AN IMPROVED UNCERTAINTY PROPAGATION METHOD FOR ROBUST I-VECTOR BASED SPEAKER RECOGNITION

Dayana Ribas

ViVoLab, Aragon Institute for
Engineering Research (I3A), Spain

Emmanuel Vincent

Université de Lorraine, CNRS,
Inria, LORIA, F-54000 Nancy, France

ABSTRACT

The performance of automatic speaker recognition systems degrades when facing distorted speech data containing additive noise and/or reverberation. Statistical uncertainty propagation has been introduced as a promising paradigm to address this challenge. So far, different uncertainty propagation methods have been proposed to compensate noise and reverberation in i-vectors in the context of speaker recognition. They have achieved promising results on small datasets such as YOHO and Wall Street Journal, but little or no improvement on the larger, highly variable NIST Speaker Recognition Evaluation (SRE) corpus. In this paper, we propose a complete uncertainty propagation method, whereby we model the effect of uncertainty both in the computation of unbiased Baum-Welch statistics and in the derivation of the posterior expectation of the i-vector. We conduct experiments on the NIST-SRE corpus mixed with real domestic noise and reverberation from the CHiME-2 corpus and preprocessed by multichannel speech enhancement. The proposed method improves the equal error rate (EER) by 4% relative compared to a conventional i-vector based speaker verification baseline. This is to be compared with previous methods which degrade performance.

Index Terms— Uncertainty propagation, speaker verification, data distortion, robustness, i-vector

1. INTRODUCTION

Uncertainty propagation has emerged as a paradigm for robust signal processing whereby the data are not treated as point estimates anymore, but as a parametric posterior distribution, typically approximated as a Gaussian. This approach provides a principled framework to deal with the loss of information due to signal distortion – epistemic uncertainty – or to the finite number of data points – aleatoric uncertainty –. The uncertainty is represented by a set of scalar variances or covariance matrices, which are first estimated on the data, and then propagated through the subsequent processing steps in order to compensate for the effect of uncertainty on the computed quantities and ultimately improve the system performance [1–4].

In speaker recognition, the uncertainty propagation approach has gained traction motivated by the uncertain nature of the system pipeline when the application scenario tends to more real-world situations. The necessity of improving the system robustness in noisy environments has inspired the development of epistemic uncertainty approaches for speaker modeling based on Gaussian mixture models [5, 6] and, more recently, i-vectors [7, 8]. Other work based on the aleatoric uncertainty concept has also given rise to uncertainty propagation approaches for speaker recognition. However, these approaches focused on the issue of computing representations with in-sufficient data, caused by utterances with different, possibly short durations [9–15].

This paper pursues the same line as our preliminary study [8], that considered an epistemic uncertainty propagation approach for noise-robust text-independent speaker verification using a system based on i-vectors [16] and probabilistic linear discriminant analysis (PLDA) [17]. Despite the recent introduction of deep learning based modules in the speaker recognition pipeline, reports of the last NIST Speaker Recognition Evaluation (SRE) in 2016 [18] and the experience in the recent campaign NIST-SRE 2018[1] evidenced that the i-vector-PLDA approach still performs among the best systems of the state-of-the-art. It is also the focus of current challenges in the field such as the 2018 Multi-target Speaker Detection and Identification Challenge Evaluation[2].

In [8], we proposed a method to estimate and propagate the residual uncertainty after multichannel speech enhancement from the acoustic features to the i-vectors. Specifically, we modified the posterior probability of each Gaussian mixture component to obtain unbiased Baum-Welch (BW) statistics. Preliminary experiments yielded good results on the Wall Street Journal (WSJ) corpus, but little or no improvement on a subset of the NIST-SRE 2008 corpus, similarly to the findings of Yu et al. [7] on the YOHO and NIST-SRE 2010 datasets. We studied the causes of this under-performance on NIST-SRE and found out that the high variability of NIST-SRE makes it intrinsically harder to obtain accurate BW statistics.

In this paper, we propose a new uncertainty propagation method that models the effect of epistemic uncertainty both in the computation of the BW statistics and in the derivation of the i-vector. This method provides a more complete strategy towards compensating for background noise in all steps of the i-vector computation process. Furthermore, this study contributes to clarifying how the i-vector speaker representations are affected by the environmental distortion. The results are evaluated on a subset of the NIST-SRE 2008 corpus mixed with real domestic noise and reverberation from the CHiME-2 corpus [19].

Section 2 recalls the i-vector computation process. Section 3 provides a novel analysis of the limitations of related previous works and Section 4 introduces the proposed method. Speaker verification experiments are conducted in Section 5. The results are reported and discussed in Section 6. Finally, the conclusions of the study and future work are presented in Section 7.

---

[1] https://www.nist.gov/itl/iad/mig/nist-2018-speaker-recognition-evaluation

[2] http://mce.csail.mit.edu/

2. I-VECTOR COMPUTATION

Front-end factor analysis [16] relies on a universal background model (UBM) that is a mixture of C Gaussian components indexed by c. Denoting by F the feature dimension, the $CF \times 1$ supervector $M(u)$ for one utterance u is expressed as

$$M(u) = m + Tw(u) + \epsilon(u) \qquad (1)$$

where m consists of the means m_c of all UBM components, T is the $CF \times D$ low-rank total variability matrix, $w(u)$ is the $D \times 1$ vector of total factors or *i-vector*, and $\epsilon(u)$ represents the residual data variability not captured by T. The i-vector is modeled as a zero-mean standard Gaussian random vector. It is obtained by computing the posterior expectation of $w(u)$ over the feature sequence $\{y_1, \ldots, y_L\}$, with L the number of time frames:

$$\mathbb{E}[w(u)] = (I + T'V^{-1}N(u)T)^{-1}T'V^{-1}\hat{F}(u). \qquad (2)$$

In this equation, $N(u)$ is a $CF \times CF$ diagonal matrix with diagonal blocks $N_c(u)I$ where $N_c(u)$ are the zeroth-order BW statistics for all components c, $\hat{F}(u)$ is a $CF \times 1$ supervector obtained by concatenating the centralized first-order BW statistics $\hat{F}_c(u)$, V is the diagonal $CF \times CF$ covariance matrix of $\epsilon(u)$, and $'$ denotes matrix transposition. The BW statistics are given by

$$N_c(u) = \sum_{t=1}^{L} \gamma_t(c) \qquad (3)$$

$$\hat{F}_c(u) = \sum_{t=1}^{L} \gamma_t(c)(y_t - m_c) \qquad (4)$$

where

$$\gamma_t(c) = \frac{\pi_c \mathcal{N}(y_t|\mu_c, \Sigma_c)}{\sum_{i=1}^{C} \pi_i \mathcal{N}(y_t|\mu_i, \Sigma_i)} \qquad (5)$$

is the posterior probability of the c-th UBM component, as obtained from its mean m_c, covariance Σ_c and weight π_c.

Note that (2) can be equivalently rewritten in terms of the "normalized" statistics $\tilde{N}(u)$ and $\tilde{F}(u)$ as

$$\mathbb{E}[w(u)] = (I + T'\tilde{N}(u)T)^{-1}T'\tilde{F}(u). \qquad (6)$$

where $\tilde{N}(u) = V^{-1}N(u)$ is a diagonal matrix with diagonal blocks $\tilde{N}_c(u) = N_c(u)V_c^{-1}$, $\tilde{F}(u) = V^{-1}\hat{F}(u)$ is obtained by concatenating $\tilde{F}_c(u) = V_c^{-1}\hat{F}_c(u)$, and V_c is the c-th diagonal block of V. The multiplication by V_c^{-1} can be distributed at each time t of the summations in (3) and (4). The purpose of this rewriting will become clear in the next section.

3. UNCERTAINTY PROPAGATION TO THE I-VECTOR

Let us assume that we now observe a corrupted speech signal x_t involving noise and/or reverberation. Using a speech enhancement algorithm together with an uncertainty estimation technique, the posterior probability of the clean speech features y_t can be modeled as [20]

$$p(y_t|x_t) = \mathcal{N}(y_t|\bar{y}_t, \bar{\Sigma}_t) \qquad (7)$$

with $\bar{y}_t$ the enhanced features and $\bar{\Sigma}_t$ the uncertainty covariance matrix at time t. In other words, $\bar{\Sigma}_t$ is the covariance of the estimation error between the enhanced features and the (unknown) clean features at a given time.

To the best of our knowledge, the studies in [7, 8] are the only ones exploiting this model for noise and reverberation robustness in i-vector based speaker recognition systems, while other works focused on earlier, now deprecated systems. They proposed two different ways to propagate the uncertainty from the enhanced features to the i-vectors.

3.1. Uncertainty propagation through the front-end factor analysis model

The authors in [7] considered the generative data model corresponding to (1). By integrating over the unknown clean features, they accounted for the impact of uncertainty on the expression of the joint posterior probability. They derived the posterior expectation of $w(u)$ over the feature sequence in a similar way to (6) as[3]

$$\mathbb{E}[w_{\mathrm{unc}}(u)] = (I + T'\tilde{N}_{\mathrm{unc}}(u)T)^{-1}T'\tilde{F}_{\mathrm{unc}}(u). \qquad (8)$$

$\tilde{N}_{\mathrm{unc}}(u)$ becomes the $CF \times CF$ block-diagonal matrix with diagonal blocks

$$\tilde{N}_{\mathrm{unc},c}(u) = \sum_{t=1}^{L} \gamma_t(c)V_{\mathrm{unc},c,t}^{-1}, \qquad (9)$$

$\tilde{F}_{\mathrm{unc}}(u)$ is the $CF \times 1$ supervector obtained by concatenating

$$\tilde{F}_{\mathrm{unc},c}(u) = \sum_{t=1}^{L} \gamma_t(c)V_{\mathrm{unc},c,t}^{-1}(\bar{y}_t - m_c), \qquad (10)$$

and $V_{\mathrm{unc},c,t}$ is the total covariance of the residual variability and the uncertainty:

$$V_{\mathrm{unc},c,t} = V_c + \bar{\Sigma}_t. \qquad (11)$$

In order to compute the posterior probability of the c-th UBM component, the authors substituted the clean features y_t in (5) by the enhanced features $\bar{y}_t$:

$$\gamma_t(c) = \frac{\pi_c \mathcal{N}(\bar{y}_t|\mu_c, \Sigma_c)}{\sum_{i=1}^{C} \pi_i \mathcal{N}(\bar{y}_t|\mu_i, \Sigma_i)} \qquad (12)$$

This expression does not account for the difference between the enhanced features and the clean features, hence the BW statistics are biased. Indeed, it is well known from the field of speech recognition that using enhanced data as inputs to an acoustic model trained on clean data often results in poor recognition performance due to the residual distortions in the enhanced data [3, 20]. We conclude that the uncertainty is not fully propagated to the i-vector domain, and therefore the obtained i-vectors remain affected by the data distortion. This may be the reason for the poor results achieved by this method on the NIST-SRE 2010 corpus, even in the situation when *oracle* (ideal) uncertainty estimates are used [7].

3.2. Uncertainty propagation through the UBM

The study in [8] presented an algorithm to compute unbiased BW statistics instead. By considering the generative data model associated with the UBM and integrating over the unknown clean features, the likelihood $p(y_t|c) = \mathcal{N}(y_t|\mu_c, \Sigma_c)$ of the c-th UBM component is classically substituted by [3]

$$p(x_t|c) = \mathcal{N}(\bar{y}_t|\mu_c, \Sigma_{\mathrm{unc},c,t}) \qquad (13)$$

---

[3]For the sake of clarity, we modified the notations in [7] for consistency with the above.

where

$$\Sigma_{\text{unc},c,t} = \Sigma_c + \bar{\Sigma}_t. \tag{14}$$

The authors considered the i-vector computation in (2) as a deterministic operation and consequently derived the i-vector as

$$\mathbb{E}[w_{\text{unc}}(u)] = (I + T'V^{-1}N_{\text{unc}}(u)T)^{-1}T'V^{-1}\hat{F}_{\text{unc}}(u) \tag{15}$$

where $N_{\text{unc}}(u)$ is a diagonal matrix with blocks $N_{\text{unc},c}(u)I$, $\hat{F}_{\text{unc}}(u)$ is the concatenation of $\hat{F}_{\text{unc},c}(u)$, and the unbiased BW statistics [6] are given by

$$\gamma_{\text{unc},t}(c) = \frac{\pi_c \mathcal{N}(\bar{y}_t | \mu_c, \Sigma_{\text{unc},c,t})}{\sum_{i=1}^{C} \pi_i \mathcal{N}(\bar{y}_t | \mu_i, \Sigma_{\text{unc},i,t})} \tag{16}$$

$$N_{\text{unc},c}(u) = \sum_{t=1}^{L} \gamma_{\text{unc},t}(c) \tag{17}$$

$$\hat{F}_{\text{unc},c}(u) = \sum_{t=1}^{L} \gamma_{\text{unc},t}(c) W_{c,t}(\bar{y}_t - m_c) \tag{18}$$

with $W_{c,t}$ the Wiener filter [21] defined as

$$W_{c,t} = \Sigma_c \Sigma_{\text{unc},c,t}^{-1}. \tag{19}$$

Despite the use of unbiased BW statistics, the resulting i-vectors are not fully compensated for the uncertainty at the input. Since the i-vector computation is not according to the modifications introduced to the speaker model due to the propagation of the uncertainty, it raises a mismatch that negatively impacts the system performance. Indeed, i-vector computation is not a deterministic operation and it must be modified to account for this uncertainty. This may partly explain the poor results reported on the NIST-SRE 2008 corpus in [8], in addition to the high variability of this corpus which makes it harder to obtain meaningful BW statistics (see study in [8]).

4. PROPOSED METHOD

The analysis of previous works on uncertainty propagation for i-vector based speaker recognition has uncovered the fact that the uncertainty is propagated either through the front-end factor analysis model or through the UBM. In order to fully compensate i-vectors for the distortion of the input data, we propose to propagate the uncertainty through both models. I-vectors are then computed in the same way as in (8), albeit with unbiased normalized statistics. Specifically, $\tilde{N}_{\text{unc}}(u)$ is now the block-diagonal matrix with diagonal blocks

$$\tilde{N}_{\text{unc},c}(u) = \sum_{t=1}^{L} \gamma_{\text{unc},t}(c) V_{\text{unc},c,t}^{-1} \tag{20}$$

and $\tilde{F}_{\text{unc}}(u)$ is obtained by concatenating

$$\tilde{F}_{\text{unc},c}(u) = \sum_{t=1}^{L} \gamma_{\text{unc},t}(c) V_{\text{unc},c,t}^{-1} W_{c,t}(\bar{y}_t - m_c), \tag{21}$$

where $\gamma_{\text{unc},t}(c)$, $V_{\text{unc},c,t}$, and $W_{c,t}$ are defined in (16), (11), and (19), respectively.

This time, the enhanced data $\bar{y}_t$ and the uncertainty $\bar{\Sigma}_t$ are both used to compute the unbiased BW statistics. Furthermore, the i-vector obtained from the computation of the posterior expectation of $w(u)$ also considers the influence of $\bar{\Sigma}_t$.

In order to assess the impact of unbiased BW statistics independently of the i-vector computation process, we compared them to biased BW statistics as follows. We chose the unnormalized $\hat{F}$ statistics as a rough representation of the speaker characteristics before i-vector computation. We computed the cosine distance between the $\hat{F}$ vectors of each trial of the dataset used in the experimental setup (see section 5.1). The results displayed in Fig. 1 show higher cosine distances for the unbiased statistics in non-target trials. The average cosine distance over all non-target trials is 0.7939 for the unbiased $\hat{F}$ statistics vs. 0.7422 for the biased $\hat{F}$ statistics. This suggests that non-target utterances can be better separated using unbiased statistics. The poor results reported on the NIST-SRE 2008 corpus in [8] using i-vector based cosine distances suggest that the benefit of unbiased BW statistics observed when considering the statistics themselves was lost when considering i-vectors instead. Overall, this suggests that unbiased BW statistics and uncertainty-aware i-vector computation are both necessary to achieve good results.
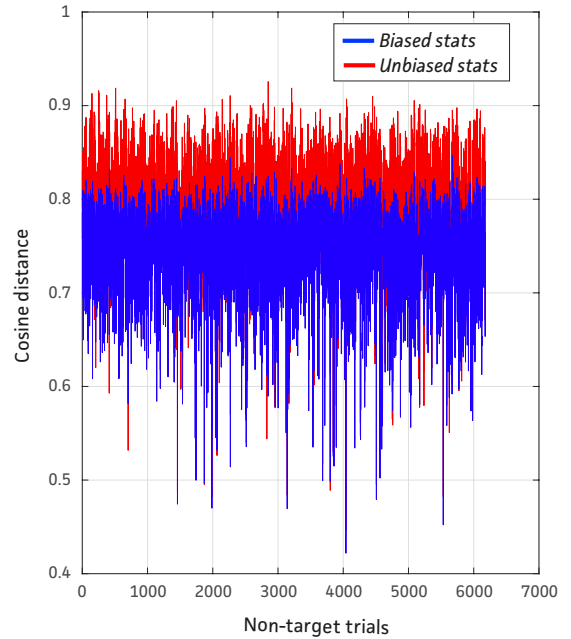

Fig. 1. *Cosine distance for the unbiased and biased $\hat{F}$ statistics in non-target trials.*

5. EXPERIMENTAL SETUP

5.1. Dataset description

We compare the proposed method with [7] and [8] using the same experimental setup as in [8]. The speech signals are male conversations in English from the NIST-SRE 2004 and 2005 corpora for the training stage (3285 speech signals from 262 speakers), and from the NIST-SRE 2008 *short2* and *short3* corpora for the evaluation stage (enrollment: 470 utterances, test: 671 utterances). A total of 6615 verifications were performed on the *det7* condition of NIST-SRE.

We used two-channel noise and reverberation from Track 1 of the 2nd CHiME Challenge [19], recorded in a real domestic environment. Each training or evaluation signal was convolved with one

of 121 two-microphone room impulse responses with a reverberation time of 0.3 seconds. Moreover, the enrollment and test signals were mixed with a random segment of real background noise including, e.g., voices, TV, game console, cutlery sounds, and footsteps. This mixing process resulted in noisy speech signals with different SNRs ranging from about -10 to +20 dB with an average of 6.1 dB.

5.2. Speech enhancement and uncertainty estimation

For speech enhancement, we applied the CHiME-2 recipe [22] of the Flexible Audio Source Separation Toolbox (FASST) [23]. This recipe was designed to reduce noise, but not reverberation. Hence we consider y_t to be the reverberated noiseless data. For uncertainty estimation, in order to assess the potential of each method, we used the oracle diagonal uncertainty covariance matrices $\bar{\Sigma}_t = \mathrm{diag}(y_t - \bar{y}_t)^2$ [20]. However, for a real application there are several uncertainty estimators [24].

5.3. Speaker recognition system

The features consist of 19 Mel frequency cepstral coefficients (MFCCs), the log-energy, and their first- and second-order derivatives, followed by voice activity detection (VAD) and cepstral mean and variance normalization (CMVN) [25]. The UBM ($C = 512$) and the T matrix ($D = 400$) were trained on either clean or reverberated noiseless data [26]. Since the chosen speech enhancement method reduces noise but not reverberation, training on reverberated speech provides better results than training on clean speech. The i-vectors are centered, whitened and length-normalized, and subsequently projected with 330-dimensional LDA. Classification relies on Gaussian PLDA [17].

6. RESULTS AND DISCUSSION

The speaker verification results expressed in terms of equal error rate (EER) are presented in Table 1. For comparison purposes, the first two rows show the results obtained when training and testing on clean (original NIST) signals or when training and testing on reverberated noiseless signals. The following rows show the results obtained when training on reverberated noiseless signals and testing on noisy or enhanced signals. The last three rows correspond to the two previous uncertainty propagation methods reviewed in Section 3 and to the proposed one in Section 4. Yu et al.'s method [7] was tested based on an implementation from the authors.

Table 1. *Speaker recognition results on NIST-SRE.*

Training set	Evaluation set	EER (%)
Clean	Clean	3.19
Reverberated	Reverberated	4.33
Reverberated	Noisy	31.85
Reverberated	Enhanced	10.48
Reverberated	Uncertainty propagation [7]	11.39
Reverberated	Uncertainty propagation [8]	10.69
Reverberated	Proposed uncertainty propagation	**10.02**

The best achievable EER on reverberated data (without performing dereverberation) is 4.33%. Background noise strongly degrades it up to 31.85%. Multichannel speech enhancement improves it down to 10.48%. Previous uncertainty propagation methods do not outperform this baseline; actually, the method in [7] significantly

degrades it. Only the proposed method manages to improve performance down to an EER of 10.02%. This corresponds to a 4% relative improvement with respect to the speech enhancement baseline without uncertainty propagation. Although this might seem limited, it is to the best of our knowledge the first time an improvement in terms of noise robustness is reported for i-vector based systems by means of uncertainty propagation on a large NIST-SRE corpus.

Figure 2 shows the histograms of scores for the three uncertainty propagation algorithms in Table 1. The scores are normalized using the maximum score value across the three sets. The scores for non-target trials obtained with the proposed method appear to be smallest among the three tested systems. Indeed the corresponding curve is the leftmost one in the zoomed view. This behavior is consistent with our findings regarding unbiased vs. biased statistics in Section 4. This indicates there is a wider separation between non-target and target trials in the system using the proposed uncertainty propagation algorithm, hence the performance improvement obtained in Table 1.
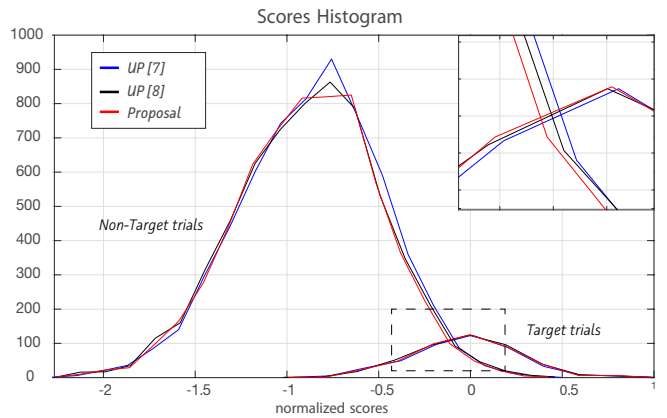

Fig. 2. *Histogram of scores for non-target (left) and target trials (right). The top right corner is a zoomed view of the dashed area.*

7. CONCLUSIONS

In this paper, we proposed a new method for the propagation of uncertainty due to speech distortion from the input data to the i-vectors in a speaker recognition framework. We analyzed the limitations of previously proposed methods, and introduced a more complete propagation method that operates both on the UBM and on the front-end factor analysis model. We evaluated the speaker verification performance on the NIST-SRE 2008 corpus mixed with real domestic noise and reverberation and obtained a 4% relative EER improvement compared to a speech enhancement baseline without uncertainty propagation, while previous methods degraded performance. Future work will target the extension of the uncertainty propagation approach to the full speaker verification pipeline including the PLDA and the joint handling of epistemic uncertainty due to signal distortion and aleatoric uncertainty due to finite signal size. Also we plan to extend the experimental setup for including female speakers.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] J. A. Arrowood and M. A. Clements, "Using observation uncertainty in HMM decoding," in *Proceedings of Interspeech*, 2002.

[2] Néstor Becerra Yoma and Miguel Villar, "Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 158–166, 2002.

[3] Li Deng, J. Wu, J. Droppo, and Alex Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 13, no. 3, pp. 412–421, 2005.

[4] Hank Liao and MJF Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proceedings of Interspeech*.

[5] Xianyu Zhao, Yuan Dong, Jian Zhao, Liang Lu, Jiqing Liu, and Haila Wang, "Variational Bayesian joint factor analysis for speaker verification," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 1032–1042, 2012.

[6] Alexey Ozerov, Mathieu Lagrange, and Emmanuel Vincent, "Uncertainty-based learning of acoustic models from noisy data," *Computer Speech and Language*, vol. 27, no. 3, pp. 874–894, 2013.

[7] C. Yu, G. Liu, S. Hahm, and John Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4017–4021.

[8] Dayana Ribas, Emmanuel Vincent, and Jose R. Calvo, "Uncertainty propagation for noise robust speaker recognition: the case of NIST-SRE," in *Proceedings of Interspeech*, Dresden, Germany, 2015, pp. 3536–3540, ISCA.

[9] Ville Hautamäki, You-Chi Cheng, Padmanabhan Rajan, and Chin-Hui Lee, "Minimax i-vector extractor for short duration speaker verification," in *Proceedings of Interspeech*, Lyon, France, 2013, pp. 3708–3712, ISCA.

[10] Themos Stafylakis, Patrick Kenny, Pierre Ouellet, J. Perez, M Kockmann, and Pierre Dumouchel, "Text dependent speaker recognition using PLDA with uncertainty propagation," in *Proceedings of Interspeech*, Lyon, France, 2013, pp. 3684–3688, ISCA.

[11] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Md. Jahangir Alan, and Pierre Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, 2013, pp. 7649–7653, IEEE.

[12] Sandro Cumani, Oldrich Plchot, and Pietro Laface, "On the use of ivector posterior distributions in Probabilistic Linear Discriminant Analysis," *IEEE Transaction on Audio Speech and Language Processing*, vol. 22, pp. 846–857, 2014.

[13] Rahim Saeidi and Paavo Alku, "Accounting for uncertainty of i-vectors in speaker recognition using uncertainty propagation and modified imputation," in *Proceedings of Interspeech*, Dresden, Germany, 2015, pp. 3546–3550, ISCA.

[14] Patrick Kenny, Themos Stafylakis, Md. Jahangir Alam, and Marcel Kockmann, "Uncertainty modeling without subspace methods for text-dependent speaker recognition," in *Speaker and Language Recognition Workshop (Odyssey)*, 2016.

[15] W.W. Lin, M.W. Mak, and J.T. Chien, "Fast scoring for plda with uncertainty propagation via i-vector grouping," *Computer, Speech and Language*, vol. 45, pp. 503–515, 2017.

[16] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[17] Daniel Garcia and Carol Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of Interspeech*, Florence, Italy, 2011, pp. 249–252, ISCA.

[18] K A Lee, V Hautamäki, T Kinnunen, Hector Delgado, Massimiliano Todisco, Nicholas Evans, and et al., "The I4U submission to the 2016 NIST speaker recognition evaluation," in *Proceedings of the NIST SRE 2016 Workshop*, San Diego, CA, 2016.

[19] Emmanuel Vincent, Jon Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHIME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 126–130.

[20] Dorothea Kolossa and Reinhold Haeb-Umbach, *Robust speech recognition of uncertain or missing data*, Springer, 2011.

[21] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, Singapore, 2006.

[22] Yann Salaün, Emmanuel Vincent, Nancy Bertin, Nathan Souviraà-Labastie, Xabier Jaureguiberry, Dung Tran, and Frédéric Bimbot, "The flexible audio source separation toolbox version 2.0," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Show and Tell*, Florence, Italy, 2014, IEEE.

[23] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118 – 1133, May 2012.

[24] Dung T Tran, Emmanuel Vincent, and Denis Jouvet, "Nonparametric uncertainty estimation and propagation for noise robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1835–1846, Jan. 2015.

[25] Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, *Springer Handbook of Speech Processing*, Springer-Verlag, Berlin, Germany, 2008.

[26] Dayana Ribas, Emmanuel Vincent, and Jose R. Calvo, "Full multicondition training for robust i-vector based speaker recognition," in *Proceedings of Interspeech*, Dresden, Germany, 2015, pp. 1057–1061, ISCA.