Short Utterance Variance Modelling and Utterance Partitioning for PLDA Speaker Verification

Ahilan Kanagasundaram, David Dean, Sridha Sridharan, Clinton Fookes, Ivan Himawan

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
`{a.kanagasundaram, d.dean, s.sridharan, c.fookes, i.himawan}@qut.edu.au`

Abstract

This paper analyses the short utterance probabilistic linear discriminant analysis (PLDA) speaker verification with utterance partitioning and short utterance variance (SUV) modelling approaches. Experimental studies have found that instead of using single long-utterance as enrolment data, if long enrolled-utterance is partitioned into multiple short utterances and average of short utterance i-vectors is used as enrolled data, that improves the Gaussian PLDA (GPLDA) speaker verification. This is because short utterance i-vectors have speaker, session and utterance variations, and utterance-partitioning approach compensates the utterance variation. Subsequently, SUV-PLDA is also studied with utterance partitioning approach, and utterance-partitioning-based SUV-GPLDA system shows relative improvement of 9% and 16% in EER for NIST 2008 and NIST 2010 truncated 10sec-10sec evaluation condition as utterance-partitioning approach compensates the utterance variation and SUV modelling approach compensates the mismatch between full-length development data and short-length evaluation data.

Index Terms: speaker verification, i-vectors, PLDA, SUV, utterance partitioning

1. Introduction

A significant amount of speech is required for speaker model enrolment and verification, especially in the presence of large intersession variability, which has limited the widespread use of speaker verification technology in everyday applications. Reducing the amount of speech required for development, training and testing while obtaining satisfactory performance has been the focus of a number of recent studies in state-of-the-art speaker verification design, including joint factor analysis (JFA), i-vectors, probabilistic linear discriminant analysis (PLDA) and support vector machines (SVM) [1, 2, 3, 4]. Continuous research on this field has been ongoing to address the robustness of speaker verification technologies under such conditions.

Previous research studies had found that long utterance i-vectors contain two source of variation: changing speaker characteristics, and changing channel (or session) characteristics [5]. Recently it was found that short utterance i-vectors vary due to speaker, session and linguistic content (utterance variation) [6]. In typical PLDA speaker verification, a single utterance is used as enrolment data. In this paper, instead of using a single long-utterance as enrolment data, long-enrolment utterances are partitioned into multiple short-enrolment utterances and the average i-vector over the short utterances is used as enrolment data to improve the short utterance speaker verification system. Recently, we have also introduced short utter-ance variance (SUV) modelling to PLDA speaker verification system to compensate the session and utterance variations [7]. Subsequently, in this paper, we also investigate the utterance-partitioning-based Gaussian PLDA (GPLDA) speaker verification with SUV modelling approach.

This paper is structured as follows: Section 2 details the i-vector feature extraction techniques. Section 3 details the short utterance variance added i-vector feature extraction approach. Section 4 explains the GPLDA based speaker verification system. The experimental protocol and corresponding results are given in Section 6 and Section 7. Section 8 concludes the paper.

2. I-vector feature extraction

I-vectors represent the GMM super-vector by a single total-variability subspace. This single-subspace approach was motivated by the discovery that the channel space of JFA contains information that can be used to distinguish between speakers [8]. An i-vector speaker and channel dependent GMM super-vector can be represented by,

$$\boldsymbol{\mu} \quad = \quad \mathbf{m} + \mathbf{Tw}, \tag{1}$$

where $\mathbf{m}$ is the same universal background model (UBM) super-vector used in the JFA approach and $\mathbf{T}$ is a low rank total-variability matrix. The total-variability factors ($\mathbf{w}$) are the i-vectors, and are normally distributed with parameters *N(0,1)*. Extracting an i-vector from the total-variability subspace is essentially a *maximum a-posteriori adaptation* (MAP) of $\mathbf{w}$ in the subspace defined by $\mathbf{T}$. An efficient procedure for the optimization of the total-variability subspace $\mathbf{T}$ and subsequent extraction of i-vectors is described Dehak *et al.* [5, 9]. In this paper, the pooled total-variability approach is used for i-vector feature extraction where the total-variability subspace (dimension = 500) is trained on telephone and microphone speech utterances together.

3. Short utterance variance

The long-length utterance i-vectors have speaker and session variations whereas short-length i-vectors have speaker, session and a lot of utterance variations. Thus, during development for SUV-PLDA, utterance variance is captured using the inner product of the difference between the full- and short-length i-vectors, and it is artificially added to full-length utterances and the simulated SUV is modelled using the PLDA approach [10]. The short utterance variance matrix, $\mathbf{S}_{SUV}$, can be calculated as follows,

$$\mathbf{S}_{SUV} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{A}^T(\mathbf{w}_n^{full} - \mathbf{w}_n^{short}))(\mathbf{A}^T(\mathbf{w}_n^{full} - \mathbf{w}_n^{short}))^T \tag{2}$$

Standard PLDA speaker verification system

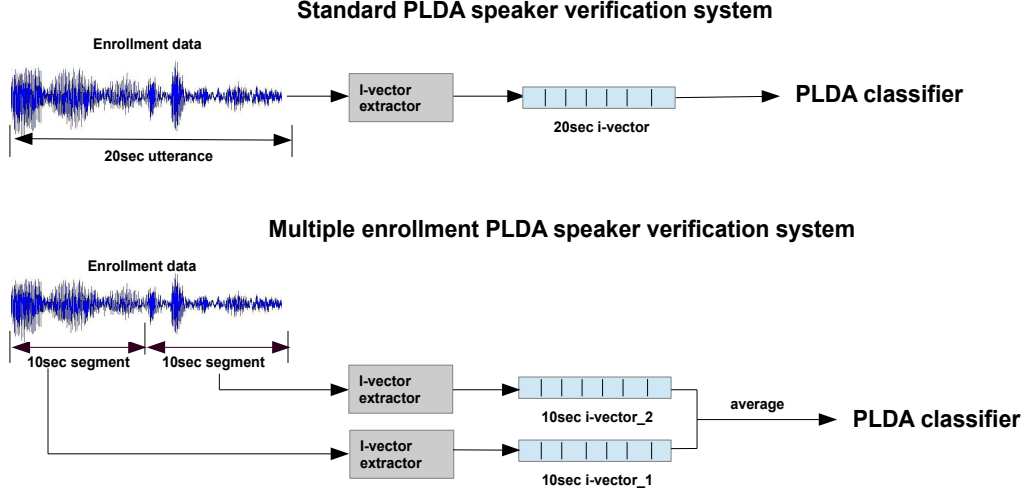**Multiple enrollment PLDA speaker verification system**

Figure 1: *Utterance-partitioning for the development of speaker verification PLDA models.*

where the estimation of the LDA matrix, $\mathbf{A}$ is detailed in our previous work [7]. For $\mathbf{S}_{SUV}$ estimation, the actual definition of what constitutes a full and/or short-length utterance needs to be established. For this research, we have defined full-length to be NIST standard utterance length, and in order to capture the SUV, full-length data is truncated into 20 sec. The SUV decorrelated matrix, $\mathbf{D}$, is calculated using the Cholesky decomposition of $\mathbf{D}\mathbf{D}^T = \mathbf{S}_{SUV}$. A random vector representing the utterance-length variability can be generated if a random normally independently distributed vector, $\mathbf{d}$, with $\mu = 0.0$ and $\sigma = 1.0$ is multiplied by the SUV decorrelated matrix, $\mathbf{D}$. The SUV-added full-length development vectors can be estimated as follows,

$$\mathbf{w} = \mathbf{w}_{full} + \mathbf{D}^T \mathbf{d} \tag{3}$$

After the SUV-added full-length i-vectors are extracted, length-normalized GPLDA model parameters are estimated as described in Section 4.

4. Length-normalized GPLDA

4.1. PLDA modelling

The length-normalization approach is detailed by Garcia-Romero *et al.* [11], and this approach is applied on development and evaluation data prior to GPLDA modelling. A speaker and channel dependent length-normalized i-vector, $\mathbf{w}_r$ can be defined as,

$$\mathbf{w}_r \;\;=\;\; \bar{\mathbf{w}} + \mathbf{U}_1 \mathbf{x}_1 + \boldsymbol{\varepsilon}_r \tag{4}$$

where for given speaker recordings $r = 1, ...R$; $\mathbf{U}_1$ is the eigenvoice matrix, $\mathbf{x}_1$ is the speaker factors and $\boldsymbol{\varepsilon}_r$ is the residuals. In the PLDA modelling, the speaker specific part can be represented as $\bar{\mathbf{w}} + \mathbf{U}_1 \mathbf{x}_1$, which represents the between speaker variability. The covariance matrix of the speaker part is $\mathbf{U}_1\mathbf{U}_1^T$. The channel specific part is represented as $\boldsymbol{\varepsilon}_r$, which describes the within speaker variability. The covariance matrix of the channel part is $\mathbf{\Lambda}^{-1}$. We assume that the precision matrix ($\mathbf{\Lambda}$) is full rank. Prior to GPLDA modelling, a standard LDA approach is applied to compensate the additional channel variations as well as to reduce the computational time.

4.2. GPLDA scoring

Scoring in GPLDA speaker verification systems is conducted using the batch likelihood ratio between a target and test i-vector [12]. Given two i-vectors, $\mathbf{w}_{target}$ and $\mathbf{w}_{test}$, the batch likelihood ratio can be calculated as follows,

$$\ln \frac{P(\mathbf{w}_{target}, \mathbf{w}_{test} \mid H_1)}{P(\mathbf{w}_{target} \mid H_0)P(\mathbf{w}_{test} \mid H_0)} \tag{5}$$

where H_1 denotes the hypothesis that the i-vectors represent the same speakers and H_0 denotes the hypothesis that they do not.

5. Utterance partitioning

A single long utterance is commonly used as enrolment data in PLDA speaker verification systems. It was previously found that short utterance i-vectors have speaker, session and utterance variations [6]. It is hypothesised that if long-duration speech data is partitioned into short utterances and if average of short utterance i-vectors is estimated, this approach would compensate the utterance variations. In this paper, in order to test this hypothesis, long enrolled-utterance are partitioned into short utterances and multiple i-vectors are extracted and the average of extracted i-vectors is used as enrolled i-vector. A block diagram of utterance-partitioning-based PLDA speaker verification is shown in Figure 1. Though truncated short utterances are extracted from the same speaker and session, every short utterance i-vectors have different behaviour due to linguistic content and the averaging of multiple short utterance i-vectors can be used to compensate the linguistic content variation.

6. Experimental methodology

The proposed methods were evaluated using the the NIST 2008 and 2010 SRE corpora. The shortened evaluation utterances were obtained by truncating the NIST 2008 *short2-short3* and NIST 2010 *core-core* conditions to the specified length of active speech for both enrolment and verification. Prior to truncation, the first 20 seconds of active speech were removed from all utterances to avoid capturing similar data across multiple utterances. For NIST 2008, the performance was evaluated using the equal error rate (EER) and the minimum decision cost

Table 1: *Utterance-partitioning-based standard GPLDA speaker verification performance on (a) NIST 2008 and (b) NIST 2010. The best performing systems are highlighted in each column.*

(a) *NIST 2008 short2-short3 truncated condition*

Evaluation utterance lengths	EER	DCF
Standard GPLDA system (Baseline)		
10sec-10sec	15.90%	0.0656
20sec-10sec	13.26%	**0.0549**
Utterance-partitioning-based GPLDA system		
10sec (2)-10sec	**12.60%**	0.0552

(b) *NIST 2010 core-core truncated condition*

Evaluation utterance lengths	EER	DCF
Standard GPLDA system (Baseline)		
10sec-10sec	15.98%	0.0693
20sec-10sec	13.85%	0.0608
Utterance-partitioning-based GPLDA system		
10sec (2)-10sec	**13.01%**	**0.0588**

Table 2: *Comparison of SUV-GPLDA and utterance-partitioning-based SUV-GPLDA speaker verification systems on the common set of the 2008 and 2010 NIST SRE truncated conditions. (a) NIST 2008 short2-short3 truncated condition (b) NIST 2010 core-core truncated condition. The best performing systems by both EER and DCF are highlighted across each row.*

(a) *NIST 2008 short2-short3 truncated condition*

Evaluation utterance lengths	EER	DCF
Standard SUV-GPLDA system		
10sec-10sec	14.58%	0.0624
20sec-10sec	12.35%	0.0523
Utterance-partitioning-based SUV-GPLDA system		
10sec (2)-10sec	**12.05%**	**0.0519**

(b) *NIST 2010 core-core truncated condition*

Evaluation utterance lengths	EER	DCF
Standard SUV-GPLDA system		
10sec-10sec	14.70%	0.0672
20sec-10sec	12.00%	0.0578
Utterance-partitioning-based SUV-GPLDA system		
10sec (2)-10sec	**11.58%**	**0.0555**

function (DCF), calculated using $C_{miss} = 10$, $C_{FA} = 1$, and $P_{target} = 0.01$ [13]. The performance for the NIST 2010 SRE was evaluated using the EER and the old minimum DCF (DCF_{old}), calculated using $C_{miss} = 10$, $C_{FA} = 1$, and $P_{target} = 0.01$, where evaluation was performed using the *telephone-telephone* condition [14].

We have used 13 feature-warped MFCC with appended delta coefficients and two gender-dependent UBMs containing 512 Gaussian mixtures throughout our experiments. The UBMs were trained on telephone and microphone speech from NIST 2004, 2005, and 2006 SRE corpora, and then used to calculate the Baum-Welch statistics before training a gender dependent total-variability subspace of dimension $R_w = 400$. The pooled total-variability representation and the GPLDA parameters were trained using telephone and microphone speech data from NIST 2004, 2005 and 2006 SRE corpora as well as Switchboard II which includes 1386 female and 1117 male speakers. We empirically selected the number of eigenvoices (N_1) equal to 120 as best value according to speaker verification performance over an evaluation set. 150 eigenvectors were selected for LDA estimation. S-normalisation was applied for experiments, and randomly selected telephone and microphone utterances from NIST 2004, 2005 and 2006 were pooled to form the S-normalisation dataset [15].

7. Results and discussion

7.1. Utterance-partitioning GPLDA

Table 1 compares the performance of utterance-partitioning against standard GPLDA on NIST 2008 and 2010 truncated conditions. In the standard GPLDA system, 10 sec and 20 sec utterances were used as enrolment utterances and 10 sec utterances were used as verification utterances. On the other hand, in the utterance partitioning approach, each 20 sec enrolment utterance was truncated into two 10sec utterances and the average of 10sec i-vectors were used as enrolment i-vector. The

utterance-partitioning approach shows an improvement over standard GPLDA system. This is because averaging multiple short enrolled-utterance i-vectors would compensate the linguistic content and session variations.

7.2. Utterance-partitioning SUV-GPLDA

In our previous studies, we found that the SUV-added GPLDA approach can effectively model the short utterance variance [7]. In this section, short utterance variance was studied with utterance-partitioning. Table 2 compares the performance of utterance partitioning and SUV approaches. The utterance-partitioning SUV-GPLDA system shows relative improvement of 9% and 16% in EER over standard GPLDA systems on NIST 2008 and NIST 2010 truncated 10sec-10sec evaluation condition as SUV models the utterance variations in the development phase and utterance partitioning approach compensates the linguistic content and session variations in the evaluation phase.

8. Conclusion

This paper studied the PLDA speaker verification approach with utterance partitioning and SUV modelling approaches. Our experimental studies have found that instead of using single long-utterance as enrolment data, if long enrolled-utterance was partitioned into multiple short utterances and the average of short utterance i-vectors was used as enrolled data, that improved the GPLDA speaker verification. This is because short utterance i-vectors have speaker, session and utterance variations, and utterance-partitioning approach compensates the utterance and session variations. SUV-GPLDA speaker was also studied with utterance-partitioning approach, and the utterance-partitioning-based SUV-GPLDA system showed relative improvement of 9% and 16% in EER for NIST 2008 and NIST 2010 truncated 10sec-10sec evaluation condition as SUV mod-

els the utterance variations in the development phase and the utterance partitioning approach compensates the linguistic content and session variations in the evaluation phase.

9. Acknowledgements

10. References

[1] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech 2008*, Brisbane, Australia, September 2008, pp. 853–856.

[2] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "Experiments in SVM-based speaker verification using short utterances," in *Proc. Odyssey Workshop*, 2010, pp. 83–90.

[3] A. Kanagasundaram, R. Vogt, B. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances," in *Proceed. of INTERSPEECH*. International Speech Communication Association (ISCA), 2011, pp. 2341–2344.

[4] M.-W. Mak and W. Rao, "Utterance partitioning with acoustic vector resampling for gmm–svm speaker verification," *Speech Communication*, vol. 53, no. 1, pp. 119–130, 2011.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, pp. 1 –1, 2010.

[6] P. Kenny, T. Stafylakis, P. Ouellet, M. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013.

[7] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, D. Ramos, and J. Gonzalez-Rodriguez, "Improving short utterance i-vector speaker recognition using utterance variance modelling and compensation techniques," in *Speech Communication*. Publication of the European Association for Signal Processing (EURASIP), 2014, pp. 69–82.

[8] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of Interspeech*, 2009, p. 1559 1562.

[9] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[10] T. Hasan, R. Saeidi, J. Hansen, and D. Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 7663 – 7667.

[11] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *International Conference on Speech Communication and Technology*, 2011, pp. 249–252.

[12] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey Speaker and Language Recogntion Workshop, Brno, Czech Republic*, 2010.

[13] "The NIST year 2008 speaker recognition evaluation plan," NIST, Tech. Rep., 2008. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2008/

[14] "The NIST year 2010 speaker recognition evaluation plan," NIST, Tech. Rep., 2010. [Online]. Available: www.itl.nist.gov/iad/mig/tests/sre/2010/index.html

[15] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," *Proc. Odyssey*, 2010.