SNR-Invariant PLDA Modeling for Robust Speaker Verification

Na LI and Man-Wai MAK

Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR
lina011779@126.com, enmwmak@polyu.edu.hk

Abstract

In spite of the great success of the i-vector/PLDA framework, speaker verification in noisy environments remains a challenge. To compensate for the variability of i-vectors caused by different levels of background noise, this paper proposes a new framework, namely SNR-invariant PLDA, for robust speaker verification. By assuming that i-vectors extracted from utterances falling within a narrow SNR range share similar SNR-specific information, the paper introduces an SNR factor to the conventional PLDA model. Then, the SNR-related variability and the speaker-related variability embedded in the i-vectors are modeled by the SNR factor and the speaker factor, respectively. Accordingly, an i-vector is represented by a linear combination of three components: speaker, SNR, and channel. During verification, the variability due to SNR and channels are marginalized out when computing the marginal likelihood ratio. Experiments based on NIST 2012 SRE show that SNR-invariant PLDA achieves superior performance when compared with the conventional PLDA and SNR-dependent mixture of PLDA.
Index Terms: i-vector, PLDA, SNR-invariant, speaker verification

1. Introduction

During the last few years, the i-vector [1] has become a popular feature representation in the speaker verification domain. Inspired by the joint factor analysis (JFA) framework [2], in the i-vector framework, both speaker and channel information was compressed into a low-dimensional subspace called the total variability subspace. Through this subspace, utterances of variable-length can be represented by fixed-length i-vectors. Such a representation greatly simplifies the modeling process in speaker verification. To suppress the channel- and session-variability embedded in i-vectors, linear discriminant analysis (LDA) [3], within-class covariance normalization (WCCN) [4], and probabilistic LDA (PLDA) [5] can be applied. Typically, LDA is applied to the i-vectors followed by the WCCN. In the verification stage, the cosine distance between target-speaker's i-vector and the i-vector of a test utterance is used as the similarity measure between the target speaker and the test speaker. Alternatively, the likelihood-ratio score of a test i-vector can be computed by marginalizing over the latent variables of a heavy-tailed PLDA model [6] or a Gaussian PLDA model. The former assumes that the i-vectors follow a Student's t distribution and the latter requires applying length-normalization [7] to the i-vectors so that the resulting i-vectors are more amenable to Gaussian PLDA modeling.

This work was in part supported by The RGC of Hong Kong SAR (Grant No. PolyU 152117/14E). An extended version of this work will appear in [27].

Several studies have shown that background noise has severe effects on the performance of speaker verification systems [8–10]. This issue can be addressed in the feature domain [11–16] and model domain [17–23]. The former attempts to find features that are more robust than the conventional MFCC, whereas the latter focuses on the training of back-end classifiers to make them more resilient to noise.

Although the conventional PLDA models are very good at suppressing session variability, their ability in modeling i-vectors derived from utterances having different signal-to-noise ratio (SNR) is limited. The reason is that when training a PLDA model, the i-vectors of the same speaker are grouped together regardless of the noise level of the corresponding utterances. The resulting model attempts to model speaker and channel subspaces, where the channel subspace also comprises the variability caused by background noise. To address this issue, several methods have been proposed to improve the robustness of i-vector/PLDA systems. In [20–23], clean and noisy utterances were pooled together to train a robust PLDA model. Garcia-Romero *et al.* [24] 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. Mak [25] proposed an adaptive multi-condition training algorithm called SNR-dependent mixture of PLDA to handle test utterances with a wide range of SNR.

Although the above methods improve the robustness of the state-of-the-art i-vector/PLDA systems under noisy conditions, they still have at least one of the following limitations: (1) when the distributions of SNR in the training set and the test set are not consistent, the system performance degrades; (2) multiple PLDA models should be trained, which increases computation complexity; and (3) the noise level of test utterances need to be estimated during verification.

To address the limitations of multi-condition training and to improve the system performance of current i-vector/PLDA framework, we propose a noise robust speaker verification framework that can deal with the mismatch caused by the variability in SNR. Our proposal is inspired by the work in [26] where the face recognition system is robust to the change in the facial features of its users when they are getting older, i.e., insensitive to age variability. Based on a similar line of thought, we attempt to make speaker verification systems more resilient to SNR variability by introducing a subspace called SNR-subspace in the PLDA model. With this new subspace, the PLDA model not only able to capture the speaker and channel variabilities embedded in the i-vectors (as in the conventional i-vector/PLDA systems), but also capable of modeling the variability caused by different noise levels. We refer to the new approach as SNR-invariant PLDA and the factors corresponding to the SNR-subspace as SNR factors. In this model, the identity

component and the SNR component live in two different subspaces which can be obtained by an expectation-maximization (EM) algorithm. During the verification stage, SNR variability and channel variability are marginalized out when the likelihood ratio is computed.

2. PLDA Modeling

Prince and Elder [28,29] proposed a probabilistic LDA (PLDA) approach to increasing the separability between the facial images of different persons, and Kenny [6] brought this idea to the speaker recognition community. In i-vector/PLDA systems, a preprocessed i-vector $\mathbf{x}_{ij}$ – which has gone through a series of transformations (LDA, WCCN, and length normalization) – is regarded as an observation generated from a PLDA model:

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

where $\mathbf{m}$ is the global mean of all preprocessed i-vectors, the columns of $\mathbf{V}$ define the bases of the speaker subspace, $\mathbf{h}_i$ is a latent identity factor with a standard normal distribution, and $\boldsymbol{\epsilon}_{ij}$ denotes the residual term which follows a Gaussian distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$. Gaussian PLDA model assumes that $\mathbf{x}_{ij}$ follows a Gaussian distribution.

3. SNR-Invariant PLDA Modeling

3.1. Generative Models

SNR-invariant PLDA is inspired by the notion of Gaussian PLDA in which i-vectors from the same speaker should share an identical latent identity factor. Similarly, we assume that i-vectors derived from utterances that fall within a narrow SNR range should share similar SNR-specific information. From a modeling standpoint, both SNR-specific and identity-specific information can be captured using latent factors. We refer to these latent factors as SNR factor and identity factor in the sequel.

Under the above assumptions, an i-vector can be regarded as an observation generated from a linear generative model that comprises three components: (1) identity component, (2) SNR component, and (3) channel variability and the remaining variability that cannot be captured by the first two components. Assume that we have a set of D-dimensional i-vectors $\mathcal{X} = \{\mathbf{x}_{ij}^k | i = 1, \ldots, S; j = 1, \ldots, H_i(k); k = 1, \ldots, K\}$ obtained from S speakers, where $\mathbf{x}_{ij}^k$ is the *j*-th i-vector from speaker *i* in the *k*-th SNR group. In SNR-invariant PLDA, $\mathbf{x}_{ij}^k$ can be expressed as:

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

where $\mathbf{m}$ is a $D \times 1$ vector representing the global offset, $\mathbf{h}_i$ is a $P \times 1$ vector denoting the latent identity factor with prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{w}_k$ is a $Q \times 1$ vector denoting the latent SNR factor with prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\epsilon}_{ij}^k$ is a $D \times 1$ vector denoting the residual with distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{V}$ is a $D \times P$ matrix whose columns span the speaker subspace, and $\mathbf{U}$ is a $D \times Q$ matrix whose columns span the SNR subspace. $\mathbf{h}_i$ and $\mathbf{w}_k$ are assumed to be statistically independent.

3.2. EM Algorithm for SNR-Invariant PLDA

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

$$\begin{aligned}\mathbf{Q}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{h},\mathbf{w}}\left\{\ln p(\mathcal{X}, \mathbf{h}, \mathbf{w}|\hat{\boldsymbol{\theta}}) \middle| \mathcal{X}, \boldsymbol{\theta}\right\} \\ &= \mathbb{E}_{\mathbf{h},\mathbf{w}}\left\{\sum\nolimits_{ijk} \ln\left[p(\mathbf{x}_{ij}^k|\mathbf{h}_i, \mathbf{w}_k, \hat{\boldsymbol{\theta}}) p(\mathbf{h}_i, \mathbf{w}_k)\right] \middle| \mathcal{X}, \boldsymbol{\theta}\right\}\end{aligned} \tag{3}$$

To maximize Eq.3, we need to estimate the posterior distributions of the latent variables given the model parameters $\boldsymbol{\theta}$. Denote $N_i = \sum_{k=1}^{K} H_i(k)$ as the number of training i-vectors from the *i*-th speaker and $M_k = \sum_{i=1}^{S} H_i(k)$ as the number of training i-vectors falling in the *k*-th SNR group. Then the E-step is as follows:

$$\mathbf{L}_i^1 = \mathbf{I} + N_i \mathbf{V}^\top \boldsymbol{\Phi}_1^{-1} \mathbf{V} \quad i = 1, \ldots, S \tag{4}$$

$$\mathbf{L}_k^2 = \mathbf{I} + M_k \mathbf{U}^\top \boldsymbol{\Phi}_2^{-1} \mathbf{U} \quad k = 1, \ldots, K \tag{5}$$

$$\langle \mathbf{h}_i | \mathcal{X} \rangle = (\mathbf{L}_i^1)^{-1} \mathbf{V}^\top \boldsymbol{\Phi}_1^{-1} \sum_{k=1}^{K} \sum_{j=1}^{H_i(k)} (\mathbf{x}_{ij}^k - \mathbf{m}) \tag{6}$$

$$\langle \mathbf{w}_k | \mathcal{X} \rangle = (\mathbf{L}_k^2)^{-1} \mathbf{U}^\top \boldsymbol{\Phi}_2^{-1} \sum_{i=1}^{S} \sum_{j=1}^{H_i(k)} (\mathbf{x}_{ij}^k - \mathbf{m}) \tag{7}$$

$$\langle \mathbf{h}_i \mathbf{h}_i^\mathsf{T} | \mathcal{X} \rangle = (\mathbf{L}_i^1)^{-1} + \langle \mathbf{h}_i | \mathcal{X} \rangle \langle \mathbf{h}_i | \mathcal{X} \rangle^\mathsf{T} \tag{8}$$

$$\langle \mathbf{w}_k \mathbf{w}_k^\mathsf{T} | \mathcal{X} \rangle = (\mathbf{L}_k^2)^{-1} + \langle \mathbf{w}_k | \mathcal{X} \rangle \langle \mathbf{w}_k | \mathcal{X} \rangle^\mathsf{T} \tag{9}$$

$$\langle \mathbf{w}_k \mathbf{h}_i^\mathsf{T} | \mathcal{X} \rangle = \langle \mathbf{w}_k | \mathcal{X} \rangle \langle \mathbf{h}_i | \mathcal{X} \rangle^\mathsf{T} \tag{10}$$

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

where

$$\boldsymbol{\Phi}_1 = \mathbf{U}\mathbf{U}^\top + \boldsymbol{\Sigma} \text{ and } \boldsymbol{\Phi}_2 = \mathbf{V}\mathbf{V}^\top + \boldsymbol{\Sigma},$$

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

Given Eq. 4–Eq. 11, the model parameters $\hat{\boldsymbol{\theta}}$ can be estimated via the M-step as follows:

$$\mathbf{m} = \frac{1}{N} \sum_{i=1}^{S} \sum_{k=1}^{K} \sum_{j=1}^{H_i(k)} \mathbf{x}_{ij}^k \tag{12}$$

$$\begin{aligned}\mathbf{V} = &\left\{ \sum_{i=1}^{S} \sum_{k=1}^{K} \sum_{j=1}^{H_i(k)} \left[(\mathbf{x}_{ij}^k - \mathbf{m}) \langle \mathbf{h}_i | \mathcal{X} \rangle - \mathbf{U} \langle \mathbf{w}_k \mathbf{h}_i^\mathsf{T} | \mathcal{X} \rangle \right] \right\} \\ &\times \left\{ \sum_{i=1}^{S} \sum_{k=1}^{K} \sum_{j=1}^{H_i(k)} \langle \mathbf{h}_i \mathbf{h}_i^\mathsf{T} | \mathcal{X} \rangle \right\}^{-1}\end{aligned} \tag{13}$$

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

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

where $N = \sum_{i=1}^{S} N_i = \sum_{k=1}^{K} M_k$. Algorithm 1 shows the

Algorithm 1 EM Algorithm for SNR-Invariant PLDA

Input:

Development data set consists of LDA- or NFA-reduced [30] i-vectors $\mathcal{X} = \{\mathbf{x}_{ij}^k | i = 1, \ldots, S; j = 1, \ldots, H_i(k); k = 1, \ldots, K\}$, with speaker labels and SNR group labels.

Initialization:

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

$\mathbf{V}, \mathbf{U} \leftarrow$ eigenvectors of PCA projection matrix obtained from data set $\mathcal{X}$;

Parameter Estimation:

1) Compute $\mathbf{m}$ via Eq. 12;
2) Compute $\mathbf{L}_i^1$ and $\mathbf{L}_k^2$ according to Eq. 4 and Eq. 5;
3) Compute the sufficient statistics using Eq. 6 to Eq. 11;
4) Update the model parameters using Eq. 13 to Eq. 15;
5) Go to step 2 until convergence;

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

procedures of applying the EM algorithm.

3.3. Likelihood Ratio Scores

Given a test i-vector $\mathbf{x}_t$ and a target-speaker i-vector $\mathbf{x}_s$, the likelihood ratio score can be computed as follows:

$$\begin{aligned} L(\mathbf{x}_s, \mathbf{x}_t) &= \ln \frac{P(\mathbf{x}_s, \mathbf{x}_t | \text{same-speaker})}{P(\mathbf{x}_s, \mathbf{x}_t | \text{different-speakers})} \\ &= \text{const} + \frac{1}{2}\mathbf{x}_s^\top \mathbf{Q}\mathbf{x}_s + \frac{1}{2}\mathbf{x}_t^\top \mathbf{Q}\mathbf{x}_t + \mathbf{x}_s^\top \mathbf{P}\mathbf{x}_t \end{aligned} \tag{16}$$

where

$$\begin{aligned} \mathbf{P} &= \boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac}(\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac})^{-1}, \\ \mathbf{Q} &= \boldsymbol{\Sigma}_{tot}^{-1} - (\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac})^{-1}, \\ \boldsymbol{\Sigma}_{ac} &= \mathbf{V}\mathbf{V}^\top, \text{ and } \boldsymbol{\Sigma}_{tot} = \mathbf{V}\mathbf{V}^\top + \mathbf{U}\mathbf{U}^\top + \boldsymbol{\Sigma}. \end{aligned}$$

4. Experiments

4.1. Speech Data and Front-End Processing

Experiments were performed on common conditions (CC) 1 and 4 of the core set of NIST 2012 Speaker Recognition Evaluation [31]. The test segments under CC1 and CC4 comprise interview conversations and telephone conversations, respectively. The microphone and telephone speech files from NIST 2005–2008 SREs were used as development data to train the gender-dependent UBMs and total variability matrices.

A two-channel voice activity detector (VAD) [32, 33] was applied to detect the speech regions of each utterance. 19 Mel frequency cepstral coefficients together with log energy plus their 1st- and 2nd-derivatives were extracted from the speech regions as detected by the VAD, followed by cepstral mean normalization [34] and feature warping [15] with a window size of 3 seconds. A 60-dim acoustic vector was extracted every 10ms, using a Hamming window of 25ms.

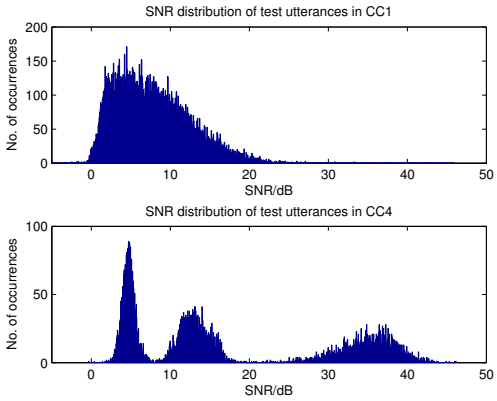

Figure 1: SNR distributions of test utterances in CC1 and CC4 of NIST 2012 SRE.

4.2. Preparation of Training Data

The telephone and microphone speech files in 2006–2010 SREs, excluding speakers with less than two utterances, were used as the training data to train the gender-dependent subspace projection matrices and all PLDA models.

The SNR distributions of test utterances in CC1 and CC4 are shown in Fig. 1. Because the SNR range of the test utterances in CC4 is large,[1] the SNR mismatch between the training and the test utterances has significant effect on the test trials in CC4. To address this issue, we added noise to the telephone training data. Specifically, for each telephone speech file, a noise waveform file was randomly selected from the 30 noise waveform files in the PRISM data set [35] and added to the speech file at a target SNR using the FaNT tool [36]. The target SNR was selected in turn from an SNR set comprising $\{6\text{dB}, 7\text{dB}, \ldots, 15\text{dB}\}$. As a result, for each original file, ten noise corrupted files with different SNRs were generated. Then, we used the voltmeter function of FaNT and the decisions of the VAD to estimate the "actual" SNR of the noise-corrupted speech files. While the actual SNR is close to the target SNR, they will not be exactly the same. The distribution of the actual SNRs (as measured by FaNT) of the noise-corrupted speech files together with the original telephone and microphone speech files is shown in the bottom panel of Fig. 2.

For experiments on CC4, 14,226 (resp. 22,356) noise corrupted files from 763 male (resp. 1030 female) speakers were combined with the original telephone and microphone utterances in 2006–2010 SREs to form the training set for training the male PLDA models. For experiments on CC1, the microphone utterances from 347 male speakers and 425 female speakers in NIST 2006-2010 SREs were used as the training set. The SNR distributions of the training sets used for male speakers in CC1 and CC4 are respectively shown in Fig. 2.

[1] In the SRE, noise was artificially added to the test segments of CC4.

Table 1: Division of male training utterances for CC1 of NIST 2012 SRE into $K = 3$ SNR sub-groups.

Sub-Group	SNR Range (dB)	No. of Utterances
1	SNR ≤ 10	4022
2	$10 <$ SNR ≤ 16	4023
3	SNR > 16	3963

Table 2: Performance of PLDA, mPLDA [25] and SNR-invariant PLDA on CC1 and CC4 of NIST 2012 SRE (core set). K is the number of SNR groups and Q is the dimension of SNR factors in SNR-invariant PLDA.

Method	Parameters		CC1				CC4			
	K	Q	Male		Female		Male		Female	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
PLDA	–	–	5.42	**0.371**	7.53	0.531	3.13	0.312	2.82	0.341
mPLDA	–	–	5.28	0.415	7.70	0.539	2.88	0.329	2.71	0.332
SNR-invariant PLDA	2	40	5.41	0.376	7.03	0.525	2.75	0.290	2.42	0.325
	3	40	5.42	0.382	6.93	0.528	2.72	0.289	2.36	0.314
	4	40	5.42	0.392	7.03	0.522	2.70	0.289	2.39	0.329
	5	40	**5.28**	0.381	**6.89**	**0.522**	2.67	0.291	2.38	0.322
	6	40	5.29	0.388	6.90	0.536	**2.63**	**0.287**	2.43	0.319
	7	30	5.48	0.385	7.03	0.533	2.63	0.294	2.32	0.316
	8	30	5.56	0.384	7.05	0.545	2.70	0.292	**2.29**	**0.313**

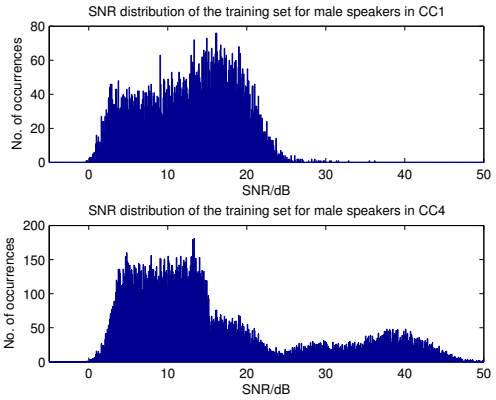

Figure 2: SNR distributions of the training utterances for male speakers in CC1 and CC4.

4.3. I-vector Preprocessing

The extraction of i-vectors was based on a gender-dependent UBM with 1024 mixtures and a total variability matrix with 500 total factors. Similar to [37], we applied within-class covariance normalization (WCCN) [4] and i-vector length normalization (LN) to the 500-dimensional i-vectors. Then nonparametric feature analysis (NFA) [30] was used to reduce intra-speaker variability and emphasize discriminative class boundary information. After this procedure, the dimension of i-vectors was reduced to 200. Then PLDA models and SNR-invariant PLDA models with 150 latent identity factors were trained. Also, the SNR-dependent mixture of PLDA in [25] was used as a comparison, which is named as mPLDA in the sequel.

4.4. SNR Sub-group Division

To train the SNR-invariant PLDA models, the training set was divided into K groups according to the measured SNRs of the utterances. The SNRs of the whole training set were divided into K SNR intervals. The k-th group comprises the i-vectors whose corresponding utterances have SNR falling in the k-th SNR interval. The numbers of the i-vectors in each sub-group should be comparable. For example, when $K = 3$, the divisions for the training set used for male speakers in CC1 and the numbers of training utterances falling in each of the sub-groups are shown in Table 1.

4.5. Results and Discussions

This section reports the performance of different systems based on equal error rate (EER) and minimum DCF (minDCF) [31].

Results on CC1 in Table 2 show that mPLDA and SNR-invariant PLDA can achieve a lower EER than PLDA for male test segments. For female test segments, SNR-invariant PLDA outperforms mPLDA and PLDA in terms of both EER and minDCF, and it achieves the best performance when the number of SNR groups was set to 5. The results suggest that SNR-invariant PLDA can address SNR mismatch under noisy conditions.

Results on CC4 in Table 2 show that mPLDA and SNR-invariant PLDA outperform PLDA, and the best result was achieved by SNR-invariant PLDA. Moreover, the performance of SNR-invariant PLDA stays stable for different numbers of SNR groups.

For SNR-invariant PLDA, it is important to determine an appropriate value of K, especially when the training samples are not abundant (such as in CC1). In particular, in the two extreme cases where K is either very small or very large (same as the number of training i-vectors), the performance gain of SNR-invariant PLDA will be limited. This is because for the former, each of the SNR factors ($\mathbf{w}_k$ in Eq. 2) will need to represent the i-vectors with a wide range of SNR. On the other hand, for the latter case, there will be so many SNR factors in Eq. 2 such that each i-vector is considered to be obtained from a distinct SNR group. This means that in such extreme situations, the SNR-invariant PLDA model reduces to the traditional Gaussian PLDA, which only considers the session variability instead of the variability caused by different levels of SNR.

5. Conclusions

In this paper, SNR-invariant PLDA was proposed to deal with the mismatch caused by different levels of background noise. By assuming that the i-vectors share the same SNR-specific information when the corresponding utterances' SNRs fall within a narrow range, we incorporated an SNR factor to the traditional Gaussian PLDA model. Experiments on the NIST SRE 2012 demonstrate the effectiveness of the proposed method.

6. References

[1] 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.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

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

[4] 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.

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

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

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

[8] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4249–4252.

[9] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and systems Magazine, IEEE*, vol. 11, no. 2, pp. 23–61, 2011.

[10] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1711–1723, 2007.

[11] Y. Shao, S. Srinivasan, and D. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Acoustics, Speech and Signal Processing (ICASSP), 2007 IEEE International Conference on*. IEEE, 2007, pp. 277–280.

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

[13] Y. Shao and D. 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.

[14] 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.

[15] 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.

[16] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 836–845, 2014.

[17] R. Saeidi, K. Lee, T. Kinnunen, T. Hasan, B. Fauve, P. Bousquet, E. Khoury, P. S. Martinez, J. Kua, C. You *et al.*, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. Interspeech*, 2013, pp. 1986–1990.

[18] S. O. Sadjadi, J. W. Pelecanos, and W. Zhu, "Nearest neighbor discriminant analysis for robust speaker recognition," in *Proc. Interspeech*, 2014, pp. 1860–1864.

[19] S. Sarkar and K. S. Rao, "A novel boosting algorithm for improved i-vector based speaker verification in noisy environments," in *Proc. Interspeech*, 2014, pp. 671–675.

[20] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6778–6782.

[21] 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.

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

[23] 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.

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

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

[26] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, "Hidden factor analysis for age invariant face recognition," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2872–2879.

[27] N. Li and M. W. Mak, "SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, to appear.

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

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

[30] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 755–761, 2009.

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

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

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

[34] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, Jun. 1974.

[35] 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."

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

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