Improving PLDA Speaker Verification using WMFD and Linear-weighted Approaches in Limited Microphone Data Conditions

Ahilan Kanagasundaram, David Dean and 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 the addition of a weighted median Fisher discriminator (WMFD) projection prior to length-normalised Gaussian probabilistic linear discriminant analysis (GPLDA) modelling in order to compensate the additional session variation. In limited microphone data conditions, a linear-weighted approach is introduced to increase the influence of microphone speech dataset. The linear-weighted WMFD-projected GPLDA system shows improvements in EER and DCF values over the pooled LDA- and WMFD-projected GPLDA systems in *interview-interview* condition as WMFD projection extracts more speaker discriminant information with limited number of sessions/ speaker data, and linear-weighted GPLDA approach estimates reliable model parameters with limited microphone data.
Index Terms: speaker verification, i-vectors, GPLDA, WMFD, linear-weighted

1. Introduction

A significant amount of speech data is required to develop a robust speaker verification system, especially in the presence of large intersession variability [1]. However, it is often difficult to acquire a sufficient number of sessions for each individual speaker for developing robust background models in many real-world environments, limiting the availability of speaker verification technology for many everyday applications. A significant example of this problem is the relative scarcity of microphone speech data available across the many NIST Speaker Recognition Evaluation (SRE) databases [2, 3], which have, at least until more recent evaluations, focussed largely on collecting large quantities of telephone speech.

Speaker verification is a data-driven research field, and it has clearly been established that the development of state-of-the-art speaker verification systems require a significant volume of speech data covering multiple sessions across a large number of speakers [1]. However, the volume of data required to adequately model the background behaviour of speaker models is not always available, particularly in new environments. Recently, we have analysed the linear discriminant analysis (LDA) projected Gaussian probabilistic linear discriminant analysis (GPLDA) speaker verification system with limited development data, and found that when the number of sessions/speaker are reduced, the speaker verification performance is considerably affected [4]. As an alternative approach to LDA projection, we have also previously introduced the median Fisher discriminator (MFD) and a weighted variant (WMFD) to show better speaker discriminative performance from limited-session development data than the mean-centroid approach of LDA [4].

In addressing the disparate microphone and telephone data sources available in the NIST evaluations, researchers have shown that pooling the telephone and microphone speech data is the best approach for the development of GPLDA [5, 6] speaker verification systems. In our recent work, we have introduced a linear-weighted approach to effectively model the GPLDA parameters proportionally from telephone and microphone speech data [7] that has shown promise in limited development session conditions.

In this paper, initially a LDA-projected GPLDA speaker verification system was analysed with limited development data to investigate the effect on speaker verification performance. This approach is then compared to the alternative, WMFD-projected linear-weighted GPLDA approach, to show an improvement in speaker verification performance for limited microphone development sessions. In our GPLDA speaker verification system, telephone speakers (of which we are developing across 1286 female and 1034 male speakers) with more than 10 sessions and a limited number of microphone speakers (100 female and 83 male speakers) with more than 15 session were used for GPLDA modelling. To demonstrate the effect of limited sessions during development, we have restricted both the telephone and microphone speech speakers to 7 sessions per speaker.

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 reduction approaches, including LDA and WMFD. Section 4 details the GPLDA model-parameter estimation techniques for scarce microphone speech. The experimental protocol and corresponding results are given in Section 5 and Section 6, and 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,

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

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 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.* [13] 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 therefore 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,

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

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, $\mathbf{x}_{1,s}$ are 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 [6]. Scoring in GPLDA speaker verification systems is conducted using the batch-likelihood ratio between a target and test i-vector [12].

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 linear discriminant analysis (LDA) followed by within-class covariance normalisation (WCCN) (WCCN[LDA]) approach is to first reduce the dimensionality using LDA and then scale the resultant space using WCCN. This WCCN[LDA] approach has been clearly explained in our previous work for the interested reader [14, 15].

3.2. Weighted 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 of estimate of the 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, as outliers can have a much bigger effect in small datasets. By taking the median as the estimator for the central tendency, instead of the mean, the WMFD approach should help to attenuate this loss, as the median tends to provide a more robust estimate of the central tendancy [6, 4]. WMFD estimation is performed by calculating the weighted between- and within-class scatter estimations using the median as the central tendency rather than the mean, $\boldsymbol{S}_w^{median}$ and $\boldsymbol{S}_b^{w-median}$, calculated as follows;

$$\mathbf{S}_w^{median} = \sum_{s=1}^{S}\sum_{i=1}^{n_s}(\mathbf{w}_{s,i} - \tilde{\mathbf{w}}_s)(\mathbf{w}_{s,i} - \tilde{\mathbf{w}}_s)^T \tag{3}$$

$$\mathbf{S}_b^{w-median} = \frac{1}{N}\sum_{p=1}^{S-1}\sum_{q=p+1}^{S} w(d_{pq}) n_p n_q (\tilde{\mathbf{w}}_p - \tilde{\mathbf{w}}_q)(\tilde{\mathbf{w}}_p - \tilde{\mathbf{w}}_q)^T, \tag{4}$$

where the median i-vectors, $\tilde{\mathbf{w}}_s$ for each speaker is defined by

$$\tilde{\mathbf{w}}_s = \mathit{Median}(\{\mathbf{w}_{s,1}, \mathbf{w}_{s,2}, \mathbf{w}_{s,3}, \ldots, \mathbf{w}_{s,n_s}\}) \tag{5}$$

where $\tilde{\mathbf{w}}_p$ and $\tilde{\mathbf{w}}_q$ are the median i-vectors of speaker p and q respectively estimated using Equation 5, 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, ${w_{(d_{pq})}}^{Euc}$,

$${w_{(d_{pq})}}^{Euc} = ((\bar{\mathbf{w}}_p - \bar{\mathbf{w}}_q)^T(\bar{\mathbf{w}}_p - \bar{\mathbf{w}}_q))^{-n}. \tag{6}$$

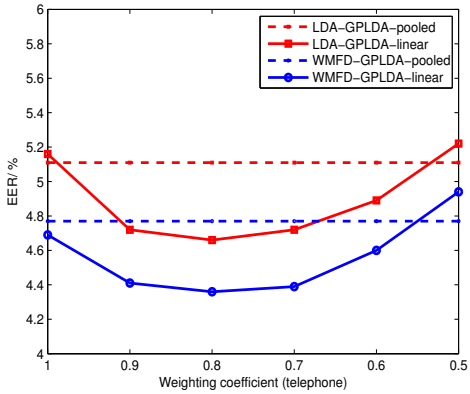
n was selected as 4 for weighting function estimation. The WMFD transformation is estimated using the same approach as the LDA transformation as detailed in paper [14].

4. GPLDA parameter estimation

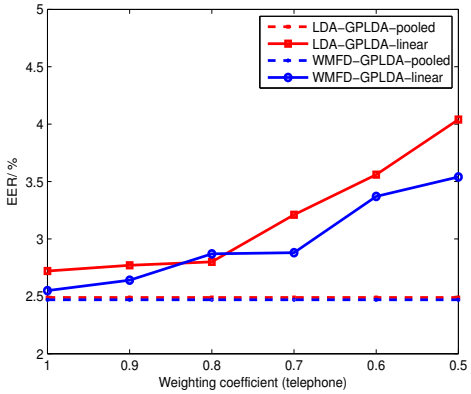
In i-vector feature domain, pooled total-variability approach was used to exploit sufficient speaker variation from telephone and microphone speech sources [5, 6]. In this section, in PLDA model domain, both pooled and linear weighted approaches are investigated to estimate the proper GPLDA model parameters from rich telephone and scarce microphone speech data [7].

4.1. Pooled approach

It is commonly believed that robust probabilistic parameters can be estimated if adequate amount of speech data is available. In the pooled subspace training approach, telephone and microphone speech is pooled together to create large development data set, and the length-normalized GPLDA parameters, including mean ($\bar{\mathbf{w}}_{telmic}$), precision matrix ($\boldsymbol{\Lambda}_{telmic}$) and eigenvoice matrix ($\mathbf{U}_{1telmic}$) are estimated using telephone and microphone pooled data.

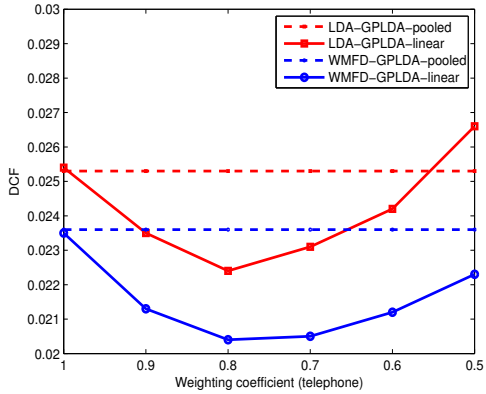


(a) Interview-interview condition

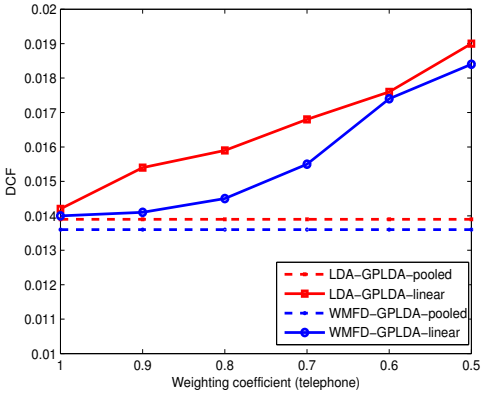


(b) Telephone-telephone condition

Figure 1: Comparison of EER values of pooled and linear-weighted based LDA- and WMFD-projected GPLDA systems at different weighting coefficients on NIST 08 short2-short3 condition. ,(a) interview-interview *and* (b) telephone-telephone



(a) Interview-interview condition



(b) Telephone-telephone condition

Figure 2: Comparison of DCF values of pooled and linear-weighted based LDA- and WMFD-projected GPLDA systems at different weighting coefficients on NIST 08 short2-short3 condition. ,(a) interview-interview *and* (b) telephone-telephone

4.2. Linear weighted approach

If a sufficient amount of telephone and microphone speech data is available, it is believed that the pooled approach would be the better approach. However, this condition is rarely met in the real world, and in the case of the limited microphone data conditions of this examination, a linear weighted approach can be used to increase the influence of microphone speech data [7]. Firstly, the GPLDA model parameters, including mean ($\bar{\mathbf{w}}_{tel}$), precision matrix ($\mathbf{\Lambda}_{tel}$) and eigenvoice matrix ($\mathbf{U}_{1tel}$) are estimated using telephone speech. Similarly, the GPLDA model parameters, including mean ($\bar{\mathbf{w}}_{mic}$), precision matrix ($\mathbf{\Lambda}_{mic}$) and eigenvoice matrix ($\mathbf{U}_{1mic}$) are also estimated using microphone speech. These two domain-specific sets of parameters can then be combined using a linear weighted approach, estimated as follows,

$$\bar{\mathbf{w}}_{telmic} = \alpha\bar{\mathbf{w}}_{tel} + (1-\alpha)\bar{\mathbf{w}}_{mic} \quad (7)$$

$$\mathbf{\Lambda}_{telmic} = \alpha\mathbf{\Lambda}_{tel} + (1-\alpha)\mathbf{\Lambda}_{mic} \quad (8)$$

$$\mathbf{U}_{1telmic} = \alpha\mathbf{U}_{1tel} + (1-\alpha)\mathbf{U}_{1mic} \quad (9)$$

5. Experimental methodology

The GPLDA experiments will be evaluated using the NIST 2008 SRE 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 $P_{target} = 0.01$ [3].

We have used 13 feature-warped MFCCs 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 $R_w = 400$, which was then used to calculate the i-vector speaker representations. 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 $\mathbf{\Lambda}$, rather than the diagonal. 150 eigenvectors were selected for standard LDA and WMFD estimation. Randomly selected telephone and microphone utterances from NIST04, 05 and 06 were pooled to form the S-normalization dataset [16].

6. Results and discussions

Experiments were carried out to identify the best length-normalized GPLDA-model parameter-estimation approach in limited development data conditions. Both LDA- and WMFD-projected GPLDA speaker verification systems were analysed with pooled and linear weighted based GPLDA modelling approaches to estimate the robust GPLDA model parameters from both telephone and microphone sourced speech. The LDA- and WMFD-projected GPLDA systems were trained using a limited number of session/ speaker data (7 sessions/ speaker) with a limited number of microphone speech speakers and a larger number of telephone speech speakers (Female (1286 tel and 100 mic speakers), Male (1034 tel and 83 mic speakers)).

Figure 1 and 2 respectively compare the EER and DCF values of pooled and linear-weighted based LDA- and WMFD-projected GPLDA systems at different weighting coefficients on the NIST 2008 short2-short3 condition. In order to increase the influence of microphone data, the linear weighted based GPLDA approach was analysed across several values of the telephone weighting parameter (α_{tel}) at 0.1 intervals. Through telephone-estimated and microphone-estimated PLDA parameters are respectively good and poor estimates, the influence of microphone can be increased by reducing the α_{tel} weights. It can be clearly seen that when the α_{tel} is selected as 0.8, the linear-weighted LDA-projected GPLDA shows improvement in EER and DCF values over the pooled LDA-projected GPLDA in *interview-interview* condition. However, as α_{tel} is further reduced, the performance is reduced in microphone speech conditions as the microphone-estimated GPLDA parameters provide a poor estimate of the true parameters due to the scarcity of microphone data.

The linear-weighted WMFD-projected GPLDA system also shows improvement in EER and DCF values over the pooled LDA- and WMFD-projected GPLDA systems in *interview-interview* condition as MFD projection extracts more speaker discriminant information with limited number of sessions/ speaker data, and linear-weighted GPLDA approach estimates reliable model parameters with limited microphone data. It is also observed that when the telephone weighting parameter (α_{tel}) is reduced, it significantly affects the performance of linear-weighted LDA- and WMFD-projected GPLDA systems in *telephone-telephone* condition as microphone-estimated GPLDA parameters are a poor estimate.

7. Conclusions

This paper proposed the addition of a WMFD projection prior to length-normalised GPLDA modelling in order to compensate the additional session variation. In limited microphone data conditions, a linear-weighted approach was introduced to increase the influence of microphone speech dataset. The linear-weighted WMFD-projected GPLDA system showed improvements in EER and DCF values over the pooled LDA- and WMFD-projected GPLDA systems in *interview-interview* condition as WMFD projection extracts more speaker discriminant information with limited number of sessions/ speaker data, and linear-weighted GPLDA approach estimates reliable model parameters with limited microphone data.

8. Acknowledgements

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

9. References

[1] 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.

[2] "The NIST year 2006 speaker recognition evaluation plan," NIST, Tech. Rep., 2006. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2006/

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

[4] 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.

[5] M. McLaren and D. van Leeuwen, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 755–766, 2012.

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

[7] A. Kanagasundaram, D. Dean, J. Gonzalez-Dominguez, S. Sridharan, D. Ramos, and J. Gonzalez-Rodriguez, "Improving the PLDA based speaker verification in limited microphone data conditions," in *Proceed. of INTERSPEECH*. International Speech Communication Association (ISCA), 2013.

[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, 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.

[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. 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, 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.

[15] 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*, 2013.

[16] 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.