

Speech Technology manuscript No.
(will be inserted by the editor)

Noise Robust Speaker Verification via the Fusion of SNR-Independent and SNR-Dependent PLDA

Received: date / Accepted: date

Abstract While i-vectors with probabilistic linear discriminant analysis (PLDA) can achieve state-of-the-art performance in speaker verification, the mismatch caused by acoustic noise remains a key factor affecting system performance. In this paper, a fusion system that combines a multi-condition SNR-independent PLDA model and a mixture of SNR-dependent PLDA models is proposed to make speaker verification systems more noise robust. First, the whole range of SNR that a verification system is expected to operate is divided into several narrow ranges. Then, a set of SNR-dependent PLDA models, one for each narrow SNR range, are trained. During verification, the SNR of the test utterance is used to determine which of the SNR-dependent PLDA models is used for scoring. To further enhance performance, the SNR-dependent and SNR-independent models are fused using linear and logistic regression fusion. The performance of the fusion system and the SNR-dependent system is evaluated on the NIST 2012 SRE for both noisy and clean conditions. Results show that a mixture of SNR-dependent PLDA models perform better in both clean and noisy conditions. It was also found that the fusion system is more robust than the conventional i-vector/PLDA systems under noisy conditions.

Keywords Speaker verification · i-vectors · probabilistic LDA · NIST 2012 SRE · noise robustness · fusion

1 Introduction

In practical situations, performance of speaker verification systems is always degraded by the variation in the acoustic environments. There has been a lot of research in compensating for the effect of these variations, [which results in](#)

Centre for Signal Processing,
Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University
E-mail: enmwmak@polyu.edu.hk

a number of methods that work in either the front-end [30, 38] or the backend [26, 37, 39, 19, 21] of the verification process. It has been found that the backend techniques is more promising, especially when joint factor analysis (JFA) [13] and i-vector/PLDA frameworks [4, 12] are employed.

The i-vector is a low-dimensional vector that represents both speaker and channel characteristics of an utterance. The low-dimensionality of i-vectors facilitates the usage of classical statistical techniques such as linear discriminant analysis (LDA) [1], within-class covariance normalization (WCCN) [11] and probabilistic linear discriminant analysis (PLDA) [31] to suppress the channel-variability [4, 34, 24]. PLDA performs factor analysis on the i-vector space by grouping the i-vectors derived from the same speaker in order to find a subspace with minimal channel variability. PLDA is one of the most promising techniques in speaker verification.

Based on the i-vector/PLDA framework, more advanced approaches have been proposed for noise-robust i-vector extraction. For example, Yu *et al.* [40] propagated the uncertainty of noisy acoustic features into the i-vector extraction process in an attempt to marginalizing out the effect of noise. This is achieved by expressing the posterior density of an i-vector in terms of the joint density of the clean and noisy acoustic features where the uncertainties of the noisy features are represented through the variances of the joint density. To account for all possible clean features, the joint density is marginalized over all possible clean acoustic features. The marginalized density is then plugged back into the posterior density of i-vectors, where the noise-robust i-vector is its posterior mean. This modified i-vector extraction method has shown potential for improving the robustness of speaker recognition especially in low SNR conditions.

Hasan and Hansen [8] proposed an acoustic factor analysis (AFA) scheme, which is essentially a mixture-dependent feature transformation that integrates dimensionality reduction, de-correlation, normalization and enhancement together. It was demonstrated that this transformation method can remove the need for hard feature clustering and avoid retraining of the [universal background model](#) (UBM) from the new features. In [9], the AFA concept was further enhanced by replacing the UBM with a mixture of factor analyzers and a new i-vector extractor was proposed.

Lei et al. [17] proposed a noise robust i-vector extractor using vector Taylor series (VTS). The method adapts the UBM to speech signals contaminated with additive and convolutive noises and then extracts the noise-compensated i-vector based on the sufficient statistics collected from the adapted UBM. To release the computational burden of the VTS approach, [18] further proposed an efficient approximation, called simplified-VTS (sVTS), which collects sufficient statistics and whitens them using the VTS-synthesized UBM. As an alternative approach to VTS, an unscented transform was presented in [23] to approximate the nonlinearities between clean and noisy speech models in the cepstral domain. It is expected that the unscented transform is more accurate than VTS when the distortions are far from locally linear.

Noting that the convolution and max-pooling operations in convolution neural networks (CNN) can reduce distortion caused by noise, McLaren et al. [25] proposed using a CNN to estimate the posterior probability of senones and used the posterior probabilities to replace the zero-order statistics extracted from the UBM. The resulting sufficient statistics are then used for estimating the i-vectors. It was found that the performance of this CNN/i-vector framework is comparable to that of UBM/i-vector framework and that fusion of these two framework is very promising.

There are also a lot of research concentrating on the backend PLDA stage. For example, McLaren et al. [25] developed a multiple system fusion approach that uses multiple streams of noise-robust features for i-vector fusion and score fusion. I-vector fusion consists of concatenating the stream-dependent i-vectors to a single vector and score fusion fuses scores obtained from the i-vector fusion system and the single-feature i-vector systems. Many systems use LDA and [within-class covariance normalization](#) (WCCN) to pre-process the length-normalized i-vectors before presenting them to the PLDA model for scoring. Noting that the actual distortion of i-vectors may not be Gaussian, Sadjjadi et al. [36] replaced LDA by non-parametric discriminant analysis (NDA) that uses nearest-neighbor rule to estimate the between- and within-speaker scatter matrices. They found that NDA is more effective than the conventional LDA under noisy and channel degraded conditions.

In [15, 16, 10, 32, 33], multi-condition training, in which a PLDA model is trained by pooling clean and noisy utterances together, was employed to enhance noise robustness. Garcia-Romero et al. [7] trained a collection of PLDA models, each for a specific condition, and found that the pooled-PLDA is more appealing due to its good performance as well as the small number of parameters.

Unlike [7] where the verification score is a convex mixture of the individual PLDA models weighted by the posterior probability of the test condition (Eq. 4 of [7]), the SNR-dependent PLDA models proposed in this paper compute the verification scores based on the SNR of test utterances. Specifically, hard-decision SNR-dependent PLDA chooses one of the SNR-dependent PLDA models based on the SNR of test utterances; soft-decision SNR-dependent PLDA calculates weights of the individual PLDA by incorporating posterior of the SNR of test utterances. Observing the performance improvement in multi-condition training, a fusion system combining a mixture of SNR-dependent PLDA models and a multi-condition PLDA model was developed in this work.

One of the challenges in speaker verification is to maintain performance under adverse acoustic condition. For the i-vector/PLDA framework, approaches such as advanced transformation [8], noise robust i-vector extraction [17], and multi-condition PLDA [15, 16, 10, 32, 33] have shown promise in improving the robustness of speaker verification systems. However, none of these methods explore the noise robustness of the SNR-dependent PLDA models. This paper aims to fill this gap by extending our earlier work on SNR dependent models [29] by the following three fronts:

1. investigating both hard- and soft-decision strategies for the SNR-dependent PLDA.
2. conducting additional experiments on clean phone call speech (Common Condition 2) and interview speech (Common Conditions 1 and 3) for both male and female speakers in NIST 2012 SRE.
3. performing more analysis on fusion systems with respect to decision thresholds, decision strategies, fusion methods, and fusion weights.

The paper is organized as follows. Section 2 outlines the i-vector/PLDA framework for speaker verification. Sections 3 and 4 describe hard- and soft-decision SNR-dependent PLDA models and fusion systems respectively. In Sections 5 and 6, we report evaluations based on NIST 2012 SRE [28]. Section 7 concludes the findings.

2 The I-Vector/PLDA Framework

2.1 I-Vector Extraction

The i-vector approach [4] defines a low-dimensional total variability space that encompasses both speaker and channel variabilities. In this space, each utterance is represented by the latent factor in a factor analysis model:

$$\mathbf{m}_x = \mathbf{m} + \mathbf{T}\mathbf{x} \quad (1)$$

where \mathbf{m} is the speaker- and channel-independent GMM-supervector formed by stacking the mean vectors of the universal background model (UBM) [35], \mathbf{m}_x is the speaker-dependent supervector, \mathbf{T} is a low-rank total variability matrix, and \mathbf{x} is the low-dimensional latent factor. Given an utterance, the posterior mean of the latent factor is the utterance’s i-vector. The training of the total variability matrix \mathbf{T} is similar to the training of the eigenvoice matrix in JFA [14], except that the speaker labels are ignored.

2.2 PLDA Model

Probabilistic linear discriminant analysis (PLDA) [6, 12, 31] considers the i-vectors of utterances as observations generated by a generative model. Specifically, assuming there are R utterances from a speaker s and denoting \mathbf{x}_{sr} ($r = 1, \dots, R$) as the collection of the corresponding i-vectors, the PLDA model decomposes i-vector \mathbf{x}_{sr} into:

$$\mathbf{x}_{sr} = \boldsymbol{\mu} + \mathbf{V}\mathbf{z}_s + \boldsymbol{\epsilon}_{sr}, \quad (2)$$

where $\boldsymbol{\mu}$ is the global offset, \mathbf{V} defines the bases of the speaker subspace, \mathbf{z}_s is the speaker factors, and $\boldsymbol{\epsilon}_{sr}$ is the residual noise assumed to follow a Gaussian distribution with zero mean and diagonal covariance $\boldsymbol{\Sigma}$. An [expectation-maximization](#) (EM) algorithm [31] is applied to estimate the parameters of the factor analyzer (Eq. 2).

Given a test i-vector \mathbf{x}_t and target-speaker's i-vector \mathbf{x}_s , a verification score can be computed as a log-likelihood ratio of two Gaussian distributions [6]:

$$\text{score} = \log \left[\frac{p(\mathbf{x}_s, \mathbf{x}_t | \mathcal{H}_1)}{p(\mathbf{x}_s | \mathcal{H}_0)p(\mathbf{x}_t | \mathcal{H}_0)} \right] \quad (3)$$

where the hypotheses \mathcal{H}_1 and \mathcal{H}_0 denote that the two i-vectors come from the same- and different-speakers, respectively. By assuming that the i-vectors (after length normalization) and the latent factor \mathbf{z} follow Gaussian distributions, the verification score is [6]:

$$\begin{aligned} \text{score} &= \log \left[\frac{\int p(\mathbf{x}_s, \mathbf{x}_t, \mathbf{z} | H_1) d\mathbf{z}}{\int p(\mathbf{x}_s, \mathbf{z}_s | H_0) d\mathbf{z}_s \int p(\mathbf{x}_t, \mathbf{z}_t | H_0) d\mathbf{z}_t} \right] \\ &= \frac{1}{2} [\mathbf{x}_s^T \mathbf{Q} \mathbf{x}_s + 2\mathbf{x}_s^T \mathbf{P} \mathbf{x}_t + \mathbf{x}_t^T \mathbf{Q} \mathbf{x}_t] + \text{const} \end{aligned} \quad (4)$$

where

$$\mathbf{Q} = \boldsymbol{\Sigma}_{tot}^{-1} - (\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac} \boldsymbol{\Sigma}_{tot}^{-1} \boldsymbol{\Sigma}_{ac})^{-1} \quad \text{and} \quad \mathbf{P} = \boldsymbol{\Sigma}_{tot}^{-1} \boldsymbol{\Sigma}_{ac} (\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac} \boldsymbol{\Sigma}_{tot}^{-1} \boldsymbol{\Sigma}_{ac})^{-1}, \quad (5)$$

where

$$\boldsymbol{\Sigma}_{tot} = \mathbf{V} \mathbf{V}^T + \boldsymbol{\Sigma} \quad \text{and} \quad \boldsymbol{\Sigma}_{ac} = \mathbf{V} \mathbf{V}^T. \quad (6)$$

3 SNR-Dependent PLDA

Classical Gaussian PLDA assumes that i-vectors follows a Gaussian distribution. However, the assumption of single Gaussian is rather limited, especially under noisy environments with a wide range of signal-to-noise ratio (SNR). In this situation, a group of SNR-dependent PLDA models, in which each model is responsible for a small range of SNR, are more suitable. Specifically, the parameters of each SNR-dependent PLDA model are estimated independently by an EM algorithm [31] using training data contaminated with different level of background noise.

3.1 Hard-Decision SNR-Dependent PLDA

For the hard-decision SNR-dependent systems, one SNR-dependent PLDA model is chosen for each test i-vector. During verification, the 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:

$$\text{If} \begin{cases} \ell_t \leq \eta_1, & \text{use 6dB PLDA and 6dB target's i-vectors} \\ \eta_1 < \ell_t \leq \eta_2, & \text{use 15dB PLDA and 15dB target's i-vectors} \\ \ell_t > \eta_2, & \text{use clean PLDA and clean target's i-vectors} \end{cases} \quad (7)$$

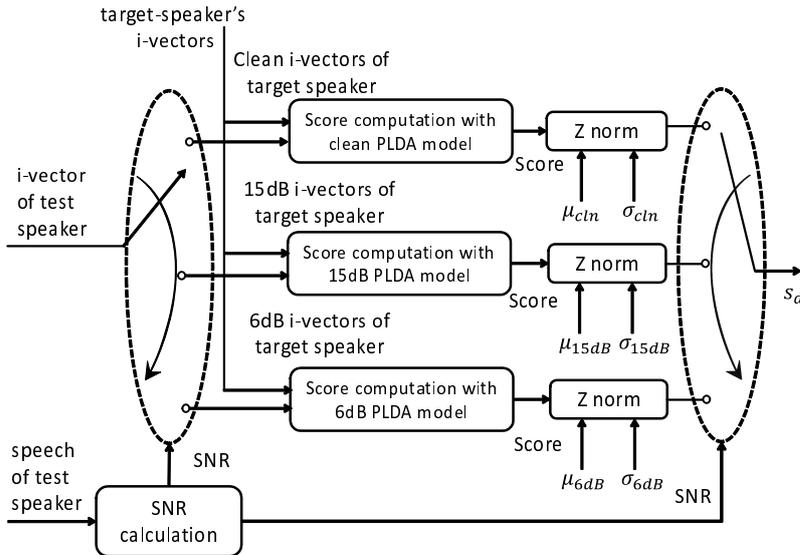


Fig. 1: The data-flow of hard-decision SNR-dependent PLDA scoring.

where ℓ_t is the SNR of the test utterance, and η_1 and η_2 are decision thresholds. A disadvantage of this hard-decision approach is that it is necessary to determine η_1 and η_2 and that their optimal values depend on the SNR of test utterances. Instead of finding their optimal values using a held-out set, in this work, we varied their values and investigated how they affect performance.

Fig. 1 illustrates the data-flow of hard-decision SNR-dependent PLDA scoring. A test i-vector is fed to one of the SNR-dependent PLDA models and is scored against the corresponding i-vectors of the target speaker. Because the three PLDA models produce scores at different ranges, the scores should be normalized before computing the [equal error rate \(EER\)](#) and [minimum decision-cost function \(minDCF\)](#). We applied SNR-dependent Z-norm to the PLDA scores, with the three sets of Z-norm parameters found independently using the training files contaminated with different level of background noise. In theory, Z-norm is not necessary if the PLDA scores are well calibrated [3]. However, we found that it is not easy to achieve perfect calibration without having a set of held-out set that has the same characteristics as the test data. Therefore, we opted for using the more conventional Z-norm in this step.

3.2 Soft-Decision SNR-Dependent PLDA

In the soft-decision SNR-dependent systems, given a test utterance, the posterior probability of SNR of the test utterance is used to combine the scores of different PLDA models. Specifically, denote the SNR of a test utterance as

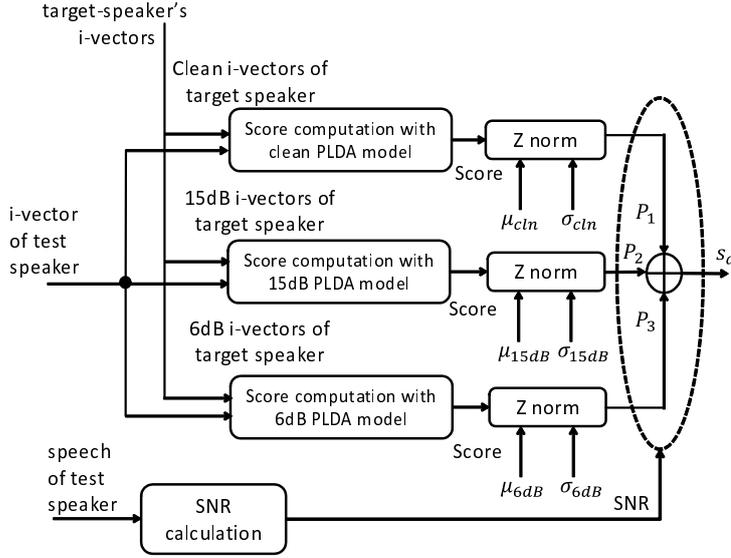


Fig. 2: The data-flow of soft-decision SNR-dependent PLDA scoring.

ℓ_t , the posterior probability of SNR is

$$P(\omega_j|\ell_t) = \frac{P(\omega_j)p(\ell_t|\mu_j, \sigma_j)}{\sum_{i=1}^3 P(\omega_i)p(\ell_t|\mu_i, \sigma_i)} \quad j = 1, 2, \text{ or } 3 \quad (8)$$

where ω_j 's are class labels corresponding to clean, 15dB or 6dB PLDA models, $P(\omega_j)$ is the prior probability of ω_j , $p(\ell_t|\mu_j, \sigma_j)$ is the probability density function of ℓ_t with mean μ_j and standard deviation σ_j . To implement Eq. 8, we trained a 1-dimensional Gaussian mixture model (GMM) with 3 mixtures using the SNR of training utterances. See Section. 5.4 for details.

In the hard-decision SNR-dependent systems, the SNR of the test utterance directly determines which SNR-dependent PLDA model is used and only one SNR-dependent PLDA model is chosen for scoring with the test i-vector, as Eq. 7 describes. However, in the soft-decision SNR-dependent systems, three SNR-dependent PLDA models and SNR-dependent i-vectors of the target speakers are all used for scoring. The posterior probability obtained in Eq. 8 determines how much the three scores derived from the three SNR-dependent PLDA models contribute to the overall score. Specifically, the overall score is:

$$s_d = P_1 s_{cln} + P_2 s_{15dB} + P_3 s_{6dB} \quad (9)$$

where s_{cln} , s_{15dB} and s_{6dB} are the normalized score from the clean, 15dB and 6dB SNR-dependent PLDA respectively, P_1 , P_2 and P_3 are $P(\text{clean}|\ell_t)$, $P(15\text{dB}|\ell_t)$ and $P(6\text{dB}|\ell_t)$ in Eq. 8, respectively, and s_d is the soft-decision SNR-dependent systems score. Fig. 2 illustrates the data-flow of soft-decision

SNR-dependent PLDA scoring. In the figure, a test i-vector is fed to three SNR-dependent PLDA models simultaneously and scored against the corresponding i-vectors of the target speaker. The three normalized scores are then linearly combined to obtain the overall score s_d , as Eq. 9 describes.

4 Fusion of SNR-Dependent PLDA

The fusion system combines the SNR-dependent system and the SNR-independent system. We investigated linear fusion and logistic regression fusion in this work. Fig. 3 illustrates the fusion system. In the figure, the upper part is the SNR-independent system whose PLDA model is trained by pooling the training data with variable noise levels. The lower part is the SNR-dependent system. It can be hard- or soft-decision SNR-dependent PLDA. The fusion block can be either linear fusion or logistic regression fusion. For linear fusion, the test scores are used to determine the best fusion weight while in logistic regression fusion, training scores are used to compute the fusion parameters. The following two subsections describe these two fusion methods.

4.1 Linear Fusion

Fig. 4 shows the test scores of imposters and true speakers obtained by SNR-dependent PLDA and SNR-independent PLDA models. Evidently, the scores from the two systems can be separated by a straight line. Therefore, a simple way to fuse the two systems is to linearly combine their scores:

$$s = ws_i + (1 - w)s_d \quad (10)$$

where s_i is the normalized score from the SNR-independent system, s_d is the normalized score from the SNR-dependent system, and w is the combination weight.

As described earlier, the Z-norm parameters represented by μ and σ in Fig. 1 and Fig. 2 were derived independently from the i-vectors used for training the PLDA models. The scores obtained from the SNR-independent system are also normalized to make sure that they are consistent with those obtained from the SNR-dependent system. The Z-norm equation is as follows:

$$s_i = \frac{\text{score} - \mu_{multi}}{\sigma_{multi}} \quad (11)$$

where s_i is the score after normalization, μ_{multi} and σ_{multi} are the normalization parameters shown in Fig. 3.

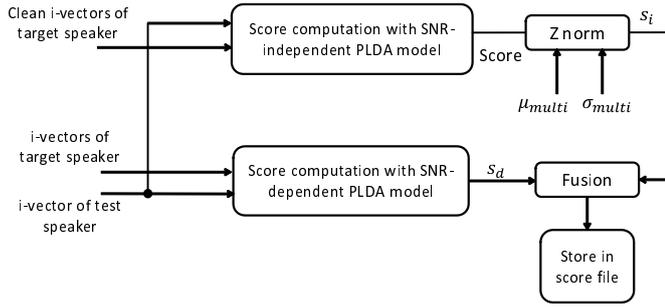


Fig. 3: The data-flow of fusing SNR-independent and SNR-dependent system. The fusion block can be either linear fusion or logistic regression fusion.

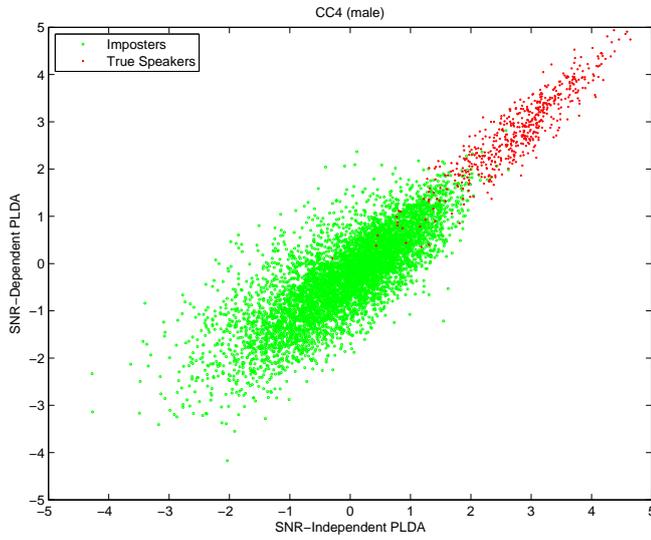


Fig. 4: Test score distributions of imposters and true speakers of SNR-independent (pooling) and hard-decision SNR-dependent PLDA in CC4 of NIST 2012 SRE (male speakers). The decision thresholds η_1 and η_2 in Eq. 7 were set to 3 and 20, respectively.

4.2 Logistic Regression Fusion

In logistic regression fusion [1, 2], the fused scores are also a linear combination of N sub-systems' scores:

$$s = \alpha_0 + \alpha_1 s_1 + \alpha_2 s_2 + \cdots + \alpha_N s_N \quad (12)$$

where $\{\alpha_1, \alpha_2, \dots, \alpha_N\}$ are the fusion weights for the corresponding subsystems and α_0 is a bias term used for calibrating the fused scores. The fusion weights $\{\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_N\}$ can be obtained by the learning algorithm of logistic regression [2]. In this work, we used the data that have been used for training the PLDA models for estimating the fusion weights in Eq. 12. Given the i-vectors of a set of training speakers, we computed the intra- and inter-speaker PLDA scores from the SNR-independent and SNR-dependent PLDA models (both soft- and hard-decisions). These scores represent the speaker and impostor scores of s_1 , s_2 , and s_3 in Eq. 12. Then, $[s_1 \ s_2 \ s_3]^T$'s are considered as 3-dimensional training vectors of a logistic regression classifier [2].

5 Experiments

5.1 Speech Data and Acoustic Features

Both the phonecall speech and the interview speech in the core set of NIST 2012 Speaker Recognition Evaluation (SRE) [28] were used for performance evaluation. Table 1 [28] summarizes the conditions of the test segments in the evaluation. In this paper, we use the term “segment” and “utterance” interchangeably. Fig. 6 shows the SNR distributions of test utterances for male speakers in the evaluation (the distributions for female speakers are similar). It shows that the noisy test utterances cover a wide range of SNR, especially for CC4. [To make the SNR distribution of training segments comparable with that of the test segments, we added babble noise from the PRISM dataset to the training files at 6dB and 15dB to create 3 SNR-dependent PLDA models: 6dB, 15dB, and clean \(using the original sound files\).](#) Fig. 5 shows the SNR distribution of telephone (tel) and microphone (mic) speech files after adding noise.

The training segments comprise phonecall speech and interview speech with variable length. We removed the 10-second segments and the summed-channel segments 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 projection matrices, PLDA models and Z-norm parameters.

Speech regions in the speech files were extracted by using a two-channel voice activity detector (VAD) [20]. For each frame, 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 [30] with a window size of 3 seconds. A 60-dim acoustic vector was extracted every 10ms, using a Hamming window of 25ms.

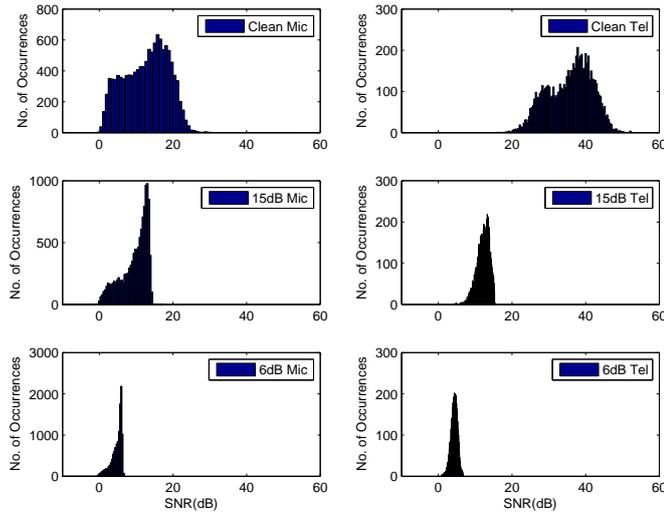


Fig. 5: SNR distributions of clean, 15dB and 6dB training utterances (male speakers). Left panels: interview speech; right panels: telephone speech.

Common Condition	Test Segment Condition
CC1	Interview speech
CC2	Phonecall speech
CC3	Interview speech with added noise
CC4	Phonecall speech with added noise
CC5	Phonecall speech intentionally collected in noisy environments

Table 1: Test segment conditions for CC1–CC5 of NIST 2012 SRE.

5.2 Creating Noise Contaminated Utterances

For each clean training file, we randomly selected one out of the 30 noise files from the PRISM dataset [5] and added the noise waveform to the file at an SNR of 6dB and 15dB using the FaNT tool.¹

To measure the “actual” SNR of speech files (including the original and noise contaminated ones), we used the voltmeter function of FaNT and the speech/non-speech decisions of our VAD [20, 41] as follows. Given a speech file, we passed the waveform to the G.712 frequency weighting filter in FaNT and then estimated the speech energy using the voltmeter function (`sv-p56.c` from the ITU-T Software Tool Library [27]). Then, we extracted the non-speech

¹ <http://dnt.kr.hsrn.de/download.html>

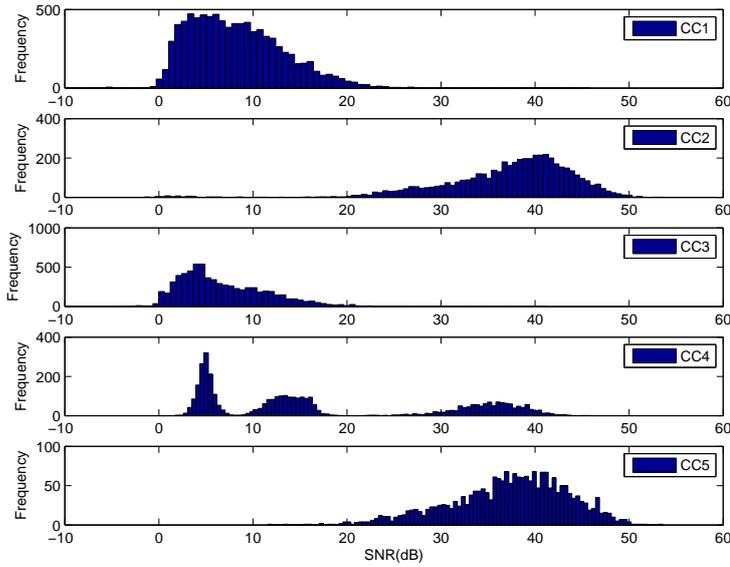


Fig. 6: SNR distributions of test utterances in CC1–5 of NIST 2012 SRE (male speakers).

segments based on the VAD’s decisions and passed the non-speech segments to the voltmeter function to estimate the noise energy. The difference between the signal and noise energies in the log domain gives the measured SNR of the file. While the measured SNR is close to the target SNR, they will not be exactly the same. This explains why we have a continuous SNR distribution in Fig. 5. Because we used our VAD (which is more robust than the VAD in `sv-p56.c`) to determine the background segments, we are able to measure the SNR even for very noisy files.

5.3 I-Vector Extraction and PLDA-Model Training

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 (without adding noise) from NIST 2005–2008 SREs were used for training the UBMs and total variability matrices. Following [24], WCCN [11] and i-vector length normalization [6] were applied to the 500-dimensional i-vectors. Then, linear discriminant analysis (LDA) [1] and WCCN were applied to reduce the dimension to 200 before training the PLDA models with 150 latent variables.

Considering that CC1 and CC3 contain interview speech and that CC2, CC4 and CC5 contain phonecall speech, their PLDA models were trained sep-

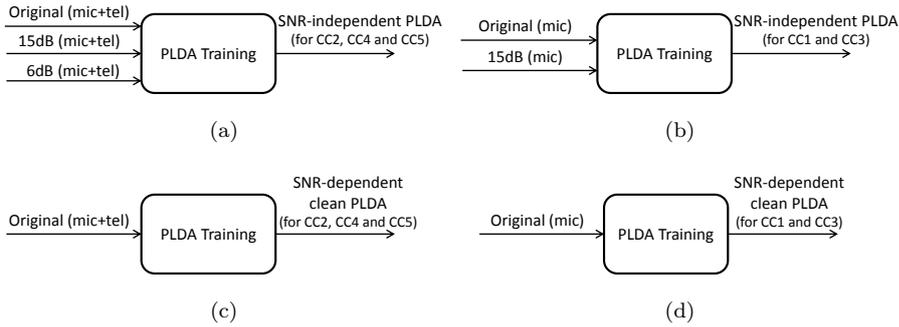


Fig. 7: Training procedure for (a) SNR-independent models for phonecall speech, (b) SNR-independent models for interview speech, (c) SNR-dependent clean models for phonecall speech (same for 15dB and 6dB models except using noisy speech), and (d) SNR-dependent models for interview speech.

arably. Specifically, for CC2, CC4 and CC5, both SNR-independent and SNR-dependent PLDA models were trained. For the former (Fig. 7a), we pooled the 6dB (mic+tel), 15dB (mic+tel), and original (mic+tel) speech files — excluding speakers with less than two utterances — into a single training set. Two SNR-independent PLDA models with 150 factors were then trained, one for each gender. For SNR-dependent PLDA, the 6dB (mic+tel), 15dB (mic+tel), and original (mic+tel) speech files were independently used to train three PLDA models, each with 150 factors (Fig. 7c).

According to the left panels in Fig. 5, SNR distribution of clean mic, 15dB mic and 6dB mic overlap each other. Therefore, for CC1 and CC3, we only used clean mic and 15dB mic to train the SNR-independent PLDA (Fig. 7b) and SNR-dependent PLDA (Fig. 7d).

5.4 PLDA scoring

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/microphone sessions of original (clean) speech files using the conventional PLDA scoring function (Eq. 4).

For hard-decision SNR-dependent PLDA, as Eq. 7 describes, one of the SNR-dependent PLDA models was chosen to score against the corresponding target's i-vectors based on the SNR of the test utterance. Fig. 6 shows the SNR distributions of male test utterances in CC1–CC5 in 2012 SRE. Based on the distributions, the decision thresholds (Eq. 7) for hard-decision SNR-dependent PLDA were set.

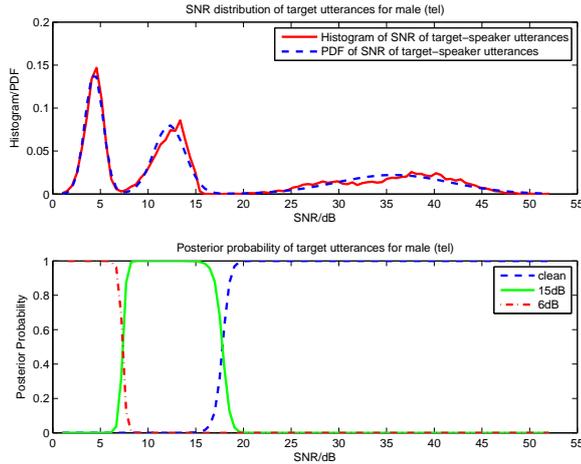


Fig. 8: Probability density function (pdf) and posterior probability of SNR derived from the training sessions (tel) of NIST 2012 SRE (male speakers).

For the soft-decision SNR-dependent system, for each test i-vector, three scores were computed using three SNR-dependent PLDA models and the corresponding target’s i-vectors independently. Then, the posterior probability of SNR of the test utterance was used to determine the weights for combining the scores produced by the SNR-dependent PLDA models, as Eq. 9 describes. The posterior probability (Eq. 8) is based on a 1-dimensional Gaussian mixture model (GMM) that models the SNR of training utterances.

For CC2, CC4 and CC5 whose test segments were from phonecall speech, the SNR of the original (tel), 15dB (tel) and 6dB (tel) utterances were used to train a 3-mixture GMM. For CC1 and CC3 whose test segments were from interview speech, the SNR of the original (mic) and 15dB (mic) training utterances were used to train a 2-mixture GMM.

The upper panels of Fig. 8 and Fig. 9 show the histogram and probability density function of SNR for male speakers of the telephone and microphone training sessions, respectively. Fig. 8 shows that the three mixtures derived from three telephone training files are well separated and they have equal mixture coefficients since the 15dB and 6dB noisy training files were derived from the original training files by adding noise at different SNR. On the other hand, Fig. 9 shows that the SNR of the original and 15dB mic training files are not well separated. The lower panels of Fig. 8 and Fig. 9 show the posterior probability of SNR obtained from Eq. 8 for male speakers, and the posterior probability of female speakers is similar with that of male speakers. During verification, test utterance’s SNR determines P_1 , P_2 and P_3 in Eq. 9 based on the posterior probability in Fig. 8 and Fig. 9.

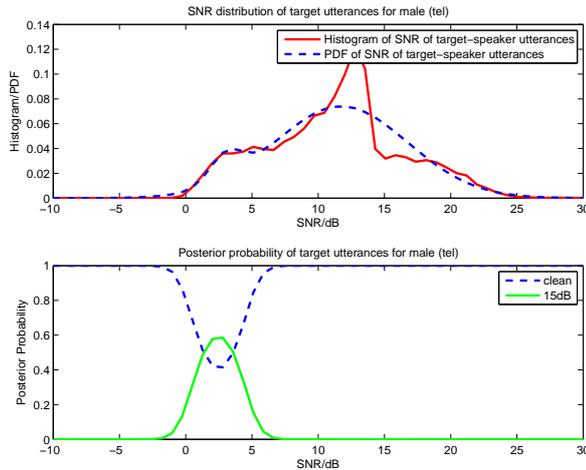


Fig. 9: Probability density function (pdf) and posterior probability of SNR derived from the training sessions (mic) of NIST 2012 SRE (male speakers).

5.5 Fusion of SNR-Dependent PLDA models

In linear fusion, the fusion scores are a linear combination of SNR-independent scores and SNR-dependent scores (Eq. 10). Either hard-decision or soft-decision SNR-dependent system can be fused with the SNR-independent system. In logistic regression fusion, according to in Eq. 12, N subsystems can be combined. In particular, the SNR-independent system can be combined with both hard-decision and soft-decision SNR-dependent systems and these three systems can also be combined. The fusion parameters $\{\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_N\}$ in the logistic regression fusion were derived from the training scores obtained from the original, 15dB and 6dB i-vectors used for training the PLDA models.

As described in Sections 3 and 4, score normalization is necessary for both SNR-independent and SNR-dependent systems. The Z-norm parameters represented by μ and σ in Fig. 1, Fig. 2 and Fig. 3 were derived independently from the i-vectors used for training the PLDA models. Specifically, μ_{cln} and σ_{cln} were derived from the original i-vectors, μ_{15dB} and σ_{15dB} were derived from both the original and 15dB i-vectors, and μ_{6dB} and σ_{6dB} were derived from the original, 15dB and 6dB i-vectors. The reason for this arrangement is to make sure that the scores produced by the three PLDA models in the SNR-dependent system have the same ranges. Besides, μ_{multi} and σ_{multi} were derived by pooling the original, 15dB and 6dB i-vectors together.

6 Results and Discussions

6.1 Performance Analysis of SNR-Dependent Systems

Table 2 shows the EER and minimum DCF ($\min C_{\text{Primary}}$ in NIST 2012 SRE [28]) achieved by SNR-independent (pooled) PLDA, and the hard- and soft-decision SNR-dependent PLDA in CC2, CC4 and CC5 for both male and female in NIST 2012 SRE. We consider the SNR-independent PLDA as the baseline. Table 3 shows the performance of the same systems in CC1 and CC3. The term ‘‘mix’’ in the tables denotes that the PLDA model was trained by both telephone and microphone data, and the terms ‘‘mic’’ denote that the PLDA models were trained by microphone data only. In the following, we refer the PLDA model trained by microphone data only to as mic PLDA. Likewise, we refer the model trained by both microphone and telephone data to as mix PLDA.

Model	I-Vectors for Training PLDA	CC2		CC4		CC5	
		EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
PLDA(baseline)	6dB(mix)+15dB(mix)+cln(mix)	2.72	0.283	3.64	0.321	3.51	0.317
Clean PLDA	cln(mix)	2.40	0.325	3.93	0.370	2.80	0.303
Hard-decision SNR-dependent PLDA	6dB(mix)+15dB(mix)+cln(mix) $\eta_1 = 3, \eta_2 = 20$	2.40	0.326	3.60	0.381	2.78	0.303
Soft-decision SNR-dependent PLDA	6dB(mix)+15dB(mix)+cln(mix)	2.40	0.325	3.95	0.372	2.79	0.303

(a) Male

Model	I-Vectors for Training PLDA	CC2		CC4		CC5	
		EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
PLDA(baseline)	6dB(mix)+15dB(mix)+cln(mix)	2.12	0.331	3.01	0.357	2.55	0.350
Clean PLDA	cln(mix)	2.10	0.326	3.76	0.424	2.63	0.351
Hard-decision SNR-dependent PLDA	6dB(mix)+15dB(mix)+cln(mix) $\eta_1 = 3, \eta_2 = 15$	2.10	0.326	2.86	0.362	2.63	0.351
Soft-decision SNR-dependent PLDA	6dB(mix)+15dB(mix)+cln(mix)	2.10	0.326	3.34	0.378	2.63	0.351

(b) Female

Table 2: Performance of SNR-independent, clean PLDA, and SNR-dependent PLDA (hard-decision and soft-decision) for (a) male and (b) female speakers in CC2, CC4 and CC5 of NIST 2012 SRE (core set). η_1 and η_2 are the decision thresholds in Eq. 7. ‘‘Mix’’ denotes that the PLDA model was trained by both telephone and microphone data. Boldface indicates the best performance for each common condition.

Model	I-Vectors for Training PLDA	CC1		CC3	
		EER(%)	minDCF	EER(%)	minDCF
PLDA(baseline)	15dB(mic)+cln(mic)	5.54	0.404	5.10	0.276
Clean PLDA	cln(mic)	5.42	0.371	5.62	0.276
Hard-decision SNR-dependent PLDA	15dB(mic)+cln(mic) $\eta_1 = -\infty, \eta_2 = 0$	5.13	0.370	5.83	0.311
Soft-decision SNR-dependent PLDA	15dB(mic)+cln(mic)	5.56	0.388	5.52	0.249

(a) Male

Model	I-Vectors for Training PLDA	CC1		CC3	
		EER(%)	minDCF	EER(%)	minDCF
PLDA(baseline)	15dB(mic)+cln(mic)	7.99	0.539	6.19	0.496
Clean PLDA	cln(mic)	7.98	0.491	6.71	0.449
Hard-decision SNR-dependent PLDA	15dB(mic)+cln(mic) $\eta_1 = -\infty, \eta_2 = 0$	7.56	0.542	6.15	0.572
Soft-decision SNR-dependent PLDA	15dB(mic)+cln(mic)	7.53	0.530	6.02	0.543

(b) Female

Table 3: Performance of SNR-independent PLDA, clean PLDA and hard- and soft-decision SNR-dependent PLDA for (a) male and (b) female speakers in CC1 and CC3 of NIST 2012 SRE (core set). η_1 and η_2 are the decision thresholds in Eq. 7. ‘‘Mic’’ denotes that the PLDA model was trained by microphone data only. Boldface indicates the best performance for each common condition.

6.1.1 Hard-Decision SNR-Dependent Systems

Different values of thresholds (η_1 and η_2 in Eq. 7) have been tried in the experiments and the best combination is reported in Table 2. It shows that hard-decision SNR-dependent PLDA with appropriate thresholds generally outperforms SNR-independent PLDA (baseline) especially in terms of EER.

Table 3 shows the performance of SNR-dependent PLDA in CC1 and CC3. As mentioned in Section 5, the SNR distribution of clean mic, 15dB mic and 6dB mic training data overlap with each other (see Fig. 5, left panels), so only clean mic and 15dB mic were used for training model in the SNR-dependent PLDA for CC1 and CC3. Therefore, we set η_1 to $-\infty$. Similar to CC2, CC4 and CC5, hard-decision SNR-dependent PLDA performs better than SNR-independent PLDA (baseline).

6.1.2 Soft-Decision SNR-Dependent System

Fig. 8 and Fig. 9 show the distribution and posterior probability of SNR of tel and mic training utterances for male speakers. Based on the posterior probability distribution of SNR of training utterances, the posterior probabilities of SNR of test utterances can be derived (Eq. 8). In particular, the posterior

probabilities obtained in Fig. 8 were used for CC2, CC4 and CC5, and those obtained from Fig. 9 were used for CC1 and CC3. Then, based on Eq. 9, the scores of soft-decision SNR-dependent PLDA can be determined.

According to Table 2, the performance of soft-decision SNR-dependent PLDA, hard-decision SNR-dependent PLDA, and clean PLDA (trained by clean mic and tel) under CC2 and CC5 are comparable. This is mainly because both soft- and hard-decision SNR-dependent PLDA under CC2 and CC5 heavily dependent on the clean PLDA model (Fig. 6, Eq. 8, Eq. 9 and Fig. 8, lower panel).

On the other hand, in CC4, the performance of soft-decision SNR-dependent PLDA is poorer than that of the hard-decision counterpart. This is mainly because the soft-decision PLDA relies on the SNR posterior distributions to determine the weights for combining the scores from the three PLDA models. According to Fig. 8 (lower panel), the posterior probabilities (i.e., combination weights) for the 6dB model and 15dB model crossover at around 7dB. Moreover, according to Fig. 6, a fairly large number of utterances in CC4 have SNR below this crossover point, which means that the scores from the 6dB model are heavily weighted for some of the noisy utterances. While the whole idea of SNR-dependent PLDA is to maximize the match between test utterances' SNR and training utterances' SNR, we conjecture that if a test utterance is not too noisy, it is more appropriate to use the 15dB PLDA model rather than the 6dB one. Unfortunately, for the soft-decision PLDA, we have no control on the crossover point, and thus the posterior probabilities. However, for hard-decision PLDA, we have full control on the decision threshold η_1 , which allows us to find a better threshold such that only very noisy utterances will use the 6dB model.

From Table 3, similar performance of soft- and hard-decision SNR-dependent PLDA are observed in CC1 and CC3 for female speakers. However, in CC1 and CC3 for male speakers, soft-decision SNR-dependent PLDA performs poorer than the hard-decision counterpart.

6.2 Performance Analysis of Fusion Systems

Fig. 10 shows the performance of linear fusion and logistic regression fusion. For the former, as described in Eq. 10, only two systems can be fused and the fusion weight w requires adjustment based on the test scores. For the logistic regression fusion, as described in Section 4.2 and Eq. 12, multiple systems can be combined and the fusion parameters $\{\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_N\}$ were derived from the training scores. The linear fusion weight w in Eq. 7 was set to 0.5. For CC2, CC4 and CC5, the decision thresholds used in the hard-decision SNR-dependent system were set to $\eta_1 = 3$ and $\eta_2 = 20$. For CC3 (male), $\eta_1 = -\infty$ and $\eta_2 = 10$. For CC1 (both gender) and CC3 (female), $\eta_1 = -\infty$ and $\eta_2 = 0$. The decision thresholds were empirically chosen according to the histogram of SNR distribution shown in Fig. 9.

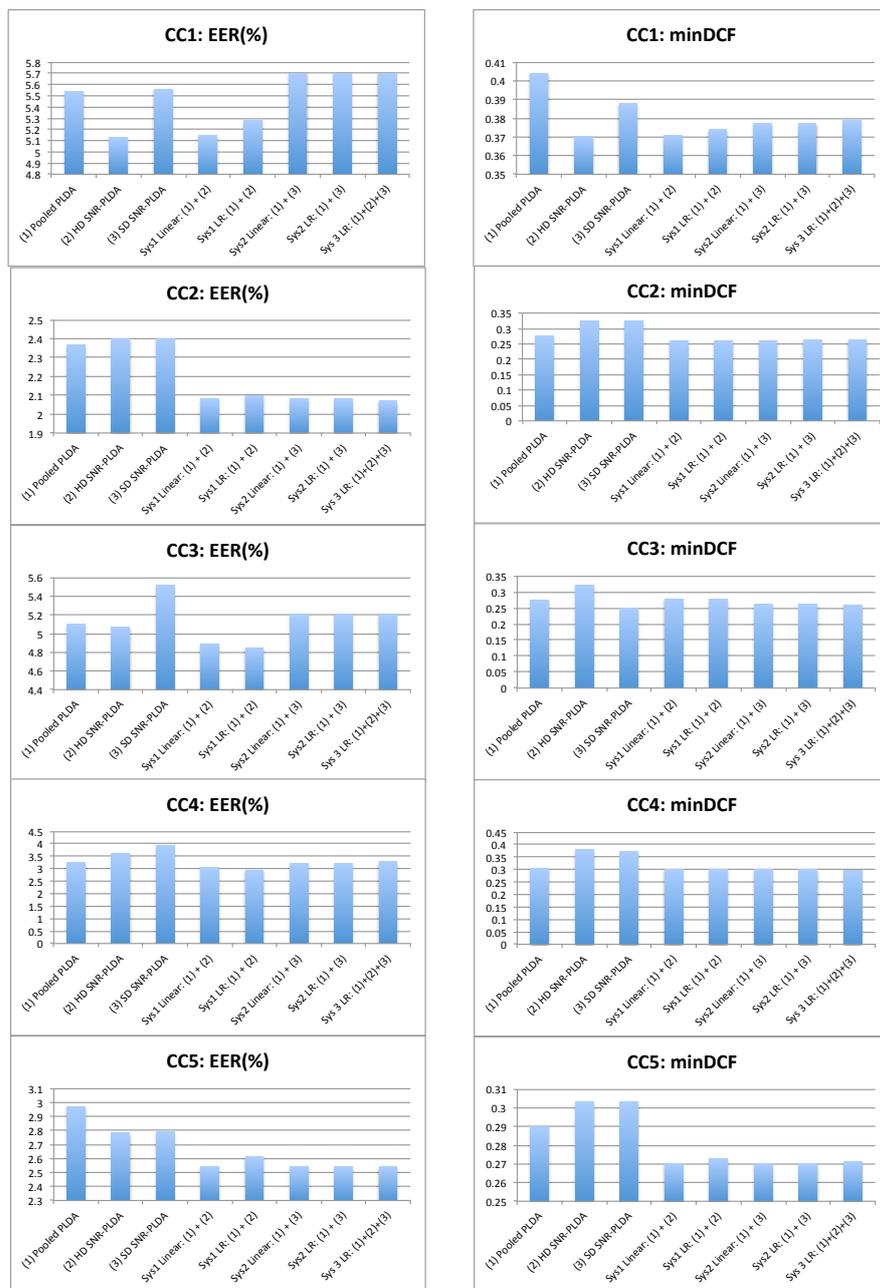


Fig. 10: Performance of (1) SNR-independent (pooled) PLDA, (2) hard-decision SNR-dependent PLDA, (3) soft-decision SNR-dependent PLDA and fusion systems in NIST 2012 SRE (core set) for male speakers (results for female speakers have similar patterns). *Linear* and *LR* denote linear fusion and logistic regression fusion described in Section 4, respectively. *Sys1*: Fusion of SNR-independent and hard-decision SNR-dependent systems. *Sys2*: Fusion of SNR-independent and soft-decision SNR-dependent systems. *Sys3*: Fusion of SNR-independent, hard- and soft-decision SNR-dependent systems.

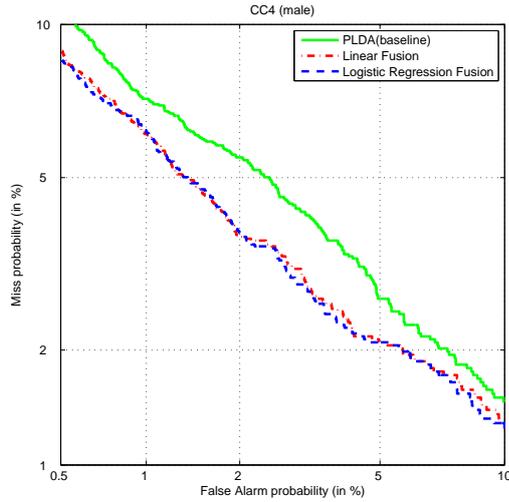


Fig. 11: DET performance of fusing SNR-independent PLDA and hard-decision SNR-dependent PLDA. The decision thresholds η_1 and η_2 in Eq. 7 were set to 3 and 20, respectively. The linear fusion weight w in Eq. 7 was set to 0.5.

Fig. 10 shows that the fusion systems perform significantly better than both the SNR-independent PLDA and the SNR-dependent PLDA. The following five subsections analyze the performance of the fusion systems with respect to fusion methods, fusion weights, decision strategies, decision thresholds and Z-norm parameters.

6.2.1 Performance with Respect to Fusion Methods

Fig. 11 shows the [detection error tradeoff](#) (DET) curves [22] for male speakers in CC4 using both linear and logistic regression fusions. The SNR-dependent PLDA subsystems in these two fusion systems are based on hard-decision. Fig. 10 and Fig. 11 suggest that the performance of logistic regression fusion is similar to that of linear fusion. However, instead of using the test data to determine the optimal fusion weights, fusion weight for logistic regression fusion were determined from development data. Therefore, logistic regression fusion is more practical.

6.2.2 Performance with Respect to Fusion Weights

Table 4 lists the performance of the fusion systems with different fusion weights under CC4. It can be observed that a large fusion weight tends to achieve better performance. On the other hand, logistic regression fusion can achieve a comparable performance without using this kind exhausted search for the fusion weight.

Fusion Method	w	Male (CC4)					
		EER(%)			minDCF		
		Sys 1	Sys 2	Sys 3	Sys 1	Sys 2	Sys 3
Linear	0.3	3.27	3.42	–	0.321	0.313	–
	0.4	3.15	3.32	–	0.309	0.303	–
	0.5	3.06	3.20	–	0.300	0.299	–
	0.6	2.95	3.05	–	0.299	0.296	–
	0.7	2.91	2.99	–	0.295	0.294	–
Logistic Regression	–	2.94	3.20	3.28	0.299	0.299	0.297

Table 4: Performance of the fusion systems with different fusion weights w under CC4 in NIST 2012 SRE (core set, male). w is the weight in Eq. 10. The decision thresholds η_1 and η_2 in Eq. 7 were set to 3 and 20, respectively. Boldface indicates the best performance.

6.2.3 Performance with Respect to Decision Strategies

Fig. 12 shows the DET curves for male speakers in CC4 using both hard- and soft-decision. The thresholds were set to $\eta_1 = 3$ and $\eta_2 = 20$ for hard-decision SNR-dependent PLDA, and the fusion weight was set to $w = 0.7$. The results in this figure are consistent with those in Table 2(a). Evidently, fusion systems outperform both two subsystems. In spite of the performance difference between the hard- and soft-decision SNR-dependent PLDA, $Sys1$ and $Sys2$ have similar performance.

6.2.4 Performance with Respect to Decision Thresholds

As shown in Tables 3 and 2, the performance of hard-decision SNR-dependent PLDA is affected by the selection of the thresholds (η_1 and η_2). This subsection is to investigate the effect of different thresholds on the fusion of SNR-independent PLDA and hard-decision SNR-dependent PLDA. The results on CC4 are shown in Table 5. It can be observed that the performance is comparable across different values of η_1 and η_2 . In Table 2, when $\eta_1 = 5$ and $\eta_2 = 25$, the hard-decision SNR-dependent PLDA in CC4 performs poorly. However, as shown in Table 5, the fusion system using $\eta_1 = 5$ and $\eta_2 = 25$ has similar performance as compared to the fusion systems using other thresholds. This suggests that the fusion operation makes the hard-decision PLDA system less sensitive to η_1 and η_2 .

6.3 Sensitivity Analysis of Z-norm Parameters

One important factor that affects the performance of the SNR-dependent systems and the fusion systems is the Z-norm parameters. An experiment was performed to investigate the sensitivity of system performance with respect to the Z-norm parameters. In the experiment, the Z-norm parameters (μ_{clin} , σ_{clin}) and ($\mu_{15\text{dB}}$, $\sigma_{15\text{dB}}$) in Fig. 1 were first obtained from the scores of test

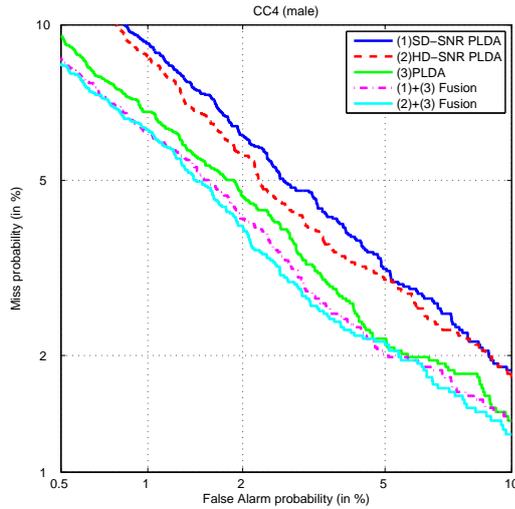


Fig. 12: DET performance of SNR-independent PLDA, hard- and soft-decision SNR-dependent PLDA, fusing SNR-independent PLDA and hard- or soft-decision SNR-dependent PLDA. The decision thresholds η_1 and η_2 in Eq. 7 were set to 3 and 20, respectively. The linear fusion weight w in Eq. 7 was set to 0.7.

	η_1	η_2	Male		Female	
			EER(%)	minDCF	EER(%)	minDCF
SNR-independent PLDA + Hard-decision SNR-dependent PLDA	3	15	2.92	0.293	2.75	0.306
	3	20	2.91	0.295	2.75	0.306
	3	25	2.92	0.287	2.75	0.306
	5	25	2.85	0.291	2.54	0.296

Table 5: Performance of fusing SNR-independent PLDA and SNR-dependent PLDA with different decision thresholds (η_1 and η_2 in Eq. 7) in CC4 of NIST 2012 SRE (core set). The fusion weight w in Eq. 10 was set to 0.7.

utterance. Then, the values of μ_{cIn} and $\mu_{15\text{dB}}$ were perturbed by $\pm 0.1\sigma_{\text{cIn}}$ and $\pm 0.1\sigma_{15\text{dB}}$, respectively. Fig. 13 shows that the performance of SNR-dependent PLDA is still better than that of SNR-independent PLDA even if the Z-norm parameters μ_{cIn} is perturbed $0.1\sigma_{\text{cIn}}$. This suggests that the fusion systems are fairly robust with respect to the deviation of the Z-norm parameters. Similar results were also obtained by perturbing $\mu_{15\text{dB}}$.

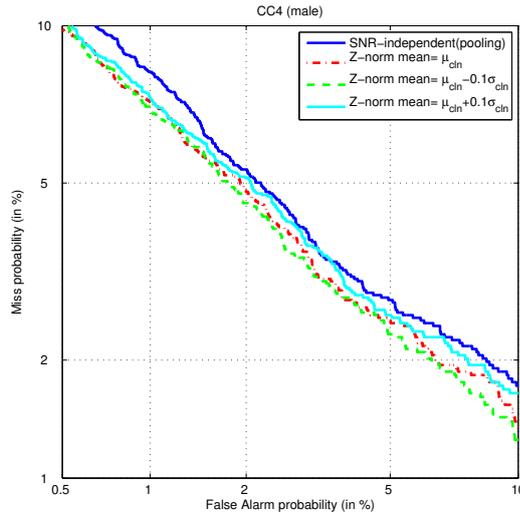


Fig. 13: DET curves of hard-decision SNR-dependent systems for male speakers with Z-norm parameters (mean) deviated from the mean obtained from test data. ($\mu_{cIn} = -86.5$, $\sigma_{cIn} = 64.8$)

7 Conclusions

In this paper, fusion of SNR-dependent PLDA models was presented. Both SNR-dependent and fusion of SNR-dependent models were evaluated on the core set of NIST 2012 SRE. Performance of the SNR-dependent PLDA model depends on the decision strategies, the decision thresholds, Z-norm parameters and the degree of mismatch between the SNR of test utterances and the SNR-dependent PLDA models. The SNR-dependent PLDA outperforms the baseline in 9 out of 10 conditions (CC1–CC5 for both male and female) especially in terms of EER.

Fusion of SNR-dependent and SNR-independent PLDA models can bring benefit regardless of the decision strategies, decision thresholds and Z-norm parameters. The fusion operation makes the hard-decision PLDA system less sensitive to η_1 and η_2 and fusion systems are fairly robust with respect to the deviation of the Z-norm parameters. Besides, while logistic regression fusion achieves a comparable performance with linear fusion, it does not require using the brute-force search for the fusion weight. Fusion of SNR-independent PLDA and soft-decision SNR-dependent PLDA with logistic regression, which is the most favoured since it does not need any prior information about the test utterances, brings benefits in 8 out of 10 conditions.

Acknowledgements This work was in part supported by The Hong Kong Research Grant Council (Grant No. PolyU 152117/14E and PolyU 152068/15E) and The Hong Kong Polytechnic University (Grant No. 4-ZZCX).

References

1. Bishop C (2006) Pattern recognition and machine learning. springer, New York
2. Brümmer N (2014) FoCal. <https://sites.google.com/site/nikobrummer/focal>
3. Brümmer N, de Villiers E (2011) The bosaris toolkit user guide: Theory, algorithms and code for binary classifier score processing. Documentation of BOSARIS toolkit
4. 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(4):788–798
5. Ferrer L, Bratt H, Burget L, Cernocky H, Glembek O, Graciarena M, Lawson A, Lei Y, Matejka P, Plchot O, et al (2011) Promoting robustness for speaker modeling in the community: The PRISM evaluation set. In: *Proc. of NIST 2011 Workshop*
6. Garcia-Romero D, Espy-Wilson C (2011) Analysis of i-vector length normalization in speaker recognition systems. In: *Proc. Interspeech*, pp 249–252
7. 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: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 4257–4260
8. Hasan T, Hansen J (2013) Acoustic factor analysis for robust speaker verification. *IEEE Trans on Audio, Speech, and Language Processing* 21(4):842–853
9. Hasan T, Hansen J (2014) Maximum likelihood acoustic factor analysis models for robust speaker verification in noise. *IEEE Trans on Audio, Speech, And Language Processing* 22(2):381–391
10. Hasan T, Sadjadi SO, Liu G, Shokouhi N, Boril H, Hansen JHL (2013) CRSS system for 2012 NIST speaker recognition evaluation. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 6783–6787
11. Hatch A, Kajarekar S, Stolcke A (2006) Within-class covariance normalization for SVM-based speaker recognition. In: *Proc. of the 9th International Conference on Spoken Language Processing, Pittsburgh, PA, USA*, pp 1471–1474
12. Kenny P (2010) Bayesian speaker verification with heavy-tailed priors. In: *Proc. of Odyssey 2010: Speaker and Language Recognition Workshop, Brno, Czech Republic*
13. 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(5):980–988

14. Kenny P, Boulianne G, Ouellet P, Dumouchel P (May 2007) Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Trans on Audio, Speech and Language Processing* 15(4):1435–1447
15. Leeuwen DA, Saeidi R (2013) Knowing the non-target speakers: The effect of the i-vector population for PLDA training in speaker recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, pp 6778 – 6782
16. Lei Y, Burget L, Ferrer L, Graciarena M, Scheffer N (2012) Towards noise-robust speaker recognition using probabilistic linear discriminant analysis. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, pp 4253 – 4256
17. Lei Y, Burget L, Scheffer N (2013) A noise robust i-vector extractor using vector Taylor series for speaker recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 6788–6791
18. Lei Y, McLaren M, Ferrer L, Scheffer N (2014) Simplified VTS-based i-vector extraction in noise-robust speaker recognition. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 4065–4069
19. Li Q, Huang Y (2010) Robust speaker identification using an auditory-based feature. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 4514–4517
20. Mak MW, Yu HB (2013) A study of voice activity detection techniques for NIST speaker recognition evaluations. *Computer, Speech and Language* 28(1):295–313
21. Mallidi S, Ganapathy S, Hermansky H (2013) Robust speaker recognition using spectro-temporal autoregressive models. In: *Proc. Interspeech*
22. Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M (1997) The DET curve in assessment of detection task performance. In: *Proc. Eurospeech'97*, pp 1895–1898
23. Martinez D, Burget L, Stafylakis T, Lei Y, Kenny P, Lleida E (2014) Unscented transform for i-vector-based noisy speaker recognition. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 4070–4074
24. McLaren M, Mandasari M, Leeuwen D (2012) Source normalization for language-independent speaker recognition using i-vectors. In: *Proc. Odyssey 2012: The Speaker and Language Recognition Workshop*, pp 55–61
25. McLaren M, Scheffer N, Graciarena M, Ferrer L, Lei Y (2013) Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 6773–6777
26. Ming J, Hazen T, Glass J, Reynolds D (July 2007) Robust speaker recognition in noisy conditions. *IEEE Trans on Audio, Speech and Language Processing* 15(5):1711–1723

27. Neto SFDC (1999) The itu-t software tool library. *International journal of speech technology* 2(4):259–272
28. NIST (2012) The NIST year 2012 speaker recognition evaluation plan. <http://www.nist.gov/itl/iad/mig/sre12cfm>
29. Pang XM, Mak MW (2014) Fusion of SNR-dependent PLDA models for noise robust speaker verification. In: *ISCSLP'2014*, pp 619–623
30. Pelecanos J, Sridharan S (2001) Feature warping for robust speaker verification. In: *Proc. Odyssey 2001: The Speaker and Language Recognition Workshop*, Crete, Greece, pp 213–218
31. Prince S, Elder J (2007) Probabilistic linear discriminant analysis for inferences about identity. In: *IEEE 11th International Conference on Computer Vision, 2007 (ICCV 2007)*., pp 1–8
32. Rajan P, Kinnunen T, Hautamäki V (2013) Effect of multicondition training on i-vector PLDA configurations for speaker recognition. In: *Proc. Interspeech*, pp 3694–3697
33. Rajan P, Afanasyev A, Hautamki V, Kinnunen T (2014) From single to multiple enrollment i-vectors: Practical plda scoring variants for speaker verification. *Digital Signal Process* p Online version: <http://dx.doi.org/10.1016/j.dsp.2014.05.001>
34. Rao W, Mak MW (2013) Boosting the performance of i-vector based speaker verification via utterance partitioning. *IEEE Trans on Audio, Speech and Language Processing* 21(5):1012–1022
35. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10(1–3):19–41
36. Sadjadi S, Pelecanos J, Zhu W (2014) Nearest neighbor discriminant analysis for robust speaker recognition. In: *Proc. Interspeech*, pp 1860–1864
37. Sadjadi SO, Hasan T, Hansen J (2012) Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition. In: *Proc. Interspeech*, pp 1696–1699
38. Saeidi R, van Leeuwen DA (2012) The Radboud University Nijmegen submission to NIST SRE-2012. In: *Proc. of the NIST Speaker Recognition Evaluation Workshop*
39. Shao Y, Wang D (2008) Robust speaker identification using auditory features and computational auditory scene analysis. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 1589–1592
40. Yu C, Liu G, Hahm S, Hansen J (2014) Uncertainty propagation in front end factor analysis for noise robust speaker recognition. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 4045–4049
41. Yu H, Mak M (2011) Comparison of voice activity detectors for interview speech in nist speaker recognition evaluation. In: *Proc. Interspeech*, pp 2353–2356