An Adaptive i-Vector Extraction for Speaker Verification with Short Utterance

Arnab Poddar[1(✉)], Md Sahidullah[2], and Goutam Saha[1]

[1] Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India
arnabpoddar@iitugp.ac.in

[2] School of Computing, University of Eastern Finland, Joensuu, Finland

Abstract. A prime challenge in automatic speaker verification (ASV) is to improve performance with short speech segments. The variability and uncertainty of intermediate model parameters associated with state-of-the-art i-vector based ASV system, extensively increases in short duration. To compensate increased variability, we propose an adaptive approach for estimation of model parameters. The pre-estimated *universal background model* (UBM) parameters are used for adaptation. The speaker models i.e., i-vectors are generated with the proposed adapted parameters. The ASV performance with the proposed approach considerably outperformed conventional i-vector based system on publicly available speech corpora, NIST SRE 2010, especially in short duration, as required in real-world applications.

Keywords: i-vector · Short utterance · Duration variability · Baum-Welch statistics

1 Introduction

Automatic speaker verification (ASV) is the process of recognizing the identity claimed by a person through speech samples. I-vector based ASV are considered as the state-of-the-art technology for its high performance, low complexity, and easy session/channel compensation. The applications of ASV in important sectors like banking, finance, forensic, defense etc., often constrain the duration of speech data [1,4,7]. The performance of ASV rapidly degrades in short duration [6,7]. To overcome the deficiency with short utterances, the work in [2] has attempted to model variability caused by short duration segments in i-vector domain. The short utterance problem is also addressed in other application of speech processing like language identification where an alternate estimation of i-vector was introduced treating all supervector dimensions with equal in the i-vector modeling [11].

The main challenge for short utterances is the increased intra-speaker variability in estimated parameters due to variability in lexicon and speech duration [3]. Utterance duration is associated with the uncertainty in i-vector point

B.U. Shankar et al. (Eds.): PReMI 2017, LNCS 10597, pp. 326–332, 2017.
https://doi.org/10.1007/978-3-319-69900-4_41

estimation [8]. In i-vector based system, the intermediate parameters *i.e*, Baum Welch statistics, totally represent the extracted features from speech [2,6]. The zero-order BW statistics, i.e., the probabilistic counts, define the co-variance matrix of the posterior distribution given the utterance [8]. An analysis on BW statistics, presented in this work, showed increased intra-speaker variability due to sparse nature of estimated parameters in limited duration. To mitigate the sparsity and uncertainty in the estimated parameters, we have proposed a method of adapting them with information from pre-estimated background model parameters. Consequently, a comparative analysis on uncertainty of conventional and proposed adapted i-vector is presented which showed reduced uncertainty of proposed adapted i-vectors in different duration conditions. Considerable improvement of performance is noted in different duration condition on speaker recognition evaluation (SRE) corpora, NIST SRE 2010. In the rest of the paper, we briefly describe i-vector GPLDA and proposed modifications system in Sects. 2 and 3 respectively. Subsequently, we describe the experimental setup, results in Sect. 4 and draw the conclusion in 5.

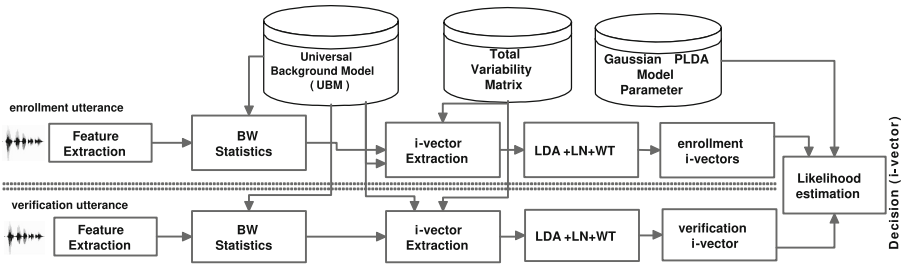

Fig. 1. Block Diagram showing i-vector GPLDA based ASV system.

2 Descriptions of i-Vector ASV System

Figure 1 shows the block diagram of i-vector based ASV system. An i-vector is a fixed-dimensional representation of a speech signal in factor analysis framework [1]. The i-vector ($\mathbf{y}$) decomposes *Gaussian mixture model* (GMM) *supervector* of $s-th$ speaker ($\mu_\mathbf{s}$) into a low-dimensional subspace [1] as, $\mu_s = \bar{\mu} + \mathbf{\Phi}\mathbf{y_s}$, where $\mathbf{\Phi}$ is a low-rank *total variability* (TV) matrix defining the speaker and channel independent space, (μ) is used from GMM universal background model (UBM). The GMM-UBM is mathematically represented as $\lambda_{\text{UBM}} = \{w_i, \bar{\mathbf{m}}_i, \bar{\mathbf{\Sigma}}_i; i = 1, 2, \ldots, C\}$ where C is the Gaussian components, w_i is the prior of i-th component (w_i satisfies $\sum_{i=1}^{C} w_i = 1$), $\mathbf{m}_i$ and $\mathbf{\Sigma}_i$ are the mean and co-variance matrix [9]. The i-vectors are estimated using zeroth and first order BW statistics N_i and $\mathbf{E}_i$, respectively, from an utterance ($\mathbf{X}$) with T frames $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ as,

$$N_i = \sum_{t=1}^{T} P(i|\mathbf{x}_t, \lambda_{UBM}),\ and\ \mathbf{E}_i(\mathbf{X}) = \frac{1}{N_i}\sum_{t=1}^{T} P(i|\mathbf{x}_t, \lambda_{UBM})\mathbf{x}_t, \qquad (1)$$

where P corresponds to the posterior probability of Gaussian component i generating the vector $\mathbf{x}_t$ [1]. The posterior distribution of $\mathbf{E}$, conditioned on the i-vector $\mathbf{y}$ is hypothesized to be $p(\mathbf{E}|\mathbf{y}) = \mathcal{N}(\mathbf{\Phi y}, \mathbf{N}^{-1}\mathbf{\Sigma})$. The i-vectors, *i.e,* MAP estimate of $\mathbf{y}$ is given by

$$\mathbb{E}(\mathbf{y}|\mathbf{E}) = (\mathbf{I} + \mathbf{\Phi}^\top \mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top \mathbf{\Sigma}^{-1}\mathbf{N}(\mathbf{E} - \bar{\mathbf{m}}) \tag{2}$$

The i-vectors are further projected on subspaces to reduce the session and channel variability. For session and channel compensation, we have used widely used Gaussian probabilistic LDA (GPLDA) to compute recognition scores as likelihood ratio [4].

3 Analysis and Proposed Modification

From Eq. 1, it can be shown that N depends on the number of speech frames (T) in the utterance, $\sum_{i=1}^{C} N_i = \sum_{i=1}^{C}\sum_{t=1}^{T} Pr(i|\mathbf{x}_t) = T$ Therefore, N is normalized with T. *Normalized zero-order Baum-Welch statistics* (NBS) for the i-th Gaussian component is shown as, $\tilde{N}_i = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} Pr(i|\mathbf{x}_t)$. We use duration independent NBS for further analysis.

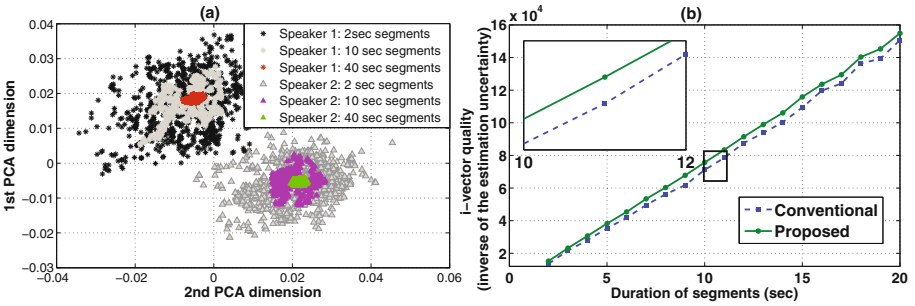

Fig. 2. (a) Scatter plot of PCA projected NBS ($\tilde{N}$) for two speakers. (b) Comparison of i-vector estimation quality as inverse of the estimation uncertainty in conventional and proposed system. The segments of different duration conditions are truncated from a long utterance of a speaker in NIST 2010 corpus.

Scatter plots of principal component (PCA) projected NBS ($\tilde{N}$) for two speakers are shown in Fig. 2(a). 1st two principal components for different truncated segments of 2sec, 10sec and 40sec are shown here. The PCA projection matrix is estimated from 1000 truncated segments from long duration segments of approximately 2.5 min of 2 male speakers from NIST 2008. It can be observed that the NBS show higher variability in short utterances. Larger variation in NBS for short duration condition incorporates higher uncertainty in i-vector

estimation. This can be explained by i-vector estimation, i.e., in Eq. 2. It is known from the theory of i-ivector that the co-variance of the estimated i-vector is defined by $(\mathbf{I} + \mathbf{\Phi}^\top \mathbf{\Sigma}^{-1} \mathbf{N} \mathbf{\Phi})^{-1}$ [5]. For short utterances, $\mathbf{N}$ becomes lower and as a consequence, the uncertainty in i-vector estimation increases. In this work, we attempt to improve the zero-order statistics $\mathbf{N}$ estimation by adapting background model parameters, estimated with sufficiently large speech data. We propose the modified NBS (N_i^{adp})as,

$$N_i^{adp} = T \times [\beta \tilde{N}_i + (1 - \beta) w_i] \, where \, 0 \leq \beta \leq 1 \tag{3}$$

where β controls the adaptation of NBS. Hence the modified i-vector extraction equation is given by:

$$\mathbb{E}(\mathbf{y}|\mathbf{E}) = (\mathbf{I} + \mathbf{\Phi}^\top \mathbf{\Sigma}^{-1} \mathbf{N^{adp}} \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{\Sigma}^{-1} \mathbf{N^{adp}} (\mathbf{E} - \bar{\mathbf{m}}) \tag{4}$$

In Fig. 2(b), we show the comparison of i-vector estimation quality for both conventional and proposed NBS adapted i-vector system in different duration condition using a quality measure based on the i-vector posterior covariance [8]. The posterior distribution of i-vector $\mathbf{y}$ is Gaussian with covariance matrix $\mathbf{y}_\Sigma = (\mathbf{I} + \mathbf{\Phi}^\top \mathbf{\Sigma}^{-1} \mathbf{N} \mathbf{\Phi})^{-1}$ [1,8]. The quality measure $Q(\mathbf{y}_\Sigma)$ is calculated as $Q(\mathbf{y}_\Sigma) = \frac{1}{tr(\mathbf{y}_\Sigma)}$, where $tr(\cdot)$ is the trace operator. Higher value of quality measure $Q(\mathbf{y}_\Sigma)$ indicates lower uncertainty and vice-versa. It compares the quality metric $Q(\mathbf{y}_\Sigma)$ of conventional and proposed i-vectors of segments in different segment duration. For this, the value of adaptation parameter β is kept at 0.5. observations from Fig. 2(b) suggests that the quality metric has improved for the proposed adapted NBS based system over the conventional i-vector based system in different duration condition.

Table 1. Summary of speech corpora used in the experiments.

Specifications	#target model	#test segments	#genuine trials	#imposter trials
NIST 2010	489	351	353	13307
Other specifications: Features and Development parameters				
MFCC	Dimension: 19+19Δ+19$\Delta\Delta$; 20 filterbank, 20ms Hamming window			
GMM-UBM	Dimension: 512; Data: NIST SRE '04, '05, Switchboard II			
TV (Φ) Matrix	Dimension: 400; Data: NIST SRE '04, '05, '06, Switchboard II			
GPLDA	Dimension: 150; Data: NIST SRE '04, '05, '06, Switchboard II			

4 Experimental Results and Discussion

In ASV experiments, we use mel-frequency cepstral coefficients (MFCC) appending delta (Δ) and double-delta ($\Delta\Delta$) coefficients. The non-speech frames are rejected using a voice activity detector (VAD) as in [10]. Subsequently, cepstral

mean and variance normalization (CMVN) is utilized as feature normalization [10]. A gender-specific UBM is trained by *expectation maximization* (EM) algorithm. We carried out the ASV experiments on NIST 2010 *core-core*[1] speaker recognition evaluation (SRE) corpus on the *telephone-telephone* part of male speakers. The summary of the databases, development parameters and features are detailed in Table 1. We truncate the long speech segments in 2 sec (200 active frames), 5 sec (500 active frames), 10 sec (1000 active frames) duration, rejecting prior 500 active speech frames after VAD to diminish phonetic similarity in initial salutation in conversation to avoid text-dependence as in [2]. The ASV performance is observed in *equal error rate* (EER) and *detection cost function* (DCF) [6,10].

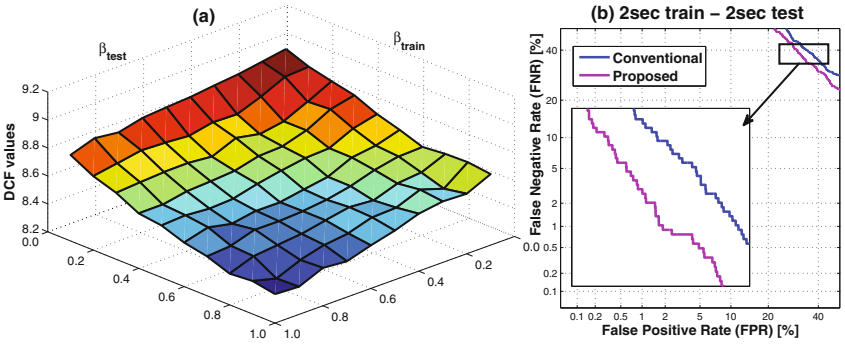

Fig. 3. (a) Surface plot of DCF obtained by varying the adaptation parameter $(\beta_{train}, \beta_{test})$ in *5 sec train - 5 sec testing* condition for in NIST SRE 2010, (b) Detection error trade-off (DET) curve for NIST 2010 *2sec train-2sec test* condition.

In this work, NBS $(\tilde{N}_i)$ is adapted with the information from UBM weight (w_i) to diminish the effect of duration variability (Eq. 3). In order to observe the effect of adaptation parameter (β) on ASV, its value is varied between 0 and 1 in steps of 0.1 for both train and test segments. In Fig. 3(a), the surface plot of the DCF values for $5sec - 5sec$ condition is presented. The blue regions denotes the lower values of DCF indicating optimal operating region of adaptation parameters β_{train} and β_{test} for a particular duration condition. The process is followed for different duration condition separately to estimate optimal β_{train} and β_{test}. In a separate experiments with *full training*$-$*truncated test* condition, the NBS of only truncated segment is adapted. For the 6 different duration conditions, the optimal value of β in Table 2 is shown along with the performance of the conventional and the proposed i-vector based system. The adaptation of NBS improves the performance in different duration conditions. In Fig. 3(b), the detection error trade off (DET) curve for 2sec train-2sec test condition is

[1] https://www.nist.gov/sites/default/files/documents/itl/iad/mig/NIST_SRE10_evalplan-r6.pdf.

Table 2. ASV performance with baseline and proposed i-vector based system on NIST 2010 (*core-core*).

Train-Test Condition	EER[%] (baseline)	EER[%] (proposed)	DCF×100 (baseline)	DCF×100 (proposed)	β_{train}	β_{test}
(a) **Truncated training - Truncated testing**						
2sec-2sec	37.67	**34.27**	9.98	**9.81**	0.7	0.9
5sec-5sec	25.95	**24.07**	9.01	**8.39**	0.8	0.8
10sec-10sec	14.44	**13.31**	6.52	**6.38**	0.3	0.4
(b) **Full training - Truncated testing**						
Full-2sec	21.81	**20.11**	8.52	**8.07**	-	0.9
Full-5sec	12.72	**12.00**	5.51	**5.36**	-	0.6
Full-10sec	7.36	**7.08**	3.72	**3.69**	-	0.3

presented for both conventional and proposed i-vector based system. The results reported in Table 2 are shown for best values of adaptation parameters computed by extensive experimentation. We observe that optimal value of β for different duration condition decreases with the increase in test segment duration. This suggests that adaptation is more effective for short utterances.

5 Conclusion

Considerable ASV performance with limited duration speech is a major requirement for real-world application. We found that the variability of zero-order Baum-Welch statistics and uncertainty associated with the i-vector increases considerably in shorter duration speech. For better estimation of i-vector, we propose adaptation of zero-order statistics using the information from pre-estimated UBM parameter. The proposed approach reduced the uncertainty associated with the i-vector computation. The performance of state-of-the-art ASV system with proposed adaptation has considerably improved especially in short duration condition.

References

1. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
2. Kanagasundaram, A., Dean, D., Sridharan, S., Gonzalez-Dominguez, J., Gonzalez-Rodriguez, J., Ramos, D.: Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. Speech Commun. **59**, 69–82 (2014)
3. Kanagasundaram, A., Vogt, R., Dean, D.B., Sridharan, S., Mason, M.W.: i-vector based speaker recognition on short utterances. In: Proceedings of INTERSPEECH, pp. 2341–2344. ISCA (2011)

4. Kenny, P.: Bayesian speaker verification with heavy-tailed priors. In: The Speaker and Language Recognition Workshop, Odyssey, p. 14. ISCA (2010)
5. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A study of interspeaker variability in speaker verification. IEEE Trans. Audio Speech Lang. Process. **16**(5), 980–988 (2008)
6. Poddar, A., Sahidullah, M., Saha, G.: Performance comparison of speaker recognition systems in presence of duration variability. In: Annual IEEE India Conference (INDICON), pp. 1–6. IEEE (2015)
7. Poddar, A., Sahidullah, M., Saha, G.: Speaker verification with short utterances: a review of challenges. trends and opportunities. In: IET Biometrics (accepted with minor) (2017)
8. Poorjam, A.H., Saeidi, R., Kinnunen, T., Hautamäki, V.: Incorporating uncertainty as a quality measure in i-vector based language recognition, Odyssey, pp. 74–80 (2016)
9. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. **10**(1), 19–41 (2000)
10. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Commun. **54**(4), 543–565 (2012)
11. Van Segbroeck, M., Travadi, R., Narayanan, S.S.: Rapid language identification. IEEE Trans. Audio Speech Lang. Process. **23**(7), 1118–1129 (2015)