Short Utterance PLDA Speaker Verification using SN-WLDA and Variance Modelling Techniques

Ahilan Kanagasundaram, David Dean, Sridha Sridharan

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

Abstract

This paper proposes a combination of source-normalized weighted linear discriminant analysis (SN-WLDA) and short utterance variance (SUV) PLDA modelling to improve the short utterance PLDA speaker verification. As short-length utterance i-vectors vary with the speaker, session variations and phonetic content of the utterance (utterance variation), a combined approach of SN-WLDA projection and SUV PLDA modelling is used to compensate the session and utterance variations. Experimental studies have found that a combination of SN-WLDA and SUV PLDA modelling approach shows an improvement over baseline system (WCCN[LDA]-projected Gaussian PLDA (GPLDA)) as this approach effectively compensates the session and utterance variations.

Index Terms: speaker verification, session variation, utterance variation, LDA, SN-WLDA

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, enrolment and verification 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, 5, 6]. Recently, Kenny *et al.* [7], have investigated how to quantify the uncertainty associated with the i-vector extraction process and propagate it into a PLDA classifier. Continuous research on this field has been ongoing to address the robustness of speaker verification technologies under such conditions.

The total-variability, or *i-vector*, approach has risen to prominence as the de-facto standard in recent state-of-the-art speaker verification systems, due to its intrinsic capability to map an utterance to a single low-dimensional i-vector, turning a complex high-dimensional speaker recognition problem into a low-dimensional classical pattern recognition one. However, i-vectors extracted from different durations should not be considered equal in reliability concerns. Moreover, long utterance i-vectors vary with speaker and session variations whereas short utterance i-vectors contain speaker, session and utterance variations, and these session and utterance variations need to be compensated in short utterance speaker verification.

As the session variability is included within the i-vector space, PLDA approach is commonly used to model speaker and session variations [8, 9]. In recent times, prior to the PLDA modelling, linear discriminant analysis (LDA) followed by within-class covariance normalization (WCCN) (WCCN[LDA]) session compensation approach is applied to compensate the additional session variation and reduce the computational complexness [10]. Recently, we have introduced the short utterance variance normalisation (SUVN) and short utterance variance (SUV) modelling to cosine similarity scoring (CSS) i-vector and PLDA speaker verification systems to compensate the session and utterance variations [11].

The main aim of this paper is to find a method that effectively compensates the session and utterance variations. Previously, we have found that source-normalised weighted LDA (SN-WLDA) followed by WCCN (WCCN[SN-WLDA])-projected Gaussian PLDA (GPLDA) system effectively compensates the session variation than standard WCCN[LDA]-projected GPLDA system in long utterance evaluation conditions [12]. Recently, it was also found that WCCN projection doesn't provide any advantage to PLDA speaker verification as PLDA models the intra-speaker variance itself [11]. In this paper, initially SN-WLDA-projected GPLDA system is studied with short utterance evaluation conditions. Subsequently, a combination of SN-WLDA and SUV modelling approach is introduced to PLDA speaker verification to effectively compensate the session and utterance variations.

This paper is structured as follows: Section 2 details the i-vector feature extraction techniques, and Section 3 explains how short utterance variance is added to i-vector features. Section 4 gives a brief details of SN-WLDA and SUV modelling approaches. Section 5 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 [13]. An i-vector speaker and channel dependent GMM super-vector can be represented by,

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

where **m** is the same universal background model (UBM) super-vector used in the JFA approach and **T** is a low rank total-variability matrix. The total-variability factors (**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 by Dehak *et al.* [8, 14]. In this paper, the pooled total-variability approach is used for i-vector feature extraction where the total-variability subspace (${R_w}^{telmic} = 500$) is trained on telephone and microphone speech utterances together.

3. Short utterance variance added i-vector features

The full-length utterance i-vectors have less utterance variation whereas short-length i-vectors have a lot of utterance variation. Thus, during development for SUV PLDA, utterance variation information is artificially added to full-length utterances, and the simulated SUV is modelled using the PLDA approach. The short utterance variance matrix , $\mathbf{S}_{SUV}$, can be calculated as the inner product of the difference between the full- and short-length i-vectors, ie:

$$\mathbf{S}_{SUV} \quad = \tfrac{1}{N}\textstyle\sum_{n=1}^{N}(\mathbf{w}_n^{full} - \mathbf{w}_n^{short})(\mathbf{w}_n^{full} - \mathbf{w}_n^{short})^T \quad (2)$$

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 a 100-sec utterance, and in order to capture the SUV, short utterance length was selected as 30 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 with utterance variation information can be generated if 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,

$$\mathbf{w} = \mathbf{w}_{full} + \mathbf{D}^T\mathbf{d} \quad (3)$$

After the SUV-added full-length i-vectors are extracted, LDA and SN-WLDA approaches are used to reduce the dimensionality and length-normalized GPLDA model parameters are estimated in as described in Sections 4 and 5.

4. LDA-/ SN-WLDA-projected i-vector features

LDA-/ SN-WLDA-projection is used to compensate the addition session variation prior to PLDA modelling, and it also significantly reduces the computational complexness as PLDA is modelled in reduced space.

4.1. LDA

The LDA is estimated based up the standard within- and between-class scatter estimations S_b and S_w, calculated as

$$\begin{aligned}\mathbf{S}_b &= \sum_{s=1}^{S} n_s(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})^T, && (4)\\ \mathbf{S}_w &= \sum_{s=1}^{S}\sum_{i=1}^{n_s}(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)^T, && (5)\end{aligned}$$

where S is the total number of speakers, n_s is number of utterances of speaker s. The mean i-vectors, $\bar{\mathbf{w}}_s$ for each speaker, and $\bar{\mathbf{w}}$ is the across all speakers are defined by

$$\bar{\mathbf{w}}_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{w}_i^s, \quad (6)$$

$$\bar{\mathbf{w}} = \frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\mathbf{w}_i^s. \quad (7)$$

where N is the total number of sessions. In the first stage, LDA attempts to find a reduced set of axes $\mathbf{A}$ through the eigenvalue decomposition of $\mathbf{S}_b\mathbf{v} = \lambda\mathbf{S}_w\mathbf{v}$.

4.2. SN-WLDA

LDA approach does not take advantage of the discriminative relationships that can be found between pairs of classes. The WLDA and SN-WLDA approaches have been used to overcome these shortcoming [15]. The weighted between-class scatter matrix, S_b^w, is defined as

$$\mathbf{S}_b^w \quad = \tfrac{1}{N}\textstyle\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} w(d_{ij})n_i n_j(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j)(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j)^T, (8)$$

where $\bar{\mathbf{w}}_i$ and $\bar{\mathbf{w}}_j$ are the mean i-vectors of speaker i and j respectively. In equation (8), the weighting function $w(d_{ij})$ is defined such that the classes that are closer to each other will be more heavily weighted. In this paper, we will be investigating the Euclidean and Mahalanobis distance weighting functions, ${w_{(d_{ij})}}^{Euc}$ and ${w_{(d_{ij})}}^{Maha}$, and these can be defined as follows,

$$\begin{aligned}{w_{(d_{ij})}}^{Euc} &= ((\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j)^T(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j))^{-n} && (9)\\ w_{(d_{ij})_{Maha}} &= ((\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j)^T(\mathbf{S}_w)^{-1}(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j))^{-n} && (10)\end{aligned}$$

The source normalized weighted between-class scatter matrix, $\mathbf{S}_b^{w_{src}}$, can be calculated as follows,

$$\mathbf{S}_b^{w_{src}} = \mathbf{S}_b^{w_{tel}} + \mathbf{S}_b^{w_{mic}}, \quad (11)$$

where the telephone-sourced dependent-weighted between-class scatter, $\mathbf{S}_b^{w_{tel}}$, and the microphone-sourced dependent-weighted between-class scatter, $\mathbf{S}_b^{w_{mic}}$, are individually calculated for telephone and microphone sources using Equation 8. In this paper, classification performance will be analyzed with several arbitrary values of n. The Euclidean and Mahalanobis distance weighting functions, are monotonically-decreasing function, so neighboring classes closer together will be heavily weighted than neighboring classes wider. The standard within-class scatter S_w and the corresponding SN-WLDA as described in section 4.1. The dimension-reduced i-vector can be calculated as follows,

$$\hat{\mathbf{w}} = \mathbf{A}^T\mathbf{w} \quad (12)$$

where $\mathbf{A}$ is LDA matrix, and dimension reduced i-vector, $\hat{\mathbf{w}}$, will be used for the GPLDA modelling in the following section.

5. Length-normalized GPLDA system

5.1. PLDA modelling

In this paper, we have chosen the length-normalized GPLDA, as it is also a simplified and computationally efficient approach [16]. The length-normalization approach is detailed by Garcia-Romero *et al.* [16], and this approach is applied on development and evaluation data prior to GPLDA modelling. A

Table 1: *Comparison of SN-WLDA-projected GPLDA against LDA-projected GPLDA on common condition of NIST 2008 truncated 10sec-10sec evaluation condition.*

System	**Interview-interview**		**Interview-telephone**		**Telephone-interview**		**Telephone-telephone**	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
LDA-GPLDA	17.84%	0.0769	20.38%	0.0843	17.72%	0.0809	15.80%	0.0664
SN-WLDA-Euclidean-GPLDA	**17.51%**	**0.0745**	20.08%	**0.0808**	17.66%	**0.0721**	15.39%	0.0669
SN-WLDA-Mahalanobis-GPLDA	17.61%	0.0752	**19.90%**	**0.0808**	**17.19%**	0.0747	**15.24%**	**0.0655**

Table 2: *Comparison of SUV SN-WLDA-projected GPLDA against SUV LDA-projected GPLDA on common condition of NIST 2008 truncated 10sec-10sec evaluation condition.*

System	**Interview-interview**		**Interview-telephone**		**Telephone-interview**		**Telephone-telephone**	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
SUV LDA-GPLDA	16.90%	0.0697	19.54%	0.0824	17.33%	0.0718	**14.35%**	**0.0629**
SUV SN-WLDA-Euclidean-GPLDA	16.86%	0.070	18.89%	**0.0796**	**16.72%**	**0.0705**	14.82%	0.0633
SUV SN-WLDA-Mahalanobis-GPLDA	**16.75%**	**0.0689**	18.62%	0.0821	16.97%	0.0720	14.41%	0.0631

speaker and channel dependent length-normalized i-vector, $\hat{\mathbf{w}}_r$ can be defined as,

$$\hat{\mathbf{w}}_r = \bar{\hat{\mathbf{w}}} + \mathbf{U}_1\mathbf{x}_1 + \boldsymbol{\varepsilon}_r \tag{13}$$

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 modeling, 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 channel part is $\boldsymbol{\Lambda}^{-1}$. We assume that precision matrix ($\boldsymbol{\Lambda}$) is full rank. Prior to GPLDA modeling, standard LDA followed by WCCN approach is applied to compensate the additional channel variations as well as reduce the computational time [4].

5.2. GPLDA scoring

Scoring in GPLDA speaker verification systems is conducted using the batch likelihood ratio between a target and test i-vector [9]. Given two i-vectors, $\mathbf{w}_{target}$ and $\mathbf{w}_{test}$, the batch likelihood ratio can be calculated as follows,

$$\ln \frac{P(\mathbf{w}_{target}, \mathbf{w}_{test} \mid H_1)}{P(\mathbf{w}_{target} \mid H_0)P(\mathbf{w}_{test} \mid H_0)} \tag{14}$$

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

6. Experimental methodology

The GPLDA based experiments were evaluated using the NIST 2008 corpora. For NIST 2008, the performance was evaluated using the equal error rate (EER) and the minimum decision cost function (DCF), calculated using $C_{miss} = 10$, $C_{FA} = 1$, and $P_{target} = 0.01$ [17].

We have used 13 feature-warped MFCC with appended delta coefficients and two gender-dependent UBM containing 512 Gaussian throughout our experiments. UBMs were trained on telephone and microphone from NIST 2004, 2005, and 2006 SRE corpora for telephone and microphone i-vector experiments. These gender-dependent UBMs were used to calculate the Baum-Welch statistics before training a gender dependent total-variability subspace of dimension $R_w = 500$. The pooled total-variability representation was trained using telephone and microphone speech data from NIST 2004, 2005 and 2006 SRE corpora as well as Switchboard II. The GPLDA parameters were trained using telephone and microphone speech data from NIST 2004, 2005 and 2006 SRE corpora as well as Switchboard II. We empirically selected the number of eigenvoices (N_1) equal to 120 as best value according to speaker verification performance. A full precision matrix was used for $\boldsymbol{\Lambda}$, rather than the diagonal. 150 eigenvectors were selected for standard LDA and SN-WLDA estimations. Randomly selected telephone and microphone utterances from NIST 2004, 2005 and 2006 were pooled to form the S-normalization dataset [18].

7. Results and discussions

7.1. LDA- and SN-WLDA-projected GPLDA systems

In this section, we have analyzed how the SN-WLDA-projected GPLDA system performs over the baseline approach, LDA-projected GPLDA system. Table 1 presents the results on the common set of the NIST SRE 2008 truncated 10sec-10sec condition. The SN-WLDA-projected GPLDA system shows improvement over LDA-projected GPLDA system as SN-WLDA projection extracts the discriminatory information between pairs of speakers as well as capturing the source variation information.

7.2. Modelling the short utterance variance using LDA- and SN-WLDA-projected GPLDA

A performance comparison of the SUV SN-WLDA-projected GPLDA approach against SUV LDA-projected GPLDA approach on NIST 2008 truncated 10sec-10sec condition is shown in Table 2. From these results, it can be observed that the SUV SN-WLDA-projected GPLDA approach is shown to provide a clear improvement over the SUV LDA-projected GPLDA approach across all conditions, except *telephone-telephone* condition, as SUV SN-WLDA-projected GPLDA approach compensates the utterance variation and extracts the discriminatory information between pairs of speakers as well as capturing the source variation information. When SUV LDA- and SN-WLDA-projected GPLDA are compared against LDA- and SN-WLDA-projected GPLDA, SUV LDA-projected GPLDA shows improvement over LDA- and SN-WLDA-projected GPLDA on *telephone-telephone*, and SUV SN-WLDA GPLDA shows improvement over SUV LDA-projected GPLDA, LDA- and SN-WLDA-projected GPLDA systems on microphone conditions.

8. Conclusion

In this paper, PLDA speaker verification was investigated with LDA and SN-WLDA approaches and found that SN-WLDA-projected GPLDA is better approach than LDA-projected GPLDA approach. Subsequently a combination of SUV and SN-WLDA approach was introduced to PLDA speaker verification and found that SUV SN-WLDA-projected GPLDA effectively compensates the session and utterance variance over other approaches.

9. Acknowledgements

This project was supported by an Australian Research Council (ARC) Linkage grant LP130100110.

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.

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

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

[4] A. Kanagasundaram, R. Vogt, D. Dean, and S. Sridharan, "PLDA based speaker recognition on short utterances," in *The Speaker and Language Recognition Workshop (Odyssey 2012)*, ISCA, 2012.

[5] A. Kanagasundaram, Dean, and S. Sridharan, "Improving PLDA speaker verification with limited development data," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2014.

[6] A. Kanagasundaram, D. Dean, J. Gonzalez-Dominguez, S. Sridharan, D. Ramos, and J. Gonzalez-Rodriguez, "Improving short utterance based i-vector speaker recognition using source and utterance-duration normalization techniques," in *Proceed. of INTERSPEECH*, International Speech Communication Association (ISCA), 2013.

[7] 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.

[8] 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. 19, no. 4, pp. 788–798, 2010.

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

[10] A. Kanagasundaram, D. Dean, S. Sridharan, and R. Vogt, "PLDA based speaker recognition with weighted LDA techniques," in *Proc. Odyssey Workshop*, 2012.

[11] 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.

[12] A. Kanagasundaram, D. Dean, S. Sridharan, M. McLaren, and R. Vogt, "I-vector based speaker recognition using advanced channel compensation techniques," in *Computer Speech and Language*, vol. 28, pp. 121–140, 2014.

[13] 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*, p. 1559 1562, 2009.

[14] 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.

[15] A. Kanagasundaram, D. Dean, R. Vogt, M. McLaren, S. Sridharan, and M. Mason, "Weighted LDA techniques for i-vector based speaker verification," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 4781–4784, 2012.

[16] 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*, pp. 249–252, 2011.

[17] "The NIST year 2008 speaker recognition evaluation plan," tech. rep., NIST, 2008.

[18] 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.