IMPROVING PLDA SPEAKER VERIFICATION WITH LIMITED DEVELOPMENT DATA

Ahilan Kanagasundaram, David Dean and Sridha Sridharan

Speech Research Laboratory, Queensland University of Technology, Australia

{a.kanagasundaram, d.dean, s.sridharan}@qut.edu.au

ABSTRACT

This paper analyses the probabilistic linear discriminant analysis (PLDA) speaker verification approach with limited development data. This paper investigates the use of the median as the central tendency of a speaker's i-vector representation, and the effectiveness of weighted discriminative techniques on the performance of state-of-the-art length-normalised Gaussian PLDA (GPLDA) speaker verification systems. The analysis within shows that the median (using a median fisher discriminator (MFD)) provides a better representation of a speaker when the number of representative i-vectors available during development is reduced, and that further, usage of the pair-wise weighting approach in weighted LDA and weighted MFD provides further improvement in limited development conditions. Best performance is obtained using a weighted MFD approach, which shows over 10% relative improvement in EER over the baseline GPLDA system on mismatched and *interview-interview* conditions.

Index Terms— Speaker verification, PLDA, WLDA, WMFD

1. INTRODUCTION

Speaker verification has traditionally required a large volume of speech during development and evaluation, particularly in the presence of high intersession variability. However, it can be hard to acquire a sufficient quantity of data in many real-world environments, limiting the suitability of speaker verification for many everyday applications. Recently, a number of interesting techniques have focussed on reducing the amount of speech required during evaluation (covering the enrolment of speaker models, and their verification), but little effort has been put into reducing the volume of speech required to develop new models for deployment into previously unseen environments.

Reducing the amount of speech required during enrolment and verification whilst maintaining satisfactory performance has been the focus in a number of recent studies focused on a variety of speaker verification technologies: joint factor analysis (JFA) [1], support vector machines (SVM) [2], i-vectors [3] and probabilistic linear discriminant analysis (PLDA) [4]. These studies have shown that, across all technologies, the verification performance degrades considerably when very-short utterances ($< 10s$) are used as evaluation data. A number of attempts to compensate for this reduction in performance have been undertaken in the literature. Kenny *et al.* [5] have investigated how to quantify the uncertainty associated with summarising various length utterances down to a constant-length i-vector and demonstrated how to propagate that into the PLDA classifier. An alternative approach was demonstrated by Hasan *et al.* [6] where they found that the duration variability can be modelled as additive noise in the i-vector space, also using a PLDA classifier.

Speaker verification is a data-driven research field, and it has clearly been established that state-of-the-art speaker verification systems require a significant volume of development data covering multiple sessions across a large number of speakers [7]. However, the volume of data required to adequately model the background behaviour of speaker models is not always available, particularly in new environments. In a recent study, the i-vector and PLDA speaker verification systems' performance were analysed when the long- and short-length utterance development data was used for speaker development, where Kanagasundaram *et al.* [4] have found that instead of using the full-length utterance development data, when short-length utterance development data is used for PLDA modelling, speaker verification systems shows a significant improvement. However, there hasn't yet been any detailed investigations on how state-of-the-art PLDA speaker verification copes with limited session development data.

In this paper, initially a LDA-projected Gaussian PLDA (GPLDA) speaker verification system's is analysed with limited development data to investigate the effect on speaker verification performance. An alternative approach to LDA projection, the median Fisher discriminator (MFD) is then introduced to show better speaker discriminative performance from limited channel development data than the mean-centroid approach of LDA. Finally, weighted approaches, where weighting the speakers that are closer to each other to reduce speaker confusion, of LDA (WLDA) and MFD (WMFD) are introduced to provide a further boost in speaker discrimination from limited development data.

This paper is structured as follows: Section 2 outlines a typical state-of-the-art GPLDA speaker verification system, and Section 3 gives a brief overview of dimensionality reduc-

tion approaches, including the MFD and weighted approaches introduced in this paper. The experimental protocol and corresponding results are given in Section 4 and Section 5. Section 7 concludes the paper.

2. GPLDA SPEAKER VERIFICATION

2.1. I-vectors

I-vectors represent a Gaussian mixture model (GMM) mean super-vector by a single total-variability subspace. This single-subspace approach was motivated by the discovery that the channel space of the earlier, related JFA technique contained valuable speaker-discriminant information [8]. An i-vector speaker-and-channel-dependent GMM super-vector $\boldsymbol{\mu}$ can be represented by,

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

where $\mathbf{m}$ is a universal background model (UBM) mean super-vector trained over a large development set 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* (MAP) adaptation of $\mathbf{w}$ in the subspace defined by $\mathbf{T}$. An efficient procedure for the optimisation of the total-variability subspace $\mathbf{T}$ and subsequent extraction of i-vectors is described by Dehak *et al.* [9, 10]. 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 to provide the best i-vector representation [11].

2.2. GPLDA modelling

When originally introduced by Kenny [12], the Gaussian (GPLDA) and Heavy-tailed PLDA (HTPLDA) approaches were introduced to directly model the speaker and channel variability directly in the i-vector space, with better performance obtained using HTPLDA at a cost of higher complexity. However, recently Garcia-Romero *et al.* have shown that a simple whitening and length-normalisation approach can bring the performance of GPLDA up to HTPLDA with a much simpler approach, and it is this length-normalised GPLDA approach that will be used in this paper. The length-normalisation approach is detailed by Garcia-Romero *et al.* [13], and this approach is applied on development and evaluation i-vectors prior to GPLDA modelling. A speaker and session-dependent length-normalised i-vector, $\mathbf{w}'_{s,i}$ can be defined as,

$$\mathbf{w}'_{s,i} = \bar{\mathbf{w}}' + \mathbf{U}_1\mathbf{x}_{1,s} + \boldsymbol{\varepsilon}_{s,i} \quad (2)$$

where for a given speaker, s, having n_S sessions $i = 1, \ldots, n_s$, $\bar{\mathbf{w}}'$ is the mean length-normalised i-vector, $\mathbf{x}_{1,s}$ is the speaker factors and $\boldsymbol{\varepsilon}_{s,i}$ is the residual for each session; Finally, $\mathbf{U}_1$ is the eigenvoice matrix trained in PLDA modelling. The speaker specific part can be represented as $\bar{\mathbf{w}}' + \mathbf{U}_1\mathbf{x}_{1,s}$, which represents the between-speaker variability and the covariance matrix of the speaker part is $\mathbf{U}_1{\mathbf{U}_1}^T$. The session-specific part is represented as $\boldsymbol{\varepsilon}_{s,i}$, which describes the within-speaker variability, and the covariance matrix of the session variability is $\boldsymbol{\Lambda}^{-1}$. We assume that the precision matrix ($\boldsymbol{\Lambda}$) is full rank.

Prior to length-normalisation and GPLDA modelling, a number of dimensional reduction techniques can be used, as outlined in Section 3, to compensate for session variation prior to GPLDA modelling as well as reducing the computational time of the modelling itself [14].

2.3. 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 length-normalised i-vectors, $\mathbf{w}'_{target}$ and $\mathbf{w}'_{test}$, the batch-likelihood ratio can be calculated as follows,

$$\ln \frac{P(\mathbf{w}'_{target}, \mathbf{w}'_{test} \mid H_1)}{P(\mathbf{w}'_{target} \mid H_0)P(\mathbf{w}'_{test} \mid H_0)} \quad (3)$$

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

3. DIMENSIONALITY REDUCTION OF I-VECTOR FEATURES

3.1. Linear discriminant analysis

Because i-vectors are calculated on a subspace covering both speaker and session variation, session compensation techniques are typically introduced after i-vector extraction and before modelling to improve the speaker discriminative ability of the i-vector subspace. A typical approach is to first reduce the dimensionality using linear discriminant analysis (LDA) and then scale the resultant space using within-class covariance normalisation (WCCN) [15]. In this paper, we will refer to this technique as WCCN of the LDA space or WCCN[LDA], and in the first stage of this process, an LDA transformation attempts to find a reduced set of axes $\mathbf{A}$ through the eigenvalue decomposition of $\mathbf{S}_b\mathbf{v} = \lambda\mathbf{S}_w\mathbf{v}$, where the standard within- and between-class scatter estimations S_b and S_w, are calculated as

$$\mathbf{S}_b = \sum_{s=1}^{S} n_s(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})^T, \quad (4)$$

$$\mathbf{S}_w = \sum_{s=1}^{S}\sum_{i=1}^{n_s}(\mathbf{w}_{s,i} - \bar{\mathbf{w}}_s)(\mathbf{w}_{s,i} - \bar{\mathbf{w}}_s)^T, \quad (5)$$

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

$$\bar{\mathbf{w}}_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{w}_{s,i}, \tag{6}$$

$$\bar{\mathbf{w}} = \frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\mathbf{w}_{s,i}. \tag{7}$$

where N is the total number of sessions.

In the second stage, the WCCN transformation matrix ($\boldsymbol{B}$) is trained on LDA-projected development i-vectors. The WCCN matrix ($\boldsymbol{B}$) is then calculated using Cholesky decomposition of $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{W}^{-1}$, where the within-class covariance matrix $\boldsymbol{W}$ is calculated using

$$\mathbf{W} = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(\mathbf{A}^T(\mathbf{w}_{s,i}-\bar{\mathbf{w}}_s))(\mathbf{A}^T(\mathbf{w}_{s,i}-\bar{\mathbf{w}}_s))^T \tag{8}$$

3.2. Median fisher discriminator

In traditional LDA, the mean i-vector of each speaker plays a major role in the definition of the between-class and within-class scatter matrices. Therefore, the accuracy estimate of mean has a substantial effect on the resulting projected directions of the LDA transformation. In this paper, as we investigate speaker verification with limited session development data, averaging these few recording could lead to a loss of speaker-discriminant information. By taking the median as the estimator for the central tendency, instead of the mean, the MFD approach should help to attenuate this loss, as the median tends to provide a more robust estimate [16]. MFD estimation is performed by calculating the between- and within-class scatter estimations using the median as the central tendency rather than the mean, $\boldsymbol{S}_w^{median}$ and $\boldsymbol{S}_b^{median}$, calculated as follows;

$$\mathbf{S}_b^{median} = \sum_{s=1}^{S} n_s(\tilde{\mathbf{w}}_s-\tilde{\mathbf{w}})(\tilde{\mathbf{w}}_s-\tilde{\mathbf{w}})^T, \tag{9}$$

$$\mathbf{S}_w^{median} = \sum_{s=1}^{S}\sum_{i=1}^{n_s}(\mathbf{w}_{s,i}-\tilde{\mathbf{w}}_s)(\mathbf{w}_{s,i}-\tilde{\mathbf{w}}_s)^T \tag{10}$$

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

$$\tilde{\mathbf{w}}_s = \textit{Median}(\{\mathbf{w}_{s,1},\mathbf{w}_{s,2},\mathbf{w}_{s,3},\ldots,\mathbf{w}_{s,n_s}\}), \tag{11}$$

$$\tilde{\mathbf{w}} = \frac{1}{N}\sum_{s=1}^{S} n_s\tilde{\mathbf{w}}_s. \tag{12}$$

where N is the total number of sessions. The MFD transformation is estimated using the same approach as the LDA transformation in the previous section.

3.3. Weighted LDA and MFD

Traditional discriminative dimensional reduction techniques focus on the scatter matrices of the development space as a whole, but recently more advanced weighted-approaches have been introduced that can calculate the scatter matrices on a pair-wise basis, taking advantage of the discriminative information contained in the relationship between individual pairs of speakers [17]. In this paper, we will investigate the effectiveness of the weighted variants of LDA and MFD (WLDA and WMFD respectively) when trained on limited session development data. For the WLDA approach, the weighted between-class scatter matrix, $\boldsymbol{S}_b^w$, is defined as

$$\mathbf{S}_b^w = \frac{1}{N}\sum_{p=1}^{S-1}\sum_{q=p+1}^{S} w(d_{pq})n_p n_q(\bar{\mathbf{w}}_p-\bar{\mathbf{w}}_q)(\bar{\mathbf{w}}_p-\bar{\mathbf{w}}_q)^T, \tag{13}$$

where $\bar{\mathbf{w}}_p$ and $\bar{\mathbf{w}}_q$ are the mean i-vectors of speaker p and q respectively, n_p and n_q the number of sessions, and $w(d_{pq})$ is a weighting function defined such that the classes that are closer to each other will have a higher weight in forming the final scatter matrix. In this paper, we will be investigating the Euclidean distance weighting function, ${w_{(d_{pq})}}^{Euc}$,

$${w_{(d_{pq})}}^{Euc} = ((\bar{\mathbf{w}}_p-\bar{\mathbf{w}}_q)^T(\bar{\mathbf{w}}_p-\bar{\mathbf{w}}_q))^{-n}. \tag{14}$$

The standard within-class scatter S_w and the corresponding WLDA and WCCN transformation matrices can be estimated as described in Section 3.1. For the WMFD estimation, a similar approach is taken, but with $\bar{\mathbf{w}}_p$ and $\bar{\mathbf{w}}_q$ replaced with $\tilde{\mathbf{w}}_p$ and $\tilde{\mathbf{w}}_q$ in Equations 13 and 14.

4. EXPERIMENTAL METHODOLOGY

The GPLDA based experiments were evaluated using the common set of NIST 2008 short2-short3 evaluation corpora. 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 $P_{target} = 0.01$ [18].

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 $R_w = 500$. 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.

Table 1. *Weighted LDA and MFD performance versus unweighted LDA performance of length-normalised GPLDA as the number of development sessions is increased.*

System	Interview-interview		Interview-telephone		Telephone-interview		Telephone-telephone	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
3 sessions/speaker								
WCCN[LDA]-GPLDA	10.85%	0.0473	11.69%	0.0526	9.51%	0.0423	4.04%	0.0188
WCCN[MFD]-GPLDA	10.52%	0.0465	11.06%	0.0492	8.82%	0.0383	4.29%	0.0172
WCCN[WLDA]-GPLDA	9.95%	0.0455	11.25%	0.0515	8.69%	0.0393	**3.71%**	0.0193
WCCN[WMFD]-GPLDA	**9.69%**	**0.0435**	**10.15%**	**0.0470**	**8.15%**	**0.0364**	3.95%	**0.0186**
5 sessions/speaker								
WCCN[LDA]-GPLDA	8.69%	0.0395	9.86%	0.0467	7.81%	0.0344	3.21%	**0.0148**
WCCN[MFD]-GPLDA	8.09%	0.0372	8.94%	0.0440	7.34%	0.0331	2.96%	0.0149
WCCN[WLDA]-GPLDA	7.94%	0.0379	9.29%	0.0451	6.79%	0.0303	2.97%	0.0157
WCCN[WMFD]-GPLDA	**7.29%**	**0.0350**	**8.11%**	**0.0402**	**6.11%**	**0.0271**	**2.72%**	0.0154
7 sessions/speaker								
WCCN[LDA]-GPLDA	8.00%	0.0361	8.29%	0.0430	7.00%	0.0307	**2.55%**	**0.0143**
WCCN[MFD]-GPLDA	7.67%	0.0348	8.29%	0.0416	7.00%	0.0303	2.88%	0.0144
WCCN[WLDA]-GPLDA	6.78%	0.0326	7.66%	0.0401	5.98%	**0.0265**	2.70%	0.0143
WCCN[WMFD]-GPLDA	**6.12%**	**0.0306**	**7.39%**	**0.0373**	**5.36%**	0.0268	2.63%	0.0149

150 eigenvectors were selected for LDA, MFD, WLDA and WMFD estimations. 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 [19].

5. RESULTS AND DISCUSSIONS

Table 1 presents the results comparing the performance of the weighted LDA and MFD systems against the baseline unweighted LDA approach. It can be observed from Table 1 that, as the number of development sessions made available to WCCN[LDA]-projected GPLDA system increases from 3 per speaker to 7, the EER drops considerably across all conditions. WCCN[MFD]-projected GPLDA system shows improvement over WCCN[LDA]-projected GPLDA system, as MFD is robust estimate when limited session data is available. The weighted LDA system shows useful improvement over the unweighted LDA approach on mismatched and *interview-interview* conditions, showing that the weighted approach effectively extracts more discriminant information from pair of speakers. This improvement is further extended by the choice of the median over the mean as the central tendency, with the WMFD system showing a further improvement of WLDA for a total of over 10% relative improvement in EER over the baseline LDA system on mismatched and *interview-interview* conditions.

6. ACKNOWLEDGEMENT

This research was funded by the Australian Research Council (ARC) Linkage Grant No: LP130100110.

7. CONCLUSION

In this paper, length-normalised PLDA speaker verification was analysed with limited session data, and it was found that when the number of sessions that is used to train the PLDA subspace and dimensionality reduction techniques is increased, it significantly affects speaker verification performance. It was found that, by taking advantage of pair-wise differences in speaker i-vectors, weighted LDA improved speaker verification performance in limited development session conditions. Further improvement was found by using the median as the measure of central tendency rather than the mean for calculating the scatter matrices. The final median-based and weighted WMFD PLDA system provided over a 10% improvement in EER over the baseline LDA GPLDA system on mismatched and *interview-interview* conditions.

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

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

[4] Ahilan Kanagasundaram, Robert J Vogt, David B Dean, and Sridha Sridharan, "PLDA based speaker recognition on short utterances," in *The Speaker and Language Recognition Workshop (Odyssey 2012)*. ISCA, 2012.

[5] 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.

[6] 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.

[7] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[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] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," *Odyssey Speaker and Language Recognition Workshop*, 2010.

[10] 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.

[11] M. McLaren and D. van Leeuwen, "Improved speaker recognition when using i-vectors from multiple speech sources," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 5460–5463.

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

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

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

[15] 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.

[16] J. Yang, J. Yang, and D. Zhang, "Median fisher discriminator: a robust feature extraction method with applications to biometrics," *Frontiers of Computer Science in China*, vol. 2, no. 3, pp. 295–305, 2008.

[17] 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*, 2012, pp. 4781–4784.

[18] NIST, "The NIST year 2008 speaker recognition evaluation plan," Tech. Rep., NIST, 2008.

[19] 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.