https://doi.org/10.1016/j.csl.2017.04.001

Discriminative Subspace Modeling of SNR and Duration Variabilities for Robust Speaker Verification

Na LI[a], Man-Wai MAK[a], Wei-Wei LIN[a], Jen-Tzung CHIEN[b]

[a] *Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR of China*
[b] *Dept. of Electrical and Computer Engineering, National Chiao Tung University, Taiwan*

Abstract

Although i-vectors together with probabilistic LDA (PLDA) have achieved a great success in speaker verification, how to suppress the undesirable effects caused by the variability in utterance length and background noise level is still a challenge. This paper aims to improve the robustness of i-vector based speaker verification systems by compensating for the utterance-length variability and noise-level variability. Inspired by the recent findings that noise-level variability can be modeled by a signal-to-noise ratio (SNR) subspace and that duration variability can be modeled as additive noise in the i-vector space, we propose to add an SNR factor and a duration factor to the PLDA model. In this framework, we assume that i-vectors derived from utterances with comparable durations share similar duration-specific information and that i-vectors extracted from utterances within a narrow SNR range have similar SNR-specific information. Based on these assumptions, an i-vector can be represented as a linear combination of four components: speaker, SNR, duration, and channel. A variational Bayes algorithm is developed to infer this latent variable model via a discriminative subspace training procedure. In the testing stage, different variabilities are compensated when computing the likelihood ratio. Experiments on Common Conditions 1 and 4 in NIST 2012 SRE show that the proposed model outperforms the conventional PLDA and SNR-invariant PLDA. Results also show that the proposed model performs better than the uncertainty-propagation PLDA (UP-PLDA) for long test utterances.

Keywords: Speaker verification, duration variation, SNR mismatch, variational Bayes, I-vector, PLDA

1. Introduction

In text-independent speaker verification, i-vectors [6] have become the most popular feature representation in recent years. Inspired by the joint factor analysis (JFA) [15, 17, 18] framework, both the speaker and undesirable information (e.g. channel, additive noise, and so on) were compressed into a low-dimensional subspace called the total variability subspace, through which utterances with variable durations can be represented as low-dimensional i-vectors of fixed-length. Such a representation converts a speaker verification problem to an ordinary biometric pattern recognition problem similar to face recognition and fingerprint recognition. Based on the i-vector representation, many statistical techniques have been applied to deal with the mismatch between the training and test utterances. For example, linear discriminant analysis (LDA) [2] followed by within-class covariance normalization (WCCN) [12] were applied to i-vectors to compensate for session variability; then cosine distance between the target speaker's i-vector and test i-vector was used as the similarity measure between the target speaker and the test speaker. More recently, probabilistic LDA (PLDA) [31] was employed to suppress the channel- and session-variability within the i-vector space. Typically, i-vectors were preprocessed by a series of transformations – WCCN, length normalization [7], and LDA – before presenting the i-vectors to a Gaussian PLDA model.

Although the i-vector/PLDA framework performs well in suppressing session variability, it still has the following limitations: (1) the ability of PLDA in modeling the variability arising from utterances of different SNRs is limited; (2) i-vectors extracted from short utterances are less reliable than those extracted from long utterances [19], leading to performance degradation when only short utterances are available.

One of the main focuses of NIST 2012 SRE is robust speaker verification in which the SNR and length of enrollment and test utterances have substantial variation. To improve the noise robustness of i-vector/PLDA systems, several methods have been proposed. In [10], clean and noisy utterances were pooled together to train a robust PLDA model. Garcia-Romero *et al.* [8] 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 [23, 24], a mixture of SNR-dependent PLDA was proposed so that each mixture focuses on a small range of SNRs. During verification, the mixtures cooperated with each other to deal with utterances of various noise levels. 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 model, resulting in SNR-invariant PLDA [21, 20]. With the added SNR subspace, the SNR-invariant PLDA can capture both speaker, noise-level, and channel variabilities embedded in the i-vectors.

The problem of duration variability in utterances has attracted attention in the community because an i-vector extracted from a short utterance should not be treated as being equally reliable as an i-vector extracted from a long utterance. The reason is that the posterior distribution of hidden variables in the i-vector extractor is a Gaussian whose covariance matrix is related to the utterance duration. The shorter the utterance is, the larger the covariance will become, leading to greater uncertainty in the estimated i-vector.

The issue of duration variability has been addressed to a certain extent in the past. For example, Sarkar *et al.* [32] investigated how duration mismatches affect the optimal choice of the duration of training utterances for estimating the parameters of i-vector systems. In [19], the uncertainty arising from the i-vector extraction process was propagated into a PLDA model. This method did not treat an i-vector as the maximum *a posteriori* point estimate, but rather as a random vector whose uncertainty was represented by the posterior covariance matrix of the latent factors. The shorter the utterance, the larger the posterior covariances. By propagating this information into PLDA and using a loading matrix to model the variability due to duration variation, the resulting PLDA model better handled the length-variability than the conventional PLDA model. Cumani et al. [5, 4] did not map an utterance to a single i-vector, but instead mapped it to the posterior distribution of i-vectors. Then, the likelihood of two speech segments coming from the same speaker was obtained by integrating out all possible i-vectors based on the i-vector posterior density.

Hasan *et al.* [11] found that duration variability could be modeled as additive noise in the i-vector space. A short-utterance variance normalization technique and a short-utterance variance modeling approach were proposed in [14] to compensate for utterance-length variability. In [34], a weight associated with the utterance's duration was added to the corresponding i-vector; then duration-weighted means, covariance matrix, and within-class scatter matrix were computed; finally, principal component analysis (PCA) and WCCN were applied using these duration-weighted terms to take utterance duration into account. Motivated by the belief that i-vectors derived from long utterances are more reliable [19] and therefore their corresponding covariances in the PLDA model should be smaller, Cai et al. [3] proposed to regularize the PLDA covariance matrix by scaling it by a duration-dependent exponential term. On top of this duration-dependent covariance regularization, Hong et al. [13] introduced a quality measure function for score calibration, which effectively compensated for the score shift due to duration mismatch. In [36], a denoising autoencoder was used to compensate for the phonetic imbalance in short utterances. Given a short utterance, the autoencoder received an i-vector and a phonetic vector (the utterance's zero-order statistics) as input and produced an output comprising an i-vector as if it were produced by a phonetically balance utterance. The autoencoder was trained by using the i-vectors and phonetic vectors derived from many short-long utterance pairs.

This paper focuses on improving the robustness of the state-of-the-art i-vector/PLDA systems when duration mismatch and SNR mismatch between the training and test utterances occur simultaneously. According to [11, 14], duration variability in the i-vectors can be modeled as additive noise in i-vector space. If the i-vector extracted from a long utterance is considered as "clean", the i-vector extracted from a short utterance can be considered as "noisy". Inspired by this observation, we propose a new method to deal with the mismatch caused by the variabilities in SNR and duration. Our proposal is motivated by the success of SNR-invariant PLDA in dealing with SNR mismatch [21, 20]. More specifically, we attempt to make the i-vector/PLDA framework more resilient to SNR and duration variabilities by introducing two discriminant subspaces – namely SNR subspace and duration subspace – to the PLDA models. These subspaces are trained discriminatively by exploiting the SNR and duration information in the training utterances. Through joint discriminative training, these subspaces enable the new PLDA models to capture not only speaker and channel variabilities, but also SNR and duration variabilities. In the proposed model, the speaker component, SNR component, and duration component live in three different subspaces which can be inferred according to the variational Bayes procedure. During the verification stage, SNR variability, duration variability, and channel variability are marginalized out when the likelihood ratio is computed.

The organization of this paper is as follows: Section 2 describes the i-vector/PLDA speaker verification. Based on different assumptions, a new method of estimating the parameters of duration-invariant PLDA and two new scoring methods are proposed in Section 3. The proposed modeling method, namely SNR- and duration-invariant PLDA, is explained for robust speaker verification in Section 4. The experimental results and analysis of the proposed framework are detailed in Section 5 and Section 6, respectively.

Finally, conclusions are drawn in Section 7.

2. I-vector/PLDA Speaker Verification

2.1. Conventional PLDA

In the conventional i-vector/PLDA framework [16], an i-vector $\mathbf{x}_{ij}$ is regarded as an observation generated from a linear model [31, 30]:

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

where $\mathbf{m}$ is the global mean of i-vectors, $\mathbf{V}$ defines the speaker subspace, $\mathbf{G}$ defines the channel subspace, $\mathbf{h}_i$ and $\mathbf{r}_{ij}$ are the 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 variation that cannot be described by $\mathbf{V}\mathbf{V}^{\mathsf{T}}$ and $\mathbf{G}\mathbf{G}^{\mathsf{T}}$.

According to [16, 7], the PLDA model in Eq. 1 can be divided into two parts: (1) the speaker part $(\mathbf{m} + \mathbf{V}\mathbf{h}_i)$ that depends on the i-th speaker only and (2) the channel part $(\mathbf{G}\mathbf{r}_{ij} + \boldsymbol{\epsilon}_{ij})$ that depends not only on the i-th speaker but also on the j-th session. As i-vectors are of sufficiently low dimension, the term $\mathbf{G}\mathbf{r}_{ij}$ can be absorbed into $\boldsymbol{\Sigma}$ if the latter is a full covariance matrix. Accordingly, the Gaussian PLDA model can be simplified as follows [33]:

$$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \boldsymbol{\epsilon}_{ij}, \tag{2}$$

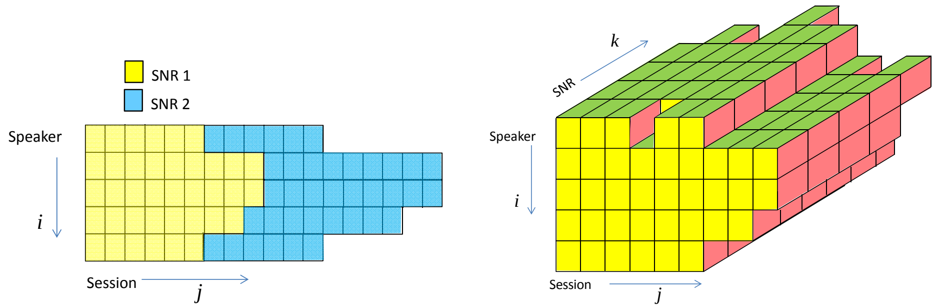
where $\boldsymbol{\epsilon}_{ij} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ being a full covariance matrix. This paper adopts this simplified model.

2.2. SNR-invariant PLDA

To enhance the robustness of i-vector/PLDA, we have recently proposed an SNR-invariant PLDA model (SI-PLDA) [21, 20] to deal with SNR mismatch. In this model, training utterances are first divided into K groups according to their SNRs. As a result, each of the training i-vectors is associated with one SNR group. Denote $\mathbf{x}_{ij}^k$ as the j-th i-vector from speaker i in the k-th SNR group. Then, $\mathbf{x}_{ij}^k$ is expressed as:

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

where $\mathbf{m}$ is the global mean of i-vectors, $\mathbf{V}$ defines the speaker subspace, $\mathbf{h}_i$ is a latent speaker factor with a standard normal prior, $\mathbf{U}$ defines the SNR subspace, $\mathbf{w}_k$ is a latent SNR factor with a standard normal prior, $\boldsymbol{\epsilon}_{ij}^k$ is a



(a) (b)

Figure 1: (a) Arrangement of training i-vectors in the multi-condition training of conventional PLDA. Each small square represents an i-vector. While the training set comprises two SNR groups, PLDA training ignores the group labels and sums over the statistics across both groups. (b) Arrangement of training i-vectors in SNR-invariant PLDA. Each small cube represents an i-vector. For the i-th speaker, there are $H_i(k)$ i-vectors from the k-th SNR group. Training in SNR-invariant PLDA considers the group labels and sums over the statistics within individual groups.

residual term with distribution $\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{\Sigma})$. In [21, 20], $\boldsymbol{\Sigma}$ 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. 3) 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. 3) depend on the SNR groups, whereas the session latent factor ($\mathbf{r}_{ij}$ in Eq. 1) depends on the speaker and session.

Fig. 1 illustrates how the labels (speaker and SNR groups) can be used in training these two types of PLDA models. As can be seen, in the conventional PLDA (Fig. 1(a), Eq. 1, and Eq. 2), the i-vectors for each speakers are treated equally regardless of which SNR group they come from. On the other hand, in SI-PLDA (Fig. 1(b) and Eq. 3), i-vectors derived from utterances of similar SNR are grouped together in a vertical slice. These extra SNR labels, together with the speaker labels, help to suppress the SNR variability in the i-vectors.

3. Duration-invariant PLDA

According to [11, 14], duration variability in the i-vectors can be modeled as additive noise in i-vector space. Inspired by the success of SI-PLDA in handling SNR variability, we propose to handle duration variability by a duration-invariant PLDA (DI-PLDA).

3.1. Generative Model and EM Formulation

Assume that we have a set of i-vectors

$$\mathcal{X} = \{\mathbf{x}_{ij}^p | i = 1, \ldots, S; j = 1, \ldots, H_i(p); p = 1, \ldots, P\}$$

obtained from S speakers, where $\mathbf{x}_{ij}^p$ is the j-th utterance from speaker i at the p-th duration group. For the i-th speaker, there are $H_i(p)$ i-vectors from the p-th duration group. Eq. 3 becomes DI-PLDA if the SNR-related term is replaced by a duration-related term, i.e.,

$$\mathbf{x}_{ij}^p = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{R}\mathbf{y}_p + \boldsymbol{\epsilon}_{ij}^p, \tag{4}$$

where $\mathbf{R}$ defines the duration subspace, $\mathbf{y}_p$ is a latent duration factor with a standard normal distribution. Other terms have the same meaning as in Eq. 3.

In [21], the latent factors $\mathbf{h}_i$ and $\mathbf{y}_p$ are assumed to be posteriorly independent. In this paper, we consider $\mathbf{h}_i$ and $\mathbf{y}_p$ are posteriorly dependent and use variational Bayes methods [2] to derive EM algorithms for training the SI-PLDA and DI-PLDA models.

Denote $N_i = \sum_{p=1}^{P} H_i(p)$ as the number of training utterances from the i-th speaker and $B_p = \sum_{i=1}^{S} H_i(p)$ as the number of the training utterances in the p-th duration group. Given an old estimate of the model parameters $\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}, \mathbf{R}, \boldsymbol{\Sigma}\}$, we aim to find a new estimate $\boldsymbol{\theta}'$ that maximizes the auxiliary function:

$$\begin{aligned} Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) &= \mathbb{E}_{q(\underline{\mathbf{h}},\underline{\mathbf{y}})} \left\{ \ln p(\mathcal{X}, \underline{\mathbf{h}}, \underline{\mathbf{y}}|\boldsymbol{\theta}') \middle| \mathcal{X}, \boldsymbol{\theta} \right\} \\ &= \mathbb{E}_{q(\underline{\mathbf{h}},\underline{\mathbf{y}})} \left\{ \sum\nolimits_{ijp} \ln \left[p(\mathbf{x}_{ij}^p | \mathbf{h}_i, \mathbf{y}_p, \boldsymbol{\theta}') p(\mathbf{h}_i, \mathbf{y}_p) \right] \middle| \mathcal{X}, \boldsymbol{\theta} \right\}, \end{aligned} \tag{5}$$

where $\underline{\mathbf{h}} = \{\mathbf{h}_1, \ldots, \mathbf{h}_s\}$, $\underline{\mathbf{y}} = \{\mathbf{y}_1, \ldots, \mathbf{y}_p\}$, and $q(\underline{\mathbf{h}}, \underline{\mathbf{y}})$ is the variational posterior density of $\underline{\mathbf{h}}$ and $\underline{\mathbf{y}}$. To maximize $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$, we differentiate $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ with respect to the model parameters $\{\mathbf{m}, \mathbf{V}, \mathbf{R}, \boldsymbol{\Sigma}\}$ and set the resulting derivatives to $\mathbf{0}$. This leads to

$$\mathbf{m} = \frac{1}{N} \sum_{i=1}^{S} \sum_{p=1}^{P} \sum_{j=1}^{H_i(p)} \mathbf{x}_{ij}^p \tag{6}$$

$$\mathbf{V}' = \left\{ \sum_{i=1}^{S} \sum_{p=1}^{P} \sum_{j=1}^{H_i(p)} \left[(\mathbf{x}_{ij}^p - \mathbf{m}) \langle \mathbf{h}_i | \mathcal{X} \rangle - \mathbf{R} \langle \mathbf{y}_p \mathbf{h}_i^\mathsf{T} | \mathcal{X} \rangle \right] \right\} \left[\sum_{i=1}^{S} \sum_{p=1}^{P} \sum_{j=1}^{H_i(p)} \langle \mathbf{h}_i \mathbf{h}_i^\mathsf{T} | \mathcal{X} \rangle \right]^{-1} \tag{7}$$

$$\mathbf{R}' = \left\{ \sum_{i=1}^{S} \sum_{p=1}^{P} \sum_{j=1}^{H_i(p)} \left[(\mathbf{x}_{ij}^p - \mathbf{m}) \langle \mathbf{y}_p | \mathcal{X} \rangle - \mathbf{V} \langle \mathbf{h}_i \mathbf{y}_p^\mathsf{T} | \mathcal{X} \rangle \right] \right\} \left[\sum_{i=1}^{S} \sum_{p=1}^{P} \sum_{j=1}^{H_i(p)} \langle \mathbf{y}_p \mathbf{y}_p^\mathsf{T} | \mathcal{X} \rangle \right]^{-1} \tag{8}$$

$$\mathbf{\Sigma}' = \frac{1}{N} \sum_{i=1}^{S} \sum_{p=1}^{P} \sum_{j=1}^{H_i(p)} \left[(\mathbf{x}_{ij}^p - \mathbf{m})(\mathbf{x}_{ij}^p - \mathbf{m})^\mathsf{T} - \mathbf{V} \langle \mathbf{h}_i | \mathcal{X} \rangle (\mathbf{x}_{ij}^p - \mathbf{m})^\mathsf{T} - \mathbf{R} \langle \mathbf{y}_p | \mathcal{X} \rangle (\mathbf{x}_{ij}^p - \mathbf{m})^\mathsf{T} \right] \tag{9}$$

where $N = \sum_{i=1}^{S} N_i = \sum_{p=1}^{P} B_p$.

Eq. 6–Eq. 9 constitute the M-step of the EM algorithm. To update the model parameters in the M-step, we need to estimate the posterior distribution of $\mathbf{h}_i$ and $\mathbf{y}_p$. These posteriors can be obtained through the variational Bayes method as explained below.

We approximate the true posterior $p(\underline{\mathbf{h}}, \underline{\mathbf{y}} | \mathcal{X})$ by a variational posterior $q(\underline{\mathbf{h}}, \underline{\mathbf{y}})$ and write the marginal likelihood of $\mathcal{X}$ as

$$\begin{aligned}
\ln p(\mathcal{X}) &= \int \int q(\underline{\mathbf{h}}, \underline{\mathbf{y}}) \ln p(\mathcal{X}) \mathrm{d}\underline{\mathbf{h}} \mathrm{d}\underline{\mathbf{y}} \\
&= \int \int q(\underline{\mathbf{h}}, \underline{\mathbf{y}}) \ln \left[\frac{p(\underline{\mathbf{h}}, \underline{\mathbf{y}}, \mathcal{X})}{p(\underline{\mathbf{h}}, \underline{\mathbf{y}} | \mathcal{X})} \right] \mathrm{d}\underline{\mathbf{h}} \mathrm{d}\underline{\mathbf{y}} \\
&= \int \int q(\underline{\mathbf{h}}, \underline{\mathbf{y}}) \ln \left[\frac{p(\underline{\mathbf{h}}, \underline{\mathbf{y}}, \mathcal{X})}{q(\underline{\mathbf{h}}, \underline{\mathbf{y}})} \right] \mathrm{d}\underline{\mathbf{h}} \mathrm{d}\underline{\mathbf{y}} + \int \int q(\underline{\mathbf{h}}, \underline{\mathbf{y}}) \ln \left[\frac{q(\underline{\mathbf{h}}, \underline{\mathbf{y}})}{p(\underline{\mathbf{h}}, \underline{\mathbf{y}} | \mathcal{X})} \right] \mathrm{d}\underline{\mathbf{h}} \mathrm{d}\underline{\mathbf{y}} \\
&= \mathcal{L}(q) + \mathcal{D}_{\mathrm{KL}}(q(\underline{\mathbf{h}}, \underline{\mathbf{y}}) \| p(\underline{\mathbf{h}}, \underline{\mathbf{y}} | \mathcal{X})).
\end{aligned} \tag{10}$$

In Eq. 10, $\mathcal{D}_{\mathrm{KL}}(q||p)$ is the KL-divergence between distributions q and p and

$$\mathcal{L}(q) = \int \int q(\underline{\mathbf{h}}, \underline{\mathbf{y}}) \ln \left[\frac{p(\underline{\mathbf{h}}, \underline{\mathbf{y}}, \mathcal{X})}{q(\underline{\mathbf{h}}, \underline{\mathbf{y}})} \right] \mathrm{d}\underline{\mathbf{h}} \mathrm{d}\underline{\mathbf{y}} \tag{11}$$

is the variational lower bound of the marginal likelihood. Since KL-divergence is non-negative, we can maximize the marginal likelihood through maximizing the lower bound with respect to $q(\underline{\mathbf{h}}, \underline{\mathbf{y}})$. The maximum occurs when $q(\underline{\mathbf{h}}, \underline{\mathbf{y}})$ equals the true posterior $p(\underline{\mathbf{h}}, \underline{\mathbf{y}} | \mathcal{X})$. Then, we assume that the approximated posterior $q(\underline{\mathbf{h}}, \underline{\mathbf{y}})$ can be factorized as follows:

$$\ln q(\underline{\mathbf{h}}, \underline{\mathbf{y}}) = \ln q(\underline{\mathbf{h}}) + \ln q(\underline{\mathbf{y}}) = \sum_{i=1}^{S} \ln q(\mathbf{h}_i) + \sum_{p=1}^{P} \ln q(\mathbf{y}_p). \tag{12}$$

By maximizing the lower bound $\mathcal{L}(q)$ in Eq. 11, we obtain [2, 35]

$$\begin{aligned} \ln q(\underline{\mathbf{h}}) &= \mathbb{E}_{q(\underline{\mathbf{y}})}\{\ln p(\underline{\mathbf{h}}, \underline{\mathbf{y}}, \mathcal{X})\} + \text{const} \\ \ln q(\underline{\mathbf{y}}) &= \mathbb{E}_{q(\underline{\mathbf{h}})}\{\ln p(\underline{\mathbf{h}}, \underline{\mathbf{y}}, \mathcal{X})\} + \text{const}, \end{aligned} \tag{13}$$

where $\mathbb{E}_{q(\underline{\mathbf{y}})}$ means taking expectation with respect to $\underline{\mathbf{y}}$ using $q(\underline{\mathbf{y}})$ as the density.

Note that $\ln q(\underline{\mathbf{h}})$ in Eq. 13 can be written as

$$\begin{aligned} &\ln q(\underline{\mathbf{h}}) = \sum\nolimits_i \ln q(\mathbf{h}_i) = \langle \ln p(\underline{\mathbf{h}}, \underline{\mathbf{y}}, \mathcal{X})\rangle_{\underline{\mathbf{y}}} + \text{const} \\ &= \langle \ln p(\mathcal{X}|\underline{\mathbf{h}}, \underline{\mathbf{y}})\rangle_{\underline{\mathbf{y}}} + \langle \ln p(\underline{\mathbf{h}}, \underline{\mathbf{y}})\rangle_{\underline{\mathbf{y}}} + \text{const} \\ &= \sum\nolimits_{ijp} \left\langle \ln \mathcal{N}(\mathbf{x}_{ij}^p | \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{R}\mathbf{y}_p, \boldsymbol{\Sigma}) \right\rangle_{\mathbf{y}_p} + \sum\nolimits_i \langle \ln \mathcal{N}(\mathbf{h}_i|\mathbf{0}, \mathbf{I})\rangle_{\underline{\mathbf{y}}} \\ &\qquad + \sum\nolimits_p \langle \ln \mathcal{N}(\mathbf{y}_p|\mathbf{0}, \mathbf{I})\rangle_{\mathbf{y}_p} + \text{const} \\ &= -\frac{1}{2} \sum\nolimits_{ijp} (\mathbf{x}_{ij}^p - \mathbf{m} - \mathbf{V}\mathbf{h}_i - \mathbf{R}\mathbf{y}_p^*)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{ij}^p - \mathbf{m} - \mathbf{V}\mathbf{h}_i - \mathbf{R}\mathbf{y}_p^*) - \frac{1}{2} \sum\nolimits_i \mathbf{h}_i^{\mathsf{T}} \mathbf{h}_i + \text{const}^1 \\ &= \sum\nolimits_i \left[\mathbf{h}_i^{\mathsf{T}} \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \sum\nolimits_{jp} (\mathbf{x}_{ij}^p - \mathbf{m} - \mathbf{R}\mathbf{y}_p^*) - \frac{1}{2} \mathbf{h}_i^{\mathsf{T}} \left(\mathbf{I} + \sum\nolimits_p H_i(p) \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{V} \right) \mathbf{h}_i^{\mathsf{T}} \right] + \text{const}. \end{aligned} \tag{14}$$

where $\mathbf{y}_p^* \equiv \langle \mathbf{y}_p | \mathcal{X} \rangle_{\mathbf{y}_p}$ is the posterior mean of $\mathbf{y}_p$ in the previous iteration and $\langle . \rangle_{\mathbf{y}_p}$ denotes the expectation with respect to $\mathbf{y}_p$.

By reading off $\mathbf{h}_i$ in Eq. 14 and comparing with $\sum_i \ln q(\mathbf{h}_i)$, we note that $q(\mathbf{h}_i)$ is a Gaussian with the following mean vector and precision matrix:

$$\mathbb{E}_{q(\mathbf{h}_i)}\{\mathbf{h}_i|\mathcal{X}\} = \langle \mathbf{h}_i|\mathcal{X}\rangle = \left(\mathbf{L}_i^{(1)}\right)^{-1} \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \sum_{p=1}^{P} \sum_{j=1}^{H_i(p)} (\mathbf{x}_{ij}^p - \mathbf{m} - \mathbf{R}\mathbf{y}_p^*)$$

---

[1] $\langle \ln \mathcal{N}(\mathbf{y}_p|\mathbf{0}, \mathbf{I})\rangle_{\mathbf{y}_p}$ is the differential entropy of normal distribution and is independent of $\mathbf{h}_i$, see Chapter 8 in [28].

and

$$\mathbf{L}_i^{(1)} \equiv \mathbf{I} + \sum_{p=1}^{P} H_i(p)\mathbf{V}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{V}.$$

As a result, the second-order moment required in the M-step can be computed as follows:

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

Similarly, the posterior mean and second-order moment of $\mathbf{y}_p$ can also be obtained by comparing the terms in $\ln q(\mathbf{y}_p)$ with a Gaussian distribution.

The M-step also requires the posterior moment $\langle \mathbf{h}_i\mathbf{y}_p^\mathsf{T}|\mathcal{X}\rangle$, which can be approximated by using variational Bayes principle:

$$p(\mathbf{h}_i, \mathbf{y}_p|\mathcal{X}) \approx q(\mathbf{h}_i)q(\mathbf{y}_p), \tag{15}$$

where both $q(\mathbf{h}_i)$ and $q(\mathbf{y}_p)$ are Gaussians. Based on the law of total expectation [1], the factorization in Eq. 15 gives

$$\begin{aligned}\langle \mathbf{y}_p\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle &\approx \langle \mathbf{y}_p|\mathcal{X}\rangle\langle \mathbf{h}_i|\mathcal{X}\rangle^\mathsf{T}\\ \langle \mathbf{h}_i\mathbf{y}_p^\mathsf{T}|\mathcal{X}\rangle &\approx \langle \mathbf{h}_i|\mathcal{X}\rangle\langle \mathbf{y}_p|\mathcal{X}\rangle^\mathsf{T}.\end{aligned}$$

Therefore, the equations for the variational E-step are as follows:

$$\mathbf{L}_i^{(1)} = \mathbf{I} + N_i\mathbf{V}^\top\boldsymbol{\Sigma}^{-1}\mathbf{V} \qquad i = 1, \ldots, S \tag{16}$$

$$\mathbf{L}_p^{(2)} = \mathbf{I} + B_p\mathbf{R}^\top\boldsymbol{\Sigma}^{-1}\mathbf{R} \qquad p = 1, \ldots, P \tag{17}$$

$$\langle \mathbf{h}_i|\mathcal{X}\rangle = (\mathbf{L}_i^{(1)})^{-1}\mathbf{V}^\top\boldsymbol{\Sigma}^{-1}\sum_{p=1}^{P}\sum_{j=1}^{H_i(p)}(\mathbf{x}_{ij}^p - \mathbf{m} - \mathbf{R}\mathbf{y}_p^*) \tag{18}$$

$$\langle \mathbf{y}_p|\mathcal{X}\rangle = (\mathbf{L}_p^{(2)})^{-1}\mathbf{R}^\top\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{S}\sum_{j=1}^{H_i(p)}(\mathbf{x}_{ij}^p - \mathbf{m} - \mathbf{V}\mathbf{h}_i^*) \tag{19}$$

$$\langle \mathbf{h}_i\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle = (\mathbf{L}_i^{(1)})^{-1} + \langle \mathbf{h}_i|\mathcal{X}\rangle\langle \mathbf{h}_i|\mathcal{X}\rangle^\mathsf{T} \tag{20}$$

$$\langle \mathbf{y}_p\mathbf{y}_p^\mathsf{T}|\mathcal{X}\rangle = (\mathbf{L}_p^{(2)})^{-1} + \langle \mathbf{y}_p|\mathcal{X}\rangle\langle \mathbf{y}_p|\mathcal{X}\rangle^\mathsf{T} \tag{21}$$

$$\langle \mathbf{y}_p\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle \approx \langle \mathbf{y}_p|\mathcal{X}\rangle\langle \mathbf{h}_i|\mathcal{X}\rangle^\mathsf{T} \tag{22}$$

$$\langle \mathbf{h}_i\mathbf{y}_p^\mathsf{T}|\mathcal{X}\rangle \approx \langle \mathbf{h}_i|\mathcal{X}\rangle\langle \mathbf{y}_p|\mathcal{X}\rangle^\mathsf{T} \tag{23}$$

Algorithm 1 shows the procedure of training a duration-invariant PLDA model.

Algorithm 1 Variational Bayes EM Algorithm for Duration-Invariant PLDA

Input:

Development data set consisting of i-vectors $\mathcal{X} = \{\mathbf{x}_{ij}^p | i = 1, \ldots, S; j = 1, \ldots, H_i(p); p = 1, \ldots, P\}$, with identity labels and duration group labels.

Initialization:

$\mathbf{y}_p^* \leftarrow \mathbf{0}$;

$\mathbf{\Sigma} \leftarrow 0.01\mathbf{I}$;

$\mathbf{V}, \mathbf{R} \leftarrow$ eigenvectors of PCA projection matrix learned using data set $\mathcal{X}$;

Parameter Estimation:

1) Compute $\mathbf{m}$ via Eq. 6;
2) Compute $\mathbf{L}_i^{(1)}$ and $\mathbf{L}_p^{(2)}$ according to Eq. 16 and Eq. 17, respectively;
3) Set $\mathbf{y}_p^*$ to the posterior mean of $\mathbf{y}_p$. Compute the posterior mean of $\mathbf{h}_i$ using Eq. 18;
4) Use the posterior mean of $\mathbf{h}_i$ computed in Step 3 to update the posterior mean of $\mathbf{y}_p$ according to Eq. 19;
5) Compute the other terms in the E-step (Eq. 20–Eq. 23);
6) Update the model parameters using Eq. 7 to Eq. 9;
7) Go to Step 2 until convergence;

Return: the parameters of the duration-invariant PLDA model $\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}, \mathbf{R}, \mathbf{\Sigma}\}$.

3.2. Likelihood Ratio Scores

If the durations of target and test utterances are not known (or not used), the likelihood ratio score can be computed in the same manner as in SI-PLDA [21]. Because the duration ℓ is usually known in practice, the likelihood ratio score can be also computed as follows:

$$S_{\mathrm{LR}}(\mathbf{x}_s, \mathbf{x}_t | \ell_s, \ell_t) = \ln \frac{p(\mathbf{x}_s, \mathbf{x}_t | \text{same-speaker}, \ell_s, \ell_t)}{p(\mathbf{x}_s, \mathbf{x}_t | \text{different-speakers}, \ell_s, \ell_t)}, \tag{24}$$

where $\mathbf{x}_s$ and $\mathbf{x}_t$ denote the target-speaker's i-vector and test i-vector, respectively, and ℓ_s and ℓ_t denote the durations of the corresponding utterances.

Based on different assumptions on the posterior density of $\mathbf{y}_p$, we propose two methods to calculate the score. They are derived in the following subsections.

3.2.1. Duration Factors with a Sharp Posterior Density

Assume that the duration ℓ of an utterance belongs to the p-th duration group and that the posterior density of $\mathbf{y}_p$ is sharp at its mean $\mathbf{y}_p^*$.[2] Then, the marginal-likelihood of i-vector $\mathbf{x}$ can be written as:

$$\begin{aligned} p(\mathbf{x}|\ell \in p\text{-th duration group}) &= \int_{\mathbf{h}} p(\mathbf{x}|\mathbf{h}, \mathbf{y}_p^*) p(\mathbf{h}) d\mathbf{h} \\ &= \int_{\mathbf{h}} \mathcal{N}(\mathbf{x}|\mathbf{m} + \mathbf{V}\mathbf{h} + \mathbf{R}\mathbf{y}_p^*, \boldsymbol{\Sigma}) \mathcal{N}(\mathbf{h}|\mathbf{0}, \mathbf{I}) d\mathbf{h} \\ &= \mathcal{N}(\mathbf{x}|\mathbf{m} + \mathbf{R}\mathbf{y}_p^*, \mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma}), \end{aligned} \tag{25}$$

where $\mathbf{y}_p^* \equiv \langle \mathbf{y}_p | \mathcal{X} \rangle$ is the posterior mean of $\mathbf{y}_p$. Given a test i-vector $\mathbf{x}_t$ and a target i-vector $\mathbf{x}_s$, we can use Eq. 25 to compute the likelihood ratio score:

$$\begin{aligned} S_{\mathrm{LR}}(\mathbf{x}_s, \mathbf{x}_t | \ell_s, \ell_t) &= \ln \frac{p(\mathbf{x}_s, \mathbf{x}_t | \text{same-speaker}, \ell_s, \ell_t)}{p(\mathbf{x}_s, \mathbf{x}_t | \text{different-speakers}, \ell_s, \ell_t)} \\ &= \ln \frac{\mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} + \mathbf{R}\mathbf{y}_{p_s}^* \\ \mathbf{m} + \mathbf{R}\mathbf{y}_{p_t}^* \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Psi} & \boldsymbol{\Sigma}_{ac} \\ \boldsymbol{\Sigma}_{ac} & \boldsymbol{\Psi} \end{bmatrix} \right)}{\mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} + \mathbf{R}\mathbf{y}_{p_s}^* \\ \mathbf{m} + \mathbf{R}\mathbf{y}_{p_t}^* \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Psi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \end{bmatrix} \right)} \\ &= \frac{1}{2} [\bar{\mathbf{x}}_s^\mathsf{T} \mathbf{Q} \bar{\mathbf{x}}_s + 2\bar{\mathbf{x}}_s^\mathsf{T} \mathbf{P} \bar{\mathbf{x}}_t + \bar{\mathbf{x}}_t^\mathsf{T} \mathbf{Q} \bar{\mathbf{x}}_t] + \text{const} \end{aligned} \tag{26}$$

where

$$\begin{aligned} \bar{\mathbf{x}}_s &= \mathbf{x}_s - \mathbf{m} - \mathbf{R}\mathbf{y}_{p_s}^* \\ \bar{\mathbf{x}}_t &= \mathbf{x}_t - \mathbf{m} - \mathbf{R}\mathbf{y}_{p_t}^* \\ \mathbf{Q} &= \boldsymbol{\Psi}^{-1} - (\boldsymbol{\Psi} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \\ \mathbf{P} &= \boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}_{ac}(\boldsymbol{\Psi} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \\ \boldsymbol{\Psi} &= \mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma}; \quad \boldsymbol{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^\mathsf{T}. \end{aligned}$$

3.2.2. Duration Factors with a Moderately Sharp Posterior

If the duration ℓ of an utterance falls on the p-th duration group and the posterior density of $\mathbf{y}_p$ is moderately sharp and follows a Gaussian

[2] This occurs when the number of training i-vectors B_p in the p-th duration group is large, as suggested by Eq. 17.

$\mathcal{N}(\mathbf{y}_p|\boldsymbol{\mu}_p^*, \boldsymbol{\Sigma}_p^*)$,[3] the marginal-likelihood of i-vector $\mathbf{x}$ is:

$$\begin{aligned}
p(\mathbf{x}|\ell \in p\text{-th duration group}) &= \int_{\mathbf{h}}\int_{\mathbf{y}_p} p(\mathbf{x}|\mathbf{h},\mathbf{y}_p)p(\mathbf{h})p(\mathbf{y}_p)d\mathbf{h}d\mathbf{y}_p \\
&= \int_{\mathbf{h}}\int_{\mathbf{y}_p} \mathcal{N}(\mathbf{x}|\mathbf{m}+\mathbf{V}\mathbf{h}+\mathbf{R}\mathbf{y}_p,\boldsymbol{\Sigma})\mathcal{N}(\mathbf{h}|\mathbf{0},\mathbf{I})\mathcal{N}(\mathbf{y}_p|\mathbf{0},\mathbf{I})d\mathbf{h}d\mathbf{y}_p \\
&= \int_{\mathbf{y}_p} \mathcal{N}(\mathbf{x}|\mathbf{m}+\mathbf{R}\mathbf{y}_p,\mathbf{V}\mathbf{V}^{\mathsf{T}}+\boldsymbol{\Sigma})\mathcal{N}(\mathbf{y}_p|\mathbf{0},\mathbf{I})d\mathbf{y}_p \\
&= \mathcal{N}(\mathbf{x}|\mathbf{m}+\mathbf{R}\boldsymbol{\mu}_p^*,\mathbf{V}\mathbf{V}^{\mathsf{T}}+\mathbf{R}\boldsymbol{\Sigma}_p^*\mathbf{R}^{\mathsf{T}}+\boldsymbol{\Sigma}),
\end{aligned} \tag{27}$$

where $\boldsymbol{\mu}_p^*$ can be computed according to Eq. 19 and $\boldsymbol{\Sigma}_p^*$ can be estimated from the inverse of $\mathbf{L}_p^{(2)}$ in Eq. 17. Given a test i-vector $\mathbf{x}_t$ and a target i-vector $\mathbf{x}_s$, the likelihood ratio score can be computed as:

$$\begin{aligned}
S_{\text{LR}}(\mathbf{x}_s,\mathbf{x}_t|\ell_s,\ell_t) &= \ln\frac{p(\mathbf{x}_s,\mathbf{x}_t|\text{same-speaker},\ell_s,\ell_t)}{p(\mathbf{x}_s,\mathbf{x}_t|\text{different-speakers},\ell_s,\ell_t)} \\
&= \ln\frac{\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s\\ \mathbf{x}_t\end{bmatrix}\middle|\begin{bmatrix}\mathbf{m}+\mathbf{R}\boldsymbol{\mu}_{p_s}^*\\ \mathbf{m}+\mathbf{R}\boldsymbol{\mu}_{p_t}^*\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_s & \boldsymbol{\Sigma}_{ac}\\ \boldsymbol{\Sigma}_{ac} & \boldsymbol{\Sigma}_t\end{bmatrix}\right)}{\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s\\ \mathbf{x}_t\end{bmatrix}\middle|\begin{bmatrix}\mathbf{m}+\mathbf{R}\boldsymbol{\mu}_{p_s}^*\\ \mathbf{m}+\mathbf{R}\boldsymbol{\mu}_{p_t}^*\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_s & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_t\end{bmatrix}\right)} \\
&= \frac{1}{2}[\bar{\mathbf{x}}_s^{\mathsf{T}}\mathbf{A}_{s,t}\bar{\mathbf{x}}_s + 2\bar{\mathbf{x}}_s^{\mathsf{T}}\mathbf{B}_{s,t}\bar{\mathbf{x}}_t + \bar{\mathbf{x}}_t^{\mathsf{T}}\mathbf{C}_{s,t}\bar{\mathbf{x}}_t] + \text{const}
\end{aligned} \tag{28}$$

where

$$\begin{aligned}
\bar{\mathbf{x}}_s &= \mathbf{x}_s - \mathbf{m} - \mathbf{R}\boldsymbol{\mu}_{p_s}^*; \quad \bar{\mathbf{x}}_t = \mathbf{x}_t - \mathbf{m} - \mathbf{R}\boldsymbol{\mu}_{p_t}^* \\
\mathbf{A}_{s,t} &= \boldsymbol{\Sigma}_s^{-1} - (\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \\
\mathbf{B}_{s,t} &= \boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Sigma}_{ac}(\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \\
\mathbf{C}_{s,t} &= \boldsymbol{\Sigma}_t^{-1} - (\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \\
\boldsymbol{\Sigma}_s &= \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{R}\boldsymbol{\Sigma}_{p_s}^*\mathbf{R}^{\mathsf{T}} + \boldsymbol{\Sigma}; \quad \boldsymbol{\Sigma}_t = \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{R}\boldsymbol{\Sigma}_{p_t}^*\mathbf{R}^{\mathsf{T}} + \boldsymbol{\Sigma}; \quad \boldsymbol{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^{\mathsf{T}}.
\end{aligned}$$

4. SNR- and Duration-invariant PLDA

This section describes a new modeling approach, namely SNR- and duration-invariant PLDA (SDI-PLDA), for robust speaker verification. Un-

[3] This occurs when the number of training i-vectors B_p in the p-th duration group is moderate, as suggested by Eq. 17.

like conventional Gaussian PLDA and SNR-invariant PLDA, the proposed model has three labeled latent factors representing speaker-specific, SNR-specific and duration-specific information, respectively.

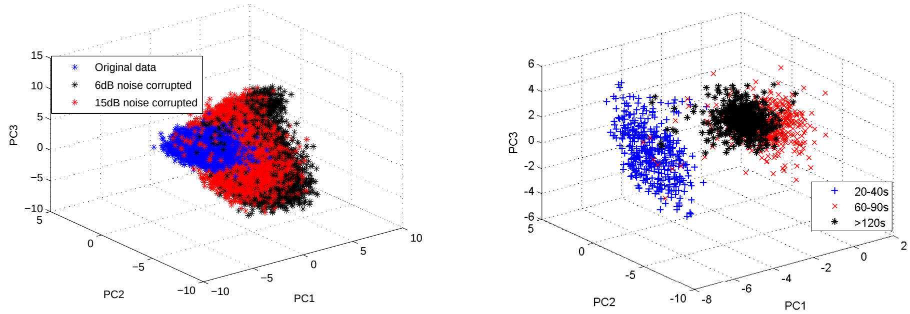
4.1. Generative Model

The SNR- and duration-invariant PLDA is inspired by the notion of Gaussian PLDA in which i-vectors from the same speaker should share a speaker latent factor. Similarly, this method is based on two hypotheses: (1) i-vectors derived from utterances that fall within a narrow SNR range should share similar SNR-specific information; and (2) i-vectors extracted from utterances with comparable durations should share similar duration-specific information.

To confirm the first hypothesis, we plotted three groups of i-vectors on the first 3 principal components in Fig. 2(a), where each group corresponds to a specific SNR-level shown in the legend. To ensure that the cluster displacement is not caused by speaker variability, each group contains the i-vectors from the same set of speakers. Evidently, the i-vectors form three clusters, one for each SNR group. To illustrate the the second hypothesis, we display three groups of i-vectors on their first 3 principal components in Fig. 2(b), where each group corresponds to one duration range shown in the legend. To ensure that the variability in i-vectors is not due to noise-level and speaker variabilities, all of the i-vectors were obtained from clean telephone conversations and each duration group comprises the same set of target speakers. Evidently, the i-vectors form three clusters and the locations of the clusters depend on the duration range.

From a modeling standpoint, both SNR-specific and duration-specific information can be captured using latent factors just like speaker factor in conventional PLDA model. We refer to these latent factors as SNR factor and duration factor in the remainder of this paper.

Under the above assumptions, an LDA- or NFA-projected i-vector [21] can be regarded as an observation generated from a linear generative model that comprises four components: (1) speaker component, (2) SNR component, (3) duration component, and (4) channel variability and the remaining variability that cannot be captured by the first three components. Assume that we have a set of D-dimensional NFA-projected i-vectors $\mathcal{X} = \{\hat{\mathbf{x}}_{ij}^{kp} | i = 1,\ldots,S; k = 1,\ldots,K; p = 1,\ldots,P; j = 1,\ldots,H_{ik}(p)\}$ obtained from S speakers, where $\hat{\mathbf{x}}_{ij}^{kp}$ is the j-th i-vector from speaker i, and k and p index to the SNR and duration groups to which the corresponding utterances belong,



(a) (b)

Figure 2: (a) Projection of i-vectors derived from utterances with different SNR-levels on their first three principal components. (b) Projection of i-vectors derived from variable-length utterances on their first three principal components.

respectively. In the proposed model, $\hat{\mathbf{x}}_{ij}^{kp}$ can be expressed as:

$$\hat{\mathbf{x}}_{ij}^{kp} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \mathbf{R}\mathbf{y}_p + \boldsymbol{\epsilon}_{ij}^{kp}, \tag{29}$$

where $\mathbf{m}$ is a $D \times 1$ vector representing the global offset, $\mathbf{h}_i$ is a $Q_1 \times 1$ vector denoting the speaker factor with prior distribution $\mathcal{N}(\mathbf{h}|\mathbf{0}, \mathbf{I})$, $\mathbf{w}_k$ is a $Q_2 \times 1$ vector denoting the latent SNR factor with a prior distribution of $\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I})$, $\mathbf{y}_p$ is a $Q_3 \times 1$ vector denoting the latent duration factor with a standard normal prior, $\boldsymbol{\epsilon}_{ij}^{kp}$ is a $D \times 1$ vector denoting the residue which follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{V}$ is a $D \times Q_1$ matrix whose columns span the speaker subspace, $\mathbf{U}$ is a $D \times Q_2$ matrix whose columns span the SNR subspace, and $\mathbf{R}$ is a $D \times Q_3$ matrix which defines the duration subspace. $\mathbf{h}_i$, $\mathbf{w}_k$, and $\mathbf{y}_p$ are assumed to be independent in their prior. Fig. 3 shows the graphical model of SDI-PLDA.

The proposed SNR- and duration-invariant PLDA is different from the conventional PLDA in that the former makes use of multiple labels (speaker IDs, SNR levels, and duration ranges) for training the loading matrices, whereas the latter only uses the speaker IDs. To exploit the duration information in the training utterances, the proposed model has an additional subspace called duration subspace, which results in an extra latent factor called duration factor. Unlike the term $\mathbf{G}\mathbf{r}_{ij}$ in Eq. 1, which is speaker- and session-dependent, the SNR component $\mathbf{U}\mathbf{w}_k$ and the duration component $\mathbf{R}\mathbf{y}_p$ in Eq. 29 depend on the SNR groups and duration groups, respectively.

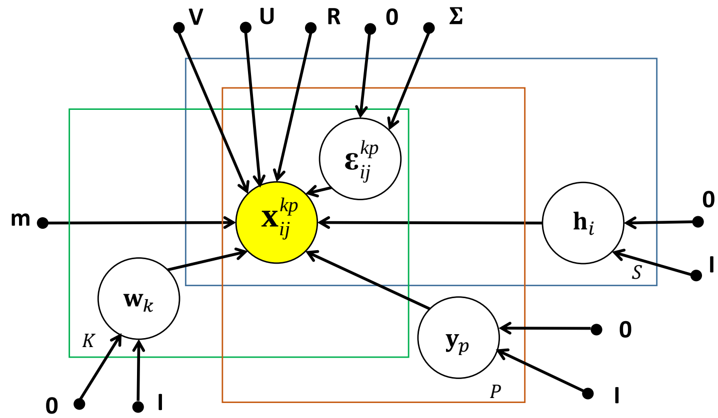


Figure 3: Probabilistic graphical model of SDI-PLDA.

4.2. Variational Bayes EM algorithm

Denote $\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}, \mathbf{U}, \mathbf{R}, \boldsymbol{\Sigma}\}$ as the parameters of the SNR- and duration-invariant PLDA model. These parameters can be learned from a training set using maximum likelihood estimation. Given an old estimate of $\boldsymbol{\theta}$, we aim to find a new estimate $\boldsymbol{\theta}'$ that maximizes the auxiliary function:

$$\begin{aligned} Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) &= \mathbb{E}_{q(\underline{\mathbf{h}},\underline{\mathbf{w}},\underline{\mathbf{y}})}\Big[\ln p(\mathcal{X}, \underline{\mathbf{h}}, \underline{\mathbf{w}}, \underline{\mathbf{y}}|\boldsymbol{\theta}')\Big|\mathcal{X}, \boldsymbol{\theta}\Big] \\ &= \mathbb{E}_{q(\underline{\mathbf{h}},\underline{\mathbf{w}},\underline{\mathbf{y}})}\Big[\sum_{ikpj} \ln\Big(p(\hat{\mathbf{x}}_{ij}^{kp}|\mathbf{h}_i, \mathbf{w}_k, \mathbf{y}_p, \boldsymbol{\theta}')p(\mathbf{h}_i, \mathbf{w}_k, \mathbf{y}_p)\Big)\Big|\mathcal{X}, \boldsymbol{\theta}\Big]. \end{aligned} \tag{30}$$

To maximize Eq.30, we need to estimate the posterior distributions of the latent variables given the model parameters $\boldsymbol{\theta}$. Denote $N_i = \sum_{kp} H_{ik}(p)$ as the number of training i-vectors from the i-th speaker, $M_k = \sum_{ip} H_{ik}(p)$ as the number of training i-vectors falling in the k-th SNR group, and $B_p = \sum_{ik} H_{ik}(p)$ as the number of training i-vectors in the p-th duration group. Similar to the derivation of duration-invariant PLDA, the variational E-step of the proposed model in Eq. 29 can be derived by the variational Bayes method as follows :

$$\mathbf{L}_i^{(1)} = \mathbf{I} + N_i \mathbf{V}^\top \boldsymbol{\Sigma}^{-1} \mathbf{V} \qquad i = 1, \ldots, S \tag{31}$$

$$\mathbf{L}_k^{(2)} = \mathbf{I} + M_k \mathbf{U}^\top \boldsymbol{\Sigma}^{-1} \mathbf{U} \qquad k = 1, \ldots, K \tag{32}$$

$$\mathbf{L}_p^{(3)} = \mathbf{I} + B_p \mathbf{R}^\top \boldsymbol{\Sigma}^{-1} \mathbf{R} \qquad p = 1, \ldots, P \tag{33}$$

$$\langle \mathbf{h}_i|\mathcal{X}\rangle = (\mathbf{L}_i^{(1)})^{-1}\mathbf{V}^\top\mathbf{\Sigma}^{-1}\sum\nolimits_{kpj}(\hat{\mathbf{x}}_{ij}^{kp} - \mathbf{m} - \mathbf{U}\mathbf{w}_k^* - \mathbf{R}\mathbf{y}_p^*) \tag{34}$$

$$\langle \mathbf{w}_k|\mathcal{X}\rangle = (\mathbf{L}_k^{(2)})^{-1}\mathbf{U}^\top\mathbf{\Sigma}^{-1}\sum\nolimits_{ipj}(\hat{\mathbf{x}}_{ij}^{kp} - \mathbf{m} - \mathbf{V}\mathbf{h}_i^* - \mathbf{R}\mathbf{y}_p^*) \tag{35}$$

$$\langle \mathbf{y}_p|\mathcal{X}\rangle = (\mathbf{L}_p^{(3)})^{-1}\mathbf{R}^\top\mathbf{\Sigma}^{-1}\sum\nolimits_{ikj}(\hat{\mathbf{x}}_{ij}^{kp} - \mathbf{m} - \mathbf{V}\mathbf{h}_i^* - \mathbf{U}\mathbf{w}_k^*) \tag{36}$$

$$\langle \mathbf{h}_i\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle = (\mathbf{L}_i^{(1)})^{-1} + \langle \mathbf{h}_i|\mathcal{X}\rangle\langle \mathbf{h}_i|\mathcal{X}\rangle^\mathsf{T} \tag{37}$$

$$\langle \mathbf{w}_k\mathbf{w}_k^\mathsf{T}|\mathcal{X}\rangle = (\mathbf{L}_k^{(2)})^{-1} + \langle \mathbf{w}_k|\mathcal{X}\rangle\langle \mathbf{w}_k|\mathcal{X}\rangle^\mathsf{T} \tag{38}$$

$$\langle \mathbf{y}_p\mathbf{y}_p^\mathsf{T}|\mathcal{X}\rangle = (\mathbf{L}_p^{(3)})^{-1} + \langle \mathbf{y}_p|\mathcal{X}\rangle\langle \mathbf{y}_p|\mathcal{X}\rangle^\mathsf{T} \tag{39}$$

$$\langle \mathbf{w}_k\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle \approx \langle \mathbf{w}_k|\mathcal{X}\rangle\langle \mathbf{h}_i|\mathcal{X}\rangle^\mathsf{T} \tag{40}$$

$$\langle \mathbf{h}_i\mathbf{w}_k^\mathsf{T}|\mathcal{X}\rangle \approx \langle \mathbf{h}_i|\mathcal{X}\rangle\langle \mathbf{w}_k|\mathcal{X}\rangle^\mathsf{T} \tag{41}$$

$$\langle \mathbf{w}_k\mathbf{y}_p^\mathsf{T}|\mathcal{X}\rangle \approx \langle \mathbf{w}_k|\mathcal{X}\rangle\langle \mathbf{y}_p|\mathcal{X}\rangle^\mathsf{T} \tag{42}$$

$$\langle \mathbf{y}_p\mathbf{w}_k^\mathsf{T}|\mathcal{X}\rangle \approx \langle \mathbf{y}_p|\mathcal{X}\rangle\langle \mathbf{w}_k|\mathcal{X}\rangle^\mathsf{T} \tag{43}$$

$$\langle \mathbf{h}_i\mathbf{y}_p^\mathsf{T}|\mathcal{X}\rangle \approx \langle \mathbf{h}_i|\mathcal{X}\rangle\langle \mathbf{y}_p|\mathcal{X}\rangle^\mathsf{T} \tag{44}$$

$$\langle \mathbf{y}_p\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle \approx \langle \mathbf{y}_p|\mathcal{X}\rangle\langle \mathbf{h}_i|\mathcal{X}\rangle^\mathsf{T} \tag{45}$$

where $\mathbf{w}_k^*$, $\mathbf{y}_p^*$, and $\mathbf{h}_i^*$ denote the posterior mean of $\mathbf{w}_k$, $\mathbf{y}_p$, and $\mathbf{h}_i$ in the previous iteration, respectively.

Given Eq. 31–Eq. 45, the model parameters $\boldsymbol{\theta}'$ can be estimated via the M-step is as follows:

$$\mathbf{m} = \frac{1}{N}\sum\nolimits_{ikpj}\hat{\mathbf{x}}_{ij}^{kp} \tag{46}$$

$$\begin{aligned}\mathbf{V}' = &\Big\{\sum\nolimits_{ikpj}\Big[(\hat{\mathbf{x}}_{ij}^{kp} - \mathbf{m})\langle \mathbf{h}_i|\mathcal{X}\rangle^\top - \mathbf{U}\langle \mathbf{w}_k\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle - \mathbf{R}\langle \mathbf{y}_p\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle\Big]\Big\}\\ &\Big[\sum\nolimits_{ikpj}\langle \mathbf{h}_i\mathbf{h}_i^\mathsf{T}|\mathcal{X}\rangle\Big]^{-1}\end{aligned} \tag{47}$$

$$\begin{aligned}\mathbf{U}' = &\Big\{\sum\nolimits_{ikpj}\Big[(\hat{\mathbf{x}}_{ij}^{kp} - \mathbf{m})\langle \mathbf{w}_k|\mathcal{X}\rangle^\top - \mathbf{V}\langle \mathbf{h}_i\mathbf{w}_k^\mathsf{T}|\mathcal{X}\rangle - \mathbf{R}\langle \mathbf{y}_p\mathbf{w}_k^\mathsf{T}|\mathcal{X}\rangle\Big]\Big\}\\ &\Big[\sum\nolimits_{ikpj}\langle \mathbf{w}_k\mathbf{w}_k^\mathsf{T}|\mathcal{X}\rangle\Big]^{-1}\end{aligned} \tag{48}$$

$$\begin{aligned}\mathbf{R}' = &\Big\{\sum\nolimits_{ikpj}\Big[(\hat{\mathbf{x}}_{ij}^{kp} - \mathbf{m})\langle \mathbf{y}_p|\mathcal{X}\rangle^\top - \mathbf{V}\langle \mathbf{h}_i\mathbf{y}_p^\mathsf{T}|\mathcal{X}\rangle - \mathbf{U}\langle \mathbf{w}_k\mathbf{y}_p^\mathsf{T}|\mathcal{X}\rangle\Big]\Big\}\\ &\Big[\sum\nolimits_{ikpj}\langle \mathbf{y}_p\mathbf{y}_p^\mathsf{T}|\mathcal{X}\rangle\Big]^{-1}\end{aligned} \tag{49}$$

$$\begin{aligned}\mathbf{\Sigma}' =& \frac{1}{N}\sum\nolimits_{ikpj}\Big[(\hat{\mathbf{x}}_{ij}^{kp}-\mathbf{m})(\hat{\mathbf{x}}_{ij}^{kp}-\mathbf{m})^{\top}-\mathbf{V}\langle\mathbf{h}_i|\mathcal{X}\rangle(\hat{\mathbf{x}}_{ij}^{kp}-\mathbf{m})^{\top} \\ &-\mathbf{U}\langle\mathbf{w}_k|\mathcal{X}\rangle(\hat{\mathbf{x}}_{ij}^{kp}-\mathbf{m})^{\top}-\mathbf{R}\langle\mathbf{y}_p|\mathcal{X}\rangle(\hat{\mathbf{x}}_{ij}^{kp}-\mathbf{m})^{\top}\Big]\end{aligned} \tag{50}$$

where $N = \sum_{i=1}^{S} N_i = \sum_{k=1}^{K} M_k$. Algorithm 2 shows the procedure of applying the variational EM algorithm.

Algorithm 2 Variational Bayes EM Algorithm for SNR- and Duration-Invariant PLDA

Input:

Development data set comprising NFA-projected i-vectors $\mathcal{X} = \{\hat{\mathbf{x}}_{ij}^{kp}|i = 1,\ldots,S; k = 1,\ldots,K; p = 1,\ldots,P; j = 1,\ldots,H_{ik}(p)\}$, with speaker labels, SNR group labels, and duration labels.

Initialization:

$\mathbf{y}_p^* \leftarrow \mathbf{0}$, $\mathbf{w}_k^* \leftarrow \mathbf{0}$;

$\mathbf{\Sigma} \leftarrow 0.01\mathbf{I}$;

$\mathbf{V}, \mathbf{U}, \mathbf{R} \leftarrow$ eigenvectors obtained from the PCA of $\mathcal{X}$;

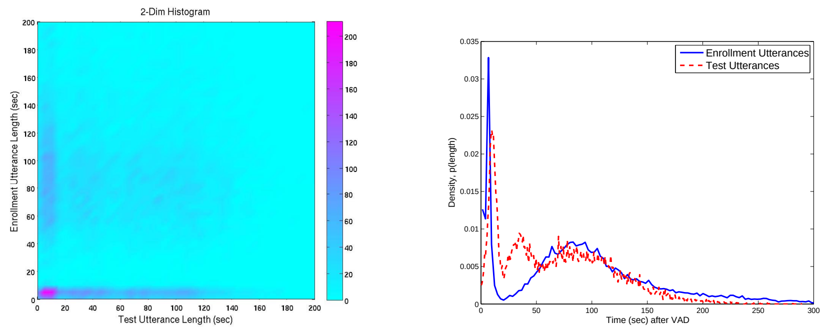
Parameter Estimation:

1) Compute $\mathbf{m}$ via Eq. 46;
2) Compute $\mathbf{L}_i^{(1)}$, $\mathbf{L}_k^{(2)}$, and $\mathbf{L}_p^{(3)}$ according to Eq. 31 to Eq. 33, respectively;
3) Compute the posterior mean of $\mathbf{h}_i$ using Eq. 34;
4) Use the posterior mean of $\mathbf{h}_i$ computed in Step 3 to update the posterior means of $\mathbf{w}_k$ and $\mathbf{y}_p$ using Eq. 35–Eq. 36;
5) Compute the other terms in the E-step (Eq. 37–Eq. 45);
6) Update the model parameters using Eq. 47 to Eq. 50;
7) Set $\mathbf{y}_p^* = \langle\mathbf{y}_p|\mathcal{X}\rangle$, $\mathbf{w}_k^* = \langle\mathbf{w}_k|\mathcal{X}\rangle$, and $\mathbf{h}_i^* = \langle\mathbf{h}_i|\mathcal{X}\rangle$;
8) Go to step 2 until convergence;

Return: the parameters of the SNR- and duration-invariant PLDA model $\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}, \mathbf{U}, \mathbf{R}, \mathbf{\Sigma}\}$.

4.3. Likelihood Ratio Scores

Assume that both the duration and SNR of target-speaker's utterance and test utterance are not known. Denote $\mathbf{x}_s$ and $\mathbf{x}_t$ as the NFA-project i-vectors of the target-speaker and test utterance, respectively, the likelihood



(a) (b)

Figure 4: Distribution of utterance duration in NIST 2012 SRE. (a) 2-D histogram showing the length distribution of all possible target-test pairs. (b) Length distributions of enrollment utterances and test utterances (after VAD).

ratio score is

$$
\begin{aligned}
S_{\mathrm{LR}}(\mathbf{x}_s, \mathbf{x}_t) &= \ln \frac{P\left(\mathbf{x}_s, \mathbf{x}_t | \text{same-speaker}\right)}{P(\mathbf{x}_s, \mathbf{x}_t | \text{different-speakers})} \\
&= \text{const} + \frac{1}{2}\bar{\mathbf{x}}_s^\top \mathbf{Q}\bar{\mathbf{x}}_s + \frac{1}{2}\bar{\mathbf{x}}_t^\top \mathbf{Q}\bar{\mathbf{x}}_t + \bar{\mathbf{x}}_s^\top \mathbf{P}\bar{\mathbf{x}}_t
\end{aligned}
\tag{51}
$$

where

$$
\begin{aligned}
\bar{\mathbf{x}}_s &= \mathbf{x}_s - \mathbf{m}, \qquad \bar{\mathbf{x}}_t = \mathbf{x}_t - \mathbf{m}, \\
\mathbf{P} &= \boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac}(\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac})^{-1}, \\
\mathbf{Q} &= \boldsymbol{\Sigma}_{tot}^{-1} - (\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac})^{-1}, \\
\boldsymbol{\Sigma}_{ac} &= \mathbf{V}\mathbf{V}^\top, \quad \text{and} \quad \boldsymbol{\Sigma}_{tot} = \mathbf{V}\mathbf{V}^\top + \mathbf{U}\mathbf{U}^\top + \mathbf{R}\mathbf{R}^\top + \boldsymbol{\Sigma}.
\end{aligned}
$$

See Appendix A for the derivation of Eq. 51. When the utterance duration and SNR are known, the scoring function can be derived using the principles in Section 3.2.

Because $\mathbf{P}$ and $\mathbf{Q}$ can be computed in advance, the computational complexity of SDI-PLDA is the same as that of Gaussian PLDA [7].

5. Experimental Setup

5.1. Evaluation Protocol and Speech Data

Experiments were performed on common conditions (CC) 1 and 4 of the core set of NIST 2012 Speaker Recognition Evaluation (SRE) [27]. We used

Table 1: Abbreviations of various PLDA models.

Abbreviation	Model Name	Formula
PLDA	Probabilistic LDA	$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \boldsymbol{\epsilon}_{ij}$ (Eq. 2)
UP-PLDA	Uncertainty propagation PLDA	$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{R}_{ij}\mathbf{y}_{ij} + \boldsymbol{\epsilon}_{ij}$ (Eq. 2 in [19])
SI-PLDA	SNR-invariant PLDA	$\mathbf{x}^k_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \boldsymbol{\epsilon}^k_{ij}$ (Eq. 3)
DI-PLDA	Duration-invariant PLDA	$\mathbf{x}^p_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{R}\mathbf{y}_p + \boldsymbol{\epsilon}^p_{ij}$ (Eq. 4)
SDI-PLDA	SNR- and duration-invariant PLDA	$\mathbf{x}^{kp}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \mathbf{R}\mathbf{y}_p + \boldsymbol{\epsilon}^{kp}_{ij}$ (Eq. 29)

data from NIST 2005–2010 for system development. The speech data were divided into the following parts:

- *Test Data*: Test utterances involved in CC1 comprise clean interview conversations. Test data in CC4 comprise noise contaminated telephone conversations with SNR ranging from 0dB to 50dB. Readers may refer to [21] for the SNR distributions of test utterances in these common conditions and the procedure for measuring SNRs.

- *Enrollment Data*: Enrollment data for CC1 comprise target-speakers' conversations recorded by using different types of microphones. For CC4, enrollment data comprises the telephone conversations of target speakers. Each target speaker has one or more conversations recorded over different telephone channels and with different durations longer than 10 seconds.

- *Development Data*: Development data were used for estimating the subspace projection matrices (WCCN and NFA) for i-vector preprocessing. They were also used for estimating the parameters of PLDA, uncertainty propagation PLDA in [19], SNR-invariant PLDA and SNR- and duration-invariant PLDA models (see Table 1 for the abbreviations of these models). For experiments on CC1, the development data consist of microphone segments in 2005–2010 SREs. For CC4, the development data comprise two parts. The first part was extracted from the telephone and microphone segments in 2005–2010 SREs and the second part was obtained by adding babble noise to the telephone segments of 2005–2010 SREs at different SNRs. The procedure of producing these noisy speech files is described in Section IV-B of [21]. For each gender, 14,000 noise corrupted files with SNR ranging from 2dB

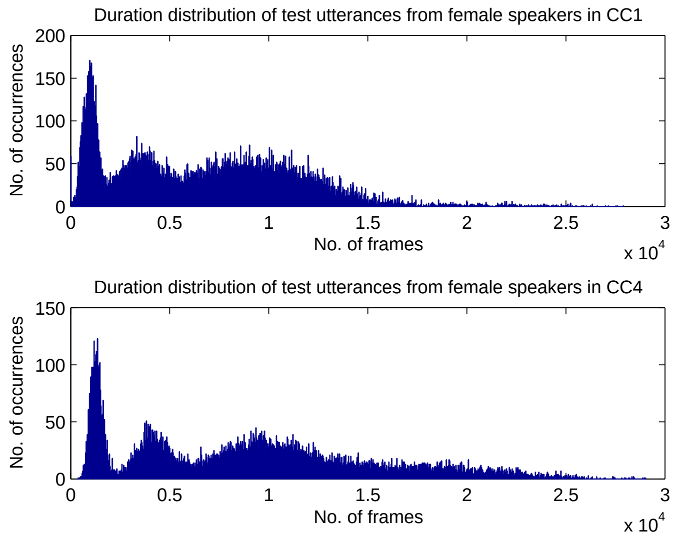


Figure 5: Duration distributions of test utterances (after VAD) from female speakers in CC1 and CC4 of NIST 2012 SRE.

to 15dB were randomly selected from all of the noise corrupted files. Speakers with less than 10 conversations were excluded. Both the microphone and telephone conversations from NIST 2005–2008 SREs were used as development data to train the gender-dependent UBMs and total variability matrices.

Fig. 4 shows the duration distribution of the enrollment and test utterances (after VAD) in NIST 2012 SRE. Evidently, there are many trials that involve short test utterances tested against long enrollment utterances or long test utterances tested against short enrollment utterances. The duration distributions of test utterances (after VAD) in CC1 and CC4 are shown in Fig. 5. It is obvious that the test utterances in CC1 and CC4 covers a wide range of durations.

5.2. Acoustic Feature Extraction

For each conversation, a two-channel voice activity detector (VAD) [25, 37] was applied to prune out silence regions. The VAD is specifically designed for NIST SREs. Special attention has been paid to address utterances with low SNR, impulsive noise, and cross talks in the interview speech files. The main idea is to apply speech enhancement as a pre-processing step to boost energy contrast between speech and non-speech regions, which facilitates the subsequence speech/non-speech decisions either by log-likelihood ratio tests or by comparing with energy-based thresholds.

The speech regions of each utterance were segmented into 25-ms Hamming windowed frames with 10-ms frame shift. For each frame, the first 19 Mel frequency cepstral coefficients (MFCC) and log energy together with their first and second derivatives were packed to form a 60-dimensional acoustic vector. Cepstral mean normalization and feature warping [29] with a window size of 3 seconds were then applied to the acoustic vectors.

5.3. I-vector Extraction and PLDA Modeling

I-vectors were extracted based on gender-dependent UBMs with 1024 mixtures and total variability matrices with 500 total factors. Similar to [26], we applied within-class covariance normalization (WCCN) [12] to whiten the i-vectors, followed by length normalization (LN) to reduce the non-Gaussian behavior of the 500-dimensional i-vectors. Then, nonparametric feature analysis (NFA) [21, 22] was applied to reduce intra-speaker variability and emphasize discriminative information. This procedure projects the i-vectors onto a 400-dimensional subspace. The NFA-projected i-vectors were then used to train PLDA models with 300 speaker factors ($Q_1 = 300$). In our experiments, we make sure that the number of speaker factors plus SNR and duration factors is no more than 400.

5.4. SNR and Duration Groups

To determine the SNR and duration subspaces in the SI-PLDA, DI-PLDA and SDI-PLDA models, the development data described in Section 5.1 were divided into K groups according to the measured SNRs and duration of the utterances, where K varied from 3 to 8.[4] Because SNR and duration are continuous variables, there will be infinite possible ways of dividing them into intervals. Therefore, we evenly divided the training utterances into K groups such that each group contains almost the same number of training i-vectors. Although this partitioning method leads to unequal SNR and duration intervals, it ensures that each group has sufficient training i-vectors for estimating the SNR and duration loading matrices reliably. Table 2 lists the SNR range and duration range when $K = 8$ and $P = 8$ in Eq. 29.

[4]To be precise, the first $K-1$ groups have $\lfloor \frac{N}{K} \rfloor$ i-vectors, whereas the last one contains $N-(K-1)\lfloor \frac{N}{K} \rfloor$, where $\lfloor x \rfloor$ is the floor of x.

Table 2: Division of SNR and duration groups for the training data of female speakers in CC4, when $K = 8$ and $P = 8$ in Eq. 29.

Group	SNR Range (dB)	Duration Range (s)
1	2.0–5.5	3–70
2	5.5–7.4	70–88
3	7.4–10.5	88–103
4	10.5–14.6	103–117
5	14.6–19.8	117–133
6	19.8–34.4	133–152
7	34.4–39.1	152–184
8	39.1–55.0	184–1215

6. Results and Analysis

We used equal error rate (EER) and minCprimary, which is the same as minimum normalized decision cost function (minDCF) defined in NIST 2012 SRE [27], to evaluate the performance of different PLDA models. Table 1 summarizes their abbreviations and formulations.

6.1. Effectiveness of SNR and Duration Factors

The first experiment aims to compare the effectiveness of SI-PLDA and DI-PLDA in compensating SNR variability and duration variability, respectively. Table 3 shows the results of SI-PLDA and DI-PLDA on CC1 and CC4 for different numbers of SNR groups and duration groups. The results show that both SI-PLDA and DI-PLDA outperform PLDA. This suggests that including the duration subspace in DI-PLDA and the SNR subspace in SI-PLDA enables these models to address mismatch caused by duration and SNR, respectively. Moreover, SI-PLDA not only outperforms DI-PLDA (in EER) in most cases, but also performs stably with respect to the number of groups K. On the other hand, the performance of DI-PLDA on CC1 (female) drops when the number of groups increases to 8.

The second experiment compares the proposed SDI-PLDA model with other PLDA models and PLDA with uncertainty propagation. Results in Table 3 show that SDI-PLDA achieves the best performance in terms of EER and minDCF on CC4. This result suggests that SDI-PLDA can compensate for SNR and duration variabilities in the i-vector space. While UP-PLDA achieves the best performance in CC1, it performs badly under CC4. The reason is that CC4 involves more test trials with long duration than CC1 (as shown in Fig. 5). As the i-vectors corresponding to utterances of long

Table 3: Performance of PLDA, UP-PLDA, SI-PLDA, DI-PLDA and SDI-PLDA in CC1 and CC4 of NIST 2012 SRE core set. K and P denote the number of SNR and duration groups, respectively. The best results are highlighted in boldface.

Model	K	P	Male				Female			
			CC1		CC4		CC1		CC4	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
PLDA	–	–	5.28	0.374	2.69	0.317	7.39	0.514	2.35	0.332
UP-PLDA	–	–	**3.89**	**0.346**	3.55	0.493	**5.47**	**0.408**	3.36	0.483
SI-PLDA	3	–	5.28	0.369	2.56	0.292	7.06	0.504	2.18	0.299
	4	–	5.28	0.368	2.56	0.287	7.12	0.499	2.18	0.287
	5	–	5.15	0.395	2.50	0.288	7.07	0.498	2.12	0.291
	6	–	5.22	0.370	2.51	0.281	7.14	0.508	2.16	0.292
	7	–	5.36	0.381	2.48	0.284	7.13	0.505	2.16	0.287
	8	–	5.23	0.369	2.48	0.288	7.03	0.506	2.13	0.287
DI-PLDA	–	3	5.42	0.368	2.60	0.287	6.98	0.512	2.25	0.287
	–	4	5.57	0.369	2.55	0.291	6.98	0.499	2.23	0.299
	–	5	5.42	0.369	2.56	0.287	7.02	0.503	2.25	0.289
	–	6	5.21	0.369	2.52	0.289	7.30	0.526	2.16	0.293
	–	7	5.36	0.381	2.55	0.287	7.38	0.534	2.18	0.302
	–	8	5.23	0.369	2.56	0.289	8.95	0.563	2.20	0.288
SDI-PLDA	3	3	5.42	0.375	2.52	0.287	6.93	0.496	2.15	**0.284**
	4	4	5.41	0.368	2.54	0.289	6.97	0.491	2.13	0.288
	5	5	5.13	0.367	2.55	0.288	6.96	0.491	2.15	0.289
	6	6	5.14	0.367	2.49	0.283	7.05	0.508	2.14	0.289
	7	7	5.42	0.373	**2.34**	**0.280**	7.13	0.505	2.13	0.289
	8	8	5.54	0.373	2.49	0.286	8.08	0.556	**2.11**	**0.284**

duration are reliable, UP-PLDA loses its advantage in handling reliable i-vectors compared to PLDA.

To confirm that SDI-PLDA really outperforms SI-PLDA and DI-PLDA, we performed McNemar's tests [9] on the differences between the EERs. For each model, the best performing configuration (by varying K and P) was used in the tests. The p-values of these tests are shown in Table 4. As the p-values between SDI-PLDA and the other two models are less than 0.05, we conclude that SDI-PLDA outperforms SI-PLDA and DI-PLDA.

We have also linearly fused the scores of the best performing DI-PLDA and SI-PLDA in CC4, with fusion weights for DI-PLDA and SI-PLDA set to 0.65 and 0.35, respectively. For male speakers, the EER after fusion is 2.41% and the minDCF is 0.282. For female speakers, the EER after fusion is 2.13% and the minDCF is 0.283. This fusion performance is comparable

Table 4: P-values of McNemar's tests [9] on the differences in EERs based on CC4 of NIST 2012 SRE core set, male speakers. For each entry, $p < 0.05$ means that the difference between the EERs is statistically significant at a confidence level of 95%.

Method	DI-PLDA	SDI-PLDA
SI-PLDA	0.002	0.013
DI-PLDA	–	0.022

Table 5: Performance of DI-PLDA on CC1 and CC4 of NIST 2012 SRE (male, core set) using different approaches to deriving the EM training algorithm and the scoring function. *EM*: EM is derived by assuming that the latent factors are posteriorly independent (Eqs. 14–24 of [21]). *VB-EM*: EM is derived by using variational Bayes and the latent factors are assumed posteriorly dependent (Eq. 6–Eq. 23). *Scoring1*: The duration is unknown during scoring (Eq. 26 in [21]). *Scoring2*: The duration is known during scoring, and the duration factor has sharp posterior (Eq. 26). *Scoring3*: The duration is known during scoring, and the duration factor has blunt posterior (Eq. 28). P in Eq. 4 was set to 6.

Method	CC1		CC4	
	EER(%)	minDCF	EER(%)	minDCF
EM + Scoring1	5.21	0.369	2.52	0.289
VB-EM + Scoring1	5.42	0.366	2.56	0.285
EM + Scoring2	5.42	0.371	2.58	0.290
EM + Scoring3	5.28	0.366	2.57	0.290

with that of SDI-PLDA, suggesting that SNR and duration variabilities can be handled either in the model domain (Eq. 29) or in the score domain. But the later requires a set of optimal fusion weights to achieve a performance comparable to that of the former.

6.2. Numbers of SNR and Duration Groups

Another observation from Table 3 is that the performance of DI-PLDA and SDI-PLDA on CC1 becomes worse when the numbers of SNR and duration groups, K and P, increase. This suggests that appropriate values of K and P are important for DI-PLDA and SDI-PLDA. Since the amount of training data for CC1 is much less than that for CC4, when K and P increase, the number of training samples in each group becomes limited, causing unreliable estimation of SNR and duration loading matrices. Hence, the values of K and P should be determined based on the amount of train-

Table 6: Performance of SDI-PLDA in CC4 of NIST 2012 SRE core set for male speakers with varying numbers of SNR and duration factors. The numbers of SNR and duration groups were fixed to 7.

Q_2 & Q_3	EER(%)	minDCF
10	2.34	0.280
20	2.33	0.281
30	2.35	0.283
40	2.37	0.286

ing data. In particular, if K and P are very large, there will be so many SNR factors and duration factors that each i-vector is considered to be obtained from a unique SNR or duration. This means that the SNR- and duration-invariant PLDA models reduce to the traditional Gaussian PLDA, which only considers the session variability instead of the variability caused by different SNRs and durations.

6.3. Combinations of Training and Scoring Methods

Table 5 shows the performance of DI-PLDA under different combinations of training methods and scoring methods. The results suggest that using the original training method (EM) and scoring method (Scoring1) in [21] achieves the best results, which assumes that the latent factors are posteriorly independent and that the SNR and duration of utterance are unknown. Although the EM algorithm derived from variational Bayes (VB) is more theoretically justifiable, VB-EM + Scoring1 in Table 5 does not outperform EM + Scoring1. This is a rather unexpected result. One possible reason is that because the number of training i-vectors in each duration group is large enough to make the posterior density of duration factors ($\mathbf{y}_p$ in Eq. 4) very sharp, causing the joint posterior density $p(\mathbf{h}_i, \mathbf{y}_p|\mathbf{x})$ to spread mainly along $\mathbf{h}_i$ instead of spreading over both $\mathbf{h}_i$ and $\mathbf{y}_p$. As a result, the assumption that the latent factors $\mathbf{h}_i$ and $\mathbf{y}_p$ are posteriorly independent becomes valid.

Comparing Rows 1, 3, and 4 in Table 5 suggests that scoring with duration information does not achieve any advantage. This may be because the EM and VB-EM have already taken duration variability into account through the duration loading matrix.

6.4. Numbers of SNR and Duration Factors

To investigate the effect of varying the number of SNR and duration factors, we set Q_2 and Q_3 to different values but keeping K and P fixed.

Table 7: Performance (EER(%)/minDCF) of (a) PLDA and (b) SDI-PLDA on CC4 of NIST 2012 SRE (male core set) under different combinations of SNR (dB) and utterance durations. The last row shows the relative increases in EER/minDCF when SNR decreases from 15dB to 6dB. The last column shows the relative increases in EER/minDCF when number of frames reduces from 2500 to 1000. For each SNR and duration combination, the one with a smaller increase in EER or minDCF is highlighted in boldface.

SNR	Number of frames					Relative Inc.
	1000	1334	1667	2000	2500	
6	6.02/0.517	5.40/0.483	4.78/0.454	4.35/0.431	3.86/0.409	0.560/**0.264**
8	5.47/0.499	4.79/0.454	4.25/0.427	3.88/0.405	3.71/0.393	**0.474**/**0.270**
10	4.68/0.460	4.32/0.431	4.00/0.405	3.72/0.388	3.41/0.367	0.372/**0.253**
12	4.58/0.462	4.07/0.416	3.53/0.399	3.49/0.374	3.23/0.362	0.418/**0.276**
15	4.57/0.441	3.87/0.412	3.71/0.389	3.48/0.363	3.15/0.351	0.451/**0.256**
Relative Inc.	0.317/**0.172**	0.395/**0.172**	0.288/**0.167**	0.250/**0.187**	0.225/0.165	–

(a) PLDA

SNR	Number of frames					Relative Inc.
	1000	1334	1667	2000	2500	
6	5.53/0.473	5.03/0.437	4.39/0.410	4.03/0.398	3.66/0.360	**0.512**/0.314
8	5.13/0.453	4.54/0.414	3.85/0.385	3.56/0.368	3.42/0.354	0.500/0.280
10	4.43/0.423	4.07/0.393	3.75/0.370	3.53/0.353	3.27/0.330	**0.355**/0.282
12	4.29/0.421	3.77/0.379	3.39/0.358	3.28/0.330	3.17/0.314	**0.353**/0.341
15	4.42/0.401	3.78/0.367	3.44/0.347	3.31/0.322	3.01/0.314	0.468/0.277
Relative Inc.	**0.251**/0.179	**0.331**/0.191	**0.276**/0.182	**0.218**/0.236	**0.216**/**0.146**	–

(b) SDI-PLDA

The effect is shown in Table 6. Although there is no obvious relation between the number of speaker factors and the number of SNR factors and duration factors, this table suggests that it is fine to set Q2 and Q3 to 10.

6.5. Robustness Against Mismatch Types

The results in Table 3 do not show which type of mismatches (SNR or duration) is more harmful to performance. To this end, we fixed one type of variability and vary the other type. The test utterances from male speakers in CC4, excluding utterances with less than 2500 frames, were used as the evaluation data. We added babble noise to the test utterances at SNR of 6dB, 8dB, 10dB, 12dB, and 15dB. Then, 2500 frames were randomly selected from each test utterance at each SNR. Finally, test utterances with different durations were created by successively discarding some frames from the 2500-frame test utterances. The SNRs and durations were set such that the percentage decrease in successive SNR is equal to the percentage decrease in successive utterance length. For example, when SNR reduces from 15dB to 12dB, the number of frames decreases from 2500 to 2000, which amount to 20% relative reduction.

Table 7 shows the results of PLDA and SDI-PLDA under different combinations of SNRs and utterance durations. The results show that system performance degrades with decreasing SNR (from 15dB to 6dB) or utterance duration (from 2500 frames to 1000 frames). Both tables show that there is almost no change in performance once the SNR is larger than or equal to 12dB, which suggests that the effect of SNR variability is small when the test utterances are not noisy.

Comparing the performance and the relative increases in EER and minDCF between PLDA and SDI-PLDA in Table 7, we can draw the following conclusions: (1) the SDI-PLDA performs better than PLDA for all combinations of utterance length and SNR; and (2) when either one the two variability types is fixed but the other is varied, the EER of SDI-PLDA is more stable (smaller relative increase) but its minDCF is less stable (larger relative increase). Despite its larger relative increases in minDCF, SDI-PLDA achieves a lower minDCF under all conditions, which provides strong evidence supporting its superiority over PLDA in tackling SNR and duration variabilities.

7. Conclusions

A new SNR- and duration-invariant PLDA model is presented. It is designed to improve the robustness of speaker verification systems under both noise-level and duration mismatches. By introducing a duration subspace to SNR-invariant PLDA, duration information can be captured and the effect of noise-level variability and duration variability can be simultaneously suppressed. Experiments on the NIST 2012 SRE demonstrate the effectiveness of the proposed method.

8. Acknowledgment

This work was supported in part by The RGC of Hong Kong SAR (Grant Nos. PolyU 152117/14E and PolyU 152068/15E) and in part by the Taiwan MOST with Grant 105-2221-E-009-137-MY2.

Appendix A.

To simplify notations, we use $\mathbf{x}_s$ and $\mathbf{x}_t$ instead of $\hat{\mathbf{x}}_s$ and $\hat{\mathbf{x}}_t$ in Eq. 51 to represent the NFA-projected i-vectors. If $\mathbf{x}_s$ and $\mathbf{x}_t$ are from the same speaker, then we have

$$\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix} + \begin{bmatrix} \mathbf{V} & \mathbf{U} & \mathbf{0} & \mathbf{R} & \mathbf{0} \\ \mathbf{V} & \mathbf{0} & \mathbf{U} & \mathbf{0} & \mathbf{R} \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \mathbf{w}_s \\ \mathbf{w}_t \\ \mathbf{y}_s \\ \mathbf{y}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_s \\ \boldsymbol{\epsilon}_t \end{bmatrix}, \tag{A.1}$$

where $\mathbf{h}$ represents the speaker factor shared by both i-vectors, $\mathbf{w}_s$ and $\mathbf{w}_t$ represent the SNR factors of the two utterances, and $\mathbf{y}_s$ and $\mathbf{y}_t$ represent the duration factors of the two utterances, respectively. Eq. A.1 can be written in a compact form:

$$\tilde{\mathbf{x}}_{st} = \tilde{\mathbf{m}} + \tilde{\mathbf{A}}\tilde{\mathbf{z}}_{st} + \tilde{\boldsymbol{\epsilon}}_{st}$$

where the tilde denotes the stacking of vectors and

$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{V} & \mathbf{U} & \mathbf{0} & \mathbf{R} & \mathbf{0} \\ \mathbf{V} & \mathbf{0} & \mathbf{U} & \mathbf{0} & \mathbf{R} \end{bmatrix}.$$

Assuming that the NFA-projected i-vectors follow a Gaussian distribution, the distribution of $\tilde{\mathbf{x}}_{st}$ can be obtained by marginalizing over all possible latent factors as follows:

$$\begin{aligned} p(\tilde{\mathbf{x}}_{st}|\text{same-speaker}) &= \int p(\tilde{\mathbf{x}}_{st}|\tilde{\mathbf{z}}_{st})p(\tilde{\mathbf{z}}_{st})\mathrm{d}\tilde{\mathbf{z}}_{st} \\ &= \int \mathcal{N}(\tilde{\mathbf{x}}_{st}|\tilde{\mathbf{m}} + \tilde{\mathbf{A}}\tilde{\mathbf{z}}_{st}, \tilde{\boldsymbol{\Sigma}})\mathcal{N}(\tilde{\mathbf{z}}_{st}|\mathbf{0}, \mathbf{I})\mathrm{d}\tilde{\mathbf{z}}_{st} \\ &= \mathcal{N}(\tilde{\mathbf{x}}_{st}|\tilde{\mathbf{m}}, \tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\mathsf{T}} + \tilde{\boldsymbol{\Sigma}}) \\ &= \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{tot} & \boldsymbol{\Sigma}_{ac} \\ \boldsymbol{\Sigma}_{ac} & \boldsymbol{\Sigma}_{tot} \end{bmatrix}\right) \end{aligned} \tag{A.2}$$

where $\tilde{\mathbf{\Sigma}} = \mathrm{diag}\{\mathbf{\Sigma}, \mathbf{\Sigma}\}$, $\mathbf{\Sigma}_{tot} = \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{U}\mathbf{U}^{\mathsf{T}} + \mathbf{R}\mathbf{R}^{\mathsf{T}} + \mathbf{\Sigma}$ and $\mathbf{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^{\mathsf{T}}$. If $\mathbf{x}_s$ and $\mathbf{x}_t$ are from the utterances of two different speakers, we have

$$\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix} + \begin{bmatrix} \mathbf{V} & \mathbf{0} & \mathbf{U} & \mathbf{0} & \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} & \mathbf{0} & \mathbf{U} & \mathbf{0} & \mathbf{R} \end{bmatrix} \begin{bmatrix} \mathbf{h}_s \\ \mathbf{h}_t \\ \mathbf{w}_s \\ \mathbf{w}_t \\ \mathbf{y}_s \\ \mathbf{y}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_s \\ \boldsymbol{\epsilon}_t \end{bmatrix} \tag{A.3}$$

which can be compactly written as

$$\tilde{\mathbf{x}}_{st} = \tilde{\mathbf{m}} + \bar{\mathbf{A}}\bar{\mathbf{z}}_{st} + \tilde{\boldsymbol{\epsilon}}_{st}.$$

The distribution of $\tilde{\mathbf{x}}_{st}$ is obtained by marginalizing over $\bar{\mathbf{z}}_{st}$:

$$\begin{aligned} p(\tilde{\mathbf{x}}_{st}|\text{diff-speaker}) &= \int p(\tilde{\mathbf{x}}_{st}|\bar{\mathbf{z}}_{st})p(\bar{\mathbf{z}}_{st})\mathrm{d}\bar{\mathbf{z}}_{st} \\ &= \int \mathcal{N}(\tilde{\mathbf{x}}_{st}|\tilde{\mathbf{m}} + \bar{\mathbf{A}}\bar{\mathbf{z}}_{st}, \tilde{\mathbf{\Sigma}})\mathcal{N}(\bar{\mathbf{z}}_{st}|\mathbf{0}, \mathbf{I})\mathrm{d}\bar{\mathbf{z}}_{st} \\ &= \mathcal{N}(\tilde{\mathbf{x}}_{st}|\tilde{\mathbf{m}}, \bar{\mathbf{A}}\bar{\mathbf{A}}^{\mathsf{T}} + \tilde{\mathbf{\Sigma}}) \\ &= \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{tot} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{tot} \end{bmatrix}\right). \end{aligned} \tag{A.4}$$

Combining Eq. A.2 and Eq. A.4, we have the log-likelihood ratio score:

$$\begin{aligned} S_{\mathrm{LR}}(\mathbf{x}_s, \mathbf{x}_t) &= \ln \frac{\mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{tot} & \mathbf{\Sigma}_{ac} \\ \mathbf{\Sigma}_{ac} & \mathbf{\Sigma}_{tot} \end{bmatrix}\right)}{\mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{tot} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{tot} \end{bmatrix}\right)} \\ &= \frac{1}{2} \begin{bmatrix} \bar{\mathbf{x}}_s^{\mathrm{T}} & \bar{\mathbf{x}}_t^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{Q} & \mathbf{P} \\ \mathbf{P} & \mathbf{Q} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_s \\ \bar{\mathbf{x}}_t \end{bmatrix} + \text{const} \\ &= \frac{1}{2}[\bar{\mathbf{x}}_s^{\mathrm{T}}\mathbf{Q}\bar{\mathbf{x}}_s + 2\bar{\mathbf{x}}_s^{\mathrm{T}}\mathbf{P}\bar{\mathbf{x}}_t + \bar{\mathbf{x}}_t^{\mathrm{T}}\mathbf{Q}\bar{\mathbf{x}}_t] + \text{const}, \end{aligned} \tag{A.5}$$

where

$$\begin{aligned} \bar{\mathbf{x}}_s = \mathbf{x}_s - \mathbf{m}, &\qquad \bar{\mathbf{x}}_t = \mathbf{x}_t - \mathbf{m}, \\ \mathbf{P} &= \mathbf{\Sigma}_{tot}^{-1}\mathbf{\Sigma}_{ac}(\mathbf{\Sigma}_{tot} - \mathbf{\Sigma}_{ac}\mathbf{\Sigma}_{tot}^{-1}\mathbf{\Sigma}_{ac})^{-1}, \\ \mathbf{Q} &= \mathbf{\Sigma}_{tot}^{-1} - (\mathbf{\Sigma}_{tot} - \mathbf{\Sigma}_{ac}\mathbf{\Sigma}_{tot}^{-1}\mathbf{\Sigma}_{ac})^{-1}. \end{aligned}$$

References

[1] Billingsley, P., 2008. Probability and Measure. John Wiley & Sons, New York.

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

[3] Cai, W., Li, M., Li, L., Hong, Q., 2015. Duration dependent covariance regularization in PLDA modeling for speaker verification, in: Proc. Interspeech, pp. 1027–1031.

[4] Cumani, S., 2015. Fast scoring of full posterior PLDA models. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23, 2036–2045.

[5] Cumani, S., Plchot, O., Laface, P., 2014. On the use of i–vector posterior distributions in probabilistic linear discriminant analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22, 846–857.

[6] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. on Audio, Speech, and Language Processing 19, 788–798.

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

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

[9] Gillick, L., Cox, S.J., 1989. Some statistical issues in the comparison of speech recognition algorithms, in: Proc. ICASSP, Glasgow, UK. pp. 532–535.

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

[11] Hasan, T., Saeidi, R., Hansen, J.H., van Leeuwen, D., et al., 2013b. Duration mismatch compensation for i-vector based speaker recognition systems, in: Proc. ICASSP, pp. 7663–7667.

[12] Hatch, A., Kajarekar, S., Stolcke, A., 2006. Within-class covariance normalization for SVM-based speaker recognition, in: Proc. ICSLP, pp. 1471–1474.

[13] Hong, Q., Li, L., Li, M., Huang, L., Wan, L., Zhang, J., 2015. Modified-prior PLDA and score calibration for duration mismatch compensation in speaker recognition system, in: Proc. Interspeech, pp. 1037–1041.

[14] Kanagasundaram, A., Dean, D., Sridharan, S., Gonzalez-Dominguez, J., Gonzalez-Rodriguez, J., Ramos, D., 2014. Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. Speech Communication 59, 69–82.

[15] Kenny, P., 2005. Joint factor analysis of speaker and session variability: Theory and algorithms. CRIM, Montreal,(Report) CRIM-06/08-13 .

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

[17] Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2006. Improvements in factor analysis based speaker verification, in: Proc. ICASSP, pp. 113–116.

[18] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P., 2008. A study of inter-speaker variability in speaker verification. IEEE Trans. on Audio, Speech and Language Processing 16, 980–988.

[19] Kenny, P., Stafylakis, T., Ouellet, P., Alam, M.J., Dumouchel, P., 2013. PLDA for speaker verification with utterances of arbitrary duration, in: Proc. ICASSP, pp. 7649–7653.

[20] Li, N., Mak, M.W., 2015a. SNR-invariant PLDA modeling for robust speaker verification, in: Proc. Interspeech, pp. 2317–2321.

[21] Li, N., Mak, M.W., 2015b. SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification. IEEE/ACM Trans. on Audio, Speech and Language Processing 23, 1648–1659.

[22] Li, Z., Lin, D., Tang, X., 2009. Nonparametric discriminant analysis for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 755–761.

[23] Mak, M.W., 2014. SNR-dependent mixture of PLDA for noise robust speaker verification, in: Proc. Interspeech, pp. 1855–1859.

[24] Mak, M.W., Pang, X.M., Chien, J.T., 2016. Mixture of PLDA for noise robust i-vector speaker verification. IEEE/ACM Trans. on Audio, Speech and Language Processing 24, 130–142.

[25] Mak, M.W., Yu, H.B., 2014. A study of voice activity detection techniques for NIST speaker recognition evaluations. Computer Speech and Language 28, 295–313.

[26] McLaren, M., Mandasari, M., Leeuwen, D., 2012. Source normalization for language-independent speaker recognition using i-vectors, in: Proc. Odyssey, pp. 55–61.

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

[28] Norwich, K.H., 1993. Information, sensation, and perception. Academic Press San Diego.

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

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

[31] Prince, S.J., Elder, J.H., 2007. Probabilistic linear discriminant analysis for inferences about identity, in: Proc. IEEE 11th International Conference on Computer Vision, IEEE. pp. 1–8.

[32] Sarkar, A.K., Matrouf, D., Bousquet, P.M., Bonastre, J.F., 2012. Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification., in: Proc. Interspeech, pp. 2662–2665.

[33] Sizov, A., Lee, K.A., Kinnunen, T., 2014. Unifying probabilistic linear discriminant analysis variants in biometric authentication, in: Structural, Syntactic, and Statistical Pattern Recognition. Springer, pp. 464–475.

[34] Vesnicer, B., Zganec-Gros, J., Dobrisek, S., Struc, V., 2014. Incorporating duration information into i-vector-based speaker-recognition systems, in: Proc. Odyssey, pp. 241–248.

[35] Watanabe, S., Chien, J.T., 2015. Bayesian Speech and Language Processing. Cambridge University Press.

[36] Yamamoto, H., Koshinaka, T., 2015. Denoising autoencoder-based speaker feature restoration for utterances of short duration, in: Proc. Interspeech, pp. 1052–1056.

[37] Yu, H., Mak, M., 2011. Comparison of voice activity detectors for interview speech in NIST speaker recognition evaluation, in: Proc. Interspeech, pp. 2353–2356.