The following publication M. Mak, X. Pang and J. Chien, "Mixture of PLDA for Noise Robust I-Vector Speaker Verification," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1, pp. 130-142, Jan. 2016 is available at https://doi.org/10.1109/TASLP.2015.2499038.

Mixture of PLDA for Noise Robust I-Vector Speaker Verification

Man-Wai Mak, *Senior Member, IEEE*, Xiaomin Pang, and Jen-Tzung Chien, *Senior Member, IEEE*

***Abstract*—In real-world environments, noisy utterances with variable noise levels are recorded and then converted to i-vectors for cosine distance or PLDA scoring. This paper investigates the effect of noise-level variability on i-vectors. It demonstrates that noise-level variability causes the i-vectors to shift, causing the noise contaminated i-vectors to form clusters in the i-vector space. It also demonstrates that optimal subspaces for discriminating speakers are noise-level dependent. Based on these observations, this paper proposes using signal-to-noise ratio (SNR) of utterances as guidance for training mixture of PLDA models. To maximize the coordination among the PLDA models, mixtures of PLDA models are trained simultaneously via an EM algorithm using the utterances contaminated with noise at various levels. For scoring, given a test i-vector, the marginal likelihoods from individual PLDA models are linearly combined by the posterior probabilities of the test utterance's SNR. Verification scores are the ratio of the marginal likelihoods. Results based on NIST 2012 SRE suggest that the SNR-dependent mixture of PLDA is not only suitable for the situations where the test utterances exhibit a wide range of SNR, but also beneficial for the test utterances with unknown SNR distribution. Supplementary materials containing full derivations of the EM algorithms and scoring functions can be found in http://bioinfo.eie.polyu.edu.hk/mPLDA/SuppMaterials.pdf.**

***Index Terms*—Speaker verification; i-vectors; probabilistic LDA; mixture of PLDA; noise robustness.**

I. Introduction

When a speaker verification system is applied in real-world scenarios, a major challenge is to make the system robust against acoustic environments with variable noise levels. Much effort has been dedicated to compensate for the effect of these variations. Some of them reduces the variability in the front-end processing stage and others focus on the backend classification stage. The former aims to (1) extract features that are less sensitive to noise [1, 2, 3], (2) develop feature transformation methods [4] that make the features more robust, and (3) suppress the noise in the original waveform through speech enhancement techniques [5]. While the effectiveness of these feature-based approaches has been demonstrated, recent studies have found that techniques that operate on the backend classification stage are more promising. Among them, the joint factor analysis (JFA) [6] and i-vector/PLDA framework [7, 8] have been by far the most successful.

In the i-vector approach, a single space called the total variability space is defined to model both the speaker and channel variability. The acoustic characteristics of an entire utterance are represented by a single low-dimension vector called the i-vector. Since the total variability space accounts for both speaker and channel (including background noise) variability, a lot of algorithms for the compensation of channel variability have been proposed. Classical statistical techniques such as linear discriminant analysis (LDA) [9] and within-class covariance normalization (WCCN) [10] have been applied [7, 11, 12]. Alternatively, by assuming that the i-vectors are produced by a generative model and that the priors on the model's latent variables follow a Gaussian distribution or Student's t distribution, the marginal likelihood ratio can be computed, leading to the Gaussian PLDA [13] and heavy-tailed PLDA [8], respectively.

A common approach to addressing noise robustness in the i-vector/PLDA framework is to use multi-condition training where clean and noisy utterances are pooled together [14, 15, 16, 17]. Alternatively, multiple PLDA models are trained, one for each condition [18]. Another idea is to model the effect of noise on i-vectors by an SNR subspace [19].

Recently, several new methods that are based on the i-vector/PLDA framework have been proposed. For example, in [20], mixture of probabilistic PCA was performed on the feature space so that the posterior means of the mixture-dependent acoustic factors can replace the MFCC acoustic vectors when computing the first-order sufficient statistics. These statistics are then plugged into an i-vector extractor. It was shown that the posterior means of acoustic factors are enhanced and normalized versions of the acoustic features, and thus improving the robustness of the i-vector extractor. In [21], the authors further enhanced the idea by replacing the UBM by a mixture of acoustic factor analyzers for i-vector extraction. In [22, 23], the mixture of factor analysers [24] 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 analysers. In [25, 26], mixture of PLDA with shared speaker space was used for verifying speakers from multiple channels.

In [27, 28], vector Taylor series (VTS) was used to adapt a clean UBM to fit noisy utterances. The resulting UBM was then used for i-vector extraction. The notion is to clean up the i-vectors so that they become independent of additive and convolutive noise. As an alternative approach to VTS, [29] used an unscented transform (UT) to approximate the nonlinearities between clean and noisy speech models in the cepstral domain. The unscented transform is expected to be more accurate than VTS when the distortions are far from locally linear.

In [30], the zero-order statistics of an i-vector extractor were

This project was in part supported by the Hong Kong RGC Grant No. PolyU 152117114E and G-YN18 and the Taiwan MOST Grant No. 103-2221-E-009-078-MY3.

© 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

replaced by the posterior probabilities of senones estimated by a convolutional neural network (CNN). The idea is based on the observation that the convolution and max-pooling operations in convolution neural networks (CNN) can reduce the distortion caused by noise. It was demonstrated that the performance of this CNN/i-vector framework is comparable to that of UBM/i-vector framework and that fusion of these two frameworks is very promising.

In a recent study [31], the uncertainty of noisy acoustic features was propagated into the i-vector extraction process in an attempt to marginalize out the effect of noise. This is achieved by expressing the posterior density of an i-vector in terms of the joint density of the clean and noisy acoustic features where the uncertainties of the noisy features are represented through the variances of the joint density. To account for all possible clean features, the joint density is marginalized over all possible clean acoustic features. The marginalized density is then plugged back into the posterior density of i-vectors, where the noise-robust i-vector is its posterior mean. This modified i-vector extraction method has shown potential for improving the robustness of speaker recognition especially in low SNR conditions

Noting that the actual distortion of i-vectors may not be Gaussian, Sadjadi *et al*. [32] and Li and Mak [19] replaced LDA by non-parametric discriminant analysis (NDA) that uses nearest-neighbor rule to estimate the between- and within-speaker scatter matrices. They found that NDA is more effective than the conventional LDA under noisy and channel degraded conditions.

Focus was shifted to noise robust speaker verification in NIST 2012 SRE [33]. Many i-vector/PLDA systems, such as [34], 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 [35], we argued that the PLDA models should focus on a small range of SNR to be effective and that they should cooperate with each other during verification. To these ends, SNR-dependent mixture of PLDA was proposed in [35]. Unlike the conventional mixture of factor analyzers [36] where the posteriors of the indicator variables depend on the data samples, in [35], the posteriors of the indicator variables depend on the SNR of the utterances. This enables the contributions of individual mixtures depend explicitly on the SNR and implicitly on the locations of the i-vectors in the i-vector space.

While the proposed method in [35] resembles multi-condition training described earlier, there are some important differences. The major difference is that its condition-dependent factor analyzers were trained simultaneously. Also, in [18], the verification scores from individual PLDA models are weighted by the posterior probability of the test condition (Eq. 4 of [18]), whereas the model in [35] computes the verification scores by incorporating the posterior of SNR of both the target-speaker's and test utterances into the marginal likelihood computation (Eq. 4 in [35]). This paper extends the SNR-dependent mixture of PLDA in [35] by the following four aspects:

1) Adding another form of mixture of PLDA in which the clusters structure is solely dependent on i-vectors.
2) Investigating the effect of noise-level variability on i-vectors to support the motivation of SNR-dependent mixture of PLDA.
3) Presenting graphical models, EM algorithms and scoring functions for SNR-independent [24] and SNR-dependent mixture of PLDA and comparing them in terms of performance on NIST 2012 SRE.
4) Testing the mixture of PLDA models using utterances with SNR different from that of the training utterances.
5) Full derivations of the EM algorithms and scoring functions for these models are provided in the supplementary materials of this paper (downloadable from the authors' website).

The paper is organized as follows. Section II justifies the motivation for SNR-dependent mixture of PLDA. Section III outlines the i-vector/PLDA framework for speaker verification. Sections IV and V describe the EM algorithms and scoring functions for SNR-independent and SNR-dependent mixture of PLDA, respectively. In Sections VI and VII, we report evaluations based on NIST 2012 SRE [33]. Sections VIII concludes the findings.

II. Motivation

In [24], clustering and dimensionality reduction were combined so that different regions of the input space were modeled by different local factor models. In [35], we applied mixture of PLDA to find multiple speaker subspaces from the i-vector space. The mixture model, however, is different from that of [24] in that the resulting clusters and speaker subspaces depend not only on the input i-vectors but also on the signal-to-noise ratio (SNR) of utterances. In essence, the SNR is used as additional information to guide the clustering and dimension reduction process so that more prominent clusters in the i-vector space can be formed. The idea of our mixture of PLDA models is based on two hypotheses on the effect of noise on i-vectors:

1) Different levels of background noise will cause the i-vectors to fall on different regions of the i-vector space (although the regions may be highly overlapped).
2) SNR variability negatively affects PLDA speaker recognition accuracy, but its effect can be mitigated by explicitly modelling the SNR-dependent speaker subspace through mixture of PLDA.

To verify these two hypotheses, we corrupted 7,156 clean telephone utterances from 763 male speakers with babble noise at 6dB and 15dB using the FaNT tool [37], which result in 3 sets of i-vectors: clean, 15dB, and 6dB. We refer to these sets as $\mathcal{X}_1$, $\mathcal{X}_2$, and $\mathcal{X}_3$, respectively. Then, we used $\mathcal{X}_k$ to find the mean vector $\boldsymbol{\tau}_k$ and covariance matrix $\boldsymbol{\Gamma}_k$ of the i-vectors in the three sets separately. A 3-mixture GMM was then constructed using these three sets of parameters. Because the number of vectors in each set is equal, the mixture coefficients are equal to 1/3.

We used partition coefficients (PC) and partition entropy coefficients (PE) [38] to quantify the cluster separability of the three groups of i-vectors. These coefficients are commonly

TABLE I
PARTITION COEFFICIENT (PC) AND PARTITION ENTROPY COEFFICIENT (PE) OF I-VECTOR CLUSTERS. NUMBER OF CLUSTERS K IS 3.

	I-vector Clusters	Ranges
Partition Coefficient (PC)	0.997	$[1/K, 1] \approx [0.333, 1]$
Partition Entropy (PE)	0.005	$[0, \log K] \approx [0, 1.099]$

used in assessing the quality of clustering results [38]. Specifically, PC and PE are defined as follows:

$$\mathrm{PC} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\zeta_{ik}^2 \tag{1}$$

$$\mathrm{PE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\zeta_{ik}\log\zeta_{ik}, \tag{2}$$

where N is the total number of i-vectors, K is the number of clusters and ζ_{ik} denotes the degree of membership of the vector $\mathbf{x}_i$ in cluster k. The degree of membership ζ_{ik} for i-vector $\mathbf{x}_i$ is the posterior probability:

$$\zeta_{ik} = \frac{\mathcal{N}(\mathbf{x}_i|\boldsymbol{\tau}_k, \boldsymbol{\Gamma}_k)}{\sum_{r=1}^{K}\mathcal{N}(\mathbf{x}_i|\boldsymbol{\tau}_r, \boldsymbol{\Gamma}_r)}.$$

The ranges of PC and PE are $[1/K, 1]$ and $[0, \log K]$, respectively. A value close to 1 for PC or a value close to 0 for PE indicates perfect clustering; on the other hand, PC closes to $1/K$ or PE closes to $\log K$ indicates the absence of clustering tendency [38].

Table I shows the partition coefficient (PC) and partition entropy coefficient (PE) of the i-vector clusters for $K = 3$. Being close to the upper bound 1.0 for the PC and close to the lower bound for the PE suggest that i-vectors with variable noise levels have clustering tendency, which means different noise levels shift the i-vectors to different positions in the i-vector space.

To verify the second hypothesis, we performed speaker identification (SID) experiments based on the three datasets $\mathcal{X}_1$, $\mathcal{X}_2$, and $\mathcal{X}_3$. Specifically, for each dataset, we used 5,717 utterances from 592 speakers to train a PLDA model, which results in three SNR-dependent PLDA models. Then, we used the remaining 1,439 utterances from 171 speakers in each group to test against the three models. This results in nine combinations of PLDA models and SNR groups, of which three are matched in training and test conditions and six are mismatched. The speakers for training the PLDA models and for SID tests are mutually exclusive. During SID tests, each of the test utterances was scored against all the other test utterances, which result in 1,438 PLDA scores. The scores corresponding to the same speakers were averaged and the speaker ID of the test utterance was identified by picking the ID corresponding to the maximum average score. Note that throughout the paper, we used a simplified variant of PLDA, commonly called Gaussian PLDA [13] or simplified PLDA [25]. See [39] for a comparison between various forms of PLDA.

TABLE II
SPEAKER IDENTIFICATION ACCURACY UNDER MATCHED (DIAGONAL) AND MISMATCHED (OFF DIAGONAL) TRAINING AND TEST CONDITIONS. FOR EACH DATASET, 80% OF THE DATA WERE USED FOR TRAINING AND THE REMAINING 20% WERE USED FOR TESTING.

		Test Data From		
		$\mathcal{X}_1$ (clean)	$\mathcal{X}_2$ (15dB)	$\mathcal{X}_3$ (6dB)
PLDA Training Data From	$\mathcal{X}_1$ (clean)	**95.6%**	83.7%	55.2%
	$\mathcal{X}_2$ (15dB)	93.9%	**93.9%**	83.7%
	$\mathcal{X}_3$ (6dB)	90.1%	93.3%	**88.5%**

The SID performances of these nine combinations are shown in Table II. Evidently, for each test group (column), the diagonal element is the largest, which is reasonable because the SNR condition of the training data matches that of the test data. The results in the first and third columns suggest that the SID accuracy gradually decreases when the SNR of the training data progressively deviates from that of the test data. More interestingly, the PLDA model trained with 6dB noisy utterances is fairly robust to the clean and 15dB test utterances, which suggests that as long as the model observes some noisy training data, it will perform reasonably well in both clean and noisy conditions. However, this is not the case when it is trained on clean data only. All of these evidences suggest that a mixture model in which each mixture component is optimized for a specific SNR condition should be able to handle more diversified test conditions.

III. THE I-VECTOR/PLDA FRAMEWORK

A. I-Vector Extraction

Unlike JFA which defines the speaker and channel spaces distinctively, the i-vector approach defines a low-dimensional total variability space that encompasses both speaker and channel variabilities. In the total variability space, each utterance is represented by an i-vector. Specifically, given the MFCCs of the t-th utterance, the speaker- and channel-dependent GMM-supervector $\boldsymbol{\mu}_t$ is written as [7]:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu} + \mathbf{T}\mathbf{x}_t \tag{3}$$

where $\boldsymbol{\mu}$ is a speaker- and channel-independent GMM-supervector formed by stacking the mean vectors of the universal background model (UBM) [40], $\mathbf{T}$ is a low-rank total variability matrix, and the posterior mean of $\mathbf{x}_t$ is the corresponding low-dimensional i-vector. The i-vector extractor is trained by approximate maximum-likelihood using all utterances in a training set without using speaker labels.

B. Generative Model of PLDA

Given a set of D-dimensional length-normalized [13] i-vectors $\mathcal{X} = \{\mathbf{x}_{ij}; i = 1, \ldots, N; j = 1, \ldots, H_i\}$ obtained from N training speakers each with H_i sessions, we estimate the latent speaker factors $\mathcal{Z} = \{\mathbf{z}_i;\ i = 1, \ldots, N\}$ and parameters $\boldsymbol{\omega} = \{\mathbf{m}, \mathbf{V}, \boldsymbol{\Sigma}\}$ of a factor analyzer [9]:

$$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{z}_i + \boldsymbol{\epsilon}_{ij}, \tag{4}$$

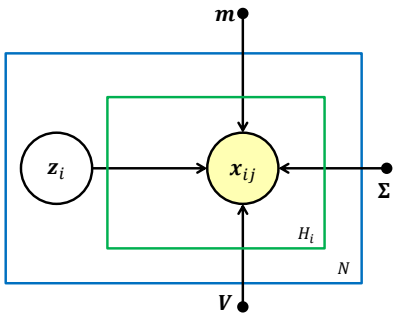


Fig. 1. Probabilistic graphical model representing PLDA with parameters $\boldsymbol{\omega} = \{\mathbf{m}, \mathbf{V}, \boldsymbol{\Sigma}\}$.

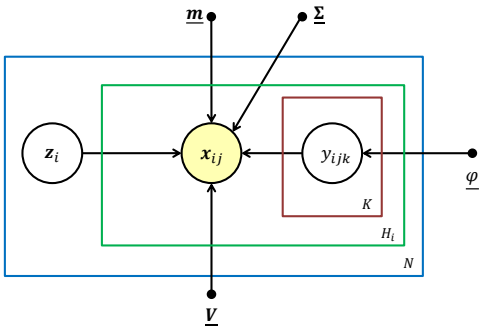


Fig. 2. Probabilistic graphical model representing SNR-independent mixture of PLDA with parameters $\underline{\boldsymbol{\omega}} = \{\varphi_k, \mathbf{m}_k, \boldsymbol{\Sigma}_k, \mathbf{V}_k\}_{k=1}^K$. In the diagram, $\underline{\mathbf{m}} = \{\mathbf{m}_k\}_{k=1}^K$, $\underline{\mathbf{V}} = \{\mathbf{V}_k\}_{k=1}^K$, $\underline{\boldsymbol{\Sigma}} = \{\boldsymbol{\Sigma}_k\}_{k=1}^K$, and $\underline{\varphi} = \{\varphi_k\}_{k=1}^K$.

where $\mathbf{V} \in \Re^{D\times M}$ is a factor loading matrix ($M < D$), $\mathbf{m} \in \Re^D$ is the global mean of $\mathcal{X}$, $\mathbf{z}_i \in \Re^M$ is the speaker factor with distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, M is the number of factors, and $\boldsymbol{\epsilon}_{ij}$'s are residual noise assumed to follow a Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ [13]. Fig. 1 shows the graphical model of PLDA with parameters $\boldsymbol{\omega} = \{\mathbf{m}, \mathbf{V}, \boldsymbol{\Sigma}\}$.

Because the i-vectors of the same speaker should share the same speaker factor in Eq. 4, we may collect the i-vectors of speaker i and rewrite Eq. 4 as

$$\tilde{\mathbf{x}}_i = \tilde{\mathbf{m}} + \tilde{\mathbf{V}}\mathbf{z}_i + \tilde{\boldsymbol{\epsilon}}_i, \tag{5}$$

where $\tilde{\mathbf{x}}_i = [\mathbf{x}_{i1}^\mathsf{T} \ \cdots \ \mathbf{x}_{iH_i}^\mathsf{T}]^\mathsf{T} \in \Re^{DH_i}$, $\tilde{\mathbf{m}} = [\mathbf{m}^\mathsf{T} \ \cdots \ \mathbf{m}^\mathsf{T}]^\mathsf{T} \in \Re^{DH_i}$, $\tilde{\mathbf{V}} = [\mathbf{V}^\mathsf{T} \ \cdots \ \mathbf{V}^\mathsf{T}]^\mathsf{T} \in \Re^{DH_i\times M}$, and $\tilde{\boldsymbol{\epsilon}}_i = [\boldsymbol{\epsilon}_{i1}^\mathsf{T} \ \cdots \ \boldsymbol{\epsilon}_{iH_i}^\mathsf{T}]^\mathsf{T} \in \Re^{DH_i}$. Eq. 5 is a factor analyzer whose parameters can be estimated via an EM algorithm [41, 42].

Given target-speaker's i-vector $\mathbf{x}_s$ and test i-vector $\mathbf{x}_t$, the likelihood ratio score is[1]

$$\begin{aligned} S_{\text{PLDA}}(\mathbf{x}_s, \mathbf{x}_t) &= \frac{p(\mathbf{x}_s, \mathbf{x}_t|\text{same-speaker})}{p(\mathbf{x}_s|\text{Spk } s)p(\mathbf{x}_t|\text{Spk } t)} \\ &= \frac{\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s^\mathsf{T} & \mathbf{x}_t^\mathsf{T}\end{bmatrix}^\mathsf{T} \middle| \begin{bmatrix}\mathbf{m}^\mathsf{T} & \mathbf{m}^\mathsf{T}\end{bmatrix}^\mathsf{T}, \hat{\mathbf{V}}\hat{\mathbf{V}}^\mathsf{T} + \hat{\boldsymbol{\Sigma}}\right)}{\mathcal{N}\left(\mathbf{x}_s|\mathbf{m}, \mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma}\right)\mathcal{N}\left(\mathbf{x}_t|\mathbf{m}, \mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma}\right)} \end{aligned} \tag{6}$$

where $\hat{\mathbf{V}} = [\mathbf{V}^\mathsf{T} \ \ \mathbf{V}^\mathsf{T}]^\mathsf{T}$ and $\hat{\boldsymbol{\Sigma}} = \text{diag}\{\boldsymbol{\Sigma}, \boldsymbol{\Sigma}\}$.

IV. SNR-Independent Mixture of PLDA

This section details a PLDA mixture model in which the posteriors of mixtures are independent of the SNR of utterances. Essentially, the model incorporates supervised learning to the mixture of factor analysers [24].

A. Generative Model

The PLDA model in Eq. 4 assumes that the length-normalized i-vectors follow a Gaussian distribution. However, to deal with cross channel tasks or tasks with varying noise and reverberation levels, the assumption of single Gaussian is rather limited. In such situations, the i-vectors will be better modeled by a mixture of K factor analyzers [24] with parameters $\underline{\boldsymbol{\omega}} = \{\varphi_k, \mathbf{m}_k, \boldsymbol{\Sigma}_k, \mathbf{V}_k\}_{k=1}^K$, where φ_k's are the mixture weights. More precisely, i-vectors are considered to be generated by a linear weighted sum of K Gaussian densities, each with its own mean vector, covariance matrix, and speaker subspace. Fig. 2 shows the graphical model of the SNR-independent mixture of PLDA (SI-mPLDA) with parameters $\underline{\boldsymbol{\omega}}$. Hereafter, we use the underline symbol to represent the set of hyper-parameters of a mixture model.

B. Likelihood Ratio Scores

Given the target-speaker's i-vector $\mathbf{x}_s$ and a test i-vector $\mathbf{x}_t$, the same-speaker marginal likelihood is

$$\begin{aligned} &p(\mathbf{x}_s, \mathbf{x}_t|\text{same-speaker}) \\ &= \sum_{k_s=1}^K \sum_{k_t=1}^K \int p(\mathbf{x}_s, \mathbf{x}_t, y_{k_s} = 1, y_{k_t} = 1, \mathbf{z}|\underline{\boldsymbol{\omega}})\text{d}\mathbf{z} \\ &= \sum_{k_s=1}^K \sum_{k_t=1}^K P(y_{k_s} = 1, y_{k_t} = 1|\underline{\boldsymbol{\omega}}) \\ &\qquad \times \int p(\mathbf{x}_s, \mathbf{x}_t|y_{k_s} = 1, y_{k_t} = 1, \mathbf{z}, \underline{\boldsymbol{\omega}})p(\mathbf{z})\text{d}\mathbf{z} \\ &= \sum_{k_s=1}^K \sum_{k_t=1}^K \varphi_{k_s}\varphi_{k_t} \int p(\mathbf{x}_s, \mathbf{x}_t|y_{k_s} = 1, y_{k_t} = 1, \mathbf{z}, \underline{\boldsymbol{\omega}})p(\mathbf{z})\text{d}\mathbf{z} \\ &= \sum_{k_s=1}^K \sum_{k_t=1}^K \varphi_{k_s}\varphi_{k_t} \\ &\quad \times \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s^\mathsf{T} & \mathbf{x}_t^\mathsf{T}\end{bmatrix}^\mathsf{T} \middle| \begin{bmatrix}\mathbf{m}_{k_s}^\mathsf{T} & \mathbf{m}_{k_t}^\mathsf{T}\end{bmatrix}^\mathsf{T}, \hat{\mathbf{V}}_{k_sk_t}\hat{\mathbf{V}}_{k_sk_t}^\mathsf{T} + \hat{\boldsymbol{\Sigma}}_{k_sk_t}\right) \end{aligned} \tag{7}$$

where $\hat{\boldsymbol{\Sigma}}_{k_sk_t} = \text{diag}\{\boldsymbol{\Sigma}_{k_s}, \boldsymbol{\Sigma}_{k_t}\}$, $\hat{\mathbf{V}}_{k_sk_t} = [\mathbf{V}_{k_s}^\mathsf{T} \ \ \mathbf{V}_{k_t}^\mathsf{T}]^\mathsf{T}$, and y_{k_s} and y_{k_t} are indicator variables indicating which of the K mixtures generates $\mathbf{x}_s$ and $\mathbf{x}_t$, respectively. Similarly, the different-speaker marginal likelihood is

$$p(\mathbf{x}_s, \mathbf{x}_t|\text{different-speaker}) = p(\mathbf{x}_s|\text{Spk } s)p(\mathbf{x}_t|\text{Spk } t),$$

where

$$\begin{aligned} p(\mathbf{x}_s|\text{Spk } s) &= \sum_{k_s=1}^K P(y_{k_s} = 1|\underline{\boldsymbol{\omega}}) \int p(\mathbf{x}_s|y_{k_s} = 1, \mathbf{z}, \underline{\boldsymbol{\omega}})\text{d}\mathbf{z} \\ &= \sum_{k_s=1}^K \varphi_{k_s}\mathcal{N}\left(\mathbf{x}_s|\mathbf{m}_{k_s}, \mathbf{V}_{k_s}\mathbf{V}_{k_s}^\mathsf{T} + \boldsymbol{\Sigma}_{k_s}\right), \end{aligned}$$

[1]We may also treat this as a kind of hypothesis test problem with the null hypothesis H_0 defined for the same speaker and the alternative hypothesis H_1 defined for different speakers.

$$S_{\text{SI-mPLDA}}(\mathbf{x}_s,\mathbf{x}_t)=\frac{\sum_{k_s=1}^{K}\sum_{k_t=1}^{K}\varphi_{k_s}\varphi_{k_t}\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s^{\mathsf{T}} & \mathbf{x}_t^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}\middle|\begin{bmatrix}\mathbf{m}_{k_s}^{\mathsf{T}} & \mathbf{m}_{k_t}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}},\hat{\mathbf{V}}_{k_sk_t}\hat{\mathbf{V}}_{k_sk_t}^{\mathsf{T}}+\hat{\mathbf{\Sigma}}_{k_sk_t}\right)}{\left[\sum_{k_s=1}^{K}\varphi_{k_s}\mathcal{N}\left(\mathbf{x}_s|\mathbf{m}_{k_s},\mathbf{V}_{k_s}\mathbf{V}_{k_s}^{\mathsf{T}}+\mathbf{\Sigma}_{k_s}\right)\right]\left[\sum_{k_t=1}^{K}\varphi_{k_t}\mathcal{N}\left(\mathbf{x}_t|\mathbf{m}_{k_t},\mathbf{V}_{k_t}\mathbf{V}_{k_t}^{\mathsf{T}}+\mathbf{\Sigma}_{k_t}\right)\right]} \tag{8}$$

and similarly for $p(\mathbf{x}_t|\text{Spk } t)$. Therefore, the likelihood ratio $S_{\text{SI-mPLDA}}$ is given by Eq. 8.

C. EM Formulation

Denote $\mathcal{Y}=\{y_{ijk}\}_{k=1}^{K}$ as the set of latent indicator variables specifying which of the K factor analyzers $\underline{\boldsymbol{\omega}}=\{\varphi_k,\mathbf{m}_k,\mathbf{\Sigma}_k,\mathbf{V}_k\}_{k=1}^{K}$ produces $\mathbf{x}_{ij}$. Specifically, $y_{ijk}=1$ if the k-th factor analyzer produces $\mathbf{x}_{ij}$, and $y_{ijk}=0$ otherwise. Then, the auxiliary function for EM is

$$\begin{aligned}
&Q(\underline{\boldsymbol{\omega}'}|\underline{\boldsymbol{\omega}})=\mathbb{E}_{\mathcal{Y},\mathcal{Z}}\{\ln p(\mathcal{X},\mathcal{Y},\mathcal{Z}|\underline{\boldsymbol{\omega}'})|\mathcal{X},\underline{\boldsymbol{\omega}}\}\\
&=\mathbb{E}_{\mathcal{Y},\mathcal{Z}}\left\{\sum\nolimits_{ijk}y_{ijk}\ln\left[p(y_{ijk}|\underline{\boldsymbol{\omega}'})p(\mathbf{x}_{ij}|\mathbf{z}_i,\boldsymbol{\omega}_k')p(\mathbf{z}_i|\underline{\boldsymbol{\omega}'})\right]\middle|\mathcal{X},\underline{\boldsymbol{\omega}}\right\}\\
&=\sum_{ijk}\mathbb{E}_{\mathcal{Y},\mathcal{Z}}\left\{y_{ijk}\ln\left[\varphi_k'\mathcal{N}(\mathbf{x}_{ij}|\mathbf{m}_k'+\mathbf{V}_k'\mathbf{z}_i,\mathbf{\Sigma}_k')\mathcal{N}(\mathbf{z}_i|\mathbf{0},\mathbf{I})\right]\middle|\mathcal{X},\underline{\boldsymbol{\omega}}\right\}.
\end{aligned} \tag{9}$$

Note that the true posterior of $\mathbf{z}$ and y are not independent, making computation slow or intractable. A practical solution is to use the variational Bayesian (VB) inference procedure where a factorized variational distribution over latent variables $\mathbf{z}_i$ and y_{ijk} can be assumed, i.e., $q(\mathbf{z}_i,y_{ijk})=q(\mathbf{z}_i)q(y_{ijk})$. VB estimates the factorized variational distribution which is closest to the true joint posterior distribution of two dependent latent variables $p(\mathbf{z}_i,y_{ijk}|\mathcal{X})$. In the VB-E step, we estimate the optimal variational distribution or variational parameters with the highest lower bound of likelihood function (also known as the variational lower bound). Given the updated variational distribution, in the VB-M step, the model parameters $\{\varphi_k,\mathbf{m}_k,\mathbf{V}_k,\mathbf{\Sigma}_k\}_{k=1}^{K}$ are estimated by further optimizing the variational lower bound. Instead of using the more complicated VB approach, however, we made a gentle assumption that the latent variable $\mathbf{z}_i$ is posteriorly independent of y_{ijk}, i.e., $p(\mathbf{z}_i,y_{ijk}|\mathcal{X}_i)=p(\mathbf{z}_i|\mathcal{X}_i)p(y_{ijk}|\mathbf{x}_{ij})$. This assumption is similar to that of traditional CDHMM [43] in which the HMM states and Gaussian mixtures are also assumed posteriorly independent.

With some mathematical manipulations, the following EM formulation can be derived:[2]

E-Step:

$$\begin{aligned}
&\langle y_{ijk}|\mathbf{x}_{ij}\rangle\equiv\mathbb{E}_{\mathcal{Y}}\left\{y_{ijk}|\mathbf{x}_{ij},\underline{\boldsymbol{\omega}}\right\}\\
&\qquad=\frac{\varphi_k\mathcal{N}(\mathbf{x}_{ij}|\mathbf{m}_k,\mathbf{V}_k\mathbf{V}_k^{\mathsf{T}}+\mathbf{\Sigma}_k)}{\sum_{k'=1}^{K}\varphi_{k'}\mathcal{N}(\mathbf{x}_{ij}|\mathbf{m}_{k'},\mathbf{V}_{k'}\mathbf{V}_{k'}^{\mathsf{T}}+\mathbf{\Sigma}_{k'})}\\
&\mathbf{L}_i=\mathbf{I}+\sum_{k=1}^{K}H_{ik}\mathbf{V}_k^{\mathsf{T}}\mathbf{\Sigma}_k^{-1}\mathbf{V}_k\\
&\langle y_{ijk}\mathbf{z}_i|\mathcal{X}_i\rangle\equiv\mathbb{E}_{\mathcal{Y},\mathcal{Z}}\left\{y_{ijk}\mathbf{z}_i|\mathcal{X}_i,\underline{\boldsymbol{\omega}}\right\}=\langle y_{ijk}|\mathcal{X}_i\rangle\langle\mathbf{z}_i|\mathcal{X}_i\rangle\\
&\langle\mathbf{z}_i|\mathcal{X}_i\rangle=\mathbf{L}_i^{-1}\sum_{k=1}^{K}\mathbf{V}_k^{\mathsf{T}}\mathbf{\Sigma}_k^{-1}\sum_{j\in\mathcal{H}_{ik}}(\mathbf{x}_{ij}-\mathbf{m}_k)\\
&\langle y_{ijk}\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X}_i\rangle=\langle y_{ijk}|\mathbf{x}_{ij}\rangle\langle\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X}_i\rangle\\
&\langle\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X}_i\rangle=\mathbf{L}_i^{-1}+\langle\mathbf{z}_i|\mathcal{X}_i\rangle\langle\mathbf{z}_i|\mathcal{X}_i\rangle^{\mathsf{T}}
\end{aligned} \tag{10}$$

[2]See the Supplementary Materials for full derivations.

M-Step:

$$\begin{aligned}
&\mathbf{m}_k'=\frac{\sum_{i=1}^{N}\sum_{j=1}^{H_i}\langle y_{ijk}|\mathbf{x}_{ij}\rangle\mathbf{x}_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{H_i}\langle y_{ijk}|\mathbf{x}_{ij}\rangle}\quad\varphi_k'=\frac{\sum_{ij}\langle y_{ijk}|\mathbf{x}_{ij}\rangle}{\sum_{ijl}\langle y_{ijl}|\mathbf{x}_{ij}\rangle}\\
&\mathbf{V}_k'=\left[\sum_{ij}(\mathbf{x}_{ij}-\mathbf{m}_k')\langle y_{ijk}\mathbf{z}_i|\mathcal{X}_i\rangle^{\mathsf{T}}\right]\left[\sum_{ij}\left\langle y_{ijk}\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X}_i\right\rangle\right]^{-1}\\
&\mathbf{\Sigma}_k'=\frac{\sum_{ij}\left[\langle y_{ij}|\mathbf{x}_{ij}\rangle\mathbf{f}_{ijk}'\mathbf{f}_{ijk}'^{\mathsf{T}}-\mathbf{V}_k'\langle y_{ijk}\mathbf{z}_i|\mathcal{X}_i\rangle\mathbf{f}_{ijk}'^{\mathsf{T}}\right]}{\sum_{ij}\langle y_{ijk}|\mathbf{x}_{ij}\rangle}\\
&\mathbf{f}_{ijk}'=\mathbf{x}_{ij}-\mathbf{m}_k',
\end{aligned} \tag{11}$$

where $\mathcal{H}_{ik}$ comprises the indexes of speaker i's i-vectors that aligned to mixture k and H_{ik} is the number of elements in $\mathcal{H}_{ik}$.

The posterior expectations $\langle\mathbf{z}_i|\mathcal{X}_i\rangle$ in Eq. 10 implies that given H_i i-vectors from speaker i, only H_{ik} of them are generated by the k-th mixture component. This property will be particularly important when the training i-vectors are derived from utterances with a wide range of SNR, because these i-vectors tend to fall on different regions of the i-vector space (see Hypothesis 1 in Section II). For example, in our experiments, training i-vectors were derived from utterances of three noise levels: clean, 15dB, and 6dB. As a result, when $K=3$, Mixtures 1, 2, and 3 will be responsible for generating i-vectors of clean, 15dB, and 6dB, respectively. This property also makes our mixture of PLDA models different from that of [23]. Specifically, in [23], given a set of i-vectors from a speaker, a mixture component is first chosen; then the selected mixture is responsible for generating all of the i-vectors from that speaker. Obviously, this structure is limited to the case where the i-vectors from a speaker are derived from similar acoustic environments with comparable noise level.

V. SNR-Dependent Mixture of PLDA

This section explains a new PLDA model which constitutes the main contribution of this work. A key difference between this model and the one described in Section IV is that the posteriors of mixtures depend on the SNR of utterances instead of the i-vectors. The SNR information enables the EM algorithm to find more distinct clusters in the i-vector space.

A. Generative Model

Based on the observation in Section II that different noise levels shift the i-vectors to different regions of the i-vector space, the i-vectors are better modeled by a mixture of SNR-dependent mixture of PLDA (SD-mPLDA) with parameters $\underline{\boldsymbol{\theta}}=\{\underline{\boldsymbol{\lambda}},\underline{\boldsymbol{\omega}}\}=\{\boldsymbol{\lambda}_k,\boldsymbol{\omega}_k\}_{k=1}^{K}=\{\pi_k,\mu_k,\sigma_k,\mathbf{m}_k,\mathbf{V}_k,\mathbf{\Sigma}_k\}_{k=1}^{K}$, where $\boldsymbol{\lambda}_k=\{\pi_k,\mu_k,\sigma_k\}$ contains the prior probability, mean and standard deviation of the SNR in the k-th group. In this model, clustering of i-vectors is guided by the SNR of the corresponding utterances and i-vectors are considered to be

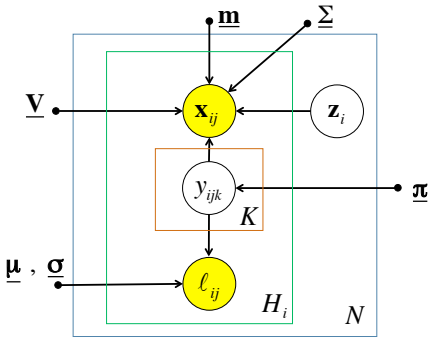


Fig. 3. Probabilistic graphical model representing SNR-dependent mixture of PLDA with parameters $\underline{\boldsymbol{\theta}} = \{\underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\omega}}\} = \{\boldsymbol{\lambda}_k, \boldsymbol{\omega}_k\}_{k=1}^K = \{\pi_k, \mu_k, \sigma_k, \mathbf{m}_k, \mathbf{V}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$. Refer to the caption of Fig. 2 for the definition of the symbol _.

generated by a linear combination of Gaussian densities in which the combination weights are the posteriors of utterances' SNR.

Denote ℓ as the SNR of the utterance whose i-vector is $\mathbf{x}$. Denote y_k's as the indicator variables specifying which of the factor analyzers is responsible for generating $\mathbf{x}$. Then, the posterior probability of y_k is

$$\gamma_\ell(y_k) \equiv P(y_k = 1|\ell, \underline{\boldsymbol{\lambda}}) = \frac{\pi_k \mathcal{N}(\ell|\mu_k, \sigma_k^2)}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(\ell|\mu_{k'}, \sigma_{k'}^2))}. \quad (12)$$

Therefore, unlike the SI-mPLDA described in Section IV, the alignment of i-vectors in SD-mPLDA is based on the posterior probabilities of SNR rather than the posterior probabilities of i-vectors.

Fig. 3 shows the graphical model of the SNR-dependent mixture of PLDA, where the subscripts i, j and k denote speaker, session, and mixture, respectively. The indicator variable y_{ijk} connects the i-vector $\mathbf{x}_{ij}$ and the corresponding SNR ℓ_{ij}. It indicates which of the factor analyzers $\boldsymbol{\omega}_k$ generates $\mathbf{x}_{ij}$. Note that unlike SI-mPLDA, y_{ijk} determines not only $\mathbf{x}_{ij}$ but also the SNR ℓ_{ij} which is modeled by a Gaussian mixture model with parameters $\underline{\boldsymbol{\lambda}} = \{\boldsymbol{\lambda}_k\}_{k=1}^K = \{\pi_k, \mu_k, \sigma_k\}_{k=1}^K$.

B. Likelihood Ratio Scores

Given target-speaker's i-vector $\mathbf{x}_s$ and test i-vector $\mathbf{x}_t$ and the SNR ℓ_s and ℓ_t (in dB) of the corresponding utterances, the same-speaker marginal likelihood is

$$\begin{aligned}
&p(\mathbf{x}_s, \mathbf{x}_t, \ell_s, \ell_t|\text{same-speaker}) \\
&= p(\ell_s)p(\ell_t)p(\mathbf{x}_s, \mathbf{x}_t|\ell_s, \ell_t, \text{same-speaker}) \\
&= p_{st} \sum_{k_s=1}^K \sum_{k_t=1}^K \int p(\mathbf{x}_s, \mathbf{x}_t, y_{k_s} = 1, y_{k_t} = 1, \mathbf{z}|\underline{\boldsymbol{\theta}}, \ell_s, \ell_t)\mathrm{d}\mathbf{z} \\
&= p_{st} \sum_{k_s=1}^K \sum_{k_t=1}^K \gamma_{\ell_s,\ell_t}(y_{k_s}, y_{k_t}) \\
&\qquad \times \int p(\mathbf{x}_s, \mathbf{x}_t|y_{k_s} = 1, y_{k_t} = 1, \mathbf{z}, \underline{\boldsymbol{\omega}})p(\mathbf{z})\mathrm{d}\mathbf{z} \\
&= p_{st} \sum_{k_s=1}^K \sum_{k_t=1}^K \gamma_{\ell_s,\ell_t}(y_{k_s}, y_{k_t}) \\
&\qquad \times \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s^\mathsf{T} & \mathbf{x}_t^\mathsf{T}\end{bmatrix}^\mathsf{T} \Big| \begin{bmatrix}\mathbf{m}_{k_s}^\mathsf{T} & \mathbf{m}_{k_t}^\mathsf{T}\end{bmatrix}^\mathsf{T}, \hat{\mathbf{V}}_{k_s k_t}\hat{\mathbf{V}}_{k_s k_t}^\mathsf{T} + \hat{\boldsymbol{\Sigma}}_k\right)
\end{aligned}$$

where $p_{st} = p(\ell_s)p(\ell_t)$, $\hat{\mathbf{V}}_{k_s k_t} = [\mathbf{V}_{k_s}^\mathsf{T} \ \mathbf{V}_{k_t}^\mathsf{T}]^\mathsf{T}$, $\hat{\boldsymbol{\Sigma}}_k = \text{diag}\{\boldsymbol{\Sigma}_{k_s}, \boldsymbol{\Sigma}_{k_t}\}$ and

$$\begin{aligned}
&\gamma_{\ell_s,\ell_t}(y_{k_s}, y_{k_t}) \equiv P(y_{k_s} = 1, y_{k_t} = 1|\ell_s, \ell_t, \underline{\boldsymbol{\lambda}}) \\
&= \frac{\pi_{k_s}\pi_{k_t}\mathcal{N}([\ell_s \ \ \ell_t]^\mathsf{T}|[\mu_{k_s} \ \ \mu_{k_t}]^\mathsf{T}, \text{diag}\{\sigma_{k_s}^2, \sigma_{k_t}^2\})}{\sum_{k'_s=1}^K \sum_{k'_t=1}^K \pi_{k'_s}\pi_{k'_t}\mathcal{N}([\ell_s \ \ \ell_t]^\mathsf{T}|[\mu_{k'_s} \ \ \mu_{k'_t}]^\mathsf{T}, \text{diag}\{\sigma_{k'_s}^2, \sigma_{k'_t}^2\})}.
\end{aligned}$$

Similarly, the different-speaker marginal likelihood is

$$\begin{aligned}
&p(\mathbf{x}_s, \mathbf{x}_t, \ell_s, \ell_t|\text{different-speaker}) \\
&\qquad = p(\mathbf{x}_s, \ell_s|\text{Spk } s)p(\mathbf{x}_t, \ell_t|\text{Spk } t),
\end{aligned}$$

where

$$\begin{aligned}
&p(\mathbf{x}_s, \ell_s|\text{Spk } s) = p(\ell_s) \sum_{k_s=1}^K \int p(\mathbf{x}_s, y_{k_s} = 1, \mathbf{z}|\underline{\boldsymbol{\theta}}, \ell_s)\mathrm{d}\mathbf{z} \\
&= p(\ell_s) \sum_{k_s=1}^K \gamma_{\ell_s}(y_{k_s}) \mathcal{N}\left(\mathbf{x}_s|\mathbf{m}_{k_s}, \mathbf{V}_{k_s}\mathbf{V}_{k_s}^\mathsf{T} + \boldsymbol{\Sigma}_{k_s}\right),
\end{aligned}$$

and similarly for $p(\mathbf{x}_t, \ell_t|\text{Spk } t)$. Therefore, the likelihood ratio $S_{\text{SD-mPLDA}}$ is given by Eq. 13 at the top of next page. Some issues of implementing Eq. 13 as Eq. 17 are addressed in Appendix A.

C. EM Formulation

Denote $\mathcal{Y} = \{y_{ijk}\}_{k=1}^K$ as the set of latent indicator variables specifying which of the K factor analyzers should be selected based on the SNR of training utterances. Also, denote $\mathcal{L} = \{\ell_{ij}; i = 1, \ldots, N; j = 1, \ldots, H_i\}$ as the SNR of the training utterances. Specifically, $y_{ijk} = 1$ if the k-th factor analyzer produces $\mathbf{x}_{ij}$, and $y_{ijk} = 0$ otherwise. Then, the auxiliary function for EM is

$$\begin{aligned}
Q(\underline{\boldsymbol{\theta}}'|\underline{\boldsymbol{\theta}}) &= \mathbb{E}_{\mathcal{Y},\mathcal{Z}}\{\log p(\mathcal{X}, \mathcal{L}, \mathcal{Y}, \mathcal{Z}|\underline{\boldsymbol{\theta}}')|\mathcal{X}, \mathcal{L}, \underline{\boldsymbol{\theta}}\} \\
&= \mathbb{E}_{\mathcal{Y},\mathcal{Z}}\Big\{\sum\nolimits_{ijk} y_{ijk} \log[p(\ell_{ij}|y_{ijk} = 1)p(y_{ijk}) \\
&\qquad \times p(\mathbf{x}_{ij}|\mathbf{z}_i, y_{ijk} = 1, \boldsymbol{\omega}'_k)p(\mathbf{z}_i)]\Big|\mathcal{X}, \mathcal{L}, \underline{\boldsymbol{\theta}}\Big\} \\
&= \sum_{i=1}^N \sum_{j=1}^{H_i} \sum_{k=1}^K \mathbb{E}_{\mathcal{Y},\mathcal{Z}}\{y_{ijk} \log[\mathcal{N}(\ell_{ij}|\mu'_k, \sigma'_k)\pi'_k \\
&\qquad \times \mathcal{N}(\mathbf{x}_{ij}|\mathbf{m}'_k + \mathbf{V}'_k\mathbf{z}_i, \boldsymbol{\Sigma}'_k)\mathcal{N}(\mathbf{z}_i|\mathbf{0}, \mathbf{I})]|\mathcal{X}, \mathcal{L}, \underline{\boldsymbol{\theta}}\}.
\end{aligned} \quad (14)$$

where $\pi'_k \equiv P(y_{ijk} = 1)$ is the prior probability of the k-th factor analyzer. Maximizing Eq. 14 leads to the following EM formulations:[3]

E-Step:

$$\begin{aligned}
&\langle y_{ijk}|\mathcal{L}\rangle \equiv \mathbb{E}_{\mathcal{Y}}\{y_{ijk}|\mathcal{L}, \underline{\boldsymbol{\lambda}}\} = \frac{\pi_k\mathcal{N}(\ell_{ij}|\mu_k, \sigma_k^2)}{\sum_{r=1}^K \pi_r\mathcal{N}(\ell_{ij}|\mu_r, \sigma_r^2)} \\
&\mathbf{L}_i = \mathbf{I} + \sum_{k=1}^K H_{ik}\mathbf{V}_k^\mathsf{T}\boldsymbol{\Sigma}_k^{-1}\mathbf{V}_k \\
&\langle y_{ijk}\mathbf{z}_i|\mathcal{X}, \mathcal{L}\rangle = \langle y_{ijk}|\mathcal{L}\rangle\langle\mathbf{z}_i|\mathcal{X}\rangle \\
&\langle\mathbf{z}_i|\mathcal{X}\rangle = \mathbf{L}_i^{-1}\sum_{k=1}^K \sum_{j\in\mathcal{H}_{ik}} \mathbf{V}_k^\mathsf{T}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_{ij} - \mathbf{m}_k) \\
&\langle y_{ijk}\mathbf{z}_i\mathbf{z}_i^\mathsf{T}|\mathcal{X}, \mathcal{L}\rangle = \langle y_{ijk}|\mathcal{L}\rangle\langle\mathbf{z}_i\mathbf{z}_i^\mathsf{T}|\mathcal{X}\rangle \\
&\langle\mathbf{z}_i\mathbf{z}_i^\mathsf{T}|\mathcal{X}\rangle = \mathbf{L}_i^{-1} + \langle\mathbf{z}_i|\mathcal{X}\rangle\langle\mathbf{z}_i|\mathcal{X}\rangle^\mathsf{T}
\end{aligned} \quad (15)$$

[3]See Appendix B and Supplementary Materials for full derivations.

$$S_{\text{SD-mPLDA}}(\mathbf{x}_s, \mathbf{x}_t) = \frac{\sum_{k_s=1}^{K}\sum_{k_t=1}^{K}\gamma_{\ell_s,\ell_t}(y_{k_s}, y_{k_t})\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s^{\mathsf{T}} & \mathbf{x}_t^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} \middle| \begin{bmatrix}\mathbf{m}_{k_s}^{\mathsf{T}} & \mathbf{m}_{k_t}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}, \hat{\mathbf{V}}_{k_sk_t}\hat{\mathbf{V}}_{k_sk_t}^{\mathsf{T}} + \hat{\mathbf{\Sigma}}_{k_sk_t}\right)}{\left[\sum_{k_s=1}^{K}\gamma_{\ell_s}(y_{k_s})\mathcal{N}\left(\mathbf{x}_s|\mathbf{m}_{k_s}, \mathbf{V}_{k_s}\mathbf{V}_{k_s}^{\mathsf{T}} + \mathbf{\Sigma}_{k_s}\right)\right]\left[\sum_{k_t=1}^{K}\gamma_{\ell_t}(y_{k_t})\mathcal{N}\left(\mathbf{x}_t|\mathbf{m}_{k_t}, \mathbf{V}_{k_t}\mathbf{V}_{k_t}^{\mathsf{T}} + \mathbf{\Sigma}_{k_t}\right)\right]} \tag{13}$$

M-Step:

$$\begin{aligned}
\mathbf{m}'_k &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{H_i}\langle y_{ijk}|\mathcal{L}\rangle\mathbf{x}_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{H_i}\langle y_{ijk}|\mathcal{L}\rangle}; \pi'_k = \frac{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle}{\sum_{ijl}\langle y_{ijl}|\mathcal{L}\rangle} \\
\mu'_k &= \frac{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle\ell_{ij}}{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle}; \quad \sigma_k^{2\prime} = \frac{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle(\ell_{ij}-\mu'_k)^2}{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle} \\
\mathbf{V}'_k &= \left[\sum_{ij}\mathbf{f}'_{ijk}\langle y_{ijk}\mathbf{z}_i|\mathcal{X},\mathcal{L}\rangle^{\mathsf{T}}\right]\left[\sum_{ij}\left\langle y_{ijk}\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X},\mathcal{L}\right\rangle\right]^{-1} \\
\mathbf{\Sigma}'_k &= \frac{\sum_{ij}\left[\langle y_{ijk}|\mathcal{L}\rangle\mathbf{f}'_{ijk}\mathbf{f}'_{ijk}{}^{\mathsf{T}} - \mathbf{V}'_k\langle y_{ijk}\mathbf{z}_i|\mathcal{X},\mathcal{L}\rangle\mathbf{f}'_{ijk}{}^{\mathsf{T}}\right]}{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle} \\
\mathbf{f}'_{ijk} &= \mathbf{x}_{ij} - \mathbf{m}'_k.
\end{aligned} \tag{16}$$

Note that to be more precise, the posterior mean $\langle y_{ijk}|\mathcal{L}\rangle$ should be $\langle y_{ijk}|\mathbf{x}_{ij}, \ell_{ij}\rangle$. However, to ensure that the clustering processing is driven by the SNR of utterances rather than the i-vectors, we make a gentle assumption that y_{ijk} is posteriorly independent of $\mathbf{x}_{ij}$. Also, to avoid using the complicated VB approach to approximating the true posterior of y_{ijk} and $\mathbf{z}_i$, we have assumed that y_{ijk} and $\mathbf{z}_i$ are posteriorly independent (see Section IV-C for discussion).

Readers are suggested to compare the differences between PLDA (Fig 1), SNR-independent mPLDA (Fig 2) and SNR-dependent mPLDA (Fig 3) through their scoring functions (Eq. 6, Eq. 8, and Eq. 13.)

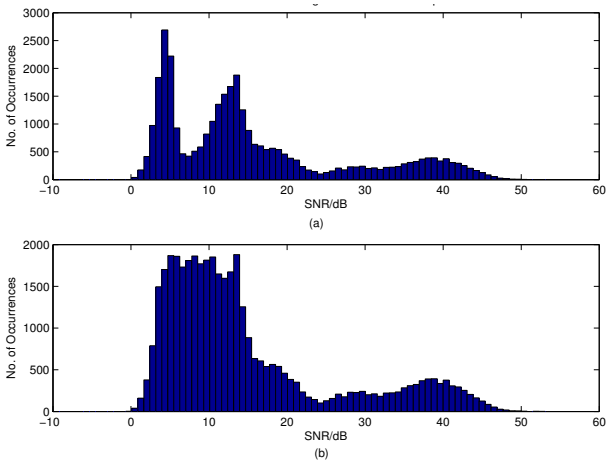


Fig. 4. SNR distribution of male training data for PLDA, SI-mPLDA, and SD-mPLDA models. *Top*: Set I. *Bottom*: Set II.

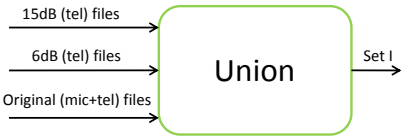


Fig. 5. Flow chart for obtaining Set I training data.

VI. Experimental Setup

A. Speech Corpora and Acoustic Features

The phonecall speech in the core set of NIST 2012 Speaker Recognition Evaluation (SRE) [33] was used for performance evaluation. In the evaluation dataset, noise was added to the test segments of common condition 4 and the test segments in common condition 5 were collected in noisy environments. Therefore, this paper focuses on these two common conditions. The training segments comprise conversations with variable length. We removed the 10-second utterances and the summed-channel utterances from the training segments but ensured that all target speakers have at least one clean utterance for enrollment. The speech files in NIST 2005–2010 SREs were used as development data for training gender-dependent UBMs, total variability matrices, LDA-WCCN, and PLDA models.

Speech regions in the speech files were extracted by using a two-channel VAD [44]. 19 MFCCs together with energy plus their 1st and 2nd derivatives were extracted from the speech regions, followed by cepstral mean normalization and feature warping [4] with a window size of 3 seconds. A 60-dim acoustic vector was extracted every 10ms, using a Hamming window of 25ms. For each clean training file, we randomly select one out of the 30 noise files from the PRISM dataset [45] and added to the speech file at a target SNR using the FaNT tool [37]. The target SNR was selected in turn from an SNR set comprising $\{6\text{dB}, 7\text{dB}, \ldots, 15\text{dB}\}$. As a result, for each original file, ten noise corrupted files with different SNRs were generated.

B. Preparation of Training Data

The i-vector systems are based on gender-dependent UBMs with 1024 mixtures and total variability matrices with 500 total factors. Microphone and telephone utterances from NIST 2005–2008 SREs were used for training the UBMs and total variability matrices. Following [12], within-class covariance normalization (WCCN) [10] and i-vector length normalization [13] were applied to the 500-dimensional i-vectors. Then, linear discriminant analysis (LDA) [9] and WCCN were applied to reduce the dimension to 200 before training the PLDA and mixture of PLDA models with 150 latent variables.

Two sets of training data were used for training the PLDA models. Fig. 5 and Fig. 6 show the flow charts for obtaining the training data. As shown in Fig. 5, Set I comprises 6dB (tel), 15dB (tel), and original (tel+mic) speech files in 2006–2010 SRE—excluding speakers with less than two utterances. The SNR distribution for Set I is shown in the upper panel of Fig. 4. It can be observed that the SNRs in the figure are not

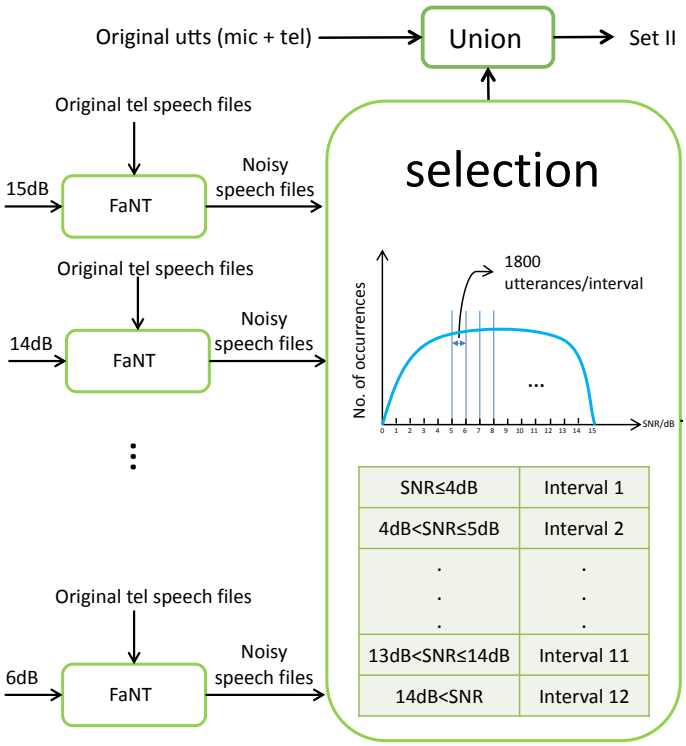


Fig. 6. Flow chart for obtaining Set II training data. The horizontal axis represents SNR ranging from 0 to 15dB.

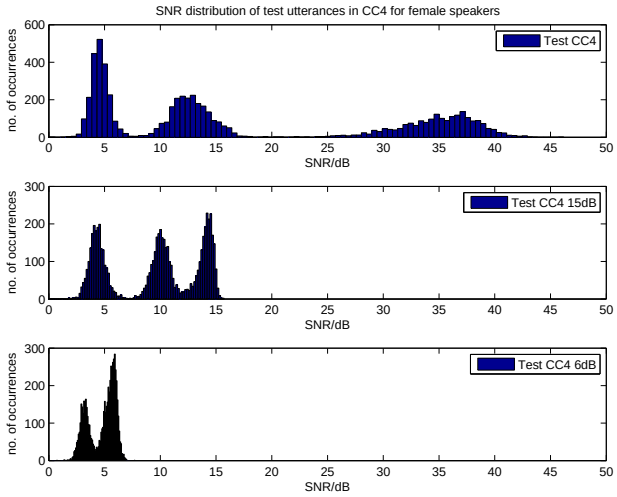


Fig. 7. SNR distribution of test utterances in CC4 for female speakers. *Top*: Original utterances. *Middle*: After adding noise at 15dB SNR. *Bottom*: After adding noise at 6dB SNR.

exactly 6dB or 15dB, instead they are only close to 6dB or 15dB because the figure shows the "actual" SNRs measured by using the VAD decisions and the voltmeter function of FaNT. Readers may refer to [19] for the details of SNR measurements.

The procedure for obtaining Set II is shown in Fig 6. All noise corrupted files were put together first and then 1800 utterances were randomly selected from each of the SNR intervals. The selected noisy utterances and the original

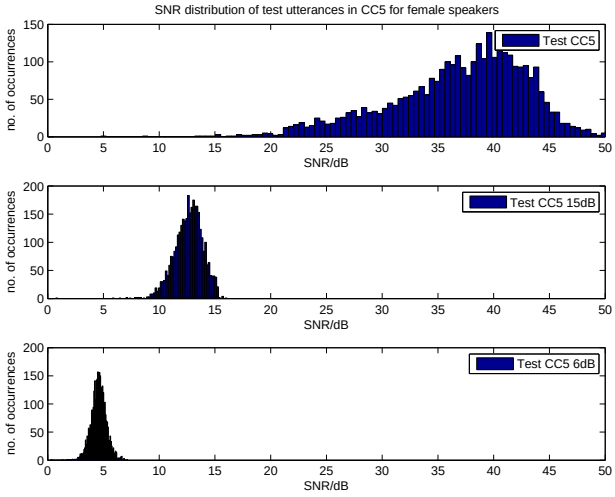


Fig. 8. SNR distribution of test utterances in CC5 for female speakers. *Top*: Original utterances. *Middle*: After adding noise at 15dB SNR. *Bottom*: After adding noise at 6dB SNR.

(tel+mic) utterances were combined to obtain Set II. The lower panel in Fig. 4 shows the SNR distribution of Set II. The reason of using Set II as training data is that this set of data is more general and is more suitable for the practical situations where no prior information about the test utterances is available.

C. Enrollment and Scoring

I-vectors derived from the original tel, 6dB tel and 15dB tel utterances were used for enrollment for all models. As a result, the number of enrollment utterances is three times the number of original telephone utterances, and each speaker will have multiple enrollment utterances. In our experiments, each enrollment utterance is represented by one target-speaker's i-vector. During scoring, a test i-vector was scored against each of the target-speaker's i-vectors and the scores were averaged. In other words, we used the score averaging approach.

VII. Results and Discussions

A. SNR-Independent vs. SNR-Dependent Mixture of PLDA

Gender-dependent PLDA, SI-mPLDA, and SD-mPLDA models with 150 factors were trained using Set I and Set II. The EM algorithms described in Sections IV and V were used to train SI-mPLDA andSD-mPLDA models with $K = 2, 3$ and 4.

The performance of PLDA, SNR-independent mixture of PLDA and SNR-dependent mixture of PLDA are shown in Table III, Table IV and Fig. 9. "Set I" and "Set II" in Table III refer to the two sets of training data described in Section VI-B.

Table III shows that SI-mPLDA performs better than PLDA in terms of minDCF for male speakers but its performance is poorer than that of the baseline for female speakers. In SI-mPLDA, the prior probability φ_k in Eq. 8 was determined by the posterior probability of the indicator variable given the training data (see Eq. 11). These prior probabilities are used as the prior for the K mixtures. This means that the

mixture weights for combining the PLDA scores in Eq. 8 are independent of the test utterances. In other words, the same combination weights will be used regardless of the characteristics of the test utterances. This leads to a very inflexible mixture of PLDA. On the other hand, in SNR-dependent mixture of PLDA (SD-mPLDA), the posterior probabilities of the indicator variables y_{ijk}'s given the SNR of test utterances are used to determine the combination weights in Eq. 13.

Figures 10(a) and 10(b) show the alignment of test i-vectors to SNR-independent and SNR-dependent mixture of PLDA models, respectively. In the figures, each point represents one i-vector in CC4 with the aligned Cluster ID and SNR shown on the two axes. The figures clearly show that for SD-mPLDA, i-vectors that are aligned to the same mixture component have similar SNR, whereas for SI-mPLDA, the i-vectors with a wide range of SNR could align to the same mixture. This suggests that it is beneficial to use the extra information available in the SNR to perform the alignment. The good alignment in SD-mPLDA increases the match between the mixture components and the test i-vectors, resulting in better verification performance.

From Table III, it can be observed that SD-mPLDA performs better than SI-mPLDA in most cases when Set I was used for training the mPLDA models. However, the situation is reverse when Set II was used for training. This is because the three SNR mixtures in Set I are more distinguishable than those in Set II (see Fig. 4), which helps SD-mPLDA to perform correct alignments of i-vectors to the PLDA mixtures.

Table III and Fig. 11 also show that PLDA and mixture of PLDA trained by Set II performs worse than those trained trained by Set I, which is caused by the mismatch between the training and the test data. However, it is more practical to use Set II to train PLDA models because in practice prior knowledge about the SNR distribution of test utterances is usually not available.

B. SNR as Features

SNRs can also be considered as a feature and appended to the i-vectors for PLDA modelling and scoring. To ensure that the SNRs have the same range as the other elements in the i-vectors, we applied Z-norm to the SNRs before appending them to the i-vectors. The results are shown in the rows labelled "PLDA (iVec+SNR)" in Table III. Evidently, when Set I was used for training the PLDA models, this approach can help improve performance. However, the SNR feature does not bring much advantage to the PLDA models when Set II was used for training. This is possibly because the mismatch between the SNR of training and test utterances is more severe in Set II than in Set I. In fact, this approach should be used with caution because SNR is speaker independent and i-vectors are speaker-dependent. It is difficult to avoid the PLDA models from performing SNR verification rather than speaker verification.

C. Robustness to SNR Mismatch

In NIST 2012 SRE, the amount of noise to be added to the test segments is given. Typically, researchers make use of this information to design multi-condition training strategies to maximize the match between training data and test data. However, in practical situations, we may not have the prior knowledge of the noise (i.e. the SNR) of the test utterances. To investigate whether the proposed mixture of PLDA can deal with such situation, we deliberately added crowd noise to the test segments in CC4 and CC5 of NIST 2012 SRE to make the SNR distribution of test segments different from that of the training segments. Specifically, noise waveforms from the PRISM dataset was added to the waveform files of the test utterances in CC4 and CC5 at 15dB and 6dB using the FaNT tool.

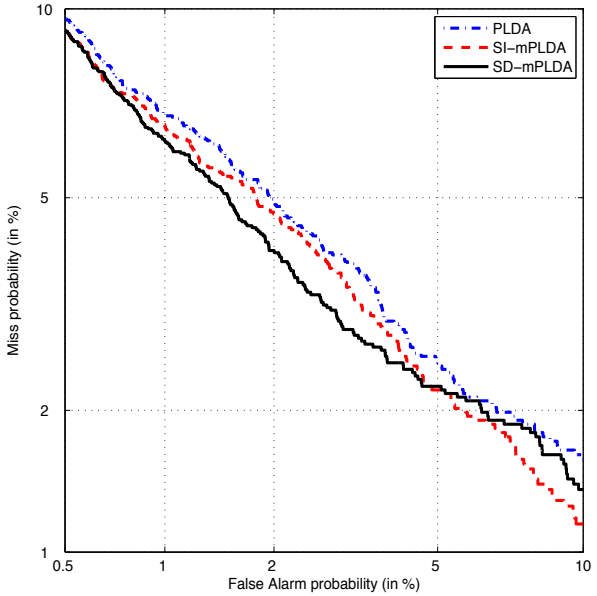


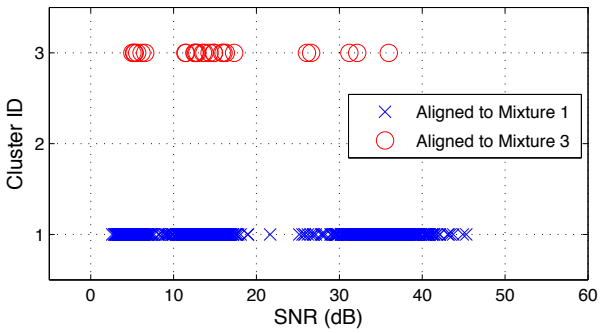
Fig. 9. DET performance of PLDA, SNR-independent mPLDA (SI-mPLDA), and SNR-dependent mPLDA (SD-mPLDA) in CC4 of NIST 2012 SRE (core set, male speakers) using Set I as training data. The number of mixtures for both mPLDA is 3.

The middle and bottom panels of Fig. 7 and Fig. 8 show the measured-SNR distributions of the noise contaminated test segments in CC4 and CC5, respectively. A comparison between these histograms with those in Fig. 4 reveals that there is substantial difference between the SNR of training and test segments. Note that the FaNT tool has the property that if the measured SNR is lower than the target SNR, no noise will be added. While the test utterances in CC5 were recorded in a noisy environment, the top panel of Fig. 8 suggests that the SNR is still fairly high. As a result, noise will be added to the test utterances in CC5 even if the target SNR is 15dB. On the other hand, as shown in the top panel of Fig. 7, some of the utterances in CC4 have SNR lower than 15dB or even 5dB. As a result, the SNR distribution is compressed and shifted to the left after adding noise, as shown in the middle and lower panel of Fig. 7.

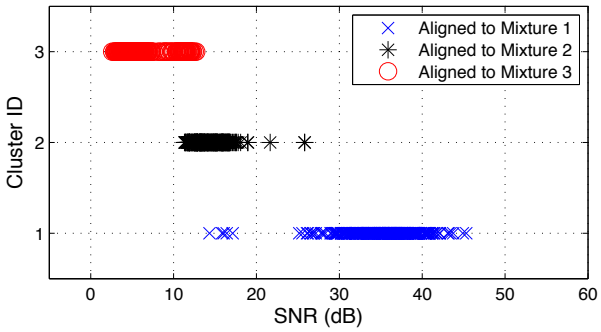
Table IV shows the EER and minimum DCF for different sets of test data. From the table, we can observe that even though the SNR distributions of training and test utterances are very different, mixture of PLDA still works better than single PLDA. From Table III, Table IV and Fig. 11, we can conclude

TABLE III
PERFORMANCE OF PLDA, SNR-INDEPENDENT MIXTURE OF PLDA, AND SNR-DEPENDENT MIXTURE OF PLDA WITH 2, 3 AND 4 MIXTURES TRAINED BY SET I AND SET II TRAINING DATA IN CC4 AND CC5 OF NIST 2012 SRE (CORE SET). "PLDA (IVEC+SNR)" MEANS THAT NORMALIZED SNR WAS APPENDED TO I-VECTORS FOR PLDA MODELLING AND SCORING.

Training data	Method		Male				Female			
			CC4		CC5		CC4		CC5	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
Set I	PLDA		3.49	0.308	**2.86**	**0.286**	3.13	0.353	2.47	0.343
	PLDA (iVec+SNR)		3.19	0.306	2.93	0.294	3.10	0.341	2.38	0.331
	SNR-independent mPLDA	2 mixtures	3.41	0.303	2.99	0.317	3.08	0.350	**2.34**	0.347
		3 mixtures	3.24	0.310	2.94	0.306	2.98	0.351	2.55	0.356
		4 mixtures	3.19	**0.291**	2.91	0.296	3.16	0.357	2.36	0.343
	SNR-dependent mPLDA	2 mixtures	3.28	0.307	2.97	0.307	3.10	0.359	2.59	0.355
		3 mixtures	**2.94**	0.315	**2.86**	0.295	**2.60**	**0.332**	2.59	**0.332**
		4 mixtures	3.11	0.313	2.90	0.307	2.84	0.333	2.74	0.349
Set II	PLDA		3.32	0.318	3.09	0.315	2.94	0.352	2.64	0.355
	PLDA (iVec+SNR)		3.24	0.315	3.23	0.314	2.98	0.353	2.72	0.331
	SNR-independent mPLDA	2 mixtures	3.32	0.318	3.09	0.336	3.06	0.357	2.67	0.352
		3 mixtures	**3.13**	0.315	3.21	0.311	**2.82**	0.352	2.59	**0.341**
		4 mixtures	3.37	**0.303**	3.13	**0.304**	2.86	0.345	**2.52**	0.344
	SNR-dependent mPLDA	2 mixtures	3.33	0.312	**3.00**	0.315	2.90	**0.349**	2.64	0.348
		3 mixtures	3.38	0.313	**3.00**	0.315	2.90	0.352	2.67	0.352
		4 mixtures	3.35	0.313	3.08	0.310	3.09	0.358	2.86	0.369



(a) SI-mPLDA



(b) SD-mPLDA

Fig. 10. Alignment of test i-vectors in CC4 of NIST 2012 SRE to (a) SNR-independent mixture of PLDA and (b) SNR-dependent mixture of PLDA. For both cases, $K = 3$. Note that in (a), none of the i-vectors aligned to Mixture 2.

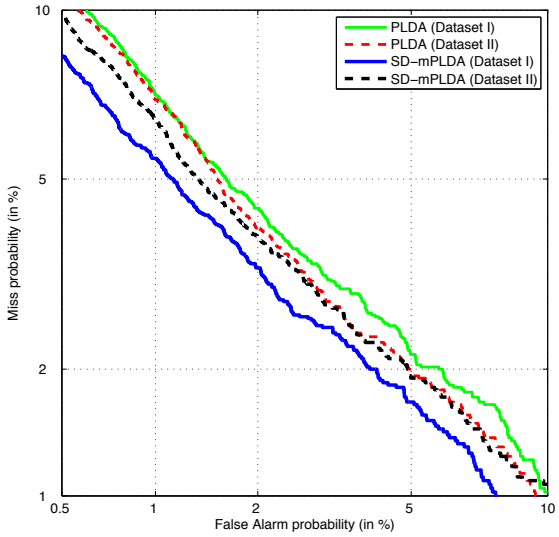


Fig. 11. DET performance of PLDA and SNR-dependent mixture of PLDA (SD-mPLDA) in CC4 of NIST 2012 SRE (core set, female speakers) using Dataset I and Dataset II for training. The number of mixtures K in SD-mPLDA is 3.

that the SNR-dependent mixture of PLDA is beneficial even though the SNR distribution of the test utterances is unknown.

VIII. CONCLUSIONS

To enhance the noise robustness of speaker verification systems, this paper applies an SNR-dependent mixture of PLDA. Two sets of data were used for training PLDA models and evaluation was performed on the latest NIST SRE. Unlike SNR-independent mixture of PLDA, the SNR of utterances is incorporated into both training and verification phases in SNR-dependent mixture of PLDA. The use of SNR in the verification phase leads to more meaningful combination of the mixtures, which makes the SNR-dependent mixture of PLDA more flexible, as evident by the promising performance in most

TABLE IV
PERFORMANCE OF PLDA , SNR-INDEPENDENT AND SNR-DEPENDENT MIXTURE OF PLDA (3 MIXTURES) WITH DIFFERENT TRAINING SETS IN CC4 AND CC5 FOR FEMALE SPEAKERS. THE SNR DISTRIBUTIONS FOR CC4(15DB), CC4(6DB), CC5(15DB) AND CC5(6DB) ARE SHOWN IN FIG. 7 AND FIG. 8.

Training data	Method	CC4 (15dB)		CC4 (6dB)		CC5 (15dB)		CC5 (6dB)	
		EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
Set I	PLDA	2.79	0.360	3.12	**0.378**	3.38	0.410	5.73	0.569
	SNR-independent mPLDA	2.70	0.362	**2.98**	0.380	**3.28**	**0.394**	**5.63**	**0.566**
	SNR-dependent mPLDA	**2.55**	**0.347**	3.02	0.406	3.37	0.405	6.11	0.587
Set II	PLDA	2.81	0.364	3.18	0.396	3.50	0.411	6.08	0.580
	SNR-independent mPLDA	**2.69**	**0.361**	3.08	**0.391**	3.49	**0.406**	**5.94**	**0.579**
	SNR-dependent mPLDA	2.91	0.382	**3.06**	0.400	**3.44**	0.421	5.98	0.583

situations. It was also found that mixture of PLDA performs better than the conventional PLDA even if there is a severe mismatch between the SNR of training and test utterances.

APPENDIX A: IMPLEMENTATION ISSUES

Eq. 8 and Eq. 13 are likely to cause numerical problems if they are evaluated directly because the determinant of $\hat{\mathbf{V}}_{k_s}\hat{\mathbf{V}}_{k_s}^{\mathsf{T}} + \hat{\mathbf{\Sigma}}_{k_s}$ could exceed the double-precision representation. This problem, however, can be avoided by computing the logarithm of determinant and noting the identity: $|\alpha\mathbf{A}| = \alpha^D|\mathbf{A}|$, where α is a scalar and $\mathbf{A}$ is a $D\times D$ matrix. Thus, we can rewrite Eq. 13 as Eq. 17, where $\hat{\mathbf{\Lambda}}_{k_s k_t} = \hat{\mathbf{V}}_{k_s}\hat{\mathbf{V}}_{k_t}^{\mathsf{T}} + \hat{\mathbf{\Sigma}}_{k_s k_t}$, $\mathbf{\Lambda}_{k_s} = \mathbf{V}_{k_s}\mathbf{V}_{k_s}^{\mathsf{T}} + \mathbf{\Sigma}_{k_s}$, $\hat{\mathbf{\Sigma}}_{k_s k_t} = \text{diag}\{\mathbf{\Sigma}_{k_s}, \mathbf{\Sigma}_{k_t}\}$, and $\mathcal{D}(\mathbf{x}\|\mathbf{y})$ is the Mahalanobis distance between $\mathbf{x}$ and $\mathbf{y}$, i.e., $\mathcal{D}(\mathbf{x}\|\mathbf{y}) = (\mathbf{x}-\mathbf{y})^{\mathsf{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})$, where $\mathbf{S} = \text{cov}(\mathbf{x},\mathbf{x})$. In this work, $\alpha = 5$. Note that α is to avoid taking exponential of very large negative numbers in Eq. 17, causing zero-divided-by-zero error. Because $\log|\alpha\mathbf{A}| = D\log\alpha + \log|\mathbf{A}|$, when $\log|\mathbf{A}| \ll 0$, the term $D\log\alpha$ can make the overall sum less negative. The same numerical technique can also be applied to Eq. 8.

Eq. 17 changes the computation of $|\mathbf{A}|$ in the Gaussian densities to $\log|\alpha\mathbf{A}|$. The latter can be easily computed using the Cholesky decomposition. More specifically, we have $\log|\alpha\mathbf{A}| = 2\sum_{i=1}^{D}\log b_{ii}$, where b_{ii}'s are the diagonal elements of $\mathbf{B}$ and $\mathbf{B}$ is the Cholesky decomposition of $\alpha\mathbf{A}$.

As pointed out by a reviewer of this paper, the numerical problems in computing the determinant in Eq. 13 can also be solved by using the matrix determinant lemma, and the inversion of the covariance matrices can be done using Cholesky factorization and the Woodbury formula [46].

APPENDIX B: EM DERIVATIONS

This appendix derives the EM formulations of SD-mPLDA. The EM formulations of SI-mPLDA can be similarly derived. For full derivations, readers may refer to Supplementary Materials available from the authors' website.

Denote $y_{i\cdot\cdot}$ as the indicator variables for all possible sessions and mixture components for speaker i. For the E-step, we start with the joint posterior density:

$$\begin{aligned}
&p(\mathbf{z}_i, y_{i\cdot\cdot}|\mathcal{X}_i,\mathcal{L}_i) \propto p(\mathcal{X}_i,\mathcal{L}_i|\mathbf{z}_i, y_{i\cdot\cdot}=1)p(\mathbf{z}_i, y_{i\cdot\cdot})\\
&= p(\mathcal{X}_i|\mathbf{z}_i, y_{i\cdot\cdot}=1)p(\mathcal{L}_i|y_{i\cdot\cdot}=1)p(y_{i\cdot\cdot})p(\mathbf{z}_i)\\
&= \prod_{j=1}^{H_i}\prod_{k=1}^{K}\left[\pi_k p(\mathbf{x}_{ij}|y_{ijk}=1,\mathbf{z}_i)p(\ell_{ij}|y_{ijk}=1)\right]^{y_{ijk}} p(\mathbf{z}_i)\\
&= \left\{\prod_{j=1}^{H_i}\prod_{k=1}^{K}\left[\pi_k\mathcal{N}(\ell_{ij}|\mu_k,\sigma_k^2)\right]^{y_{ijk}}\right\}\\
&\quad \underbrace{\cdot\, p(\mathbf{z}_i)\left\{\prod_{j=1}^{H_i}\prod_{k=1}^{K}\left[\mathcal{N}(\mathbf{x}_{ij}|\mathbf{m}_k+\mathbf{V}_k\mathbf{z}_i,\mathbf{\Sigma}_k)\right]^{y_{ijk}}\right\}}_{\propto\, p(\mathbf{z}_i|\mathcal{X}_i)} \qquad (18)
\end{aligned}$$

where we have used the fact that y_{ijk} is determined by ℓ_{ij}. The 2nd line of Eq. 18 makes use of the assumption that $\mathbf{z}_i$ and y_{ijk} are independent and the 3rd line makes use of Eq. 9.38 of [9].

To find the posterior of $\mathbf{z}_i$, we extract the terms dependent on $\mathbf{z}_i$ from Eq. 18 as follows:

$$\begin{aligned}
p(\mathbf{z}_i|\mathcal{X}_i) &\propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{H_i}\sum_{k=1}^{K} y_{ijk}(\mathbf{x}_{ij}-\mathbf{m}_k-\mathbf{V}_k\mathbf{z}_i)^{\mathsf{T}}\mathbf{\Sigma}_k^{-1}\right.\\
&\qquad\left.\times\,(\mathbf{x}_{ij}-\mathbf{m}_k-\mathbf{V}_k\mathbf{z}_i) - \frac{1}{2}\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}\right\}\\
&= \exp\left\{\mathbf{z}_i^{\mathsf{T}}\mathbf{V}_k^{\mathsf{T}}\sum_{k=1}^{K}\sum_{j\in\mathcal{H}_{ik}}\mathbf{\Sigma}_k^{-1}(\mathbf{x}_{ij}-\mathbf{m}_k) -\right.\\
&\qquad\left.\frac{1}{2}\mathbf{z}_i^{\mathsf{T}}\left(\mathbf{I}+\sum_{k=1}^{K}\sum_{j\in\mathcal{H}_{ik}}\mathbf{V}_k^{\mathsf{T}}\mathbf{\Sigma}_k^{-1}\mathbf{V}_k\right)\mathbf{z}_i\right\} \qquad (19)
\end{aligned}$$

where $\mathcal{H}_{ik}$ comprises the indexes of speaker i's i-vectors that aligned to mixture k. Comparing Eq. 19 with the standard Gaussian, $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z,\mathbf{C}_z) \propto \exp\left\{\mathbf{z}^{\mathsf{T}}\mathbf{C}_z^{-1}\boldsymbol{\mu}_z - \frac{1}{2}\mathbf{z}^{\mathsf{T}}\mathbf{C}_z^{-1}\mathbf{z}\right\}$, we obtain the posterior mean $\langle\mathbf{z}_i|\mathcal{X}_i\rangle$, posterior moment $\langle\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X}_i\rangle$, and posterior precision $\mathbf{L}_i$ in Eq. 15. The posterior

$$= \frac{\sum_{k_s=1}^{K}\sum_{k_t=1}^{K}\gamma_{\ell_s,\ell_t}(y_{k_s},y_{k_t})\exp\left\{-\frac{1}{2}\log|\alpha\hat{\mathbf{\Lambda}}_{k_sk_t}|-\frac{1}{2}\mathcal{D}\left(\left[\mathbf{x}_s^\mathsf{T}\ \ \mathbf{x}_t^\mathsf{T}\right]^\mathsf{T}\middle\|\left[\mathbf{m}_{k_s}^\mathsf{T}\ \ \mathbf{m}_{k_t}^\mathsf{T}\right]^\mathsf{T}\right)\right\}}{\left[\sum_{k_s=1}^{K}\gamma_{\ell_s}(y_{k_s})\exp\left\{-\frac{1}{2}\log|\alpha\mathbf{\Lambda}_{k_s}|-\frac{1}{2}\mathcal{D}\left(\mathbf{x}_s\|\mathbf{m}_{k_s}\right)\right\}\right]\left[\sum_{k_t=1}^{K}\gamma_{\ell_t}(y_{k_t})\exp\left\{-\frac{1}{2}\log|\alpha\mathbf{\Lambda}_{k_t}|-\frac{1}{2}\mathcal{D}\left(\mathbf{x}_t\|\mathbf{m}_{k_t}\right)\right\}\right]} \quad (17)$$

of y_{ijk} can be computed using the Bayes rule:

$$\begin{aligned}\langle y_{ijk}|\mathcal{L}\rangle &= \langle y_{ijk}|\ell_{ij}\rangle = P(y_{ijk}=1|\ell_{ij},\underline{\boldsymbol{\lambda}})\\ &= \frac{P(y_{ijk}=1)p(\ell_{ij}|y_{ijk}=1,\underline{\boldsymbol{\lambda}})}{\sum_{r=1}^{K}P(y_{ijr}=1)p(\ell_{ij}|y_{ijr}=1,\underline{\boldsymbol{\lambda}})}\\ &= \frac{\pi_k\mathcal{N}(\ell_{ij}|\mu_k,\sigma_k^2)}{\sum_{r=1}^{K}\pi_r\mathcal{N}(\ell_{ij}|\mu_r,\sigma_r^2)}.\end{aligned}$$

For the M-step, we write Eq. 14 as follows:

$$\begin{aligned}Q(\underline{\boldsymbol{\theta}}) &= \sum_{ijk}\langle y_{ijk}|\mathcal{L}\rangle\left[-\log\sigma_k-\frac{1}{2}\sigma_k^{-2}(\ell_{ij}-\mu_k)^2+\log\pi_k\right]\\ &+\sum_{ijk}\langle y_{ijk}|\mathcal{L}\rangle\left[-\frac{1}{2}\log|\mathbf{\Sigma}_k|-\frac{1}{2}(\mathbf{x}_{ij}-\mathbf{m}_k)^\mathsf{T}\mathbf{\Sigma}_k^{-1}(\mathbf{x}_{ij}-\mathbf{m}_k)\right]\\ &+\sum_{ijk}(\mathbf{x}_{ij}-\mathbf{m}_k)^\mathsf{T}\mathbf{\Sigma}_k^{-1}\mathbf{V}_k\langle y_{ijk}\mathbf{z}_i|\mathcal{X},\mathcal{L}\rangle\\ &-\frac{1}{2}\left[\sum_{ijk}\mathrm{tr}\left\{\left(\mathbf{V}_k^\mathsf{T}\mathbf{\Sigma}_k^{-1}\mathbf{V}_k+\mathbf{I}\right)\left\langle y_{ijk}\mathbf{z}_i\mathbf{z}_i^\mathsf{T}|\mathcal{X},\mathcal{L}\right\rangle\right\}\right].\end{aligned}$$

Then, we differentiate $Q(\underline{\boldsymbol{\theta}})$ with respect to π_k (subject to $\sum_k\pi_k=1$), μ_k, σ_k, $\mathbf{m}_k$, $\mathbf{V}_k$, and $\mathbf{\Sigma}_k^{-1}$, and set the derivatives to zero to obtain the M-step formulations in Eq. 16.

REFERENCES

[1] S. O. Sadjadi, T. Hasan, and J. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition." in *Proc. of Interspeech*, 2012, pp. 1696–1699.

[2] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2008, pp. 1589–1592.

[3] Q. Li and Y. Huang, "Robust speaker identification using an auditory-based feature," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 4514–4517.

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

[5] R. Saeidi and D. A. van Leeuwen, "The Radboud University Nijmegen submission to NIST SRE-2012," in *Proc. of the NIST Speaker Recognition Evaluation Workshop*, 2012.

[6] 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, 2008.

[7] 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, May 2011.

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

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

[10] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. of International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, Sep. 2006, pp. 1471–1474.

[11] W. Rao and M. W. Mak, "Boosting the performance of i-vector based speaker verification via utterance partitioning," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 1012–1022, 2013.

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

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

[14] D. A. Leeuwen and R. Saeidi, "Knowing the non-target speakers: The effect of the i-vector population for PLDA training in speaker recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*, Vancouver, BC, Canada, May 2013, pp. 6778 – 6782.

[15] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 4253 – 4256.

[16] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. H. L. Hansen, "CRSS system for 2012 NIST speaker recogntion evaluation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6783–6787.

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

[18] 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. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4257–4260.

[19] 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.

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

[21] ——, "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.

[22] K. Simonchik, T. Pekhovsk, A. Shulipa, and A. Afanasyev, "Supervised mixture of PLDA models for cross-channel speaker verification," in *Proc. of Interspeech*, 2012, pp. 1684–1687.

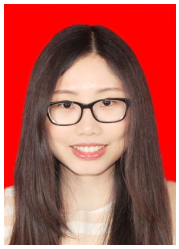
[23] T. Pekhovsky and A. Sizov, "Comparison between supervised and unsupervised learning of probabilistic linear discriminant analysis mixture models for speaker verification," *Pattern Recognition Letters*, vol. 34, no. 11, pp. 1307–1313, 2013.

[24] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," Department of Computer Science, University of Toronto, Technical Report CRG-TR-96-1, 1996.

[25] J. Villalba and E. Lleida, "Handling I-vectors from diferent recording condistions using multi-channel simplified PLDA in

speaker recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6763–6767.
[26] J. Villalba, E. Lleida, A. Ortega, and A. Miguel, "The I3A speaker recognition system for NIST SRE12: Post-evaluation analysis," in *Proc. Interspeech*, 2013, pp. 3689–3693.
[27] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector Taylor series for speaker recognition." in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6788–6791.
[28] Y. Lei, M. Mclaren, L. Ferrer, and N. Scheffer, "Simplified VTS-based i-vector extraction in noise-robust speaker recognition." in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4065–4069.
[29] D. Martinez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny, and E. Lleida, "Unscented transform for i-vector-based noisy speaker recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4070–4074.
[30] M. McLaren, Y. Lei, N. Scheffer, and L. Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions," in *Proc. of Interspeech*, 2014, pp. 686–690.
[31] C. Yu, G. Liu, S. Hahm, and J. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4045–4049.
[32] S. Sadjadi, J. Pelecanos, and W. Zhu, "Nearest neighbor discriminant analysis for robust speaker recognition," in *Proc. of Interspeech*, 2014, pp. 1860–1864.
[33] NIST, "The NIST year 2012 speaker recognition evaluation plan," *http://www.nist.gov/itl/iad/mig/sre12.cfm*, 2012.
[34] 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. of Interspeech*, 2013, pp. 1986–1990.
[35] M. W. Mak, "SNR-dependent mixture of PLDA for noise robust speaker verification," in *Proc. of Interspeech*, Singapore, Sept. 2014, pp. 1855–1859.
[36] G. McLachlan and D. Peel, *Finite Mixture Models*. John Wiley and Sons, 2000, ch. Mixtures of factor analyzers, pp. 238–256.
[37] "http://dnt.kr.hsnr.de/download.html."
[38] Y. B. M. Halkidi and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2/3, pp. 107–145, 2001.
[39] A. Sizov, K. A. Lee, and T. Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," in *Structural, Syntactic, and Statistical Pattern Recognition*. Springer, 2014, pp. 464–475.
[40] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000.
[41] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.
[42] D. Rubin and D. Thayer, "EM algorithms for ML factor analysis," *Psychometrika*, vol. 47, no. 1, pp. 69–76, 1982.
[43] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall International, Inc., 1993.
[44] 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 2013.
[45] L. Ferrer, et al., "Promoting robustness for speaker modeling in the community: The PRISM evaluation set," in *Proc. NIST Speaker Recognition Analysis Workshop (SRE11)*, 2011, pp. 1–7.
[46] W. H. Press, *Numerical recipes: The art of scientific computing*, 3rd ed. Cambridge university press, 2007.

Man-Wai Mak (M'93–SM'15) received a PhD in Electronic Engineering from the University of Northumbria in 1993. He joined the Department of Electronic and Information Engineering at The Hong Kong Polytechnic University in 1993 and is currently an Associate Professor in the same department. He has authored more than 160 technical articles in speaker recognition, machine learning, and bioinformatics. Dr. Mak also coauthored a postgraduate textbook *Biometric Authentication: A Machine Learning Approach, Prentice Hall, 2005* and a research monograph *Machine Learning for Protein Subcellular Localization Prediction, De Gruyter, 2015*. He served as a member of the IEEE Machine Learning for Signal Processing Technical Committee in 2005-2007. He has served as an associate editor of IEEE Trans. on Audio, Speech and Language Processing. He is currently an editorial board member of Journal of Signal Processing Systems and Advances in Artificial Neural Systems. He also served as Technical Committee Members of a number of international conferences, including ICASSP and Interspeech. Dr. Mak's research interests include speaker recognition, machine learning, and bioinformatics.



Xiao-Min Pang received a BEng(Hons) degree with first class honours in Electronic Engineering and an MSc degree with distinction in Electronic and Information Engineering from The Hong Kong Polytechnic University in 2013 and 2014, respectively. Her research interests include speaker recognition and machine learning.

Jen-Tzung Chien (S'97-A'98-M'99-SM'04) received his Ph.D. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan in 1997. During 1997-2012, he was with the National Cheng Kung University, Tainan, Taiwan. Since 2012, he has been with the Department of Electrical and Computer Engineering and the Department of Computer Science, National Chiao Tung University, Hsinchu, where he is currently an University Chair Professor. He held the Visiting Researcher position with the IBM T. J. Watson Research Center, Yorktown Heights, NY, in 2010. His research interests include machine learning, speaker recognition, speech recognition, face recognition, and blind source separation.

Dr. Chien served as the associate editor of the IEEE Signal Processing Letters in 2008-2011, the guest editor of the IEEE Transactions on Audio, Speech, and Language Processing in 2012, and the tutorial speaker of the Interspeech 2013 and the ICASSP 2012 and 2015. He received the Best Paper Award of the IEEE Automatic Speech Recognition and Understanding Workshop in 2011 and the Distinguished Research Award from the Ministry of Science and Technology, Taiwan in 2006, 2010 and 2014. He has published extensively, including the book "Bayesian Speech and Language Processing", Cambridge University Press, 2015. He currently serves as an elected member of the IEEE Machine Learning for Signal Processing Technical Committee.