An improved uncertainty propagation method for robust i-vector based speaker recognition

Dayana Ribas, Emmanuel Vincent

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 insufficient 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

$$M(u) = m + Tw(u) + \epsilon(u) \tag{1}$$

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:

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

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

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

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

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 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

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

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]

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

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]

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

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

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

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

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

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

$$V_{\text{unc},c,t} = V_c + \bar{\Sigma}_t. \tag{11}$$

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 $\bar{y}_t$:

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

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]

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

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