UNCERTAINTY PROPAGATION IN FRONT END FACTOR ANALYSIS FOR NOISE ROBUST SPEAKER RECOGNITION

Chengzhu Yu, Gang Liu, Seongjun Hahm, John H. L. Hansen

Center for Robust Speech Systems (CRSS)
University of Texas at Dallas, Richardson, TX 75080
{chengzhu.yu, gang.liu, seongjun.hahm, john.hansen}@utdallas.edu

ABSTRACT

In this study, we explore the propagation of uncertainty in the state-of-the-art speaker recognition system. Specifically, we incorporate the uncertainty associated with observation features into the i-Vector extraction framework. To prove the concept, both the oracle and practically estimated uncertainty are used for evaluation. The oracle uncertainty is calculated assuming the knowledge of clean speech features, while the estimated uncertainties are obtained using SPLICE and joint-GMM based methods. We evaluate the proposed framework on both YOHO and NIST 2010 Speaker Recognition Evaluation (SRE) corpora by artificially introducing noise at different SNRs. In the speaker verification experiments, we confirmed that the proposed uncertainty based i-Vector extraction framework shows significant robustness against noise.

Index Terms— robust speaker recognition, uncertainty propagation, i-Vector

1. INTRODUCTION

Speaker recognition systems based on i-Vector extraction and PLDA classifier are able to obtain relatively high accuracy under clean as well as channel mismatched conditions [1–8]. However, performance degrades dramatically in the presence of background noise [9–12]. As the real world applications of speaker recognition system often involve various degree of environmental noise [13], developing noise robust speaker recognition systems are of great importance.

In a conventional i-Vector extraction framework, the i-Vector of the given test utterance is computed as the conditional expectation of i-Vector distribution given the observation features [14]. When those observation features are corrupted by noise, the extracted i-Vectors become unreliable as well. A number of studies have been proposed to compensate this effect either by removing the noise or estimating the clean features prior to the i-Vector extraction [15, 16]. While these enhancement methods could improve the robustness of the speaker recognition system in noise, the perfect estimation of clean speech features are always not achievable. Moreover, the estimation errors of those enhancement methods are unequally distributed, causing certain features to be less reliable than others. Therefore, it is beneficial to quantify the reliability associated with those features and incorporate them in the speaker recognition system.

Previous studies have investigated the incorporation of acoustic feature uncertainty in traditional GMM-UBM based systems [17, 18]. The uncertainty associated with i-Vector representations have also been studied for propagating in back-end classifiers [19–22]. However, to the best of our knowledge no studies have yet attempted the propagation of uncertainty in the front end i-Vector extracting process. The purpose of this study is to derive an uncertainty modified i-Vector extraction framework to (1) make the i-Vector extraction system focusing on the reliable or reliably enhanced features, and (2) to further deliver the uncertainty of features to the back-end classifier which has not been achievable due to the front end factor analysis process.

In Sec. 2, we present a short overview of the conventional i-Vector extraction framework. Sec. 3 contains the derivation of the proposed uncertainty modified i-Vector extraction system. In Sec. 4, we present results to show the effectiveness of the proposed framework.

2. FRONT END FACTOR ANALYSIS

In this section, we present a short overview of the conventional i-Vector extraction framework. In a conventional i-Vector extraction framework, speaker and channel dependent GMM supervector is modeled as follows:

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

where m is the supervector obtained from the universal background model (UBM), T is the low rank total variability matrix representing the basis of reduced total variability space, and w is the low rank factor loadings referred to as i-Vectors.

The estimation of the total variability matrix T employs expectation maximization (EM) method as described in [14]. After training the total variability matrix, the i-Vector of given speech utterance is extracted as the conditional expectation of i-Vector distribution given observation features.

$$w_s^* = E[P(w_s|X_s)], \tag{2}$$

where w_s^* is the i-Vector of the given speech utterance s, X_s is the clean observation features, $P(w_s|X_s)$ is the conditional distribution of the i-Vector given observation features, and $E[\cdot]$ indicates the expectation. Finally, the i-Vector of the given speech utterance can be represented using the Baum-Welch zeroth (N_s) and centralized first (F_s) order statistics,

$$w_s^* = (T' N_s \Sigma^{-1} T + I)^{-1} T \Sigma^{-1} F_s, \tag{3}$$

where Σ is the covariance matrix obtained from UBM model and I is the identity matrix.

3. UNCERTAINTY PROPAGATION

3.1. Uncertainty modified i-Vector extraction

In this section, we will derive the propagation of feature uncertainty in i-Vector extraction process. The training of total variability matrix is the same as in conventional i-Vector extraction framework. When the testing speech utterances are corrupted by noise, the conventional i-Vector extraction framework can be written as follow:

$$w_s^* = E[P(w_s|Y_s)], \tag{4}$$

where Y_s is the noise corrupted features.

We can rewrite the conditional distribution of i-Vector in Eq. 4 as,

$$\begin{aligned} P(w_s|Y_s) &\propto P(y_1, y_2, ..., y_n|w_s)P(w_s), \\ &= \left\{ \prod_{t=1}^{n} P(y_t|w_s) \right\} \mathcal{N}(w_s; 0, I), \end{aligned} \tag{5}$$

where y_t is the corrupted feature vector at each time frame of t, $P(w_s)$ is the prior distribution of i-Vector assumed to be $\mathcal{N}(w_s; 0, I)$, n is the total frame number of given speech utterance, and $P(y_t|w_s)$ is the posterior probability of the corrupted feature at tth time frame.

As the i-Vectors of enrollment model are extracted from clean features x_t, mismatch occurs when the testing i-Vectors are extracted from corrupted features y_t. The classical methods for compensating this mismatch are to perform the point estimate of the clean features using either noise removal or feature compensation methods. The estimated clean speech features are then used for i-Vector extraction with the simple assumption that the enhanced features $\hat{x}_t$ are equal to its clean correspondence x_t.

However, the noise removal or feature compensation methods can never be perfect. Therefore, a more rigorous approach is to generate the joint probability of clean and observed features and then marginalizing over all possible hidden clean speech features [23]. Hence, $P(y_t|w_s)$ in Eq. 5 can be rewritten as,

$$P(y_t|w_s) = \int_{-\infty}^{+\infty} P(y_t, x|w_s)dx, \tag{6}$$

$$\approx \int_{-\infty}^{+\infty} P(y_t|x)P(x|w_s)dx, \tag{7}$$

where,

$$P(y_t|x) = \mathcal{N}(x; \hat{x}_t, \sigma_t^2), \tag{8}$$

$$P(x|w_s) = \mathcal{N}(x; m + Tw_s, \Sigma). \tag{9}$$

In Eq. 8, $\hat{x}_t$ can be interpreted as the conventionally enhanced speech feature, and σ_t^2 as the uncertainty associated with those features. Whereas various speech enhancement algorithms can be modified to output σ_t^2 along with $\hat{x}_t$, only SPLICE [23] and Joint-GMM based [24] uncertainty estimation (Sec. 3.3) is used in this study. Note that Eq. 9 is referred directly from Eq. 1 of conventional i-Vector derivation.

After replacing Eq. 7 with Eqs. 8 and 9, the posterior probability of observed features can be written as follow:

$$\begin{aligned} &P(y_t|w_s) \\ &= \int_{-\infty}^{+\infty} \mathcal{N}(x; \hat{x}_t, \sigma_t^2)\mathcal{N}(x; m + Tw_s, \Sigma)dx \\ &= \int_{-\infty}^{+\infty} \mathcal{N}(\hat{x}_t; x, \sigma_t^2)\mathcal{N}(x; m + Tw_s, \Sigma)dx \\ &= \mathcal{N}(\hat{x}_t; m + Tw_s, \Sigma + \sigma_t^2). \end{aligned} \tag{10}$$

Finally, the conditional distribution of i-Vector given observation features can be derived as,

$$\begin{aligned} P(w_s|Y_s) &\propto \left\{ \prod_{t=1}^{n} P(y_t|w_s) \right\} \mathcal{N}(w_s; 0, I), \\ &= \left\{ \prod_{t=1}^{n} \mathcal{N}(\hat{x}_t; m + Tw_s, \Sigma + \sigma_t^2) \right\} \mathcal{N}(w_s; 0, I) \\ &= \mathcal{N}(w_s; U, V), \end{aligned} \tag{11}$$

where,

$$U = (T'PT + nI)^{-1} T \Sigma^{-1} Q, \tag{12}$$

$$V = (T'PT + nI)^{-1} T, \tag{13}$$

with

$$P = \sum_t N_t (\Sigma + \sigma_t^2)^{-1}, \tag{14}$$

$$Q = \sum_t N_t (\Sigma + \sigma_t^2)^{-1} (\hat{x}_t - m), \tag{15}$$

and N_t as the diagonal concatenation of the posterior probability of each mixture component $P(c|\hat{x}_t)$. In this work, we use the conditional expectation of i-Vector distribution as the point estimate of the true i-Vector in a back-end classification system. However, a more rigorous approach will be using the above derived i-Vector distribution for the classification purpose as in [2].

3.2. Relation to conventional i-Vector system

As a special case, if the input uncertainty values of all features are zero such that $\sigma_t^2 = 0$ in Eqs. 14 and 15, then the uncertainty modified i-Vector system becomes exactly the same as the conventional i-Vector system. Alternatively, if the uncertainties of certain features are higher than others, then the contribution of that feature in calculating the i-Vector distribution is reduced.

3.3. Uncertainty estimation

The proposed uncertainty modified i-Vector extraction framework requires the information of feature uncertainty as an input. While this information is not directly obtainable, a number of noise removal and feature compensation methods [23, 25] have been modified to output the uncertainty of features along with enhanced features. Those uncertainty estimation methods combined with uncertainty decoding [23] or modified imputation [26] have achieved much success in the area of robust automatic speech recognition (ASR). As the focus of this study is the propagation of uncertainty rather than estimation, we borrowed two well known uncertainty estimation methods: SPLICE uncertainty estimation [23] and joint-GMM based uncertainty estimation [24], for evaluating the proposed uncertainty propagation framework for practical applications.

3.3.1. SPLICE Uncertainty Estimation

With SPLICE uncertainty estimation, the conditional distribution $P(y_t|x)$ in above Eqs. 7 and 8 is obtained as follow:

$$P(y|x) = \mathcal{N}(x; \hat{x}_k, \sigma^2_{\hat{x}_k}), \tag{16}$$

where

$$\hat{x}_k = \frac{\bar{\Sigma}_x^2(y + r_k) - \Gamma_k^2 \bar{\mu}_x}{\bar{\Sigma}_x^2 - \Gamma_k^2}, \tag{17}$$

$$\sigma^2_{\hat{x}_k} = \frac{\bar{\Sigma}_x^2 \Gamma_k^2}{\bar{\Sigma}_x^2 - \Gamma_k^2}, \tag{18}$$

where k is the index of Gaussian component having the highest posterior probability, cepstral compensation vector r_k and its covariance Γ_k are trained from stereo data using a maximum likelihood criterion, and $P(x)$ is obtained with the assumption of a single Gaussian distribution $\mathcal{N}(\bar{\mu}_x, \bar{\Sigma}_x)$. The obtained $P(y|x)$ in Eq. 16 will be used in the proposed uncertainty modified i-Vector extraction.

3.3.2. Joint Uncertainty Estimation

In joint uncertainty estimation, the conditional distribution $P(y_t|x)$ in Eqs. 7 and 8 is derived from the joint distribution of clean and noisy features,

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} \right). \tag{19}$$

The derived conditional distribution can be expressed as follows:

$$P(y|x) \approx \mathcal{N}(A^{(k)}y + b^{(k)}, \Sigma_b^{(k)}), \tag{20}$$

where

$$A^{(k)} = \Sigma_x^{(k)} {\Sigma_{yx}^{(k)}}^{-1}, \tag{21}$$

$$b^{(k)} = \mu_x^{(k)} - A^{(k)} \mu_y^{(k)}, \tag{22}$$

$$\Sigma_b^{(k)} = A^{(k)} \Sigma_y^{(k)} A'^{(k)} - \Sigma_x^{(k)}, \tag{23}$$

and k denotes the Gaussian component having the maximum posterior probability. As in SPLICE uncertainty estimation, the $P(y|x)$ obtained in Eq. 20 will be used in the proposed uncertainty modified i-Vector extraction.

4. EXPERIMENTS AND RESULTS

We evaluate the proposed uncertainty modified i-vector extraction method on noisy database created by artificially adding noise to YOHO and male part of NIST SRE 2010 telephone condition (condition 5).

4.1. System description

The baseline systems for all experiments are composed of 36-dimension feature vectors (12 MFCC + Δ + $\Delta\Delta$) extracted using a 25 ms window with 10 ms shift and normalized using a 3-s sliding window. In all experiments, voice activity detection (VAD) is applied before noise addition in order to evaluate the proposed system independent of the VAD quality. For experiments on the noised NIST SRE 2010 corpus, we used 1024-component diagonal covariance universal background models (UBM), 400 dimension i-Vector trained from Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST 2004, 2005, 2006 SRE enrollment data. The dimensionality is reduced to 200 by LDA, followed by length normalization and PLDA. For experiments on the noised YOHO database, a 512-component UBM and 80 dimension i-Vector extractor are used. The UBM and total variability matrix was trained with the YOHO database.

4.2. Benchmarking with oracle uncertainty

In order to see how much improvement we can expect by using uncertainty modified i-Vector extraction, we performed experiments using the oracle uncertainty on the noised NIST SRE 2010 database. This noisy version of NIST SRE2010 database is created by artificially adding babble noise at different SNRs assuming that the original NIST SRE10 data is clean. The babble noise is taken from the NOISEX database. In this experiment, we define the oracle uncertainty as the magnitude-squared error between noisy observation features and its clean correspondence $\sigma_t^2 = (y_t - x_t)^2$ [23]. The oracle uncertainty of features is passed into the proposed i-Vector extraction system along with unprocessed noisy features. Our purpose is to evaluate the proposed i-vector extraction method independent of any enhancement and uncertainty estimation strategy.

The results of the experiment are listed in Table 1. The results indicate that the uncertainty modified i-Vector extraction framework has the potential of achieving up to 10∼20% relative improvement over the state-of-the-art i-Vector based speaker recognition system.

NIST SRE 2010					
	clean	10 dB	5 dB	0 dB	-5 dB
i-Vector	2.07	5.70	12.08	22.17	35.45
i-Vector_U	-	4.56	10.15	20.11	32.05

Table 1. Comparison of the performance (EER %) of conventional i-vector extraction system (i-Vector) and the proposed uncertainty modified i-Vector extraction system with oracle uncertainty (i-Vector_U).

4.3. Experiment with estimated uncertainty

In the above experiment, an oracle uncertainty is used for benchmarking the proposed uncertainty propagation framework. However, in real applications the oracle uncertainty is not achievable. In order to see the performance of the proposed methods in more practical situations, we use the uncertainty estimated from SPLICE and joint-GMM based method for propagation. The noisy version of YOHO database is used in this experiment. The YOHO database consists of 138 speakers with 30 female and 108 male speakers. For each speaker, there are 4 enrollment sessions (each contains 24 phrases) and 10 verification sessions. The noisy version of YOHO database is created by artificially adding babble noise at different SNRs.

In this experiment, we consider the i-Vector extraction system without any feature processing, the i-Vector extraction system with SPLICE and joint-GMM based feature enhancement, and the uncertainty modified i-Vector extraction system using SPLICE and joint-GMM based uncertainty estimation. The result is listed in Table 2. The result shows that the uncertainty modified i-Vector extraction system consistently performs better than the conventional i-Vector systems with both unprocessed and enhanced features. This confirms the viability of incorporating uncertainty propagation in modeling for i-Vector SID systems.

5. DISCUSSION

In this study, we considered the propagation of acoustic feature uncertainty in the state-of-the-art i-Vector extraction system. The proposed uncertainty modified i-Vector extraction framework was tested using both oracle and practically estimated uncertainty. Our experiment on the noisy version

YOHO				
	clean	10 dB	5 dB	0 dB
UN	1.61	7.62	16.05	25.02
SPLICE	-	6.51	14.53	24.07
SPLICE_U	-	5.82	13.15	23.01
Joint-GMM	-	4.71	9.83	18.47
Joint-GMM_U	-	4.20	8.25	17.65

Table 2. Performance (EER %) of speaker recognition system without any feature processing (UN), using SPLICE for feature enhancement (SPLICE), using SPLICE for proposed uncertainty propagation (SPLICE_U), using joint-GMM method for feature enhancement (Joint-GMM), using joint-GMM method for proposed uncertainty propagation (Joint-GMM_U) .

of NIST SRE 2010 using oracle uncertainty indicated that the proposed method has a great potential for improving the robustness of speaker recognition especially in low SNR conditions. In addition, the experiment using SPLICE and joint uncertainty estimation methods showed that the proposed method achieves improved recognition results in practical situations as well.

In all experiments, we used a clean VAD in order to evaluate the proposed algorithm independent of VAD performance. However, it was observed that features from non-speech segment have significant amount of uncertainty. As the proposed method could effectively reduce the contribution of those unreliable features on extracted i-Vectors, we expect that the proposed uncertainty modified i-Vector extraction method could show further improvement over conventional i-Vector extraction framework when VAD quality drops in realistic noisy environments.

The proposed uncertainty propagation framework does not have any reliance on SNR level when oracle uncertainty is used as in Sec 3.2. However, the results in Table 2 shows higher relative improvements when SNR is higher. This is due to the fact that the quality of estimated uncertainty drops as the SNR gets lower [24].

Whereas the proposed uncertainty propagation have shown its robustness to noise on both NIST SRE and YOHO database, the stereo based uncertainty estimation methods employed in this study does not provide acceptable uncertainty estimation on noisy version of NIST SRE 2010 according to our initial experiment. We attribute this to their strong reliance on high quality clean speech for training, We are currently investigating the uncertainty estimation methods that are suitable for large and complex corpora such as NIST SRE.

Acknowledgments

This research was supported by National Science Foundation (NSF) under Grant 1219130.

6. REFERENCES

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

[2] Patrick Kenny, "Bayesian speaker verification with heavy tailed priors," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[3] Taufiq Hasan, Seyed Omid Sadjadi, Gang Liu, Navid Shokouhi, Hynek Boril, and John HL Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation," in *Proc. of ICASSP*, 2013, pp. 6784–6786.

[4] Gang Liu, Taufiq Hasan, H Boril, and JH Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *Proc. of ICASSP*, 2013, pp. 7755–7759.

[5] Taufiq Hasan and John HL Hansen, "Factor analysis of acoustic features using a mixture of probabilistic principal component analyzers for robust speaker verification," *Proc. Odyssey, Singapore*, 2012.

[6] R Saeidi, KA Lee, T Kinnunen, T Hasan, B Fauve, PM Bousquet, E Khoury, PL Sordo Martinez, JMK Kua, CH You, et al., "I4u submission to nist sre 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. Interspeech*, 2013.

[7] Jun-Won Suh, Seyed Omid Sadjadi, Gang Liu, Taufiq Hasan, Keith W Godin, and John HL Hansen, "Exploring hilbert envelope based acoustic features in i-vector speaker verification using htplda," in *Proc. of NIST 2011 Speaker Recognition Evaluation Workshop*, 2011.

[8] Taufiq Hasan, Gang Liu, Seyed Omid Sadjadi, Navid Shokouhi, H Bořil, A Ziaei, Abhinav Misra, KW Godin, and JHL Hansen, "Utd-crss systems for 2012 nist speaker recognition evaluation," in *Proc. NIST SRE Workshop*, 2012.

[9] Miranti Indar Mandasari, Mitchell McLaren, and David A van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *Proc. of ICASSP*, 2012, pp. 4249–4252.

[10] Yun Lei, Lukáš Burget, and Nicolas Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *Proc. of ICASSP*, 2013, pp. 6788–6791.

[11] Daniel Garcia-Romero, Xinhui Zhou, and Carol Y Espy-Wilson, "Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. of ICASSP*, 2012, pp. 4257–4260.

[12] Gang Liu, Yun Lei, and John HL Hansen, "Robust feature front-end for speaker identification," in *Proc. of ICASSP*, 2012, pp. 4233–4236.

[13] Abhijeet Sangwan, Lakshmish Kaushik, Chengzhu Yu, John HL Hansen, and Douglas W Oard, "Houston, We have a solution: Using NASA Apollo Program to advance Speech and Language Processing Technology," in *Proc. of INTERSPEECH*, 2013, pp. 1135–1139.

[14] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[15] Keith W Godin, Seyed Omid Sadjadi, and John HL Hansen, "Impact of noise reduction and spectrum estimation on noise robust speaker identification," in *Proc. of INTERSPEECH*, 2013, pp. 3656–3660.

[16] Seyed Omid Sadjadi and John HL Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions.," in *Proc. of INTERSPEECH*, 2010, pp. 2138–2141.

[17] Yang Shao, Soundararajan Srinivasan, and DeLiang Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. of ICASSP*, 2007.

[18] Alexey Ozerov, Mathieu Lagrange, and Emmanuel Vincent, "Uncertainty-based learning of acoustic models from noisy data," *Computer Speech & Language*, 2012.

[19] Padmanabhan Rajan, Tomi Kinnunen, and Ville Hautamäki, "Effect of multicondition training on i-vector PLDA configurations for speaker recognition," in *Proc. of INTERSPEECH*, 2013, pp. 3694–3698.

[20] Ville Hautamäki, You-Chi Cheng, Padmanabhan Rajan, and Chin-Hui Lee, "Minimax i-vector extractor for short duration speaker verification," in *Proc. of INTERSPEECH*, 2013, pp. 3708–3712.

[21] T Stafylakis, P Kenny, P Ouellet, J Perez, M Kockmann, and P Dumouchel, "Text-dependent speaker recognition using plda with uncertainty propagation," in *Proc. of INTERSPEECH*, pp. 3684–3688.

[22] Xianyu Zhao and Yuan Dong, "Variational bayesian joint factor analysis models for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 1032–1042, 2012.

[23] Jasha Droppo, Alex Acero, and Li Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proc. of ICASSP*, 2002.

[24] Hank Liao and MJF Gales, "Joint uncertainty decoding for noise robust speech recognition.," in *Proc. of INTERSPEECH*, 2005, pp. 3129–3132.

[25] Ramon Fernandez Astudillo, Dorothea Kolossa, Philipp Mandelartz, and Reinhold Orglmeister, "An uncertainty propagation approach to robust ASR using the ETSI advanced front-end," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 824–833, 2010.

[26] Dorothea Kolossa, Aleksander Klimas, and Reinhold Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, 2005, pp. 82–85.