SNR-INVARIANT PLDA WITH MULTIPLE SPEAKER SUBSPACES

Na LI and Man-Wai MAK

Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR

lina011779@126.com, enmwmak@polyu.edu.hk

ABSTRACT

To deal with the mismatch between the enrollment and test utterances caused by noise with different signal-to-noise ratios (SNR), we have recently proposed an SNR-invariant PLDA model for robust speaker verification. In the model, SNR-specific information were separated from speaker-specific information through marginalizing out the SNR factors during the scoring process. However, this modeling approach assumes that speaker variabilities can be captured by a single speaker subspace regardless of the noise level of the utterances. We will show in this paper that i-vectors extracted from utterances with different noise levels will shift to different regions of the i-vector space and that i-vectors extracted from utterances having similar SNR tend to cluster together. In view of this observation, we propose introducing multiple speaker subspaces to the SNR-invariance PLDA model and use multiple covariance matrices to represent SNR-dependent channel variability. Through NIST 2012 SRE, this paper demonstrates that this finer and more precise modeling of speaker and SNR variabilities leads to better performance when compared with the conventional PLDA and SNR-invariant PLDA.

***Index Terms*—** i-vectors; SNR-invariant PLDA; speaker subspaces; SNR subspaces; speaker verification.

1. INTRODUCTION

How to develop speaker verification systems that can handle different levels of background noise has become a research focus in the speaker verification domain [1]. Many strategies have been proposed. Among them, some aim to derive more robust features from i-vectors [2–4], while others attempt to make the back-end classifiers more robust to noise [5–7].

While i-vector [8] extraction followed by probabilistic LDA (PLDA) [9] is very effective in addressing channel variability, the performance degrades rapidly in the presence of background noise with a wide range of SNR [10]. To improve the robustness of i-vector/PLDA systems, several methods have been proposed. In [11], clean and noisy utterances were pooled together to train a robust PLDA model. Garcia-Romero *et al.* [12] employed multi-condition training to train multiple PLDA models, one for each condition. A robust system was then constructed by combining all of the PLDA models according to the posterior probability of each condition. In [13], mixture of probabilistic PCA was performed on the feature space so that the posterior means of the mixture-dependent acoustic factors become the enhanced and normalized version of MFCC acoustic vectors. It was found that using these enhanced features to compute the first-order sufficient statistics in an i-vector extractor can improve the robustness of the resulting i-vectors. The idea is further enhanced in [14], where the UBM was replaced by a mixture of acoustic factor analyzers for i-vector extraction.

Focus was shifted to noise robust speaker verification in NIST 2012 SRE [15]. Many i-vector/PLDA systems, such as [16], perform very well in the evaluation. However, many of them use a single PLDA model to handle all of the test utterances regardless of their noise level. In [17], a mixture of SNR-dependent PLDA is proposed so that each mixture focuses on a small range of SNR. During verification, the mixtures cooperate with each other to deal with utterances of various noise levels. Unlike the conventional mixture of factor analyzers [18] where the posteriors of the indicator variables depend on the data samples, in [17], the posteriors of the indicator variables depend on the utterances' SNR. In [19, 20], mixture of PLDA with shared speaker space was used for verifying speakers from multiple channels. In [21, 22], the mixture of factor analyzers [23] is extended to mixture of PLDA in which the stacked i-vectors from multiple sessions of a speaker are assumed to be generated from a mixture of factor analyzers.

By assuming that i-vectors derived from utterances falling within a narrow SNR range should share similar SNR-specific information, we have recently proposed to add an SNR-subspace to the conventional PLDA models, resulting in SNR-invariant PLDA [6, 24]. With the SNR-subspace, the SNR-invariant PLDA can capture both speaker, noise-level, and channel variabilities embedded in the i-vectors. A limitation of SNR-invariant PLDA is that all i-vectors are assumed to live in the same region of the i-vector space, regardless of the noise level of utterances. In essence, the method assumes that all variabilities in the i-vectors due to noise-level variability occur *exclusively* in the SNR subspace and that all variabilities due to speaker differences occur in a *single* speaker subspace. This is certainly undesirable if noise-level variability not only causes the i-vectors of the same speaker to vary in a subspace but also shifts the i-vectors to different regions of the i-vector space. We will show in this paper that noise can shift the mean of i-vectors, with the degree of shift depends on the noise level. This phenomenon clearly violates the assumptions of SNR-invariant PLDA. Inspired by this mean-shift phenomenon, this paper proposes to incorporate multiple SNR-dependent speaker subspaces and SNR-dependent residue terms (representing channel variability) into the SNR-invariant PLDA.

The paper is organized as follows. Section 2 introduces the conventional PLDA and the SNR-subspace in the SNR-invariant PLDA. The notion of multiple speaker subspaces is explained in Section 3. The training and scoring algorithms of the resulting model will al-

This work was in part supported by The RGC of Hong Kong SAR, Grant No. PolyU 152068/15E.

so be described. The advantages of using multiple speaker subspace are demonstrated in Sections 4 and 5 through experiments on NIST 2012 SRE. Finally, Section 6 provides some concluding remarks and possible extensions of the proposed method.

2. BACKGROUND

2.1. Conventional PLDA

In the conventional i-vector/PLDA framework [25], an i-vector $\mathbf{x}_{ij}$ – the *j*-th i-vector from speaker *i* – is regarded as an observation generated from a linear model [26, 27]:

$$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{G}\mathbf{r}_{ij} + \boldsymbol{\epsilon}_{ij} \tag{1}$$

where $\mathbf{m}$ is the global i-vector mean, $\mathbf{V}$ defines the speaker subspace, $\mathbf{G}$ defines the channel subspace, $\mathbf{h}_i$ and $\mathbf{r}_{ij}$ are latent factors depending on the speaker and session respectively, and $\boldsymbol{\epsilon}_{ij}$ denotes a residual term which follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{\Sigma})$. Typically, $\boldsymbol{\Sigma}$ is a diagonal matrix aiming to model any remaining variabilities that cannot be described by $\mathbf{V}\mathbf{V}^{\mathsf{T}}$ and $\mathbf{G}\mathbf{G}^{\mathsf{T}}$.

2.2. SNR-invariant PLDA

To enhance the robustness of i-vector/PLDA, we have recently proposed an SNR-invariant PLDA model [6, 24] to deal with SNR mismatch. In this model, training utterances are first divided into K groups according to their SNR. As a result, each of the training i-vectors is associated with one SNR subgroup. Denote $\mathbf{x}_{ij}^k$ as the *j*-th i-vector from speaker *i* in the *k*-th SNR subgroup. Then, $\mathbf{x}_{ij}^k$ is expressed as:

$$\mathbf{x}_{ij}^k = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \boldsymbol{\epsilon}_{ij}^k, \quad k = 1, \ldots, K \tag{2}$$

where $\mathbf{m}$ is the global mean of i-vectors, $\mathbf{V}$ defines the speaker subspace, $\mathbf{h}_i$ is a latent speaker factor, $\mathbf{U}$ defines the SNR subspace, $\mathbf{w}_k$ is a latent SNR factor with a standard normal distribution, $\boldsymbol{\epsilon}_{ij}^k$ is a residual term with distribution $\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{\Sigma})$. In [6, 24], $\boldsymbol{\Sigma}$ is a full covariance matrix aiming to model the channel variability.

The key difference between the conventional PLDA (Eq. 1) and SNR-invariant PLDA (Eq. 2) is that the former uses a channel subspace ($\mathbf{G}$) to model channel variability, whereas the latter uses an SNR subspace ($\mathbf{U}$) to capture the variability due to noise level differences. As a result, the SNR latent factors ($\mathbf{w}_k$ in Eq. 2) depend on the SNR subgroups, whereas the session latent factors ($\mathbf{r}_{ij}$ in Eq. 1) depend on the speaker and sessions.

3. PROPOSED FRAMEWORK

3.1. SNR-dependent I-vectors

This paper is based on the hypothesis that i-vectors extracted from utterances of different SNRs locate in different regions of the i-vector space. To validate this hypothesis, we added babble noise to 7156 utterances from NIST 2005–2008 SREs at 6dB and 15dB. I-vectors were then extracted from the original (clean) utterances and the noise contaminated utterances. Fig. 1 displays the three groups of i-vectors on the first 3 principal components. Evidently, the i-vectors form three clusters and the locations of the clusters depend on the SNR level. In particular, the 6dB cluster (black) is further away from the clean cluster (blue) than the less noisy cluster (red). Moreover, the cluster shapes are also not identical, meaning that there exist multiple speaker subspaces within the i-vector space.

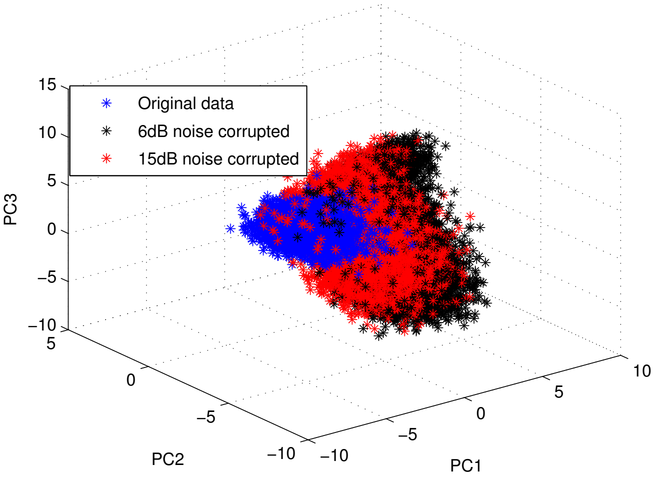

Fig. 1. Illustration of the mean-shift effect of i-vectors caused by different levels of background noise in the corresponding utterances.

3.2. Multiple Speaker Subspaces

Combining the mean-shift effect and SNR-dependent cluster shapes shown in Fig. 1, we propose to extend the SNR-invariant PLDA model as follows:

$$\mathbf{x}_{ij}^k = \mathbf{m}_k + \mathbf{V}_k\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \boldsymbol{\epsilon}_{ij}^k \quad k = 1, \ldots, K, \tag{3}$$

where $\mathbf{m}_k$ is to address the mean-shift effect and $\mathbf{V}_k$ represents the speaker subspace of the k-th SNR group. Moreover, unlike the models in Eq. 1 and Eq. 2, the covariance of the residual term also depends on the SNR group, i.e., ϵ_{ij}^k follows $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$.

3.3. EM Algorithm

To fully exploit the capability of the PLDA model defined in Eq. 3, it is necessary to divide the training i-vectors into multiple SNR groups so that each of the speaker subspaces can be estimated by more relevant i-vectors. Instead of using a manual division of SNR as in our earlier work [6], here, we propose to use a GMM to model the density of SNR of the training utterances and use the posterior probability of SNR given by the GMM to divide the training i-vectors into SNR groups. More specifically, given a K-mixture GMM, the k-th i-vector subgroup comprises the i-vectors whose corresponding SNRs have the highest posterior probability with respect to the k-th mixture in the GMM.

Denote $\boldsymbol{\theta} = \{\mathbf{m}_k, \mathbf{V}_k, \mathbf{U}, \boldsymbol{\Sigma}_k\}_{k=1}^K$ as the parameters of the proposed model. These parameters can be learned from a training set $\mathcal{X}$ using maximum likelihood estimation. Given an initial value $\boldsymbol{\theta}$, we aim to find a new estimate $\hat{\boldsymbol{\theta}}$ that maximizes the auxiliary function:

$$\begin{aligned}\mathbf{Q}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{h},\mathbf{w}}\Big[\sum\nolimits_{ikj} \ln\Big(p(\mathbf{x}_{ij}^k|\mathbf{h}_i, \mathbf{w}_k, \hat{\boldsymbol{\theta}})p(\mathbf{h}_i, \mathbf{w}_k)\Big)\Big|\mathcal{X}, \boldsymbol{\theta}\Big] \\ &= \mathbb{E}_{\mathbf{h},\mathbf{w}}\Big[\sum\nolimits_{ikj}\Big(\ln\mathcal{N}(\mathbf{x}_{ij}^k|\mathbf{m}_k + \mathbf{V}_k\mathbf{h}_i + \mathbf{U}\mathbf{w}_k, \boldsymbol{\Sigma}_k) \\ &\quad + \ln\mathcal{N}(\mathbf{h}_i|\mathbf{0}, \mathbf{I}) + \ln\mathcal{N}(\mathbf{w}_k|\mathbf{0}, \mathbf{I})\Big)\Big|\mathcal{X}, \boldsymbol{\theta}\Big].\end{aligned} \tag{4}$$

To maximize Eq. 4, we need to estimate the posterior distributions of the latent variables given the model parameters $\boldsymbol{\theta}$. Denote $H_i(k)$

as the number of training i-vectors from speaker i in the k-th i-vector subgroup, S as the number of training speakers, and $M_k = \sum_{i=1}^{S} H_i(k)$ as the number of training i-vectors falling in the k-th i-vector subgroup. Then, the E-step is as follows:[1]

$$\mathbf{L}_i^1 = \mathbf{I} + \sum_{k=1}^{K} H_i(k)\mathbf{V}_k^\top \boldsymbol{\Phi}_k^{-1}\mathbf{V}_k \tag{5}$$

$$\mathbf{L}_k^2 = \mathbf{I} + M_k\mathbf{U}^\top \boldsymbol{\Psi}_k^{-1}\mathbf{U} \tag{6}$$

$$\langle \mathbf{h}_i|\mathcal{X}\rangle = (\mathbf{L}_i^1)^{-1}\sum_{k=1}^{K}\mathbf{V}_k^\top \boldsymbol{\Phi}_k^{-1}\sum_{j=1}^{H_i(k)}(\mathbf{x}_{ij}^k - \mathbf{m}_k) \tag{7}$$

$$\langle \mathbf{w}_k|\mathcal{X}\rangle = (\mathbf{L}_k^2)^{-1}\mathbf{U}^\top \boldsymbol{\Psi}_k^{-1}\sum_{i=1}^{S}\sum_{j=1}^{H_i(k)}(\mathbf{x}_{ij}^k - \mathbf{m}_k) \tag{8}$$

$$\langle \mathbf{h}_i\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle = (\mathbf{L}_i^1)^{-1} + \langle \mathbf{h}_i|\mathcal{X}\rangle\langle \mathbf{h}_i|\mathcal{X}\rangle^\mathsf{T} \tag{9}$$

$$\langle \mathbf{w}_k\mathbf{w}_k^\mathsf{T}|\mathcal{X}\rangle = (\mathbf{L}_k^2)^{-1} + \langle \mathbf{w}_k|\mathcal{X}\rangle\langle \mathbf{w}_k|\mathcal{X}\rangle^\mathsf{T} \tag{10}$$

$$\langle \mathbf{w}_k\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle = \langle \mathbf{w}_k|\mathcal{X}\rangle\langle \mathbf{h}_i|\mathcal{X}\rangle^\mathsf{T} \tag{11}$$

$$\langle \mathbf{h}_i\mathbf{w}_k^\mathsf{T}|\mathcal{X}\rangle = \langle \mathbf{h}_i|\mathcal{X}\rangle\langle \mathbf{w}_k|\mathcal{X}\rangle^\mathsf{T} \tag{12}$$

where

$$\boldsymbol{\Phi}_k = \mathbf{U}\mathbf{U}^\top + \boldsymbol{\Sigma}_k \text{ and } \boldsymbol{\Psi}_k = \mathbf{V}_k\mathbf{V}_k^\top + \boldsymbol{\Sigma}_k,$$

and $\langle\cdot\rangle$ denotes expectation.

Given Eq. 5–Eq. 12, the model parameters $\boldsymbol{\theta}$ can be estimated via the M-step as follows:

$$\mathbf{m}_k = \frac{1}{M_k}\sum_{i=1}^{S}\sum_{j=1}^{H_i(k)}\mathbf{x}_{ij}^k \tag{13}$$

$$\begin{aligned}\mathbf{V}_k = &\left\{\sum_{i=1}^{S}\sum_{j=1}^{H_i(k)}\left[(\mathbf{x}_{ij}^k - \mathbf{m}_k)\langle \mathbf{h}_i|\mathcal{X}\rangle^\top - \mathbf{U}\langle \mathbf{w}_k\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle\right]\right\}\\ &\left\{\sum_{i=1}^{S}\sum_{j=1}^{H_i(k)}\langle \mathbf{h}_i\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle\right\}^{-1}\end{aligned} \tag{14}$$

$$\begin{aligned}\mathbf{U} = &\left\{\sum_{i=1}^{S}\sum_{k=1}^{K}\sum_{j=1}^{H_i(k)}\left[(\mathbf{x}_{ij}^k - \mathbf{m}_k)\langle \mathbf{w}_k|\mathcal{X}\rangle^\top - \mathbf{V}_k\langle \mathbf{h}_i\mathbf{w}_k^\mathsf{T}|\mathcal{X}\rangle\right]\right\}\\ &\left\{\sum_{i=1}^{S}\sum_{k=1}^{K}\sum_{j=1}^{H_i(k)}\langle \mathbf{w}_k\mathbf{w}_k^\mathsf{T}|\mathcal{X}\rangle\right\}^{-1}\end{aligned} \tag{15}$$

$$\begin{aligned}\boldsymbol{\Sigma}_k = &\frac{1}{M_k}\sum_{i=1}^{S}\sum_{j=1}^{H_i(k)}\Big[(\mathbf{x}_{ij}^k - \mathbf{m}_k)(\mathbf{x}_{ij}^k - \mathbf{m}_k)^\top\\ &- \mathbf{V}_k\langle \mathbf{h}_i|\mathcal{X}\rangle(\mathbf{x}_{ij}^k - \mathbf{m}_k)^\top - \mathbf{U}\langle \mathbf{w}_k|\mathcal{X}\rangle(\mathbf{x}_{ij}^k - \mathbf{m}_k)^\top\Big]\end{aligned} \tag{16}$$

3.4. Likelihood Ratio Scores

Given a test i-vector $\mathbf{x}_t$ and a target-speaker i-vector $\mathbf{x}_s$, we need to determine to which SNR groups they belong before calculating the verification score. If the SNRs of the test and target-speaker utterances are known, we may use the K-mixture GMM to determine the SNR groups. In case the SNRs are unknown, we can compare the Euclidean distances between the test/target i-vector with the K i-vector means $\mathbf{m}_k$, $k = 1, \ldots, K$, in Eq. 3.[2] Specifically, the test and target-speaker i-vectors are respectively classified to the k_t-th and k_s SNR groups if they are closest to $\mathbf{m}_{k_t}$ and $\mathbf{m}_{k_s}$ among the K i-vector means. Then, the likelihood ratio score can be computed as follows:

$$\begin{aligned}S_{\text{LR}}(\mathbf{x}_s, \mathbf{x}_t) &= \ln\frac{p\left(\mathbf{x}_s, \mathbf{x}_t|\text{same-speaker}\right)}{p(\mathbf{x}_s, \mathbf{x}_t|\text{different-speakers})}\\ &= \ln\frac{\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s\\ \mathbf{x}_t\end{bmatrix}\middle|\begin{bmatrix}\mathbf{m}_{k_s}\\ \mathbf{m}_{k_t}\end{bmatrix}, \begin{bmatrix}\mathbf{A}_{k_s} & \mathbf{B}_{k_sk_t}\\ \mathbf{B}_{k_sk_t}^\top & \mathbf{A}_{k_t}\end{bmatrix}\right)}{\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s\\ \mathbf{x}_t\end{bmatrix}\middle|\begin{bmatrix}\mathbf{m}_{k_s}\\ \mathbf{m}_{k_t}\end{bmatrix}, \begin{bmatrix}\mathbf{A}_{k_s} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{k_t}\end{bmatrix}\right)}\end{aligned} \tag{17}$$

where

$$\begin{aligned}\mathbf{A}_{k_s} &= \mathbf{V}_{k_s}\mathbf{V}_{k_s}^\top + \mathbf{U}\mathbf{U}^\top + \boldsymbol{\Sigma}_{k_s},\\ \mathbf{A}_{k_t} &= \mathbf{V}_{k_t}\mathbf{V}_{k_t}^\top + \mathbf{U}\mathbf{U}^\top + \boldsymbol{\Sigma}_{k_t}, \text{ and}\\ \mathbf{B}_{k_sk_t} &= \mathbf{V}_{k_s}\mathbf{V}_{k_t}^\top.\end{aligned} \tag{18}$$

4. EXPERIMENTS

4.1. Evaluation Protocol and Speech Data

Experiments were performed on common conditions (CC) 4 and 5 of the core set of NIST 2012 Speaker Recognition Evaluation (SRE) [15]. We used data from NIST 2005–2010 for system development. The speech data were divided into the following parts:

- *Test Data*: CC4 in NIST 2012 SRE comprises noise-contaminated test utterances with SNR ranges from 0dB to 50dB, and the telephone utterances in CC5 were recorded in noisy environments with SNR ranges from 10dB to 50dB. Readers may refer to [6] for the SNR distributions of test utterances in these common conditions and the procedures for measuring the SNRs.
- *Enrollment Data*: Enrollment data comprises the telephone segments of target speakers. Each target speaker has one or more segments recorded over different channels and with different durations longer than 10 seconds.[3]
- *Development Data*: Development data were used for training the subspace projection matrices (LDA and WCCN), PLDA, SNR-invariant PLDA (S-PLDA) and the proposed model. The data comprise two parts. One part includes the telephone and microphone segments in 2005–2010 SREs. The other part comprises noise-corrupted telephone segments with different SNRs. The details of how to obtain these noisy speech can be found in Section IV-B in [6]. There are totally 14226 (resp. 22356) noise corrupted files from 763 male (resp. 1030 female) speakers in the development data. The "actual" SNRs of the training data were estimated using the voltmeter function of FaNT and the speech/non-speech decisions of our VAD [28, 29]. The microphone and telephone

[1] Full derivations of Eq. 5 to Eq. 18 can be found in http://bioinfo.eie.polyu.edu.hk/SI-PLDA/SuppMaterials.pdf.

[2] The Euclidean distances are based on the i-vectors and mean i-vectors before any i-vector processing such as whitening and length normalization.

[3] We have excluded enrollment utterances shorter than 10 seconds but ensure that every target speaker has at least one enrollment utterance.

Method	No. of SNR Groups (K)	CC4				CC5			
		Male		Female		Male		Female	
		EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
PLDA (Eq. 1)	–	3.39	0.325	3.10	0.354	2.80	0.303	2.34	0.331
S-PLDA (Eq. 2)	3	3.20	**0.300**	2.95	**0.327**	2.80	0.302	2.37	**0.319**
Proposed (Eq. 3)	2	3.31	0.302	3.09	0.333	**2.74**	**0.276**	2.36	0.350
	3	**3.06**	0.309	2.88	0.332	2.80	0.278	2.31	0.325
	4	3.12	0.316	**2.84**	0.339	2.79	0.284	**2.26**	0.321

Table 1. Performance of PLDA, SNR-invariant PLDA (S-PLDA), and the SNR-invariant PLDA with multiple speaker subspace (proposed) on CC4 and CC5 of NIST 2012 SRE (core set).

Model	Model Parameters	EER(%)	minDCF
1	$\boldsymbol{\theta}_1 = \{\mathbf{m}, \mathbf{V}, \mathbf{U}, \boldsymbol{\Sigma}\}$	3.20	**0.300**
2	$\boldsymbol{\theta}_2 = \{\mathbf{m}_k, \mathbf{V}_k, \mathbf{U}, \boldsymbol{\Sigma}_k\}$	3.06	0.309
3	$\boldsymbol{\theta}_3 = \{\mathbf{m}_k, \mathbf{V}, \mathbf{U}, \boldsymbol{\Sigma}\}$	3.30	0.305
4	$\boldsymbol{\theta}_4 = \{\mathbf{m}, \mathbf{V}_k, \mathbf{U}, \boldsymbol{\Sigma}_k\}$	3.15	0.308
5	$\boldsymbol{\theta}_5 = \{\mathbf{m}_k, \mathbf{V}, \mathbf{U}, \boldsymbol{\Sigma}_k\}$	3.57	0.319
6	$\boldsymbol{\theta}_6 = \{\mathbf{m}_k, \mathbf{V}_k, \mathbf{U}, \boldsymbol{\Sigma}\}$	**2.81**	0.332

Table 2. Performance of different SNR-invariant PLDA models on CC4 of NIST 2012 SRE (core set, male speakers). In the 2nd column, $k = 1, \ldots, K$, where K is the number of SNR groups.

segments from NIST 2005–2008 SREs were used as development data to train the gender-dependent UBMs and total variability matrices.

4.2. Acoustic Feature Extraction

For each speech segment, a two-channel VAD [28,29] was applied to prune out silence regions. Then the speech regions were segmented into 25-ms Hamming windowed frames with 10-ms frame shift. The first 19 Mel frequency cepstral coefficients (MFCC) with log energy were calculated with their first and second derivatives to form a 60-dimensional acoustic vector, followed by cepstral mean normalization and feature warping [30] with a window size of 3 seconds.

4.3. I-vector Extraction

The i-vector extractor is based on a gender-dependent UBM with 1024 mixtures and a total variability matrix with 500 total factors. Similar to [31], we applied within-class covariance normalization (WCCN) and i-vector length normalization (LN) to the 500-dimensional i-vectors. Then, linear discriminant analysis (LDA) [32] and WCCN were used to further reduce intra-speaker variability and reduce the dimension to 200. Then PLDA models, SNR-invariant PLDA models, and the proposed model with 150 latent identity factors were trained.

5. RESULTS AND DISCUSSIONS

We used equal error rate (EER) and minimum decision cost function (minDCF) defined in NIST 2012 SRE to evaluate the performance of different systems. Unless stated otherwise, the number of SNR groups for SNR-invariant PLDA models was set to 3 and the dimension of SNR factors in SNR-invariant PLDA and the proposed model was set to 40.

Table 1 shows that both the proposed model and SNR-invariant PLDA are superior to the conventional PLDA trained by pooling the original and noisy training data (multi-condition training). Comparing the proposed model and SNR-invariant PLDA in Table 1, we can see that the proposed model achieves a lower EER, while SNR-invariant PLDA achieves a lower minDCF in most situations.

To demonstrate that it is important to model the mean-shift effect in i-vectors, we implemented an SNR-invariant PLDA model that uses a global mean $(\mathbf{m})$ but with multiple speaker subspaces $(\mathbf{V}_k$'s) and compared its performance against the one that uses multiple mean i-vectors $(\mathbf{m}_k)$ and multiple speaker subspaces $(\mathbf{V}_k$'s). The results are shown in Table 2. Comparing Model 2 with Model 4 reveals that the multiple means $\mathbf{m}_k$'s are important because they can reduce the EER by 2.9% with only an insignificant increase (0.3%) in minDCF. The poor performance of Model 5 when compared with Model 2 suggests that once we have used multiple i-vector means to model the mean-shift effect, it is also necessary to use multiple speaker subspaces.

We also made the SNR-invariant PLDA model to share the same covariance matrix. Surprisingly, if minimizing EER is the goal, the result achieved by Model 6 in Table 2 suggests that it is not necessary to use SNR-dependent covariances $\boldsymbol{\Sigma}_k$'s. Because $\boldsymbol{\Sigma}$ aims to model channel variability, we conjecture that channel variability is not SNR dependent, and therefore it makes more sense to pool all data to estimate a single $\boldsymbol{\Sigma}$. This is an interesting area that requires further research.

6. CONCLUSIONS AND POSSIBLE EXTENSIONS

A new SNR-invariant PLDA model is presented. It is designed to improve the robustness of speaker verification systems when the test utterances exhibit a wide range of SNR. By introducing multiple eigenvoice matrices to SNR-invariant PLDA, speaker information can be captured and the effect of noise-level variability and channel variability can be largely suppressed. Experiments on the NIST 2012 SRE demonstrate the effectiveness of the proposed method.

While both SNR-invariant PLDA (S-PLDA) and SNR-dependent mixture of PLDA (mPLDA) [17, 33] aim to address SNR variability, they achieve this goal through different means. Specifically, the former uses an SNR subspace to capture SNR variability, whereas the latter uses multiple PLDA models to capture the SNR-dependent variabilities so that each PLDA model focuses on a narrow SNR range. Therefore, possible future work includes incorporating the SNR subspace into the mixture models in mPLDA.

7. REFERENCES

[1] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *Proc. ICASSP*, 2012, pp. 4249–4252.

[2] W. Ben Kheder, D. Matrouf, J.-F. Bonastre, M. Ajili, and P.-M. Bousquet, "Additive noise compensation in the i-vector space for speaker recognition," in *Proc. ICASSP*, 2015, pp. 4190–4194.

[3] K. K. George, C. S. Kumar, and A. Panda, "Cosine distance features for robust speaker verification," in *Proc. Interspeech*, 2015, pp. 234–238.

[4] C. Yu, A. Ogawa, M. Delcroix, T. Yoshioka, T. Nakatani, and J. H. Hansen, "Robust i-vector extraction for neural network adaptation in noisy environment," in *Proc. Interspeech*, 2015, pp. 2854–2857.

[5] P. Rajan, T. Kinnunen, and V. Hautamäki, "Effect of multicondition training on i-vector PLDA configurations for speaker recognition," in *Proc. Interspeech*, 2013, pp. 3694–3697.

[6] N. Li and M. W. Mak, "SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1648–1659, 2015.

[7] M. McLaren, Y. Lei, N. Scheffer, and L. Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions," in *Proc. Interspeech*, 2014, pp. 686–690.

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

[9] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.

[10] C. Yu, G. Liu, S. Hahm, and J. H. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Proc. ICASSP*, 2014, pp. 4017–4021.

[11] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. H. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation," in *Proc. ICASSP*, 2013, pp. 6783–6787.

[12] D. Garcia-Romero, X. Zhou, and C. Espy-Wilson, "Multicondition training of gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. ICASSP*, 2012, pp. 4257–4260.

[13] T. Hasan and J. Hansen, "Acoustic factor analysis for robust speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 842–853, 2013.

[14] ——, "Maximum likelihood acoustic factor analysis models for robust speaker verification in noise," *IEEE Transactions on Audio, Speech, And Language Processing*, vol. 22, no. 2, pp. 381–391, 2014.

[15] NIST, "The NIST year 2012 speaker recognition evaluation plan," *http://www.nist.gov/itl/iad/mig/sre12.cfm*, 2012.

[16] R. Saeidi, K. Lee, T. Kinnunen, T. Hasan, B. Fauve, P. Bousquet, E. Khoury, P. S. Martinez, J. Kua, C. You *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.

[17] M. W. Mak, "SNR-dependent mixture of PLDA for noise robust speaker verification," in *Interspeech'2014*, 2014, pp. 1855–1859.

[18] G. McLachlan and D. Peel, "Mixtures of factor analyzers," *Finite Mixture Models*, pp. 238–256, 2000.

[19] J. Villalba and E. Lleida, "Handling I-vectors from diferent recording condistions using multi-channel simplified PLDA in speaker recognition," in *Proc. ICASSP*, 2013, pp. 6763–6767.

[20] J. A. Villalba, M. Diez, A. Varona, and E. Lleida, "Handling recordings acquired simultaneously over multiple channels with PLDA." in *INTERSPEECH*, 2013, pp. 2509–2513.

[21] K. Simonchik, T. Pekhovsk, A. Shulipa, and A. Afanasyev, "Supervised mixture of plda models for cross-channel speaker verification," in *Interspeech'2012*, 2012.

[22] T. Pekhovsky and A. Sizov, "Comparison between supervised and unsupervised learning of probabilistic linear discriminant analysis mixture models for speaker verification," *Pattern Recognition Letters*, vol. 34, no. 11, pp. 1307–1313, 2013.

[23] Z. Ghahramani, G. E. Hinton *et al.*, "The em algorithm for mixtures of factor analyzers," Technical Report CRG-TR-96-1, University of Toronto, Tech. Rep., 1996.

[24] N. Li and M. W. Mak, "SNR-invariant PLDA modeling for robust speaker verification," in *Proc. Interspeech*, 2015, pp. 2317–2321.

[25] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. of Odyssey: Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

[26] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.

[27] S. J. Prince, *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, 2012.

[28] M. W. Mak and H. B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," *Computer, Speech and Language*, vol. 28, no. 1, pp. 295–313, Jan 2014.

[29] H. Yu and M. Mak, "Comparison of voice activity detectors for interview speech in NIST speaker recognition evaluation," in *Interspeech*, 2011, pp. 2353–2356.

[30] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Crete, Greece, Jun. 2001, pp. 213–218.

[31] M. McLaren, M. Mandasari, and D. Leeuwen, "Source normalization for language-independent speaker recognition using i-vectors," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012, pp. 55–61.

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

[33] M. W. Mak, "Fast scoring for mixture of PLDA in i-vector/plda speaker verification," in *Proc. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA ASC)*, Hong Kong, Dec. 2015.