



Accounting For Uncertainty of i-vectors in Speaker Recognition Using Uncertainty Propagation and Modified Imputation

Rahim Saeidi and Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

{rahim.saeidi, paavo.alku}@aalto.fi

Abstract

One of the biggest challenges in speaker recognition is incomplete observations in test phase caused by availability of only short duration utterances. The problem with short utterances is that speaker recognition needs to be handled by having information from only limited amount of acoustic classes. By considering limited observations from a test speaker, the resulting i-vector as a representative of short utterance will be uncertain; the shorter the duration, the higher the uncertainty. In recent studies, an uncertainty decoding technique has been employed in probabilistic linear discriminant analysis (PLDA) modeling in order to account for uncertain i-vectors. In this paper, we propose to extend uncertainty handling using simplified PLDA scoring and modified imputation. We experiment with a state-of-the-art speaker recognition system focusing on uncertainty caused by controlled utterance duration. The uncertainties after i-vector extraction are being propagated through pre-processing steps and both uncertainty decoding and modified imputation are considered. Our experimental results indicate improved equal error rate and detection cost attained by using uncertainty-of-observation techniques in dealing with short duration utterances.

Index Terms: speaker verification, duration, uncertainty decoding, modified imputation

1. Introduction

Automatic speaker verification, the task of accepting or rejecting an identity claim given an utterance of a speaker, has received lots of attention in the last 20 years [1]. One of the main reasons is the support of the National Institute of Standards and Technology (NIST) by organizing series of benchmarks, the speaker recognition evaluations (SREs) [2] starting in 1996. For each SRE, the task, the data and the evaluation metrics are supplied by NIST and after submission of recognition scores by participating sites, researchers share thoughts in a follow up workshop. Before NIST SRE'12, the task was solely defined as *speaker detection*, whereas in SRE'12, the performance metric and evaluation condition resembles an *open set* speaker recognition task [3, 4].

Most of the research in the speaker recognition area is devoted to find robust modeling techniques, capable of handling channel and inter-session variability [5–7]. The state-of-the-art method is now using a low-rank vector, the so-called *i-vector*, to represent an utterance based on *total variability* subspace modeling [7] and *probabilistic linear discriminant analysis* (PLDA) [8] to obtain a likelihood ratio in comparing an enrollment and test utterance. The acoustic feature distribution is captured by a *universal background model* (UBM) [9] and the subspace modeling techniques developed in the joint factor analysis approach [10] are utilized.

The uncertainty in i-vector extraction is a result of several factors, such as having noisy or very short speech signals. The effect of noise on acoustic features can be modeled by an uncertainty of short-time features which is then propagated to the following modeling stages [11–13]. The problem of *incomplete observations* in speaker recognition is posed as dealing with variable (short) utterance duration and several techniques has been proposed recently to compensate for this factor in the context of i-vector extraction [14–19]. The *i-vectors* extracted using a sufficient amount speech follow a standard normal distribution. However, as shown recently in [15, 16], this is not the case when a considerable amount of uncertainty exists in the i-vector extraction.

Considering utterance duration as the source of uncertainty in i-vectors, we assume that a posterior distribution of i-vectors is provided. The i-vector extractor can produce i-vectors as the posterior distribution mean along with a covariance structure indicating the uncertainty in the extraction process [16]. This uncertainty can be propagated through the post-processing steps and integrated into PLDA using the *uncertainty decoding* (UD) technique [16, 20, 21]. Missing data marginalization and imputation has been successfully employed in robust automatic speech recognition in dealing with uncertain spectrogram components [22, 23]. In this paper, we propose to use *modified imputation* (MI) [22] before PLDA model evaluation and compare the recognition performance under controlled utterance duration conditions with UD.

2. State-of-the-art system for Speaker Recognition

The speaker recognition system used in this paper follows the i-vector framework that was proposed in [7]. The i-vector is a compact representation of the speech utterance in a low-dimensional space. This space contains both speaker and channel/session variability so that a speaker- and session-dependent Gaussian mean supervector \mathbf{M} can be modeled as:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}$$

where \mathbf{m} is the speaker- and session-independent mean supervector of the UBM, \mathbf{T} is a low-rank matrix that defines the low-dimensional space, and \mathbf{w} is the identity vector or so-called i-vector. The speaker- and session-dependent mean supervector in the i-vector speech representation is very similar to the *joint factor analysis* (JFA) speaker representation [24]. The main difference between the i-vector and JFA modeling is that JFA defines separate speaker and session subspaces, while these factors of variability are combined in a single space \mathbf{T} in the i-vector representation.

Probabilistic linear discriminant analysis is a probabilistic approach that models the i-vector distribution with a Gaussian

assumption [8]. Computed scores from the PLDA model are directly in the form of a ratio of the likelihoods that the enrollment and test i-vectors come from the same speaker and different speakers, respectively. PLDA models the distribution of i-vectors as the sum of Gaussians for the speaker-dependent term, $\boldsymbol{\mu} + \Phi \mathbf{y}_k$ and an utterance dependent term ϵ_r with $r = 1, \dots, R$ utterances for a speaker k [8, 25]. The overall mean of the training vectors is denoted by $\boldsymbol{\mu}$ and Φ is composed of the bases for the between-speaker subspace. \mathbf{y}_k is positioning the i-vector in between-speaker subspace, and ϵ_r is a Gaussian residual error term with full covariance Σ . A PLDA model q can be regarded as

$$P(\mathbf{w}|q) \sim \mathcal{N}(\mathbf{w}; \boldsymbol{\mu} + \Phi \mathbf{y}, \Sigma),$$

$$P(\mathbf{y}) \sim \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathbf{I}).$$

In the context of PLDA model, the hypothesis testing becomes the evaluation of the probabilities if the two i-vectors \mathbf{w}_1 and \mathbf{w}_2 , traditionally named as template and test, are generated by the same speaker, H_1 , or by different speakers, H_2 . This can be formulated as:

$$s = \frac{P(\mathbf{w}_1, \mathbf{w}_2|H_1)}{P(\mathbf{w}_1, \mathbf{w}_2|H_2)},$$

$$P(\mathbf{w}_1, \mathbf{w}_2|H_1) = \int_{\mathbf{y}} \prod_{i=1,2} P(\mathbf{w}_i|q)P(\mathbf{y})d\mathbf{y},$$

$$P(\mathbf{w}_1, \mathbf{w}_2|H_2) = \prod_{i=1,2} \int_{\mathbf{y}} P(\mathbf{w}_i|q)P(\mathbf{y})d\mathbf{y}.$$

It is shown in [26] and [25] that the likelihoods can be computed analytically as:

$$s = \frac{\mathcal{N}(\mathbf{w}_{12}; \boldsymbol{\mu}_2, \Sigma_p)}{\mathcal{N}(\mathbf{w}_{12}; \boldsymbol{\mu}_2, \Sigma_d)}, \quad (1)$$

where \mathbf{w}_{12} is formed by stacking i-vectors \mathbf{w}_1 and \mathbf{w}_2 and $\boldsymbol{\mu}_2$ by stacking $\boldsymbol{\mu}$ twice, and the covariance matrices for the same and different speakers are obtained by using the matrix expressions:

$$\Sigma_p = \begin{bmatrix} \Phi\Phi^T + \Sigma & \Phi\Phi^T \\ \Phi\Phi^T & \Phi\Phi^T + \Sigma \end{bmatrix}$$

$$\Sigma_d = \begin{bmatrix} \Phi\Phi^T + \Sigma & \mathbf{0} \\ \mathbf{0} & \Phi\Phi^T + \Sigma \end{bmatrix}$$

In practice, because of centring stage in the pre-processing of i-vectors, the global mean of i-vectors is zero and the conventional PLDA scoring can be written as

$$s_{\text{PLDA}} = \frac{\mathcal{N}(\mathbf{w}_{12}; \mathbf{0}, \begin{bmatrix} \Phi\Phi^T + \Sigma & \Phi\Phi^T \\ \Phi\Phi^T & \Phi\Phi^T + \Sigma \end{bmatrix})}{\mathcal{N}(\mathbf{w}_{12}; \mathbf{0}, \begin{bmatrix} \Phi\Phi^T + \Sigma & \mathbf{0} \\ \mathbf{0} & \Phi\Phi^T + \Sigma \end{bmatrix})}$$

3. Uncertainty Handling

As shown in Figure 1, the uncertainty in an i-vector \mathbf{w} could be a result of several factors which collectively result in deviating from optimal representation \mathbf{w}^* . Uncertainty decoding technique [20] is the state-of-the-art in handling uncertainty

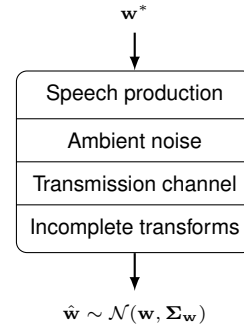


Figure 1: Assuming there exists an optimal \mathbf{w}^* to represent a speaker in low dimensional space, there are several factors leading to arrive at uncertain i-vector.

within PLDA framework and in this paper, we propose modified imputation [22] as an alternative approach. In the UD technique, the enrollment utterance is considered to be accurate (without uncertainty) and the uncertainty in test utterance $\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}, \Sigma_w)$, represented by Σ_w , is added to the noise term covariance in evaluating the PLDA model [20]. This means that we use Equation 1 and in calculating Σ_p and Σ_d we use $\Sigma + \Sigma_w$ for each test utterance. In this way, a fast scoring would not be possible because the noise term would change for every test utterance.

$$\begin{aligned} P^{\text{UD}}(\hat{\mathbf{w}}|q) &= \int P(\hat{\mathbf{w}}|\mathbf{w})P(\mathbf{w}|q)d\mathbf{w} \\ &= \int \mathcal{N}(\mathbf{w}, \Sigma_w)\mathcal{N}(\mathbf{w}; \boldsymbol{\mu} + \Phi \mathbf{y}, \Sigma)d\mathbf{w} \\ &= \mathcal{N}(\mathbf{w}; \boldsymbol{\mu} + \Phi \mathbf{y}, \Sigma + \Sigma_w) \end{aligned}$$

This implies that by assuming \mathbf{w}_1 as enrollment (without uncertainty) and \mathbf{w}_2 as test i-vector (with respective uncertainty Σ_{w_2}), the UD scoring becomes

$$s_{\text{PLDA}}^{\text{UD}} = \frac{\mathcal{N}(\mathbf{w}_{12}; \mathbf{0}, \begin{bmatrix} \Phi\Phi^T + \Sigma & \Phi\Phi^T \\ \Phi\Phi^T & \Phi\Phi^T + \Sigma + \Sigma_{w_2} \end{bmatrix})}{\mathcal{N}(\mathbf{w}_{12}; \mathbf{0}, \begin{bmatrix} \Phi\Phi^T + \Sigma & \mathbf{0} \\ \mathbf{0} & \Phi\Phi^T + \Sigma + \Sigma_{w_2} \end{bmatrix})}$$

In order to perform modified imputation, we need to have access to the mean vector of a speaker model. Since the speaker factors are being marginalized in the conventional PLDA scoring, it is not practical to apply modified imputation in the framework of conventional PLDA scoring. Inspired by formulations in [15, 16], we employ the following simplified PLDA scoring method dubbed as *relative scoring* (RS);

$$s_{\text{RS}} = \frac{P(\mathbf{w}_2|\mathbf{w}_1, q)}{P(\mathbf{w}_2|q)} = \frac{\mathcal{N}(\mathbf{w}_2; \mathbf{w}_1, \Sigma)}{\mathcal{N}(\mathbf{w}_2; \mathbf{0}, \Phi\Phi^T + \Sigma)},$$

In UD, the score s then calculated as

$$s_{\text{RS}}^{\text{UD}} = \frac{\mathcal{N}(\mathbf{w}_2; \mathbf{w}_1, \Sigma + \Sigma_{w_2})}{\mathcal{N}(\mathbf{w}_2; \mathbf{0}, \Phi\Phi^T + \Sigma + \Sigma_{w_2})}.$$

For *modified imputation* (MI), we can modify [28] the test i-vector \mathbf{w}_2 as

$$\mathbf{w}_2^{\text{MI}} = (\Sigma_{w_2} + \Sigma)^{-1}(\Sigma \mathbf{w}_1 + \Sigma_{w_2} \mathbf{w}_2),$$

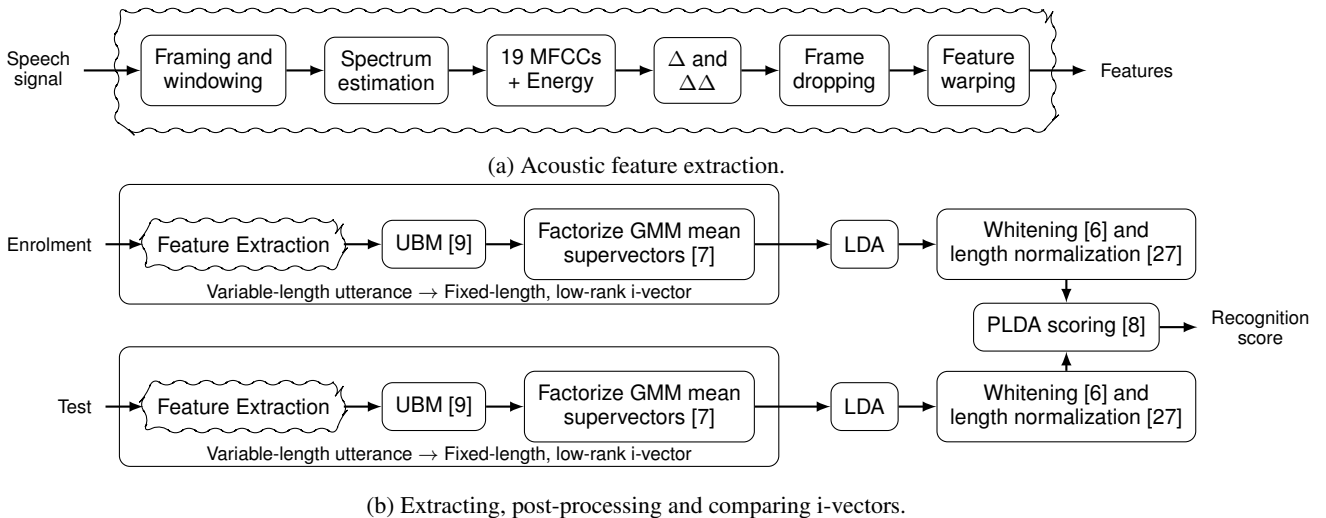


Figure 2: Block diagram of the i-vector based speaker recognition system in our experiments.

and calculate the recognition score as

$$s_{RS}^{MI} = \frac{\mathcal{N}(\mathbf{w}_2^{MI}; \mathbf{w}_1, \Sigma)}{\mathcal{N}(\mathbf{w}_2^{MI}; \mathbf{0}, \Phi\Phi^T + \Sigma)}.$$

4. Speaker Recognition Experiments

4.1. Experimental Setup

A schematic block diagram of the state-of-the-art speaker recognition system as used for experiments in this paper is shown in Figure 2. 19-dimensional Mel-frequency cepstral coefficients are extracted from frames of 20 ms windowed speech every 10 ms, appended with the frame energy and concatenated with Δ and $\Delta\Delta$ coefficients, resulting in 60-dimensional feature vectors [29, 30]. The *speech activity detector* in [31] is employed to discard non-speech frames and *feature warping* [32] is applied on the final features. In the truncation process for the experiments in this paper, feature warping is applied after truncation.

A gender-dependent *universal background model* (UBM) [9] with 2048 components is trained using a subset of NIST SRE 2004–2006, Switchboard cellular phase 1 and 2, and Fisher English corpora. To factorize the GMM mean supervectors, the total variability space [7] is trained with the same data as for the UBM with 400-dimensions. In post-processing the utterance-level i-vectors, we use a linear discriminant analysis projection to enhance separability of classes (speakers) and to reduce the i-vector dimension to 200. Prior to PLDA modeling, we remove the mean, perform whitening using within-class covariance normalization (WCCN) [6] and normalize the length of i-vectors to lie on unit sphere [27]. We use multi-condition training in finding LDA and PLDA parameters with multiple duration (truncated versions) of the same utterance [4, 14].

The enrollment i-vectors are not truncated. We average over multiple i-vectors per speaker in enrolment in order to produce the speaker template. The truncation is applied by using the first N features of each test segment to produce effective duration of 5, 10, 20 and 40 seconds. We employed the file lists of I4U developed in preparations during the NIST SRE’12 evaluation period [29]. The file lists are modified by excluding the segments that are less than 40 seconds long, in order to fit the truncation experiments in this paper. For the sake of tractability,

the experiments in this paper are performed on male speakers as the statistics is provided in Table 1.

4.2. Uncertainty Estimation

As mentioned earlier, the i-vector extractor can supply with a measure of uncertainty along with calculating the posterior mean. In order to avoid the question of accuracy of the uncertainty estimation of i-vector extractor, we provide a proof of concept study in this paper and define an *oracle uncertainty* for each i-vector. We define the oracle uncertainty value for each i-vector \mathbf{w} as

$$\Sigma_{\mathbf{w}} = \text{diag} \left[(\mathbf{w} - \mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)^T \right], \quad (2)$$

where \mathbf{w}^* is the optimal realization of the current i-vector \mathbf{w} . Such operation implies that we take into account the uncertainty in mean statistics estimation where \mathbf{w}^* would be the estimated mean statistics with enough many samples and the source of uncertainty in \mathbf{w} , with limited observations, will be the *incomplete data*. In the speaker recognition paradigm, a long recording (3 minutes) of speech presents the underlying speaker much more accurately than a short recording (5 seconds). As in the truncation process we have access to the full duration of speech, the uncertainty caused by truncation is estimated as in Equation 2. We consider duration of the provided speech for authentication as the source of uncertainty in our experiments. However, using the same principal we can analyse the effect of noise where the optimal realization of an observation measured in noise could be the measurement made in *clean* condition.

4.3. Uncertainty Propagation

The set of post-processing steps of i-vectors including LDA and WCCN are linear transformations and as it is indicated in Fig-

Table 1: Number of speakers, speech segments and trials in the modified I4U file list for male speakers in development set.

Num. of speakers		Num. of segments		Num. of trials	
Train	Test	Train	Test	True	False
680	828	5475	6501	4801	4415879

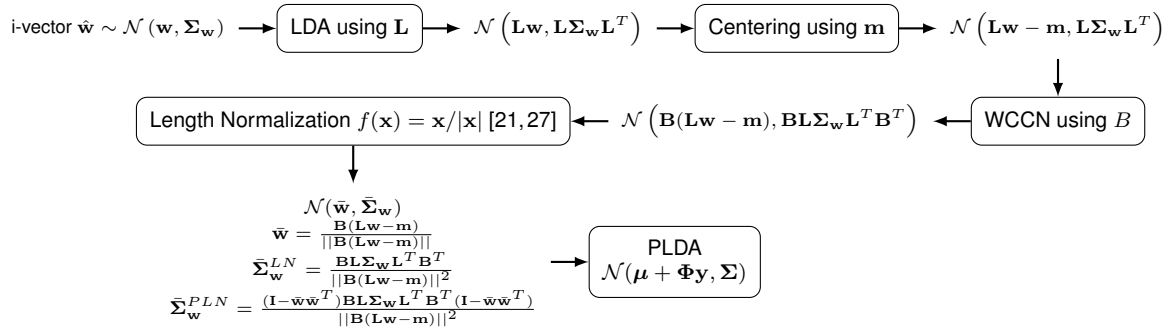


Figure 3: Propagating uncertainty of i-vectors through the post-processing steps.

Table 2: Speaker recognition results in terms of $E_{=}$ (reported in %) for uncertainty caused by duration.

	Cosine [7]	PLDA				Relative Scoring			
		Conv.	+UD		+UD		+MI		
		LN	PLN	LN	PLN	LN	PLN	LN	PLN
5	5.7	3.8	3.4	4.1	4.8	3.6	3.9	4.1	3.7
10	3.1	2.0	1.8	2.0	2.5	2.0	1.9	2.2	1.9
20	1.7	1.2	1.1	1.1	1.4	1.4	1.1	1.5	1.2
40	1.2	.8	.8	.8	1.0	1.1	.8	1.3	.9
full	.9	.6	.6	.6	.7	.7	.7	.7	.7
Avg.	2.5	1.7	1.5	1.7	2.1	1.8	1.7	2.0	1.7

Table 3: Speaker recognition results in terms of *normalized* C_{det} (Equation 3) for uncertainty caused by duration.

	Cosine [7]	PLDA				Relative Scoring			
		Conv.	+UD		+UD		+MI		
		LN	PLN	LN	PLN	LN	PLN	LN	PLN
5	.79	.60	.65	.66	.67	.57	.58	.69	.57
10	.57	.38	.40	.42	.39	.32	.35	.41	.33
20	.40	.24	.25	.26	.23	.20	.21	.27	.20
40	.28	.17	.18	.18	.14	.15	.14	.24	.15
full	.21	.13	.13	.13	.10	.10	.10	.10	.10
Avg.	.45	.30	.32	.33	.31	.27	.27	.34	.27

Figure 3, it is straightforward to transform the uncertainty through these blocks. However, the last step of post-processing is a non-linear transformation. Length normalization $f(\mathbf{x}) = \mathbf{x}/|\mathbf{x}|$ [27] is a simplified version of radial Gaussianization [33] which allows using Gaussian PLDA [34] in modeling i-vectors instead of dealing with heavy-tailed PLDA [35]. A Taylor series expansion of length normalization is suggested in [21] to enable passing the uncertainties through. A simplified version of first-order Taylor series expansion around the posterior distribution mean results in uncertainties that are scaled by the inverse of squared length. This is simply called *length normalization* (LN). The first-order approximation of transformation was called *projected LN* (PLN) as the respective operations presented in Figure 3.

4.4. Experimental Results

The results for speaker verification are presented in terms of equal error rate ($E_{=}$) and minimum of decision cost function C_{det}^{\min} . The detection cost function C_{det} is computed using

$$C_{det} = C_{miss} \times P_{tar} \times P_{miss} + C_{fa} \times (1 - P_{tar}) \times P_{fa} \quad (3)$$

with $C_{miss} = C_{fa} = 1$ and $P_{tar} = 1/1000$ as used in NIST SRE'10. In Equation 3, C_{miss} , C_{fa} and P_{tar} stand for cost of a miss, false alarm and prior probability of a target trial, respectively. The prior probability of a target trial determines target speaker presence in system evaluation phase. For the equal error rate ($E_{=}$) is that point on the *receiver operating characteristic* (ROC) curve where the probabilities of missed detection P_{miss} and false alarm P_{fa} become equal. The C_{det} and $E_{=}$ values are computed using the BOSARIS toolkit [36] via Bayes error rate computation.

The experimental results are presented in Tables 2 and 3. Employing uncertainty decoding in context of conventional PLDA

provides superior $E_{=}$ compared to other approaches. However, this comes at the cost of marginal increase in C_{det}^{\min} . The relative scoring scheme provides better performance in terms of C_{det}^{\min} . In light of our experiments, the choice of LN or PLN approximation for length normalization transform provides better performance for UD and MI, respectively. The results on full duration test utterances present no change across different techniques because of zero uncertainty according to our specific definition of oracle uncertainty in Equation 2.

5. Conclusions

The application of modified imputation in context of relative scoring is introduced for speaker recognition. The recognition performance is evaluated on NIST SRE corpora using I4U file lists and controlled duration is applied in test phase. Despite different mechanisms of operation, modified imputation (modifying input) performs on par with uncertainty decoding (modifying model). The *significance decoding* [28] as a more general case of modified imputation and uncertainty decoding can be considered in future research. The results in this paper provide a proof of concept for the efficiency of uncertainty handling using oracle uncertainties in modified imputation approach. The underlying effects of LN and PLN on choice of uncertainty handling technique are not clear at this point.

6. Acknowledgement

This work was supported by Academy of Finland (project numbers 256961, 284671). We acknowledge the computational resources provided by the Aalto Science-IT project.

7. References

- [1] T Kinnunen and H Li. An overview of text-independent speaker recognition: from features to supervectors. *Speech Comm.*, 52(1):12–40, January 2010.
- [2] NIST speaker recognition evaluations. <http://www.nist.gov/itl/iad/mig/sre12.cfm>.
- [3] NIST 2012 SRE, October 2012. <http://www.nist.gov/itl/iad/mig/sre12.cfm>.
- [4] 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. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013.
- [5] A Solomonoff, W Campbell, and I Boardman. Advances in channel compensation for SVM speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pages 629–632, Philadelphia, USA, March 2005.
- [6] A. O Hatch, S Kajari, and A Stolcke. Within-class covariance normalization for SVM-based speaker recognition. In *Proc. Interspeech 2006 (ICSLP)*, pages 1471–1474, Pittsburgh, Pennsylvania, USA, September 2006.
- [7] N Dehak, P J Kenny, R Dehak, P Dumouchel, and P Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 19(4):788–798, May 2011.
- [8] S. J. D Prince and J. H Elder. Probabilistic linear discriminant analysis for inferences about identity. In *11th International Conference on Computer Vision*, pages 1–8, 2007.
- [9] D Reynolds, T Quatieri, and R Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, January 2000.
- [10] P Kenny, G Boulianne, P Ouellet, and P Dumouchel. Speaker and session variability in GMM-based speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 15(4):1448–1460, May 2007.
- [11] C Yu, G Liu, S Hahm, and J. H. L Hansen. Uncertainty propagation in front end factor analysis for noise robust speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, 2014.
- [12] Y Shao, S Srinivasan, and D Wang. Incorporating auditory feature uncertainties in robust speaker identification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, volume 4, pages IV–277, 2007.
- [13] X Zhao, Y Wang, and D Wang. Robust speaker identification in noisy and reverberant conditions. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(4):836–845, April 2014.
- [14] T Hasan, R Saeidi, J. H. L Hansen, and D. A van Leeuwen. Duration mismatch compensation for i-vector based speaker recognition systems. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013.
- [15] P Kenny, T Stafylakis, P Ouellet, J Alam, and P Dumouchel. PLDA for speaker verification with utterances of arbitrary duration. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013.
- [16] T Stafylakis, P Kenny, P Ouellet, J Perez, M Kockmann, and P Dumouchel. Text-dependent speaker recognition using PLDA with uncertainty propagation. In *Proc. Interspeech 2013*, 2013.
- [17] V Hautamäki, Y.-C Cheng, P Rajan, and C.-H Lee. Minimax i-vector extractor for short duration speaker verification. In *Proc. Interspeech 2013*, 2013.
- [18] A Kanagasundaram, D Dean, J Gonzalez-Dominguez, S Sridharan, D Ramos, and J Gonzalez-Rodriguez. Improving short utterance based i-vector speaker recognition using source and utterance-duration normalization techniques. In *Proc. Interspeech 2013*, pages 2465–2469, 2013.
- [19] A Kanagasundaram, D Dean, S Sridharan, J Gonzalez-Dominguez, J Gonzalez-Rodriguez, and D Ramos. Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. *Speech Communication*, 59(0):69 – 82, 2014.
- [20] S Cumani, O Plchot, and P Laface. On the use of i-vector posterior distributions in probabilistic linear discriminant analysis. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(4):846–857, April 2014.
- [21] S Cumani, O Plchot, and P Laface. Probabilistic linear discriminant analysis of i-vector posterior distributions. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, pages 7644–7648, May 2013.
- [22] B Raj, M. L Seltzer, and R. M Stern. Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43(4):275 – 296, 2004.
- [23] J Gemmeke, H Van Hamme, B Cranen, and L Boves. Compressive sensing for missing data imputation in noise robust speech recognition. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):272–287, April 2010.
- [24] P Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal, (Report) CRIM-06/08-13*, 2005.
- [25] P. M Bousquet, A Larcher, D Matrouf, J. F Bonastre, and O Plchot. Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis. In *Proc. Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2012)*, 2012.
- [26] S. J. D Prince. *Computer vision: models, learning, and inference*. Cambridge University Press, 2011.
- [27] D Garcia-Romero and C. Y Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Proc. Interspeech 2011*, pages 249–252, 2011.
- [28] A Abdelaziz, S Zeiler, D Kolossa, V Leutnant, and R Haeb-Umbach. Gmm-based significance decoding. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, pages 6827–6831, May 2013.
- [29] R Saeidi, et al. I4U submission to NIST SRE 2012: a large-scale collaborative effort for noise-robust speaker verification. In *INTERSPEECH*, pages 1986–1990, 2013.
- [30] M. I Mandasari, R Saeidi, M McLaren, and D. A van Leeuwen. Quality measure functions for calibration of speaker recognition systems in various duration conditions. *IEEE Trans. Audio, Speech, and Language Processing*, 21:2425–2438, 2013.
- [31] M McLaren and D. A van Leeuwen. A simple and effective speech activity detection algorithm for telephone and microphone speech. In *in Proc. NIST SRE 2011 Workshop*, 2011.
- [32] J Pelecanos and S Sridharan. Feature warping for robust speaker verification. In *Proc. Odyssey Workshop*, pages 213–218, 2001.
- [33] S Lyu and E. P Simoncelli. Nonlinear extraction of independent components of natural images using radial gaussianization. *Neural Computing*, 21(6):1485–1519, 2009.
- [34] 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 *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, 2012.
- [35] P Kenny. Bayesian speaker verification with heavy-tailed priors. In *Proc. Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2010)*, Brno, Czech Republic, June 2010.
- [36] N Brummer. Bosaris toolkit. <https://sites.google.com/site/bosaristoolkit/>.