# Uncertainty propagation for noise robust speaker recognition: the case of NIST-SRE

Dayana Ribas, Emmanuel Vincent, José Ramon Calvo

# Uncertainty propagation for noise robust speaker recognition: the case of NIST-SRE

*Dayana Ribas*[1], *Emmanuel Vincent*[2], *José Ramón Calvo*[1]

[1]Advanced Technologies Application Center (CENATAV), Habana, Cuba
[2]Inria, Villers-lès-Nancy, F-54600, France

{dribas,jcalvo}@cenatav.co.cu, emmanuel.vincent@inria.fr

## Abstract

Uncertainty propagation is an established approach to handle noisy and reverberant conditions in automatic speech recognition (ASR), but it has little been studied for speaker recognition so far. Yu et al. recently proposed to propagate uncertainty to the Baum-Welch (BW) statistics without changing the posterior probability of each mixture component. They obtained good results on a small dataset (YOHO) but little improvement on the NIST-SRE dataset, despite the use of oracle uncertainty estimates. In this paper, we propose to modify the computation of the posterior probability of each mixture component in order to obtain unbiased BW statistics. We show that our approach improves the accuracy of BW statistics on the Wall Street Journal (WSJ) corpus, but yields little or no improvement on NIST-SRE again. We provide a theoretical explanation for this that opens the way for more efficient exploitation of uncertainty on NIST-SRE and other large datasets in the future.

**Index Terms**: speaker recognition, robustness, uncertainty propagation, i-vector.

## 1. Introduction

A current challenge in the field of speaker recognition is to migrate automatic systems developed in the lab to real world environments. The distortion of speech by environmental noise and reverberation jointly with channel mismatch provokes a variability that degrades considerably the high accuracy reached in the lab. Several methods for compensating noise have been developed in the field of automatic speech recognition (ASR) [1–3] and speaker recognition [4]. These methods operate on the input features, on the model parameters, or both.

Among the latter category, uncertainty propagation (UP) has emerged as a new paradigm whereby the data are not treated as point estimates anymore but as a Gaussian posterior distribution that is propagated through the subsequent processing steps. The variance of that distribution quantifies the loss of information due to the finite number of data points (*aleatoric uncertainty*) or to some distortion of the data (*epistemic uncertainty*). The development of UP has mainly focused on ASR [5–12], but a few works have also been presented on speaker recognition in order to handle noisy environments [13–15], signals of short duration [16], or signals of different duration [17, 18].

Table 1 presents an overview of previous works on UP for speaker identification (SI) or speaker verification (SV). The systems used include Gaussian mixture models (GMM) [19], joint factor analysis (JFA) [20], or i-vectors [21], and they may be text-independent (TI) or text-dependent (TD). The location of uncertainty in the system is described by data on which is was estimated (Origin) and the data or the model to which it was propagated (Focus), e.g., the universal background model (UBM) or the probabilistic linear discriminant analysis (PLDA). From this table, it can be seen that [15] is the only study on UP for noise robustness in the state-of-the-art i-vector PLDA framework so far. In this study, the uncertainty on the input features was propagated to the Baum-Welch (BW) statistics and the i-vectors without changing the posterior probability of each mixture component computed by the UBM. This approach gave improved results on a small dataset (YOHO), but it could not obtain good results over the NIST-SRE dataset, even though *oracle* (ideal) uncertainty estimates were used.

In this paper, we describe how to propagate the uncertainty from a speech enhancement system into the i-vectors. Based on the previous work [14] in the older GMM framework, we modify the computation of the posterior probability of each mixture component in order to obtain unbiased BW statistics. Preliminary experiments yield good results on a noisy version of the Wall Street Journal (WSJ) corpus, but little or no improvement on a noisy version of the NIST-SRE 2008 corpus similarly to [15]. We perform an analysis of the UBM and the BW statistics on both datasets which provides a theoretical explanation for the results and suggests some future research directions towards addressing UP on NIST-SRE and other large datasets.

Section 2 recalls the i-vector computation process. Section 3 presents the proposed method for computing the i-vectors taking into account the uncertainty due to noise on the features. Speaker verification experiments are carried out in Section 4. Section 5 reports the results obtained and explores the underlying causes. Finally the conclusions of the study and future work are discussed in Section 6.

## 2. I-vector computation

Let $M$ be the supervector for one utterance $u$. In the front-end factor analysis model [21], this supervector is expressed as

$$M = m + Tw \tag{1}$$

where $m$ is the UBM mean, $T$ is the low-rank total variability matrix, and $w$ is the vector of standard normal random total factors or *i-vector*. The i-vector is obtained by computing the posterior expectation of $w$ over the feature sequence $\{y_1, \ldots, y_L\}$, with $L$ the number of time frames:

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

In this equation, $N(u)$ is a diagonal matrix obtained by concatenating the zeroth order BW statistics $N_c(u)$ for all Gaussian components $c$, $\hat{F(u)}$ is a supervector obtained by concatenating

Table 1: *Overview of uncertainty propagation approaches applied to robust speaker recognition (see Section 1 for abbreviations).*

| System approach | Work | Corpus | Task | Goal | Uncertainty Origin and Focus | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Features | UBM | BW stats | JFA, $T$ mat. | i-vector | PLDA |
| **GMM** | [14] | CHiME | TI SI | Noise | Origin | Focus | | | | |
| **i-vector** | [18] | NIST 2010 | TI SV | Duration | | | | | Origin | Focus |
| | [17] | RSR 2015 | TD SV | Duration | | | | | Origin | Focus |
| | [16] | NIST 2010 | TI SV | Duration | | | Origin-Focus | | | |
| | [15] | YOHO | TI SV | Noise | Origin | | | | Focus | |

the centralized-first order BW statistics $\hat{F}_c(u)$, $\Sigma$ is the diagonal covariance matrix of the front-end factor analysis model, and $.'$ denotes matrix transposition. The BW statistics are given by

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

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

where

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

is the posterior probability of the $c$-th Gaussian component, as obtained from its mean $m_c$, covariance $\Sigma_c$ and weight $\pi_c$.

## 3. Uncertainty propagation to the i-vector

The study in [14] introduced a modified BW algorithm in order to train a GMM from uncertain data by maximizing the uncertainty decoding objective in [6]. This algorithm relies on the computation of unbiased BW statistics. In this section, we extend this algorithm to the computation of i-vectors.

Let us assume that the observed features $x_t$ now result from the addition of clean speech $y_t$ and noise. Using a speech enhancement algorithm together with an uncertainty estimation technique, e.g., [9, 12], the posterior probability of clean speech can be approximated as

$$p(y_t|x_t) = \mathcal{N}(y_t|\bar{y}_t, \bar{\Sigma}_t) \tag{6}$$

with $\bar{y}_t$ the estimated clean speech and $\bar{\Sigma}_t$ the uncertainty covariance matrix at time $t$. By integrating over the unknown clean features, the likelihood of the $c$-th Gaussian component is modified as follows [6]:

$$p(x_t|c) = \mathcal{N}(\bar{y}_t|\mu_c, \Sigma_c + \bar{\Sigma}_t) \tag{7}$$

Consequently the BW statistics must be modified as [14]

$$\gamma_{\mathrm{unc},t}(c) = \frac{\pi_c \mathcal{N}(\bar{y}_t|\mu_c, \Sigma_c + \bar{\Sigma}_t)}{\sum_{i=1}^{C} \pi_i \mathcal{N}(\bar{y}_t|\mu_i, \Sigma_i + \bar{\Sigma}_t)} \tag{8}$$

$$N_{\mathrm{unc},c}(u) = \sum_{t=1}^{L} \gamma_{\mathrm{unc},t}(c) \tag{9}$$

$$\hat{F}_{\mathrm{unc},c}(u) = \sum_{t=1}^{L} \gamma_{\mathrm{unc},t}(c) W_{c,t}(\bar{y}_t - m_c) \tag{10}$$

where $W_{c,t}$ is the Wiener filter [22] defined as

$$W_{c,t} = \Sigma_c[\Sigma_c + \bar{\Sigma}_t]^{-1}. \tag{11}$$

We obtain the i-vector from the unbiased BW statistics as

$$\mathbb{E}[w_{\mathrm{unc}}(u)] = (I + T'\Sigma^{-1} N_{\mathrm{unc},c}(u) T)^{-1} T'\Sigma^{-1} \hat{F}_{\mathrm{unc},c}(u). \tag{12}$$

## 4. Experiments and Results

Due to the current lack of a dataset for the evaluation of speaker verification in real, noisy, multi-microphone conditions, we adopt the usual approach of distorting clean speech [23–26]. In the following, we use noise and reverberation from Track 1 of the 2nd CHIME Challenge [27] which received significant attention in the robust ASR community [28]. This dataset was recorded in a real domestic environment and it stands out by the attention brought to the realism of the sound scenes.

The speech signals are male conversations in English. For the training stage, 3285 speech signals of 262 speakers from NIST-SRE 2004 and 2005 were used. For the evaluation stage, the *short2* and *short3* datasets of NIST-SRE 2008 were employed, including 470 speech signals for enrollment and 671 speech signals for test. A total of 6615 verifications were performed on the *det7* condition of NIST-SRE.

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.

### 4.1. Speaker recognition system

The speaker recognition front-end consists of 19 Mel frequency cepstral coefficients (MFCCs), the log-energy, and their first and second order derivatives, followed by frame selection using voice activity detection (VAD) and cepstral mean and variance normalization (CMVN) [29]. The UBM consists of 512 Gaussians. For the extraction of i-vectors, a $T$ matrix of dimension 400 is used. The i-vectors are centered, whitened and length normalized, and subsequently projected with 330-dimensional LDA. Classification relies on Gaussian PLDA [30].

### 4.2. Speech enhancement and uncertainty estimation

For speech enhancement, we use the Flexible Audio Source Separation Toolbox (FASST) [31], which has shown state-of-the-art performance on the CHiME data [32]. For uncertainty estimation, similarly to [15], we use oracle uncertainty estimates represented by full uncertainty covariance matrices (UPF) $\bar{\Sigma}_t = (y_t - \bar{y}_t)(y_t - \bar{y}_t)'$ or diagonal uncertainty covariance matrices (UPD) $\bar{\Sigma}_t = \mathrm{diag}(y_t - \bar{y}_t)^2$ [14].

### 4.3. Speaker recognition results

Table 2 shows speaker verification results expressed in equal error rate (EER) and minimum value of the NIST detection cost function (mDCF) [33]. For comparison purposes, the first two rows present the results obtained when training and testing

on clean (original NIST) signals or when training and testing on reverberated signals. The following rows show the results obtained when training on reverberated signals and testing on noisy or enhanced signals. Note that the chosen speech enhancement method reduces noise but not reverberation, hence training on reverberated speech provides better results than training on clean speech.

Table 2: *Speaker recognition results on NIST-SRE.*

| Training set | Evaluation set | EER (%) | mDCF |
|---|---|---|---|
| Clean | Clean | 3.19 | 0.0180 |
| Reverberated | Reverberated | 4.33 | 0.0289 |
| Reverberated | Noisy | 31.85 | 0.0982 |
| Reverberated | Enhanced | 10.48 | 0.0512 |
| Reverberated | UPD | 10.69 | 0.0586 |
| Reverberated | UPF | 10.31 | 0.0511 |

Noise strongly degrades the system performance, reaching 31.85% EER. Multichannel speech enhancement improves the EER to 10.477%, getting closer to reverberated speech. However UP does not exhibit the expected behavior. Oracle full UP improves the EER only by 0.17%, while oracle diagonal UP does not outperform the result obtained without uncertainty.

The lack of improvement brought by UP on NIST-SRE is consistent with [15]. Considering the significant improvement brought by UP on smaller datasets such as YOHO [15] and CHiME-Grid [14], this result is shocking.

### 4.4. Analysis of posterior probabilities and BW statistics

To explain this fact, we evaluate the proposed UP approach independently of the final classification stage so as find at what processing stage the problem arises. We stop the speaker recognition procedure just after the propagation of uncertainty and we measure the impact of UP on the accuracy of the UBM posterior probabilities $\gamma$ and the zeroth order BW statistics $N$. For this, we compare the probabilities $\gamma_{\text{enh}}$ and the statistics $N_{\text{enh}}$ computed from noisy data with enhancement (with or without UP) with the ground-truth probabilities $\gamma_{\text{rev}}$ and statistics $N_{\text{rev}}$ computed from reverberated noiseless data using the following two error metrics:

$$E_\gamma(u) = \frac{\sum_{t=1}^{L}\sum_{c=1}^{C}|\gamma_{\text{enh},t}(c) - \gamma_{\text{rev},t}(c)|}{L} \quad (13)$$

$$E_N(u) = \sqrt{\frac{\sum_{c=1}^{C}(N_{\text{rev},c}(u) - N_{\text{enh},c}(u))^2}{C}} \quad (14)$$

For comparison, we perform this experiment both on the NIST-SRE dataset and on another dataset obtained by distorting clean speech from the WSJ corpus in the same way as above. This dataset involves 7138 utterances from 83 speakers for training and 740 utterances from 18 speakers for evaluation. We did not perform the entire speaker recognition process on WSJ because we just need to check whether UP can denoise the quantities used to compute the i-vectors, such that we can confirm whether the NIST-SRE dataset is the cause of the problem or not.

Table 3 shows the average value of $E_\gamma(u)$ and $E_N(u)$ over all utterances $u$ in the NIST-SRE 2008 *short2* set and in the WSJ development set. As expected, the application of UP reduces the distortion over both $\gamma$ and $N$ on WSJ. The decrease is moderate for UPD and large for UPF, as previously observed in the field of ASR. On NIST-SRE, the distortion over $\gamma$ is comparable to WSJ and it decreases with UP. However, the distortion over $N$ is significantly larger and it is increases with UP.

We conclude that UP consistently improves the estimation of $\gamma$, as previously observed in ASR, but that it does not always improve the estimation of $N$, as required for speaker recognition. This might come as a surprise since $N$ is the result of the summation of $\gamma$ over time, hence improved estimation of $\gamma$ should result in improved derivation of $N$. In the next section, we analyze the cause of this surprising result.

Table 3: *$E_\gamma$ and $E_N$ for different UBM decoding strategies.*

| | NIST-SRE | | | WSJ | | |
|---|---|---|---|---|---|---|
| | **Enh.** | **UPD** | **UPF** | **Enh.** | **UPD** | **UPF** |
| $E_\gamma$ | 1.19 | 1.10 | 0.73 | 1.53 | 1.32 | 0.67 |
| $E_N$ | 5.29 | 7.24 | 5.62 | 2.45 | 1.74 | 0.65 |

## 5. The case of NIST-SRE

Figure 1 shows the average value of $N(u)$ over all utterances of the NIST-SRE 2008 *short2* set and the WSJ development set, for three types of speech: reverberated speech (ground truth), enhanced speech, and enhanced speech with diagonal UP. In order to better see the behavior of the curves, the figure also shows a zoomed section of the plot.

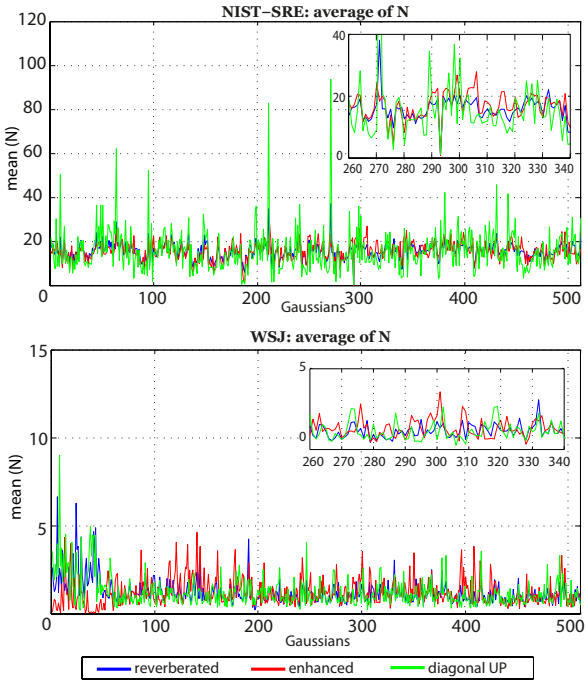


Figure 1: *Average of $N(u)$ over all test utterances.*

### 5.1. Analysis for individual Gaussian components

On the one hand, it is noticeable from the global plot that the UP curve for NIST-SRE and the curves for WSJ exhibit several narrow peaks and valleys. Since $N(u)$ consists of the accumulation of the Gaussian posterior probabilities over all time frames, these peaks/valleys indicate that the corresponding Gaussians

are highly probable/improbable. This may reflect a *hubness* effect [34]. Hubness is a fundamental issue in machine learning due to the concentration of distances in high dimensional spaces. This implies that certain points called "hubs" have a small distance to many other points while certain points called "anti-hubs" are far from all other points. This phenomenon has been mathematically studied for $l^p$ norms [35, 36] and empirically observed for other forms of "distances" including Gaussian likelihoods [37]. One possible explanation for the opposite behavior of $E_\gamma(u)$ and $E_N(u)$ on NIST-SRE may therefore be that the acoustical distortion systematically biases the Gaussian posterior probabilities toward certain Gaussian components.

On the other hand, note how in the zoomed plot the UP curve for NIST-SRE exhibits significant deviations below and above the other two curves, while on WSJ the three curves follow each other more closely. Another possible explanation may therefore be that the acoustical distortion biases to a similar extent the Gaussian posterior probabilities of all Gaussian components.

In order to find out which of these two explanations is correct, we quantify the hubness effect and analyze the internal structure of the UBM for both NIST-SRE and WSJ.

### 5.2. Analysis of hubness effect

For quantifying the hubness, we compute for each Gaussian the number of frames in which it appears among the $k$ best and we measure the skewness of the resulting histogram [38]. The larger the skewness, the stronger the hubness. Table 4 confirms that the hubness increases when UP is applied, but this increase is observed for both datasets. Furthermore, the hubness is larger in WSJ than in NIST-SRE, even without UP. Hence hubness is not the main cause of the increase of the distortion on $N$ due to UP on NIST-SRE.

Table 4: *Measured hubness.*

| | **NIST-SRE** | | | **WSJ** | | |
|---|---|---|---|---|---|---|
| $k$ | **Rev.** | **Enh.** | **UPD** | **Rev.** | **Enh.** | **UPD** |
| 1 | 1.748 | 0.986 | 3.5817 | 4.081 | 2.559 | 3.8914 |
| 3 | 1.219 | 0.618 | 2.8015 | 3.428 | 1.975 | 3.2286 |
| 5 | 0.869 | 0.421 | 2.2121 | 2.996 | 1.742 | 2.8450 |

### 5.3. Analysis of UBM overlap

For discovering the organization of the Gaussians inside the UBM, we used a measure of the amount of overlap between two normal distributions, by means of the Bhattacharyya distance ($bt$) [22]. It is equal to 0 when the distributions are equal and tends to grow up when the distributions move apart. We computed $bt$ for all unique pairs of Gaussians $(c_1, c_2)$ inside the UBM with $c_1 > c_2$. Figure 2 shows the histogram of $bt$ obtained for each dataset. The range of values, as represented by the 5-th and the 95-th percentile to avoid outliers, is also depicted.

We can see that $bt$ is much smaller on NIST-SRE than on WSJ. Moreover is remarkable that for small values of $bt$, for example $bt \leq 5$, the NIST-SRE histogram already contains several pairs of Gaussians, in the order of $10^4$, while that of WSJ just has a hundred pairs. Overall, this indicates that the Gaussians in the UBM learned from NIST-SRE are strongly overlapped, while the Gaussians in the UBM learned from WSJ are further apart.

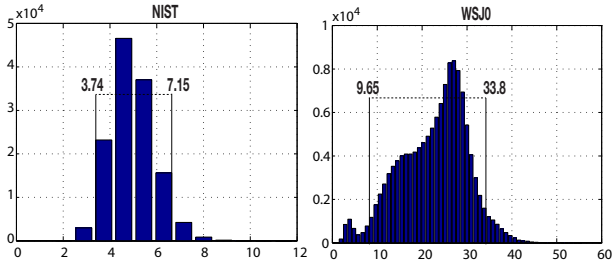


Figure 2: *Distribution of the overlap between Gaussians.*

Table 5: *Degree of overlap between Gaussians.*

| | **Dataset** | **Range of** $bt$ | **Pairs** $bt \leq 2$ |
|---|---|---|---|
| **Enh.** | **NIST-SRE** | 3.74 - 7.15 | 0.0232% |
| | **WSJ** | 9.65 - 33.8 | 0.0079% |
| **UPD** | **NIST-SRE** | 2.26 - 5.14 | 0.1958% |
| | **WSJ** | 5.47 - 19.64 | 0.0173% |

Table 5 compares the degree of overlap before and after UP. In order to measure the degree of overlap after UP, the uncertainty covariance is added to all UBM Gaussians in each time frame as in (7), the Bhattacharyya distance is measured and averaged over all time frames. As overlap indicators, the 5-th and 95-th percentiles and the percentage of Gaussians pairs with high overlap ($bt \leq 2$) are employed.

Table 5 shows that UP increases the Gaussian overlap for NIST-SRE, because the average distance is decreased and the amount of Gaussians with high overlap is increased. UP also increases the Gaussian overlap for WSJ, but ultimately results in much less overlap than NIST-SRE. Since the UBM of NIST-SRE is originally very overlapped, when the feature uncertainty is propagated to the posterior probability of each mixture component, the overlap becomes huge and causes numerous errors in the posterior probability matrix. By contrast, the UB of WSJ is reasonably sparse from the start, such that when UP is applied the overlap does not become a problem.

## 6. Conclusions and future work

In this paper we proposed a new method for propagating the uncertainty due to noise from the acoustic features to the i-vectors in a speaker recognition framework. Preliminary experiments showed that the proposed method yielded little or no improvement on NIST-SRE, confirming the previous attempt of [15]. Studying the results, we found that there is a huge overlap between the Gaussians of the UBM created from NIST-SRE. This overlap is increased by UP, which adversely affects the performance of the system. This phenomenon does not arise for the smaller WSJ dataset. We attribute it to the big size and the high speaker variability of the NIST-SRE dataset. This study suggests that addressing the noise robustness problem on NIST-SRE and other large datasets will require solving the Gaussian overlap problem first, e.g., by renormalizing the Gaussian log-likelihoods. Beyond our study, further analysis of the causes and the solutions to this problem must be conducted.

## 7. Acknowledgements

This work has been partly realized thanks to the support of the Région Lorraine and the CPER MISN TALC project.

# 8. References


[1] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Wiley, 2009.

[2] L. Deng, "Front-end, back-end, and hybrid techniques for noise-robust speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*. Springer, 2011, pp. 67–99.

[3] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[4] K. S. Rao and S. Sarkar, *Robust Speaker Recognition in Noisy Environments*. Springer Science+Business Media, 2014.

[5] J. A. Arrowood and M. A. Clements, "Using observation uncertainty in HMM decoding," in *Interspeech*, Denver, Colorado, 2002, pp. 1561–1564.

[6] L. Deng, J. Wu, J. Droppo, and A. 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.

[7] M. Delcroix, T. Nakatani, and S. Watanabe, "Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 324–334, Jan 2009.

[8] R. F. Astudillo, "Integration of short-time Fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition," Ph.D. dissertation, Von der Fakultät IV Elektrotechnik und Informatik der Technischen Universität Berlin, 2010.

[9] D. Kolossa and R. Haeb-Umbach, *Robust speech recognition of uncertain or missing data*. Springer, 2011.

[10] A. H. Abdelaziz, S. Zeiler, D. Kolossa, V. Leutnant, and R. Haeb-Umbach, "GMM-based significance decoding," in *ICASSP*, Vancouver, Canada, 2013, pp. 6827–6831.

[11] D. Tran, E. Vincent, and D. Jouvet, "Extension of uncertainty propagation to dynamics MFCCs for noise robust ASR," in *ICASSP*. Florence, Italy: IEEE, 2014, pp. 5507–5511.

[12] ——, "Fusion of multiple uncertainty estimators and propagators for noise robust ASR," in *ICASSP*. Florence, Italy: IEEE, 2014, pp. 5512–5516.

[13] X. Zhao, Y. Dong, J. Zhao, L. Lu, J. Liu, and H. 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.

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

[15] C. Yu, G. Liu, S. Hahm, and J. H. L. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *ICASSP*. Florence, Italy: IEEE, 2014.

[16] V. Hautamäki, Y.-C. Cheng, P. Rajan, and C.-H. Lee, "Minimax i-vector extractor for short duration speaker verification," in *Interspeech*. Lyon, France: ISCA, 2013, pp. 3708–3712.

[17] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text dependent speaker recognition using PLDA with uncertainty propagation," in *Interspeech*. Lyon, France: ISCA, 2013, pp. 3684–3688.

[18] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alan, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *ICASSP*. Vancouver, BC: IEEE, 2013, pp. 7649–7653.

[19] D. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Processing Letters*, vol. 2, no. 3, p. 4648, 1995.

[20] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, p. 14351447, 2007.

[21] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

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

[23] A. El-Solh, A. A. Cuhadar, and R. A. Goubran, "Evaluation of speech enhancement techniques for speaker identification in noisy environments," in *Proc. ISMW*, 2007.

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

[25] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *ICASSP*, Kyoto, Japan, 2012, pp. 4253–4256.

[26] D. Martinez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny, and E. Lleida, "Unscented transform for ivector-based noisy speaker recognition," in *ICASSP*. Florence, Italy: IEEE, 2014, pp. 4042–4046.

[27] E. Vincent, J. Barker, S. Watanabe, J. L. Roux, F. Nesta, and M. Matassoni, "The second 'CHIME' speech separation and recognition challenge: Datasets, tasks and baselines," in *ICASSP*. IEEE, 2013, pp. 126–130.

[28] ——, "The second 'CHIME' speech separation and recognition challenge: An overview of challenge systems and outcomes," in *ASRU*, 2013, pp. 162–167.

[29] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Berlin, Germany: Springer-Verlag, 2008.

[30] D. Garcia and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*. Florence, Italy: ISCA, 2011, pp. 249–252.

[31] 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.

[32] Y. Salaün, E. Vincent, N. Bertin, N. Souviraà-Labastie, X. Jaureguiberry, D. Tran, and F. Bimbot, "The flexible audio source separation toolbox version 2.0," in *Show and Tell of ICASSP*. Florence, Italy: IEEE, 2014.

[33] NIST: National Institute of Standards and Technology. [Online]. Available: http://www.nist.gov/speech

[34] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in space: Popular nearest neighbors in high dimensional data," *Journal of Machine Learning Research*, vol. 11, pp. 2487–2531, 2010.

[35] C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *ICDT*, 2001, pp. 420–434.

[36] D. François, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 873–886, 2007.

[37] D. Schnitzer, A. Flexer, and J. Schlüter, "The relation of hubs to the Doddington zoo in speaker verification," in *EUSIPCO*, 2013, pp. 1–5.

[38] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer, "Local and global scaling reduce hubness in space," *Journal of Machine Learning Research*, vol. 13, pp. 2871–2902, 2012.