SNR-Dependent Mixture of PLDA for Noise Robust Speaker Verification

Man-Wai MAK

Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR
enmwmak@polyu.edu.hk

Abstract

This paper proposes a mixture of SNR-dependent PLDA models to provide a wider coverage on the i-vector spaces so that the resulting i-vector/PLDA system can handle test utterances with a wide range of SNR. To maximise the coordination among the PLDA models, they are trained simultaneously via an EM algorithm using utterances contaminated with noise at various levels. The contribution of a training i-vector to individual PLDA models is determined by the posterior probability of the utterance's SNR. Given a test i-vector, the marginal likelihoods from individual PLDA models are linear combined based on the the posterior probabilities of the test utterance and the target-speaker's utterance. Verification scores are the ratio of the marginal likelihoods. Results based on NIST 2012 SRE suggest that this soft-decision scheme is particularly suitable for the situations where the test utterances exhibit a wide range of SNR.

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

10.21437/Interspeech.2014-421

1. Introduction

To deploy a speaker verification system in real-world scenarios, it is imperative to ensure that the system is robust to acoustic environments with variable noise levels. A number of strategies have been proposed to compensate for the variability due to background noise. Typically, these approaches reduce the variability in either the front-end processing stage or the back-end classification stage. The former aims to (1) develop 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 researches have found that techniques that operate on the back-end 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.

The i-vector approach defines a single space called the total variability space to represent the holistic variability of both speakers and channels. The acoustic characteristics of an entire utterance are represented by a single low-dimension vector called the i-vector. Because the i-vector space accounts for both speaker and channel (including background noise) variability, a second stage of dimension reduction and normalization is required to suppress the channel effects. To this end, 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.

More recent methods are typically built on top of the i-vector/PLDA framework. For example, [14, 15, 16, 17] apply multi-condition training where clean and noisy utterances are pooled together to train a PLDA model so that it becomes more robust to noisy test utterances. In [18], multiple PLDA models are trained, one for each condition. Observing that the leading eigen-directions of acoustic features contain most of the speaker-dependent information, Hasan and Hansen [19] performed mixture of probabilistic PCA on feature space so that the posterior means of the mixture-dependent acoustic factors can be incorporated into an i-vector extractor. It was shown that integrating feature dimension reduction and i-vector extraction not only removes the need to perform hard feature clustering, but also performs feature normalization and enhancement. This idea has been further enhanced by replacing the UBM by a mixture of acoustic factor analyzers for i-vector extraction [20]. Recently, Lei et al. [21] proposed adapting a clean UBM to noisy utterances using vector Taylor series. I-vectors are then extracted based on the noise-adapted UBM. The idea is to clean up the i-vectors so that they become independent of additive and convolutive noise.

In NIST 2012 SRE [22], focus was shifted to noise robust speaker verification. While i-vector/PLDA systems [23] perform very well even under noisy conditions, many of them use a single PLDA model to handle all of the test utterances regardless of their noise level. This paper argues 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. A mixture of SNR-dependent PLDA models is proposed to achieve this goal. Unlike the conventional mixture of factor analyzers [24] where the posteriors of the indicator variables depend on the data samples, the posteriors of the indicator variables in the proposed method depend on the SNR of the utterances. As a result, the contributions of individual mixtures depend explicitly on the SNR and implicitly depend on the locations of the i-vectors in the i-vector space.

While the proposed method resembles multi-condition training described earlier, there are some important differences. The major difference is that our condition-dependent factor analyzers are trained simultaneously even though their parameters are not tied. 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 our proposed model computes the verification scores by incorporating the posterior of SNR of both the target-speaker's and test utterances into the

This work was in part supported by The Hong Polytechnic University G-YN18 and Motorola Solutions Foundation 7186445.

marginal likelihood computation (Eq. 4 of this paper).

2. Mixture of PLDA

2.1. Generative Model of PLDA

Given a set of D-dim 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 variables $\mathcal{Z} = \{\mathbf{z}_i;\ i = 1, \ldots, N\}$ and parameters $\boldsymbol{\omega} = \{\mathbf{m}, \mathbf{V}, \boldsymbol{\Sigma}\}$ of a factor analyzer [9]:

$$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{z}_i + \boldsymbol{\epsilon}_{ij}; \tag{1}$$

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]. Eq. 1 suggests that the generative model of PLDA obeys

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = \int \mathcal{N}(\mathbf{x}|\mathbf{m} + \mathbf{V}\mathbf{z}, \boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})\mathrm{d}\mathbf{z} \\ &= \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma}). \end{aligned}$$

Because the i-vectors of the same speaker should share the same speaker factor in Eq. 1, we may collect the i-vectors of speaker i and rewrite Eq. 1 as

$$\tilde{\mathbf{x}}_i = \tilde{\mathbf{m}} + \tilde{\mathbf{V}}\mathbf{z}_i + \tilde{\boldsymbol{\epsilon}}_i;\ \tilde{\mathbf{x}}_i, \tilde{\mathbf{m}} \in \Re^{DH_i},\ \tilde{\mathbf{V}} \in \Re^{DH_i\times M},\ \tilde{\boldsymbol{\epsilon}}_i \in \Re^{DH_i}, \tag{2}$$

where $\tilde{\mathbf{x}}_i = [\mathbf{x}_{i1}^\mathsf{T}\ \cdots\ \mathbf{x}_{iH_i}^\mathsf{T}]^\mathsf{T}$, $\tilde{\mathbf{m}} = [\mathbf{m}^\mathsf{T}\ \cdots\ \mathbf{m}^\mathsf{T}]^\mathsf{T}$, $\tilde{\mathbf{V}} = [\mathbf{V}^\mathsf{T}\ \cdots\ \mathbf{V}^\mathsf{T}]^\mathsf{T}$, and $\tilde{\boldsymbol{\epsilon}}_i = [\boldsymbol{\epsilon}_{i1}^\mathsf{T}\ \cdots\ \boldsymbol{\epsilon}_{iH_i}^\mathsf{T}]^\mathsf{T}$. Eq. 2 is a factor analyzer whose parameters can be estimated via an EM algorithm [25, 26].

2.2. Generative Model for Mixture of PLDA

The PLDA model assumes that the same loading matrix $\mathbf{V}$ and covariance matrix $\boldsymbol{\Sigma}$ can be applied to all i-vectors regardless of the SNR of the utterances. To deal with varying noise and reverberation levels, the i-vectors are better modelled by a mixture of SNR-dependent factor analyzers [24] with parameters $\underline{\boldsymbol{\theta}} = \{\underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\omega}}\} = \{\boldsymbol{\lambda}_k, \boldsymbol{\omega}_k\}_{k=1}^K = \{\pi_k, \mu_k, \sigma_k, \mathbf{m}_k, \mathbf{V}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, 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. With these factor analyzers, the generative model follows a mixture distribution:

$$\begin{aligned} &p(\mathbf{x}, \ell) = p(\ell)p(\mathbf{x}|\ell) \\ &= p(\ell)\sum\nolimits_{k=1}^K \int P(y_k = 1|\ell, \underline{\boldsymbol{\lambda}})p(\mathbf{x}|\ell, \mathbf{z}, y_k = 1, \boldsymbol{\omega}_k)p(\mathbf{z})\mathrm{d}\mathbf{z} \\ &= p(\ell)\sum\nolimits_{k=1}^K \gamma_\ell(y_k)\mathcal{N}(\mathbf{x}|\mathbf{m}_k, \mathbf{V}_k\mathbf{V}_k^\mathsf{T} + \boldsymbol{\Sigma}_k), \end{aligned} \tag{3}$$

where ℓ represents the SNR of the utterance whose i-vector is $\mathbf{x}$, y_k's are indicator variable specifying which of the factor analyzers is responsible for generating $\mathbf{x}$, and $\gamma_\ell(y_k)$ is the posterior probability:

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

Note that we have assumed that the speaker variability is modeled by $\mathbf{V}_k\mathbf{V}_k^\mathsf{T}$ and that the channel and noise variability are modeled by $\boldsymbol{\Sigma}_k$.

2.3. 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

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

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

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

Similarly, the different-speaker marginal likelihood is

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

where

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

and similarly for $p(\mathbf{x}_t, \ell_t|\text{Spk } t)$. Therefore, the likelihood ratio is given by Eq. 4 at the bottom of next page.

Note that Eq. 4 is likely to cause numerical problems if it is evaluated directly because the determinant of $\hat{\mathbf{V}}_{k_s}\hat{\mathbf{V}}_{k_s}^\mathsf{T} + \hat{\boldsymbol{\Sigma}}_{k_s}$ could exceed the double-precision representation. This problem, however, can be avoided by 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. 4 as Eq. 5 shown at the bottom of next page, where $\hat{\boldsymbol{\Lambda}}_{k_sk_t} = \hat{\mathbf{V}}_{k_s}\hat{\mathbf{V}}_{k_t}^\mathsf{T} + \hat{\boldsymbol{\Sigma}}_{k_sk_t}$, $\boldsymbol{\Lambda}_{k_s} = \mathbf{V}_{k_s}\mathbf{V}_{k_s}^\mathsf{T} + \boldsymbol{\Sigma}_{k_s}$, $\hat{\boldsymbol{\Sigma}}_{k_sk_t} = \mathrm{diag}\{\boldsymbol{\Sigma}_{k_s}, \boldsymbol{\Sigma}_{k_t}\}$, and $\mathcal{D}(\cdot\|\cdot)$ is the Mahalanobis distance. In this work, $\alpha = 5$.

2.4. EM for Mixture of PLDA

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

$$
\begin{aligned}
&Q(\underline{\boldsymbol{\theta}}'|\underline{\boldsymbol{\theta}}) = \mathbb{E}_{\mathcal{Y},\mathcal{Z}}\{\ln p(\mathcal{X},\mathcal{Y},\mathcal{Z}|\underline{\boldsymbol{\theta}}')|\mathcal{X},\mathcal{L},\underline{\boldsymbol{\theta}}\} \\
&= \mathbb{E}\left\{\sum_{ijk} y_{ijk}\ln\left[P(y_{ijk}=1)p(\mathbf{x}_{ij}|\mathbf{z}_i,\boldsymbol{\omega}'_k)p(\mathbf{z}_i)\right]\Big|\mathcal{X},\mathcal{L},\underline{\boldsymbol{\theta}}\right\} \\
&= \sum_{ijk}\mathbb{E}\left\{y_{ijk}\ln\left[\pi'_k\mathcal{N}(\mathbf{x}_{ij}|\mathbf{m}'_k+\mathbf{V}'_k\mathbf{z}_i,\boldsymbol{\Sigma}'_k)\mathcal{N}(\mathbf{z}_i|\mathbf{0},\mathbf{I})\right]\Big|\mathcal{X},\mathcal{L},\underline{\boldsymbol{\theta}}\right\},
\end{aligned}
\tag{6}
$$

where $\pi'_k \equiv P(y_{ijk}=1)$ is the prior probability of the k-th factor analyzer. Maximizing Eq. 6 leads to the following EM formulations:

$$
\begin{aligned}
&\textbf{E-Step: } \langle y_{ijk}|\mathcal{L}\rangle = \frac{\pi_k\mathcal{N}(\ell_{ij}|\mu_k,\sigma_k^2)}{\sum_{r=1}^K\pi_r\mathcal{N}(\ell_{ij}|\mu_r,\sigma_r^2)} \\
&\quad \mathbf{L}_i = \mathbf{I} + H_i\sum_{k=1}^K\mathbf{V}_k^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}\mathbf{V}_k \\
&\quad \langle y_{ijk}\mathbf{z}_i|\mathcal{X},\mathcal{L}\rangle = \langle y_{ijk}|\mathcal{L}\rangle\langle\mathbf{z}_i|\mathcal{X}\rangle \\
&\quad \langle\mathbf{z}_i|\mathcal{X}\rangle = \mathbf{L}_i^{-1}\sum_{j=1}^{H_i}\sum_{k=1}^K\mathbf{V}_k^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_{ij}-\mathbf{m}_k) \\
&\langle y_{ijk}\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X},\mathcal{L}\rangle = \langle y_{ijk}|\mathcal{L}\rangle\langle\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X}\rangle \\
&\quad \langle\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X}\rangle = \mathbf{L}_i^{-1} + \langle\mathbf{z}_i|\mathcal{X}\rangle\langle\mathbf{z}_i|\mathcal{X}\rangle^{\mathsf{T}}
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
&\textbf{M-Step: } \mathbf{m}'_k = \frac{\sum_{i=1}^N\sum_{j=1}^{H_i}\langle y_{ijk}|\mathcal{L}\rangle\mathbf{x}_{ij}}{\sum_{i=1}^N\sum_{j=1}^{H_i}\langle y_{ijk}|\mathcal{L}\rangle};\ \pi'_k = \frac{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle}{\sum_{ijl}\langle y_{ijl}|\mathcal{L}\rangle} \\
&\mu'_k = \frac{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle\ell_{ij}}{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle};\ \ \sigma_k^{2\prime} = \frac{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle(\ell_{ij}-\mu'_k)^2}{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle} \\
&\mathbf{V}'_k = \left[\sum_{ij}\mathbf{f}'_{ijk}\langle y_{ijk}\mathbf{z}_i|\mathcal{X},\mathcal{L}\rangle^{\mathsf{T}}\right]\left[\sum_{ij}\left\langle y_{ijk}\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X},\mathcal{L}\right\rangle\right]^{-1} \\
&\boldsymbol{\Sigma}'_k = \frac{\sum_{ij}\left[\langle y_{ij}|\mathcal{L}\rangle\mathbf{f}'_{ijk}\mathbf{f}'_{ijk}{}^{\mathsf{T}} - \mathbf{V}'_k\langle y_{ijk}\mathbf{z}_i|\mathcal{X},\mathcal{L}\rangle\mathbf{f}'_{ijk}{}^{\mathsf{T}}\right]}{\sum_{ij}\langle y_{ijk}|\mathcal{L}\rangle} \\
&\mathbf{f}'_{ijk} = \mathbf{x}_{ij} - \mathbf{m}'_k
\end{aligned}
\tag{8}
$$

where $\langle y_{ijk}|\mathcal{L}\rangle \equiv \mathbb{E}_{\mathcal{Y}}\{y_{ijk}=1|\mathcal{L},\underline{\boldsymbol{\lambda}}\}$, $\langle y_{ijk}\mathbf{z}_i|\mathcal{X},\mathcal{L}\rangle \equiv \mathbb{E}_{\mathcal{Y},\mathcal{Z}}\{y_{ijk}=1,\mathbf{z}_i|\mathcal{X},\mathcal{L},\underline{\boldsymbol{\theta}}\}$, and $\langle y_{ijk}\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X},\mathcal{L}\rangle \equiv \mathbb{E}_{\mathcal{Y},\mathcal{Z}}\{y_{ijk}=1,\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X},\mathcal{L},\underline{\boldsymbol{\theta}}\}$.

3. Experiments

3.1. Speech Data and Acoustic Features

The phonecall speech in the core set of NIST 2012 Speaker Recognition Evaluation (SRE) [22] 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 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 [27]. 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 [28] and added the noise waveform to the file at an SNR of 6dB and 15dB using the FaNT tool [29]. Because the SNR-dependent PLDA requires the SNR of test utterances, we used the speech voltmeter function in FaNT and the VAD decisions to estimate the SNR of the test files.

3.2. PLDA and Mixture of PLDA

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.

Both SNR-independent and SNR-dependent PLDA and mixture of PLDA models were trained. For the former, we pooled the 6dB (tel), 15dB (tel), and original (tel+mic) speech files in 2006–2010 SRE—excluding speakers with less than two utterances—into a single training set. Gender-dependent, SNR-independent PLDA models with 150 factors were then trained. The SNR-dependent models are further divided into two types: hard-decision mixture of PLDA (HD-mPLDA) and soft-decision mixture of PLDA (SD-mPLDA). For HD-mPLDA, the 6dB, 15dB, and original speech files were independently used to train three PLDA models, each with 150 factors and corresponds to one of the three noise levels (6dB, 15dB, and clean). For SD-mPLDA, these files were pooled together and the EM algorithm specified in Eqs. 7 and 8 were used to train a mixture PLDA model with $K=3$ and $M=150$, one for each gender. We have varying K and found that $K=3$ gives the best compromise between EER and minDCF.

The scoring procedures for SNR-independent and SNR-dependent models are different. For SNR-independent PLDA models, each of the test i-vectors was scored against the target-speakers' i-vectors derived from the telephone sessions of 6dB, 15dB, and original (clean) speech files using the conventional PLDA scoring function [13].

For hard-decision mixture of PLDA (HD-mPLDA), the

$$
S_{\text{mLR}}(\mathbf{x}_s,\mathbf{x}_t) = \frac{\sum_{k_s=1}^K\sum_{k_t=1}^K\gamma_{\ell_s,\ell_t}(y_{k_s},y_{k_t})\mathcal{N}\left(\left[\mathbf{x}_s^{\mathsf{T}}\ \mathbf{x}_t^{\mathsf{T}}\right]^{\mathsf{T}}\middle|\left[\mathbf{m}_{k_s}^{\mathsf{T}}\ \mathbf{m}_{k_t}^{\mathsf{T}}\right]^{\mathsf{T}},\hat{\mathbf{V}}_{k_sk_t}\hat{\mathbf{V}}_{k_sk_t}^{\mathsf{T}}+\hat{\boldsymbol{\Sigma}}_{k_sk_t}\right)}{\left[\sum_{k_s=1}^K\gamma_{\ell_s}(y_{k_s})\mathcal{N}\left(\mathbf{x}_s|\mathbf{m}_{k_s},\mathbf{V}_{k_s}\mathbf{V}_{k_s}^{\mathsf{T}}+\boldsymbol{\Sigma}_{k_s}\right)\right]\left[\sum_{k_t=1}^K\gamma_{\ell_t}(y_{k_t})\mathcal{N}\left(\mathbf{x}_t|\mathbf{m}_{k_t},\mathbf{V}_{k_t}\mathbf{V}_{k_t}^{\mathsf{T}}+\boldsymbol{\Sigma}_{k_t}\right)\right]} \tag{4}
$$

$$
= \frac{\sum_{k_s=1}^K\sum_{k_t=1}^K\gamma_{\ell_s,\ell_t}(y_{k_s},y_{k_t})\exp\left\{-\frac{1}{2}\log|\alpha\hat{\boldsymbol{\Lambda}}_{k_sk_t}| - \frac{1}{2}\mathcal{D}\left(\left[\mathbf{x}_s^{\mathsf{T}}\ \mathbf{x}_t^{\mathsf{T}}\right]^{\mathsf{T}}\middle\|\left[\mathbf{m}_{k_s}^{\mathsf{T}}\ \mathbf{m}_{k_t}^{\mathsf{T}}\right]^{\mathsf{T}}\right)\right\}}{\left[\sum_{k_s=1}^K\gamma_{\ell_s}(y_{k_s})\exp\left\{-\frac{1}{2}\log|\alpha\boldsymbol{\Lambda}_{k_s}| - \frac{1}{2}\mathcal{D}\left(\mathbf{x}_s\|\mathbf{m}_{k_s}\right)\right\}\right]\left[\sum_{k_t=1}^K\gamma_{\ell_t}(y_{k_t})\exp\left\{-\frac{1}{2}\log|\alpha\boldsymbol{\Lambda}_{k_t}| - \frac{1}{2}\mathcal{D}\left(\mathbf{x}_t\|\mathbf{m}_{k_t}\right)\right\}\right]} \tag{5}
$$

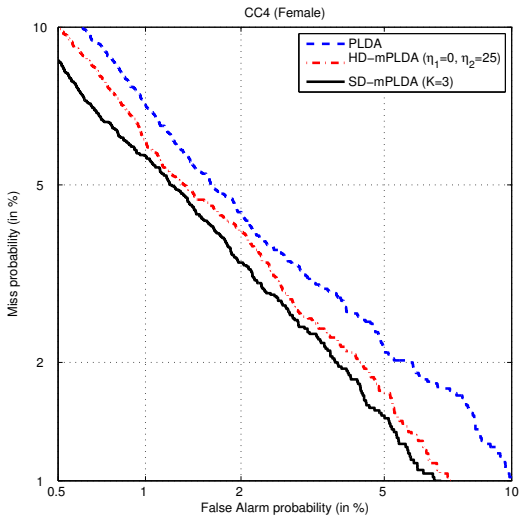

Figure 1: DET performance of PLDA, hard-decision mixture of PLDA (HD-mPLDA), and soft-decision mixture of PLDA (SD-mPLDA) in CC4 of NIST 2012 SRE (core set, female speakers).

SNR of the test utterance determines which of the SNR-dependent PLDA models and which category (6dB, 15dB or clean) of target-speaker's i-vectors should be used for scoring:

$$\text{If} \begin{cases} \ell_t \leq \eta_1, & \text{use 6dB PLDA and target's i-vectors} \\ \eta_1 < \ell_t \leq \eta_2, & \text{use 15dB PLDA and target's i-vectors} \\ \ell_t > \eta_2, & \text{use clean PLDA and target's i-vectors} \end{cases} \tag{9}$$

where ℓ_t is the SNR of the test utterance, and η_1 and η_2 are decision thresholds. Because the three PLDA models produce scores at different ranges, the scores should be normalized before computing the EER and minDCF. We applied SNR-dependent Z-norm to the PLDA scores, with the three sets of Z-norm parameters found independently using the same set of i-vectors used for training the SNR-dependent PLDA models. Again, these Z-norm parameters are gender-dependent.

For soft-decision mPLDA, we pooled the i-vectors derived from the 6dB, 15dB, and original (clean) telephone speech files of the target speakers and applied the mixture of PLDA scoring function in Eq. 5. Because the mean vectors, factor loadings and covariance matrices ($\mathbf{m}_k$, $\mathbf{V}_k$ and $\mathbf{\Sigma}_k$) in the mixture model are estimated simultaneously in the EM algorithm and the score function in Eq. 5 combines the effect of the mixtures probabilistically, SD-mPLDA does not require score normalization. This property is a key advantage of SD-mPLDA.

4. Results and Discussions

Table 1 shows the EER and minimum DCF (minC_{Primary}) achieved by multi-condition PLDA (baseline),[1] hard-decision mixture of PLDA and soft-decision mixture of PLDA in CC4 and CC5 of NIST 2012 SRE. Evidently, under CC4, using a mixture of PLDA is superior to using a single PLDA trained by pooling the clean and noisy training data. Results also show that HD-mPLDA is more sensitive to the decision threshold η_1 and η_2 under CC4 than under CC5. This is because the SNR range of the test segments in CC4 is wider than that in CC5. Specifically, the SNR range in CC4 is -0.3dB $\sim$ 46.2dB, whereas the SNR range in CC5 is 11.6dB $\sim$ 53.6dB. The high lower-limit of the SNR in CC5 suggests that η_1 is irrelevant as long as it is less than 11.6dB.

[1] We did not use the PLDA trained by clean utterances as the baseline because it has been shown [15, 30] that using clean utterances exclusively for training leads to poorer performance.

Method	η_1	η_2	CC4		CC5	
			EER(%)	minDCF	EER(%)	minDCF
PLDA	–	–	3.49	0.308	2.86	0.286
HD-mPLDA	0	15	3.10	0.398	2.79	0.290
	0	25	3.03	0.378	2.98	0.302
	0	30	3.13	0.380	3.10	0.326
	5	25	3.41	0.419	2.98	0.302
SD-mPLDA	–	–	2.88	0.329	2.80	0.287

(a) Male Speakers

Method	η_1	η_2	CC4		CC5	
			EER(%)	minDCF	EER(%)	minDCF
PLDA	–	–	3.13	0.353	2.47	0.343
HD-mPLDA	0	15	2.82	0.374	2.62	0.354
	0	25	2.81	0.376	2.85	0.360
	0	30	2.82	0.376	3.14	0.391
	5	25	3.17	0.417	2.85	0.360
SD-mPLDA	–	–	2.71	0.332	2.46	0.342

(b) Female Speakers

Table 1: Performance of PLDA, hard-decision mixture of PLDA (HD-mPLDA), and soft-decision mixture of PLDA (SD-mPLDA) in CC4 and CC5 of NIST 2012 SRE (core set). η_1 and η_2 are the decision threshold in Eq. 9.

Table 1 and Fig. 1 show that soft-decision mixture of PLDA performs significantly better than hard-decision mixture of PLDA, which in turn performs better than the simple PLDA. As compared to its hard-decision counterpart, soft-decision mPLDA is more practical because it combines the marginal likelihoods of K PLDA models probabilistically based on the posterior probability of the SNR of the test utterance and target-speaker's utterances, thereby eliminating the need to set decision thresholds (η_1 and η_2) for selecting a suitable model for scoring. Another important advantage of SD-mPLDA is that the parameters of the K PLDA models ($\{\mathbf{m}_k, \mathbf{V}_k, \mathbf{\Sigma}_k\}_{k=1}^K$) are estimated simultaneously, with the contributions of i-vectors at different SNR guided by the posterior expectation $\langle y_{ijk}|\mathcal{L}\rangle$. This property ensures that the PLDA scores from the K PLDA models fall on the same range, thus eliminating the need to perform score normalisation.

5. Conclusions

To enhance the noise robustness of speaker verification systems, a more flexible probabilistic model with a wider coverage of the i-vector space is needed. This paper proposes a mixture PLDA model trained by using both clean and noisy utterances. During verification, the contribution of the mixtures are combined probabilistically based on the posterior probabilities of the SNR of the test and target-speaker's utterances. The proposed model was evaluated on the latest NIST SRE, and results demonstrate that the mixture PLDA model outperforms the conventional multi-condition PLDA model. Natural extensions of this work include increasing the variety of noise types, e.g., traffic, office, restaurants, etc., for training the mixture models. Another direction is to apply the technique to tackle reverberant speech where each reverberant environment is handled by one PLDA model.

6. References

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

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

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

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

[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 Trans. 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 Trans. 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.M. 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 the 9th 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 Trans. on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 1012–1022, 2013.

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

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

[14] D. A.van Leeuwen and R. Saeidi, "Knowing the non-target speakers: The effect of the i-vector population for PLDA training in speaker recognition," in *Proc. 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. ICASSP 2012*, 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 *2013 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. Interspeech*, 2013, pp. 3694–3697.

[18] D. Garcia-Romero, X. Zhou, and C.Y. Espy-Wilson, "Multicondition training of gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4257–4260.

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

[20] T. Hasan and J.H.L. Hansen, "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.

[21] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector Taylor series for speaker recognition.," in *ICASSP*, 2013, pp. 6788–6791.

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

[23] R Saeidi, KA Lee, T Kinnunen, T Hasan, B Fauve, PM Bousquet, E Khoury, PL Sordo Martinez, JMK Kua, CH You, et al., "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. Interspeech*, 2013, pp. 1986–1990.

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

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

[26] D.B. Rubin and D.T. Thayer, "EM algorithms for ML factor analysis," *Psychometrika*, vol. 47, no. 1, pp. 69–76, 1982.

[27] 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.

[28] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, et al., "Promoting robustness for speaker modeling in the community: The PRISM evaluation set," .

[29] "http://dnt.kr.hsnr.de/download.html," .

[30] W. Rao and M. W. Mak, "Construction of discriminative kernels from known and unknown non-targets for PLDA-SVM scoring," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.