Fusion of SNR-Dependent PLDA Models for Noise Robust Speaker Verification

Xiaomin Pang and Man-Wai Mak

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

Abstract

The i-vector representation and probabilistic linear discriminant analysis (PLDA) have shown state-of-the-art performance in many speaker verification systems. However, in real-world environments, additive and convolutive noise cause mismatches between training and recognition conditions, degrading the performance. In this paper, a fusion system that combines a multi-condition PLDA model and a mixture of SNR-dependent PLDA models is proposed to make the verification system noise robust. The SNR of test utterances is used to determine the best SNR-dependent PLDA model to score against the target-speaker's i-vectors. The performance of the fusion system is demonstrated on NIST 2012 SRE. Results show that the SNR-dependent PLDA models can reduce EER and that the fusion system is more robust than the conventional i-vector/PLDA systems under noisy conditions. It is also found that the SNR-dependent PLDA models are insensitive to Z-norm parameters.
Index Terms: Speaker verification; i-vectors; probabilistic LDA; NIST 2012 SRE; noise robustness.

1. Introduction

When a speaker verification system is applied in real world environments, the system performance can be degraded by the presence of environmental noise. A lot of research has been conducted to compensate for the effect of environmental noise. Among them, much effort has been put into the front-end processing stage, in which noise robust features are extracted [1, 2, 3], features are transformed [4] so that they become more resilience to noise, and noisy speech signals are enhanced [5] to mitigate noise degradation. Other researches focused on the backend classification stage and found that this kind of techniques is more promising, especially when joint factor analysis (JFA) [6] and i-vector/PLDA framework [7, 8] are employed.

In the i-vector approach, a single low-dimension vector called i-vector is defined to represent the acoustic characteristics (including both speaker and channel) of an utterance. The low-dimensionality of i-vectors facilitates the usage of classical statistical techniques such as linear discriminant analysis (LDA) [9], within-class covariance normalization (WCCN) [10] and PLDA [11] to suppress the channel-variability [7, 12, 13]. PLDA performs factor analysis on the i-vector space by grouping the i-vectors derived from the same speakers 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 example, [14] 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 UBM from the new features. In [15], 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. [16] adapted a vector Taylor series approach that integrates additive and convolutive noises into the i-vector extraction framework. In [17, 18, 19, 20], multi-condition training, in which a PLDA model is trained by pooling clean and noisy utterances together, was employed to enhance noise robustness. [21] trained a collection of individual PLDA models for each 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 [21] 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 [21]), the SNR-dependent PLDA models in this paper compute the verification scores by choosing one of the SNR-dependent PLDA models based on 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.

The paper is organized as follows. Section 2 describes the SNR-dependent PLDA models and the fusion system. In Sections 3 and 4, we report evaluations based on NIST 2012 SRE [22]. Section 5 concludes the findings.

2. Fusion of SNR-Dependent Systems

2.1. SNR-Dependent Systems

The basic PLDA model considers the i-vector $\mathbf{x}$ as an observed variable following the generative model [23]:

$$\mathbf{x} = \mathbf{m} + \mathbf{V}\mathbf{z} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{m}$ is the global offset; the columns of $\mathbf{V}$ define the basis of the speaker subspace; $\mathbf{z}$ is the latent variable and $\boldsymbol{\epsilon}$ is the residual noise assumed to follow a Gaussian distribution with zero mean and full covariance $\boldsymbol{\Sigma}$. An EM algorithm [11] is applied to estimate the parameters of the factor analyzer.

Classical Gaussian PLDA assumes that $\mathbf{x}$ 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 [11] using training data contaminated with different level of background noise.

In this paper, we evaluate the idea of SNR-dependent PLDA via common conditions 4 and 5 of NIST 2012 SRE. Because

978-1-4799-4219-0/14/$31.00

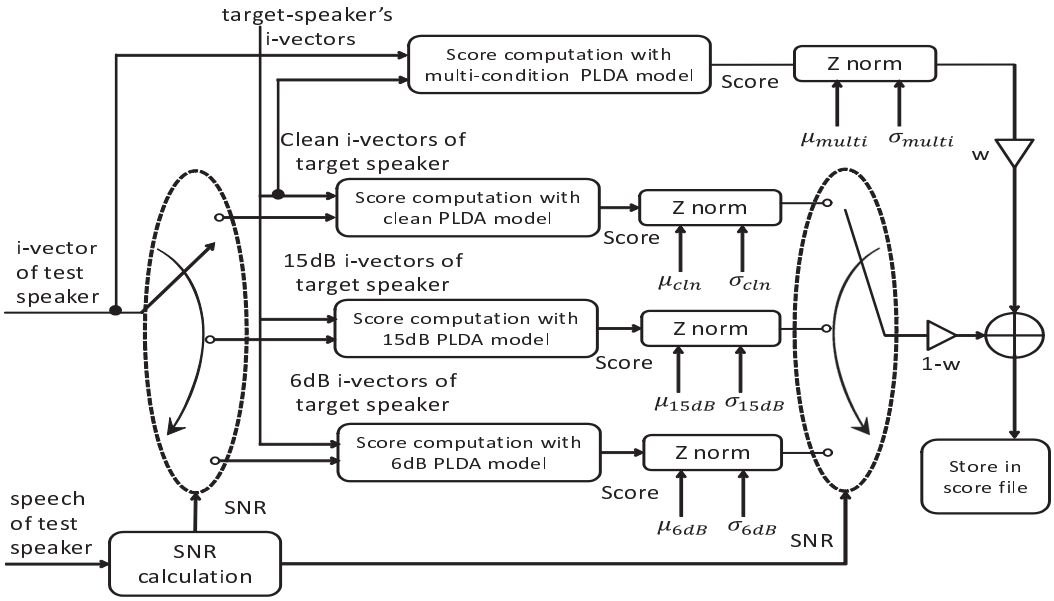

Figure 1: The dataflow of SNR-dependent PLDA scoring and linear score fusion.

noise at 6dB and 15dB have been added to some of the sound files in the test segments of the SRE, we added noise to the training files at the same level to create 3 SNR-dependent PLDA models: 6dB, 15dB, and clean (using the original sound files). 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:

$$\text{If} \begin{cases} \ell_t \le \eta_1, & \text{use 6dB PLDA and target's i-vectors} \\ \eta_1 < \ell_t \le \eta_2, & \text{use 15dB PLDA and target's i-vectors} \\ \ell_t > \eta_2, & \text{use clean PLDA and target's i-vectors} \end{cases} \tag{2}$$

where ℓ_t is the SNR of the test utterance, and η_1 and η_2 are decision thresholds. Fig. 1 shows the dataflow of SNR-dependent PLDA scoring.

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 training files contaminated with different level of background noise.

2.2. Fusion of SNR-Dependent System

The fusion system combines the SNR-dependent system and the SNR-independent system. In Fig. 1, 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 observed from the figure that in the SNR-dependent system the 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. The fusion system linearly combines the SNR-independent system and the SNR-dependent system:

$$s = ws_i + (1-w)s_d \tag{3}$$

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

As described earlier, the Z-norm parameters represented by μ and σ in Fig. 1 are 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.

Besides the linear fusion described in Fig. 1 and Eq. 3, logistic regression fusion [9, 24] also can be employed:

$$s = \alpha_0 + \alpha_1 s_i + \alpha_2 s_d \tag{4}$$

where α_0 is an offset, and α_1 and α_2 are the fusion weights for s_i and s_d, respectively. The only difference between logistic regression fusion and linear fusion is that fusion parameters α_0, α_1 and α_2 in the former are derived from development data.

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, PLDA models and Z-norm parameters.

Speech regions in the speech files were extracted by using a two-channel VAD [25]. 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 [26] and added the noise waveform to the file at an SNR of 6dB and 15dB using the FaNT tool [27]. Because the SNR-dependent PLDA requires the SNR of test utterances, we used the speech voltmeter

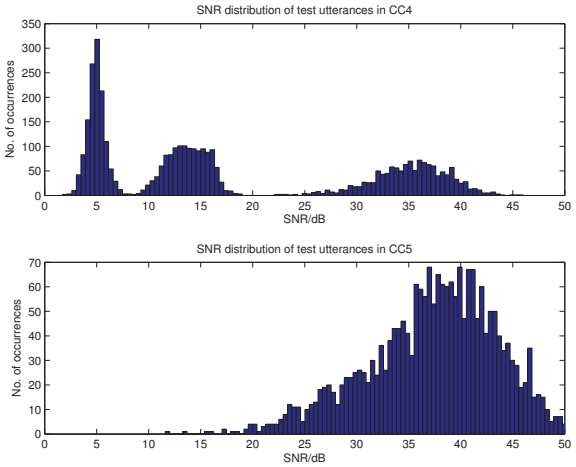

Figure 2: SNR distributions of test utterances in CC4 and CC5 of NIST 2012 SRE.

function in FaNT and the VAD decisions to estimate the SNR of the test files.

3.2. I-Vector Extraction and PLDA Models

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 [13], within-class covariance normalization (WCCN) [10] and i-vector length normalization [23] were applied to the 500-dimensional i-vectors. Because noise at 6dB and 15dB has been added to some of the sound files in the test segments of the SRE, we added noise to the training files at the same level to create 3 SNR-dependent PLDA models: 6dB, 15dB, and clean (using the original sound files). PLDA and SNR-dependent PLDA models with 150 latent variables were trained using the clean and noise contaminated i-vectors.

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. An SNR-independent PLDA model with 150 factors was then trained. For the SNR-dependent PLDA, the 6dB, 15dB, and original speech files were independently used to train three PLDA models, each with 150 factors.

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 original (clean) speech files using the conventional PLDA scoring function [23]. For SNR-dependent PLDA models, as Eq. 2 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.

3.3. Fusion of SNR-Dependent PLDA models

As mentioned in Section 2.2, both a predefined fusion weight w in the linear fusion and the fusion parameters (α_0, α_1 and α_2) in the logistic regression fusion can be used for score fusion. For the latter, the fusion parameters in Eq. 4 were derived from the PLDA scores obtained from the original, 15dB and 6dB i-vectors used for training the PLDA models.

Method	η_1	η_2	CC4		CC5	
			EER(%)	minDCF	EER(%)	minDCF
PLDA	–	–	3.42	0.33	3.30	0.32
SNR-PLDA	3	15	3.57	0.47	2.93	0.29
	3	20	3.34	0.46	3.00	0.29
	3	25	3.33	0.46	3.03	0.29
	5	25	4.14	0.52	3.03	0.29

Table 1: Performance of PLDA (pooling) and SNR-dependent PLDA for male speakers in CC4 and CC5 of NIST 2012 SRE (core set). η_1 and η_2 are the decision threshold in Eq. 2.

Fusion Method	w	CC4		CC5	
		EER(%)	minDCF	EER(%)	minDCF
Linear Fusion	0.3	2.95	0.39	2.87	0.28
	0.4	2.91	0.37	2.94	0.28
	0.5	2.94	0.34	2.96	0.28
	0.6	2.99	0.33	3.00	0.29
	0.7	3.03	0.32	3.06	0.29
Logistic Regression	–	3.06	0.32	2.93	0.29

Table 2: Performance of the fusion system in CC4 and CC5 of NIST 2012 SRE (core set). w is the weight in Eq. 3. Fusion parameters α_0, α_1 and α_2 in Eq. 4 are -16.99, 7.01 and 1.24 for CC4 and -12.89, 2.61 and 3.30 for CC5. The decision thresholds η_1 and η_2 in Eq. 2 were set to 3 and 20, respectively.

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

4. Results and Discussions

4.1. Performance Analysis of the SNR-Dependent System

Fig. 2 shows the SNR distributions of test utterances in CC4 and CC5 in 2012 SRE. Based on the distributions, the decision thresholds for SNR-dependent PLDA were set. Table 1 shows the EER and minimum DCF (minC_{Primary}) achieved by multi-condition PLDA (baseline),[1] and the SNR-dependent PLDA in CC4 and CC5 of NIST 2012 SRE with different thresholds. From the table, it can be observed that in CC4, when fixing $\eta_1 = 3$, which means the number of test utterances employing 6dB PLDA model is fixed, increasing η_2 can slightly improve performance. When fixing η_2, say at 25dB, in Table 1, increasing η_1 degrades the performance, which implies that the 6dB PLDA model is too noisy for CC4. When appropriate thresholds (say $\eta_1 = 3$ and $\eta_2 = 20$) are selected for CC4, SNR-dependent PLDA performs better than the multi-condition PLDA in terms of EER, but with the expense of minimum DCF. For CC5, when fixing $\eta_1 = 3$, increasing η_2 increases EER slightly and when fixing $\eta_2 = 25$, $\eta_1 = 3$ and $\eta_1 = 5$ give the

[1] We did not use the PLDA trained by clean utterances as the baseline because it has been shown [18, 28] that using clean utterances exclusively for training leads to poorer performance.

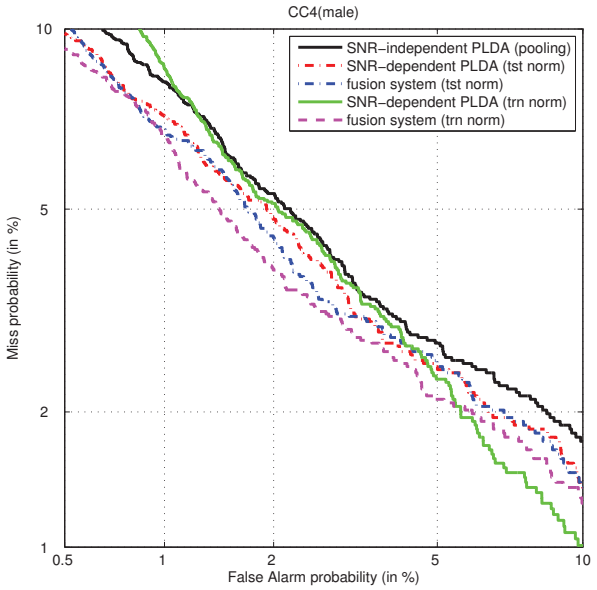

Figure 3: DET curves of the pooling, SNR-dependent and fused systems with different normalization parameters.

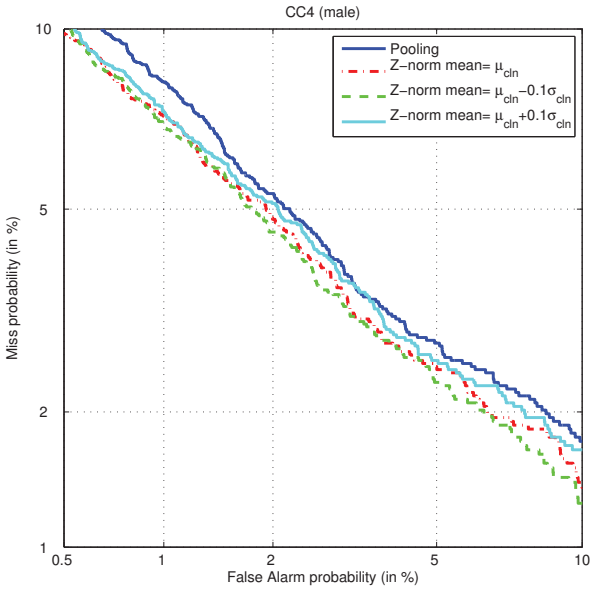

Figure 4: DET curves of SNR-dependent systems normalized by different parameters ($\mu_{cln} = -86.5, \sigma_{cln} = 64.8$)

same performance, which means that no test utterance in CC5 has SNR smaller than 5dB and this can also be observed from Fig. 2. The SNR-dependent PLDA improves performance for CC5 comparing with the baseline.

4.2. Performance Analysis of the Fusion System

Table 2 shows the performance of the fusion systems in CC4 and CC5 with different fusion weights. The decision thresholds used in the SNR-dependent system were set to $\eta_1 = 3$ and $\eta_2 = 20$. For CC5, $w = 0.3$ leads to the best performance and it is better than both the baseline and the SNR-dependent PLDA. As shown in Table 1, the SNR-dependent PLDA performs significantly better than the SNR-independent PLDA in CC5. As a result, a small fusion weight allows the SNR-dependent system to have greater contribution to the fusion system, resulting in better performance. This can be observed in CC5 of Table 2 where a small weight ($w = 0.3$) leads to the best performance. For CC4, a large fusion weight tends to achieve a lower minDCF but with a slightly increase in EER. However, when considering both EER, minDCF and DET performance, the fusion system performs better than either of SNR-dependent and SNR-independent systems, as evident in Fig. 3, Table 1, and Table 2. While the results show that logistic regression fusion is slightly inferior to the best linear fusion, it does not require using test data to determine the optimal fusion weights. Instead, its fusion weights were determined from development data.

Fig. 3 shows the DET curves for CC4 of the systems with different normalization parameters. The thresholds are $\eta_1 = 3$ and $\eta_2 = 20$ for the SNR-dependent PLDA, and $w = 0.7$ for the fusion systems. The normalization parameters in the figure are derived from test data (tst norm) and training data (trn norm). The results suggest that SNR-dependent systems perform better than the SNR-independent baseline (pooling) and that the fusion systems further improve the performance. Although the performance of the SNR-dependent system with Z-norm parameters obtained from training data (trn norm in the legend of Fig. 3) is very close to that of the baseline, combining the SNR-dependent system and the SNR-independent system (pooling) can improve the performance for a wide range of decision thresholds.

4.3. Sensitivity Analysis of Z-norm Parameters

One important factor that can affect the performance of the SNR-dependent system and the fusion system is the Z-norm parameters. It is difficult to guarantee that the Z-norm parameters derived from training data can make the three PLDA models to output scores that fall on the same range. Therefore, it is necessary to investigate the sensitivity of the system with respect to the Z-norm parameters. An experiment was performed to examine how the deviation of Z-norm parameters affects the performance of the systems. In this experiment, CC4 was used and the normalization parameters μ_{cln} and μ_{15dB} in Fig. 1 were perturbed around the respective Z-norm mean (μ_{cln} and μ_{15dB}) obtained from test data. Because the number of trials using 6dB PLDA model is too small, the effect of μ_{6dB} was not examined. Fig. 4 shows the DET curves obtained by perturbing μ_{cln}. When μ_{cln} is changed within 10% of σ_{cln}, the SNR-dependent system still performs better than the SNR-independent (pooling) system. Similar results were also obtained by perturbing μ_{15dB}.

5. 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 thresholds and the degree of match between the SNR of test utterances and the SNR-dependent PLDA models. By allowing each of the SNR-dependent PLDA model to focus on a small range of SNR, the proposed method successfully reduces the mismatch between the PLDA model, the target-speakers' i-vectors and the test i-vectors. Besides, the performance of the SNR-dependent PLDA model is affected by Z-norm parameters. Fusion of SNR-dependent and SNR-independent PLDA models can bring benefit even though the SNR-dependent models have similar performance as the baseline due to inaccurate Z-norm parameters.

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] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] "https://sites.google.com/site/nikobrummer/focal," .

[25] 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.

[26] 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," .

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

[28] 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.