adaptation scoring approach.

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

The low-rank matrix $\mathbf{T}$, referred to as the total variability matrix, contains both speaker and channel variabilities. An i-vector is then taken as the posterior mean $\phi_r = E\{\mathbf{w}_r \,|\, \mathcal{O}_r\}$ of the latent variable $\mathbf{w}_r$ representing both the speaker and channel information of a speech utterance $\mathcal{O}_r$.

### 2.2. Probabilistic LDA

In PLDA, the contribution of speaker and channel effects on an i-vector is teased apart by introducing separate subspaces. Let $\phi_{s,r}$ be an i-vector extracted from the *r*-th session of the speaker *s*. We assume that $\phi_{s,r}$ is generated from a linear Gaussian model as follows:

$$p(\phi_{s,r} \,|\, \mathbf{h}_s, \mathbf{x}_{s,r}) = \mathcal{N}(\phi_{s,r} \,|\, \boldsymbol{\mu} + \mathbf{F}\mathbf{h}_s + \mathbf{G}\mathbf{x}_{s,r}, \boldsymbol{\Sigma}) \,. \tag{2}$$

The modeling capability of PLDA relies on the latent variables $\mathbf{h}_s$ and $\mathbf{x}_{s,r}$, referred to as the speaker and channel factors, respectively. The vector $\boldsymbol{\mu}$ denotes the global mean of all i-vectors, $\mathbf{F}$ and $\mathbf{G}$ are the speaker and channel loading matrices, respectively, while the covariance matrix $\boldsymbol{\Sigma}$ models the remaining variability not accounted for by the loading matrices. This could be seen more clearly by examining the marginal density

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

To arrive at (3), the latent variables $\mathbf{h}_s$ and $\mathbf{x}_{s,r}$ are integrated out, assuming a standard normal prior for both variables. We refer to the set $\theta_{\mathrm{PLDA}} = \{\boldsymbol{\mu}, \mathbf{F}, \mathbf{G}, \boldsymbol{\Sigma}\}$ 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 about training procedure used in this paper can be found in [15, 16]. It is worth mentioning that a PLDA model, as shown in (3), is essentially a Gaussian distribution with a structured covariance matrix in the i-vector space.

## 3. MULTI-SESSION PLDA SCORING

### 3.1. Likelihood-ratio test

Consider a speaker verification task, where each target speaker has multiple training utterances. Let $\{\phi_{s,r}\}_{r=1}^{R}$ be the set of i-vectors extracted from these training utterances. The task is to determine whether the set of training i-vectors $\{\phi_{s,r}\}_{r=1}^{R}$ and a given test i-vector $\phi_{\mathrm{t}}$ are from the same target speaker or not. This question gives rise to the following hypotheses:

$\mathcal{H}_0 : \phi_{\mathrm{t}}$ and $\{\phi_{s,r=1,\ldots,R}\}$ are from the same speaker

$\mathcal{H}_1 : \phi_{\mathrm{t}}$ and $\{\phi_{s,r=1,\ldots,R}\}$ are from different speakers

The likelihoods of the two hypotheses can be evaluated using the models as shown in Fig. 1. More specifically, we compute their log-likelihood ratio, as follows

$$l(\phi_{\mathrm{t}}, \phi_{s,r=1,\ldots,R}) = \log \frac{p(\phi_{\mathrm{t}}, \phi_{s,r=1,\ldots,R} \,|\, \mathcal{H}_0)}{p(\phi_{\mathrm{t}}, \phi_{s,r=1,\ldots,R} \,|\, \mathcal{H}_1)} = \log \frac{p(\phi_{\mathrm{t}}, \phi_{s,r=1,\ldots,R})}{p(\phi_{\mathrm{t}})\, p(\phi_{s,r=1,\ldots,R})} \tag{4}$$

where each of the likelihood terms in the numerator and denominator is evaluated using (3). This is commonly referred to as the *by-the-book* multi-session PLDA scoring in the community.

### 3.2. Speaker model adaptation

One key signature of the PLDA scoring function in (4) is that no speaker model is involved. Detection scores are computed by comparing the training and test i-vectors through the use of the PLDA model in (3). In [17], it was shown that some redundant computation could be avoided, especially when multiple i-vectors are available for enrollment, by replacing $p(\phi_{\mathrm{t}}, \phi_{s,r=1,\ldots,R})$ in the numerator with $p(\phi_{\mathrm{t}} \,|\, \phi_{s,r=1,\ldots,R})\, p(\phi_{s,r=1,\ldots,R})$, which leads to

$$l(\phi_{\mathrm{t}}, \phi_{r=1,\ldots,R}) = \log \frac{p(\phi_{\mathrm{t}} \,|\, \phi_{s,r=1,\ldots,R})}{p(\phi_{\mathrm{t}})} \,. \tag{5}$$

The numerator in (5) is now given by

$$p(\phi_{\mathrm{t}} \,|\, \phi_{s,r=1,\ldots,R}) = \mathcal{N}(\phi_{\mathrm{t}} \,|\, \boldsymbol{\mu} + \mathbf{F}\mathbf{m}_s, \mathbf{F}\mathbf{L}_s^{-1}\mathbf{F}^{\mathrm{T}} + \mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}) \,. \tag{6}$$

Here, $\mathbf{m}_s$ and $\mathbf{L}_s^{-1}$ are the posterior mean and covariance of the latent speaker factor $\mathbf{h}_s \sim \mathcal{N}(\mathbf{h} \,|\, \mathbf{m}_s, \mathbf{L}_s^{-1})$ estimated using the set of training i-vectors of speaker *s*, as follows:

$$\mathbf{m}_s = \mathbf{L}_s^{-1} \cdot \sum_{r=1}^{R} \mathbf{F}^{\mathrm{T}} (\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma})^{-1} (\phi_{s,r} - \boldsymbol{\mu}) \,, \tag{7}$$

$$\mathbf{L}_s^{-1} = \left[ \mathbf{I} + R\mathbf{F}^{\mathrm{T}} (\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma})^{-1} \mathbf{F} \right]^{-1} \,. \tag{8}$$

Notice that the number of training sessions, *R*, could be different among speakers. For instance, *R* could be up to 100 in the context of NIST SRE'12.

The scoring functions in (4) and (5) are mathematically equivalent. Nonetheless, they provide different perspectives from which we could view a speaker detection task. Notably, the formulation in (5) brings in the notion of speaker model adaptation which is absent in (4). More specifically, (6) can be seen as the PLDA model adapted to a target speaker using the given set of training i-vectors. Comparing (6) to (3), $\boldsymbol{\mu}+\mathbf{F}\mathbf{m}_s$ and $\mathbf{F}\mathbf{L}_s^{-1}\mathbf{F}^{\mathrm{T}}+\mathbf{G}\mathbf{G}^{\mathrm{T}}+\boldsymbol{\Sigma}$ are the adapted mean vector and covariance matrix of the speaker-dependent PLDA model. The expression in (5) can then be interpreted as the log-likelihood ratio between the speaker-dependent PLDA model in (6) and the *universal* PLDA model in (3), in a way much similar to the idea of the *universal background model* (UBM) [1]. The major difference is that the PLDA model is adapted through a latent variable $\mathbf{h}_s$ in the current case. Figure 2 illustrates this idea in the form of graphical model.

## 4. MINIMUM DIVERGENCE ESTIMATION OF SPEAKER PRIOR

To adapt a PLDA model to a target speaker, we first estimate the posterior mean $\mathbf{m}_s$ and covariance $\mathbf{L}_s^{-1}$ using (7) and (8), and substitute the results into (6). Clearly, the estimation of the first and second moments of the posterior distribution constitutes an important part of speaker adaptation. One major problem with the posterior estimation in (8) is the shrinkage of the posterior covariance $\mathbf{L}_s^{-1}$ for large *R* which in turn affects the estimation of $\mathbf{m}_s$ in (7). This is particularly problematic when the training utterances are highly correlated, for instance, simultaneous multi-channel recordings, shorter duration cuts or exact replicas of other utterances. In the following, we advocate the use of minimum divergence [14] to address this problem.

### 4.1 Minimum divergence estimation

Consider the case where individual speaker has *R* enrollment utterances. We extract one i-vector $\boldsymbol{\phi}_{s,r}$ from each of these utterances. For each of the i-vectors, we compute the posterior distribution on the latent variable **h** as follows

$$p\left(\mathbf{h}\,|\,\boldsymbol{\phi}_{s,r}\right)=\mathcal{N}\left(\mathbf{h}\,|\,\mathbf{m}_{s,r},\mathbf{L}^{-1}\right),\ \text{for}\ r=1,2,\ldots,R, \tag{9}$$

where

$$\mathbf{m}_{s,r}=\mathbf{L}^{-1}\mathbf{F}^{\mathrm{T}}\left(\mathbf{G}\mathbf{G}^{\mathrm{T}}+\boldsymbol{\Sigma}\right)^{-1}\left(\boldsymbol{\phi}_{s,r}-\boldsymbol{\mu}\right)\ \text{and} \tag{10}$$

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

are the posterior mean and covariance, respectively. Given (9), we seek for another Gaussian distribution $\mathcal{N}\left(\mathbf{h}\middle|\theta_{\mathrm{MD}}\right)$ that best represents the *R* posterior distributions. Let $\theta_{\mathrm{MD}}=\left\{\mathbf{y}_s,\mathbf{P}_s^{-1}\right\}$ be its mean and covariance, the parameters could be obtained by minimizing the Kullback-Leibler (KL) divergence [9] of $\mathcal{N}\left(\mathbf{h}\middle|\theta_{\mathrm{MD}}\right)$ from the *R* posteriors $p\left(\mathbf{h}\,|\,\boldsymbol{\phi}_{s,r}\right)$, defined as follows

$$D\left(\theta_{\mathrm{MD}}\right)=\sum_{r=1}^{R}E\left\{\log\frac{\mathcal{N}\left(\mathbf{h}\,|\,\mathbf{m}_{s,r},\mathbf{L}^{-1}\right)}{\mathcal{N}\left(\mathbf{h}\,|\,\mathbf{y}_s,\mathbf{P}_s^{-1}\right)}\right\}, \tag{12}$$

where the expectation is taken with respect to $\mathcal{N}\left(\mathbf{h}\,|\,\mathbf{m}_{s,r},\mathbf{L}^{-1}\right)$. Notice that (12) is a sum of *R* KL divergence measures between normal distributions, the solution of which is given by [18]:

$$D\left(\theta_{\mathrm{MD}}\right)=\frac{1}{2}\cdot\mathrm{tr}\left[R\cdot\left(\mathbf{L}^{-1}+\mathbf{S}\right)\cdot\mathbf{P}_s\right]-\frac{R}{2}\cdot\log\left|\mathbf{P}_s\right|+K\ . \tag{13}$$

Here, *K* is constant for a given dataset, while **S** is data dependent:

$$\mathbf{S}=\frac{1}{R}\cdot\sum_{r=1}^{R}\left(\mathbf{m}_{s,r}-\mathbf{y}_s\right)\left(\mathbf{m}_{s,r}-\mathbf{y}_s\right)^{\mathrm{T}}\ . \tag{14}$$

We solve for $\theta_{\mathrm{MD}}=\left\{\mathbf{y}_s,\mathbf{P}_s^{-1}\right\}$ by differentiating (13) with respect to $\mathbf{y}_s$ and $\mathbf{P}_s$, separately, and set the derivatives to zero. In particular, the minimum divergence estimates could be expressed in closed form, as follows

$$\mathbf{y}_s=\frac{1}{R}\sum_{r=1}^{R}\mathbf{m}_{s,r}\ , \tag{15}$$

$$\mathbf{P}_s^{-1}=\mathbf{L}^{-1}+\mathbf{S}\ . \tag{16}$$

Different from that in (7) and (8), we estimate *R* number of posteriors instead of one and find the set of parameters $\theta_{\mathrm{MD}}=\left\{\mathbf{y}_s,\mathbf{P}_s^{-1}\right\}$ that best describes the posteriors with minimum KL divergence. Notice that, $\left\{\mathbf{y}_s,\mathbf{P}_s^{-1}\right\}$ can be seen as empirical estimate of mean and covariance of the speaker factor **h** in the subspace spanned by the eigenvoice matrix **F**.

### 4.2 Speaker adaptation

From the Bayesian perspective, $\mathcal{N}\left(\mathbf{h}\middle|\mathbf{y}_s,\mathbf{P}_s^{-1}\right)$ can be seen as the adapted prior of the speaker factor **h** from a non-informative one. Using $\mathcal{N}\left(\mathbf{h}\middle|\mathbf{y}_s,\mathbf{P}_s^{-1}\right)$ in place of the multi-session posterior $\mathcal{N}\left(\mathbf{h}\middle|\mathbf{m}_s,\mathbf{L}_s^{-1}\right)$, we have the adapted PLDA model for speaker *s,* as follows

$$p\left(\boldsymbol{\phi}_{\mathrm{t}}\middle|\boldsymbol{\phi}_{s,r=1,\ldots,R},\theta_{\mathrm{MD}}\right)=\mathcal{N}\left(\boldsymbol{\phi}_{\mathrm{t}}\middle|\boldsymbol{\mu}+\mathbf{F}\mathbf{y}_s,\mathbf{F}\mathbf{P}_s^{-1}\mathbf{F}^{\mathrm{T}}+\mathbf{G}\mathbf{G}^{\mathrm{T}}+\boldsymbol{\Sigma}\right). \tag{17}$$

It can be seen that, for the case when $R$ = 1, $\mathbf{P}_s^{-1}$ reduces to $\mathbf{L}^{-1}$ while $\mathbf{y}_s$ falls back to $\mathbf{m}_s$, which makes the speaker model adaptation in (17) exactly identical to that in (6).

Comparing (16) to (8), the covariance estimate $\mathbf{P}_s^{-1}$ consists of two parts – the posterior component $\mathbf{L}^{-1}$ and an empirical component $\mathbf{S}$. Both of them are free from the shrinkage problem as what will happen to $\mathbf{L}_s^{-1}$ in (8) when large numbers of enrollment sessions are available for a particular speaker. More specifically, $\mathbf{L}^{-1}$ in (11) is independent of *R*, while **S** in (14) reflects the empirical covariance of i-vectors in the speaker space. For the case when all i-vectors (or enrollment sessions) are identical, **S** becomes **0**, and $\mathbf{P}_s^{-1}$ takes the minimum value of $\mathbf{L}^{-1}$. Using (10) in (15), we arrive at

$$\mathbf{y}_s=\mathbf{L}^{-1}\mathbf{F}^{\mathrm{T}}\left(\mathbf{G}\mathbf{G}^{\mathrm{T}}+\boldsymbol{\Sigma}\right)^{-1}\left(\frac{1}{R}\sum_{r=1}^{R}\boldsymbol{\phi}_{s,r}-\boldsymbol{\mu}\right). \tag{18}$$

Note that the empirical mean $\mathbf{y}_s$ in the speaker subspace corresponds to the empirical mean $\sum_{r=1}^{R}\boldsymbol{\phi}_{s,r}/R$ in the original i-vector space. As such, the proposed solution is similar to the conventional method in estimating $\mathbf{y}_s$, except for that it has an additional empirical component **S** in the covariance estimate.

## 5. EXPERIMENTS

Experiments were carried out on the NIST SRE'10 and SRE'12 datasets. For SRE'12, we focus on the *Common Condition 2* of the core task where individual target speakers have one to over a hundred utterances for enrollment. For SRE'10, we focus on the *Common Condition 5* of the *8conv-core* task where each target speaker has eight utterances for enrollment. The performance was evaluated based on the *equal-error-rate* (EER) and the *detection cost function* (DCF) defined as $C_{\text{DET}} = P_{\text{tar}} P_{\text{miss}}(\theta) + (1 - P_{\text{tar}}) P_{\text{fa}}(\theta)$. We consider the minimum DCF at two different operation points, namely, DCF10 and DCF12. The probability of target, $P_{\text{tar}}$, is set to 0.001 and 0.01 for DCF10 and DCF12, respectively. The minimum DCF is found by sliding the threshold $\theta$ for different value miss and false alarm probabilities denoted as $P_{\text{miss}}(\theta)$ and $P_{\text{fa}}(\theta)$, respectively.

We used gender-dependent setup. The UBMs consisting of 512 Gaussians (with full covariance matrices) were trained with NIST SRE'04 dataset. The acoustic features were 57-dimensional vectors of *mel frequency cepstral coefficents* (MFCC) with first and second derivatives appended. The total variability space, with a dimension of 400, was trained with the telephone data from NIST SRE'04, 05 and 06. For PLDA, the channel variability is modeled with two channel matrices, $\mathbf{G}_{\text{tel}}$ and $\mathbf{G}_{\text{mic}}$, trained in a decoupled manner [19]. The rank of channel loading matrices $\mathbf{G} = [\mathbf{G}_{\text{tel}}, \mathbf{G}_{\text{mic}}]$ is set to 100, while the rank of speaker loading matrix $\mathbf{F}$ is 200.

We compared the performances of three approaches for speaker model adaptation:

i. *By-the-book* approach using the model defined in (6), (7), and (8);
ii. Minimum divergence (*MinDiv*) adaptation using the model defined in (15), (16), and (17);
iii. Minimum divergence adaptation without the empirical covariance $\mathbf{S}$ in (16). We refer to this method as *Mean* only.

It is worth mentioning that by dropping the empirical component $\mathbf{S}$ in (16), the minimum divergence approach reduces to the conventional solution of taking the average of i-vectors prior to PLDA scoring. This simple approach has been shown effective for multi-session scoring in many studies [12, 13, 20]. As such, it is our objective to see if the empirical covariance $\mathbf{S}$ would improve the performance with a proper covariance modeling motivated from the minimum divergence estimation perspective.

Table I and Table II show the performance of the three speaker adaptation approaches on SRE'10 and 12, respectively. Clearly, the *by-the-book* approach does not perform better than the other two approaches. The deficiency becomes much more significant in the SRE'12 core task where the number of enrollment sessions varies from one to over a hundred resulting in a much more intense covariance shrinkage than in SRE'10 where the number of enrollment sessions is fixed as eight. Comparing *MinDiv* to *Mean* only, results on SRE'12 show a clear benefit of including the empirical covariance $\mathbf{S}$ to the speaker model adaptation. However, this benefit is not significant on SRE'10. This may due to the fact that the enrollment utterances for a target speaker in SRE'12 include highly correlated segments (they might even include exact replicas of other sessions). In the 8conv-core task of SRE'10, the enrollment sessions were carefully selected to make sure that each of them is unique. Above all, the proposed *MinDiv* approach is far better than the *by-the-book* approach while the advantage over the *Mean* is marginal. For future work, further analysis on the use of the empirical covariance in the latent space will be conducted.

Table I: Comparison of three speaker adaptation approaches on CC 5 of NIST SRE'10 8conv-core task.

| | Male | | |
|---|---|---|---|
| | EER (%) | DCF10 | DCF12 |
| *By-the-book* | 0.8493 | 0.2476 | 0.1915 |
| *Mean* | 0.5194 | 0.1667 | 0.1446 |
| *MinDiv* | 0.7607 | 0.1905 | 0.1623 |
| | Female | | |
| | EER (%) | DCF10 | DCF12 |
| *By-the-book* | 2.9370 | 0.3289 | 0.2625 |
| *Mean* | 2.1379 | 0.3116 | 0.2546 |
| *MinDiv* | 2.4747 | 0.3720 | 0.3142 |

Table II: Comparison of three speaker adaptation approaches on CC 2 of NIST SRE'12 core task.

| | Male | | |
|---|---|---|---|
| | EER (%) | DCF10 | DCF12 |
| *By-the-book* | 6.8953 | 0.6015 | 0.5394 |
| *Mean* | 3.9395 | 0.4765 | 0.4065 |
| *MinDiv* | 3.5746 | 0.4238 | 0.3624 |
| | Female | | |
| | EER (%) | DCF10 | DCF12 |
| *By-the-book* | 6.4646 | 0.6338 | 0.5621 |
| *Mean* | 3.2145 | 0.5382 | 0.4440 |
| *MinDiv* | 3.0597 | 0.5235 | 0.4292 |

## 6. CONCLUSION

This paper presented an initial work on solving the multi-session PLDA scoring from the perspective of model adaptation. We showed that the PLDA scoring function could be formulated as model comparison between the speaker-dependent PLDA model and the *universal* PLDA, much similar to the classical GMM-UBM. Based on this formulation, we propose a speaker adaptation method through a minimum divergence estimate of speaker prior. Experimental results show that this speaker adaptation method is effective in handling multi-session task, especially, when large number of enrolment i-vectors is available. Notably, it is free from the covariance shrinkage problem typical to the standard *by-the-book* multi-session PLDA scoring.

## 7. ACKNOWLEDGEMENTS

The work of Liping Chen was partially funded by the National Nature Science Foundation of China (Grant No. 61273264) and the National 973 program of China (Grant No. 2012CB326405).

## 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] 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.

[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, 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.

[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] 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.

[7] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no.5, pp. 1136 - 1159, May 2013.

[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] H. Li, B. Ma, K. A. Lee, C. H. You, H. Sun, and A. Larcher, "IIR system description for the NIST 2012 speaker recognition evaluation," in *NIST SRE'12 Workshop*, Orlando, Dec. 2012.

[13] N. Brümmer, A. Swart, L Burget, S. Cumani, O. Glembek, M. Karafiát, P.Matejka, O.Plchot, M. Soufifar, J. Silovský, P. Kenny, J. Alam, P. Dumouchel, P. Ouellet, M. Senoussaoui and T. Stafylakis, "ABC system description for NIST SRE 2012," in *NIST SRE'12 Workshop*, Orlando, Dec. 2012.

[14] 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.

[15] 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, paper 198.

[16] 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.

[17] 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.

[18] N. Brummer, "EM for Probabilistic LDA," Technical Report, Feb. 2010, Available at https://sites.google.com/site/nikobrummer/.

[19] 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.

[20] R. Saeidi, K. A. Lee, T. Kinnunen, T. Hasan, and 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.