



Improving the PLDA based Speaker Verification in Limited Microphone Data Conditions

A. Kanagasundaram*, D. Dean*, J. Gonzalez-Dominguez†,
S. Sridharan*, D. Ramos†, J. Gonzalez-Rodriguez†

* Speech Research Laboratory, Queensland University of Technology, Australia

† ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain

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

† {javier.gonzalez, daniel.ramos, joaquin.gonzalez}@uam.es

Abstract

A significant amount of speech data is required to develop a robust speaker verification system, but it is difficult to find enough development speech to match all expected conditions. In this paper we introduce a new approach to Gaussian probabilistic linear discriminant analysis (GPLDA) to estimate reliable model parameters as a linearly weighted model taking more input from the large volume of available telephone data and smaller proportional input from limited microphone data. In comparison to a traditional pooled training approach, where the GPLDA model is trained over both telephone and microphone speech, this linear-weighted GPLDA approach is shown to provide better EER and DCF performance in microphone and mixed conditions in both the NIST 2008 and NIST 2010 evaluation corpora. Based upon these results, we believe that linear-weighted GPLDA will provide a better approach than pooled GPLDA, allowing for the further improvement of GPLDA speaker verification in conditions with limited development data.

Index Terms: speaker verification, i-vector, total-variability, length-normalization, Gaussian PLDA

1. Introduction

A significant amount speech data is required to develop a robust speaker verification system, especially in the presence of large intersession variability. In practice it is feasible to collect substantial amount of telephone data but microphone data is harder to acquire. For example, large amount of telephone speech data is available in the NIST Speaker Recognition Evaluation (SRE) databases; however, microphone speech data is scarce in this data set. In addressing this disparity data sources, researchers have pooled the telephone and microphone data for the development of modern state of the art speaker verification systems such as the Gaussian probabilistic linear discriminant analysis (GPLDA) [3]. In this paper we take a new approach to estimate reliable model parameters as a linear weighted model, taking more input from the large volume of available telephone data and smaller proportional input from limited microphone data.

Joint factor analysis (JFA), as originally proposed by Kenny [4], has evolved as a powerful tool in speaker verification to model the inter-speaker variability and to compensate for channel/session variability in the context of high-dimensional Gaussian mixture model (GMM) super-vectors. Dominguez *et al.* [1] have previously investigated the JFA approach with lim-

ited microphone speech data, and have analyzed several approaches, including joining matrices, pooled statistics and scaling statistics to avoid the data scarcity problem.

A few years ago, Dehak *et al.* [5] have proposed a front-end factor analysis technique, termed i-vector, which has evolved from JFA. Rather than taking the JFA approach of modeling a speaker and channel variability space separately, the i-vector approach forms a low-dimensional total-variability space that models both speaker and channel variability. Senoussaoui *et al.* [2] have extended Dehak work, where they have analyzed the i-vector speaker verification approach with microphone speech. They have introduced the concatenated total-variability approach to extract i-vector features from telephone and microphone sources, where total-variability approach is separately trained using telephone and microphone sources and concatenated to form a concatenated total-variability space [2].

Recently, Kenny introduced the PLDA approach to directly model speaker and session variability within the i-vector space [6], and Senoussaoui *et al.* [3] have analyzed the heavy-tailed PLDA (HTPLDA) approach with microphone data conditions. They applied a concatenated total-variability approach to extract useful speaker information from telephone and microphone speech data. However, there have been no investigations into how the length-normalized GPLDA model parameters can be explicitly modeled using both rich telephone and limited microphone speech data.

The main aim of this paper is to explicitly model the GPLDA model parameters using rich telephone and limited microphone sourced speech data in the PLDA model domain. Initially, in the i-vector feature domain, two different types of total-variability approaches, including pooled and concatenated approaches are analyzed to extract the speaker information from telephone microphone speech data. Subsequently, in the PLDA model domain, pooled and linear weighted approaches are investigated to effectively model the GPLDA model parameter from telephone and microphone speech data.

This paper is structured as follows: Section 2 details the i-vector feature extraction techniques. Section 3 gives a brief introduction to the length-normalized GPLDA based speaker verification system, and Section 4 details the GPLDA model parameter estimation methods in scarce microphone speech data. The experimental protocol and corresponding results are given in Section 5 and Section 6. Section 7 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 [7]. An i-vector speaker and channel dependent GMM super-vector can be represented by,

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

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 $N(0, \mathbf{I})$. 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.* [8, 4].

The total-variability subspace is responsible for defining a suitable subspace from which i-vectors are extracted. In this paper, we investigate telephone and microphone speech based length-normalized GPLDA system, and for this approach the total-variability subspace should be trained in a manner that best exploits the useful speaker variability contained in speech acquired from both telephone and microphone sources. In this paper, both the pooled and concatenated total-variability approaches will be investigated with length-normalized GPLDA speaker verification. For pooled total-variability approach, the total-variability subspace ($R_w^{telmic} = 500$) is trained on telephone and microphone speech utterances together. The major advantage is that it's simplified approach. For concatenated total-variability approach, the separate total-variability telephone-only subspace ($R_w^{tel} = 400$) and microphone-only subspace ($R_w^{mic} = 100$) are trained separately using telephone and microphone speech, then both subspace transformations are concatenated to create a single total-variability space.

3. Length-normalized GPLDA system

3.1. PLDA modeling

Kenny has introduced the PLDA approach to directly model the speaker and channel variations in one variability i-vector space [6]. He has introduced the Gaussian PLDA (GPLDA) and heavy-tailed PLDA (HTPLDA) approaches to speaker verification system, and his studies have also found that HTPLDA approach shows significant improvement over GPLDA [6] as the i-vector feature behavior is heavy-tailed distribution. Recently Garcia-Romero *et al.* have introduced length-normalized Gaussian PLDA approach to convert the i-vector feature behavior from heavy-tailed to Gaussian [9]. In this paper, we have chosen the length-normalized GPLDA, as it's a simplified and computationally efficient approach. The length-normalization approach is detailed by Garcia-Romero *et al.* [9], and this approach is applied on development and evaluation data prior to GPLDA modeling. A speaker and channel dependent length-normalized i-vector, \mathbf{w}_r , can be defined as

$$\mathbf{w}_r = \bar{\mathbf{w}} + \mathbf{U}_1 \mathbf{x}_1 + \boldsymbol{\varepsilon}_r \quad (2)$$

where for given speaker recordings $r = 1, \dots, R$; \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 [10].

3.2. GPLDA scoring

Scoring in GPLDA speaker verification systems is conducted using the batch likelihood ratio between a target and test i-vector [6]. Given two i-vectors, \mathbf{w}_{target} and \mathbf{w}_{test} , the batch likelihood ratio can be calculated as follows,

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

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

4. GPLDA parameter estimation

In i-vector feature domain, pooled and concatenated total-variability approaches were analyzed to exploit sufficient speaker variation from telephone and microphone speech sources, and we have found from experiments in Section 6.1 that pooled total-variability approach is better than concatenated total-variability approach. Hereafter pooled total-variability based i-vector feature extraction will be used for PLDA model domain investigations. In this section, in PLDA model domain, we analyze pooled and linear weighted approaches to estimate the proper GPLDA model parameters from rich telephone and scarce microphone speech data.

4.1. Pooled approach

It is commonly believed that robust probabilistic parameters can be estimated if adequate amount of speech data is available. Telephone and microphone speech is pooled together to create large development data set in pooled approach. 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.

4.2. Linear weighted approach

If sufficient amount of telephone and microphone speech data is available, we believe the pooled approach would be better approach. However, in NIST conditions, while larger amount of telephone sourced speech are available, the same does not apply for microphone sourced speech. In addition, the telephone and microphone sourced speech have different behavior and if both are pooled together, we may lose the influence of microphone sourced data against the large volume telephone sourced data. Thus, the pooled approach is unlikely to help to improve the speaker verification in microphone conditions.

We believe that a technique called linear weighted approach can be used to increase the influence of microphone speech data. Firstly, the GPLDA model parameters, including mean ($\bar{\mathbf{w}}_{tel}$), precision matrix ($\boldsymbol{\Lambda}_{tel}$) and eigenvoice matrix (\mathbf{U}_{1tel}) are estimated using telephone speech data. Similarly, the GPLDA model parameters, including mean ($\bar{\mathbf{w}}_{mic}$), precision matrix ($\boldsymbol{\Lambda}_{mic}$) and eigenvoice matrix (\mathbf{U}_{1mic}) are also estimated using microphone speech data as well. After that linear weighted approach is used to combined the both telephone

Table 1: Performance comparison of pooled and concatenated total-variability approach based GPLDA systems on NIST 08 short2-short3 and NIST 10 core-core conditions. The best performing systems by both EER and DCF are highlighted across each row.

(a) NIST 08 short2-short3 condition								
Total-variability approach	Interview-interview		Interview-telephone		Telephone-microphone		Telephone-telephone	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
Concatenated approach	5.21%	0.0266	6.27%	0.0314	4.82%	0.0240	2.87%	0.0156
Pooled approach	4.29%	0.0214	5.51%	0.0254	4.35%	0.0195	2.63%	0.0126

(b) NIST 10 core-core condition								
Total-variability approach	Interview-interview		Interview-telephone		Interview-microphone		Telephone-telephone	
	EER	DCF_{old}	EER	DCF_{old}	EER	DCF_{old}	EER	DCF_{old}
Concatenated approach	7.37%	0.0320	4.84%	0.0231	4.44%	0.0210	3.67%	0.0156
Pooled approach	6.76%	0.0292	4.41%	0.0220	4.10%	0.0196	3.41%	0.0152

and microphone based parameters. The combined parameters can be estimated as follows,

$$\bar{\mathbf{w}}_{telmic} = \alpha \bar{\mathbf{w}}_{tel} + (1 - \alpha) \bar{\mathbf{w}}_{mic} \quad (4)$$

$$\mathbf{\Lambda}_{telmic} = \alpha \mathbf{\Lambda}_{tel} + (1 - \alpha) \mathbf{\Lambda}_{mic} \quad (5)$$

$$\mathbf{U}_{1telmic} = \alpha \mathbf{U}_{1tel} + (1 - \alpha) \mathbf{U}_{1mic} \quad (6)$$

5. Experimental methodology

The GPLDA based experiments were evaluated using the NIST 2008 and 2010 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 $P_{target} = 0.01$ [11]. The performance for the NIST 2010 SRE was evaluated using the EER and the old minimum decision cost function (DCF_{old}), calculated using $C_{miss} = 10$, $C_{FA} = 1$, and $P_{target} = 0.01$ [12].

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 $R_w = 500$. Total variability subspace was used to calculate the i-vector speaker representations. For the concatenated total-variability approach, telephone sourced total-variability space ($R_w^{tel} = 400$) was trained on telephone speech data from NIST 2004, 2005 and 2006 SRE corpora as well as Switchboard II, and microphone sourced total-variability space ($R_w^{mic} = 100$) was trained on microphone speech data from 2005 and 2006 SRE corpora, and both total-variability spaces were concatenated. 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 estimation. Randomly selected telephone and microphone utterances from NIST04, 05 and 06 were pooled to form the S-normalization dataset [13].

6. Results and discussions

Initially, the pooled and concatenated total-variability approaches were analyzed with the GPLDA system to identify the better total-variability approach to extract i-vector features, which shows greater variation from telephone and microphone sourced speech. Later, based upon previous experiment results, the better total-variability approach was also analyzed with pooled and linear weighted based GPLDA modeling approaches to estimate the robust GPLDA model parameters from both telephone and microphone sourced speech.

6.1. I-vector feature domain investigations (Baseline)

In this section, the pooled and concatenated total-variability approaches based i-vector feature extraction techniques were analyzed with length-normalized GPLDA system. Table 1 presents the results comparing the performance of pooled and concatenated total-variability approaches based GPLDA system on NIST 08 short2-short3 and NIST 10 core-core conditions. The pooled total-variability approach GPLDA system has shown considerable improvement over concatenated total-variability approach GPLDA system. The results suggest that the influence of microphone speech data cannot be significantly increased by concatenated total-variability approach. Based upon this outcome, the pooled total-variability based i-vector feature extraction approach will be used for following section experiments.

6.2. GPLDA model domain investigations

The experiments were carried out to identify the best length-normalized GPLDA model parameter estimation approach. Table 2 presents the results comparing the performance of the linear weighted GPLDA system against the pooled GPLDA system on NIST 08 and 10 standard evaluation conditions. If a limited amount of microphone speech data is pooled together with rich amount of telephone speech data (Female (1286 tel and 100 mic speakers), Male (1034 tel and 83 mic speakers)), we believe the influence of microphone speech would be reduced. In order to increase the influence of microphone data, the linear weighted based GPLDA approach was analyzed several values of weights (α_{tel}) with each interval of 0.1. It can be clearly seen from the Table 2 results that when the influence of microphone speech data is increased over telephone speech by selecting the α_{tel} of 0.8, the system shows better performance in microphone speech conditions. However, as α_{tel} is further reduced, the performance is reduced in both telephone and microphone speech

Table 2: Performance comparison of pooled and linear weighted based GPLDA modeling approaches on NIST 08 short2-short3 and NIST 10 core-core conditions. The best performing systems by both EER and DCF are highlighted across each row.

(a) NIST 08 short2-short3 condition								
Weight (α)	Interview-interview		Interview-telephone		Telephone-microphone		Telephone-telephone	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
Baseline system (Pooled approach)								
-	4.29%	0.0214	5.51%	0.0254	4.35%	0.0195	2.63%	0.0126
New approach (Linear weighted approach)								
1.0	4.23%	0.0194	4.89%	0.0230	3.74%	0.0169	2.45%	0.0139
0.9	4.10%	0.0184	5.07%	0.0233	3.68%	0.0167	2.47%	0.0140
0.8	3.95%	0.0180	4.97%	0.0235	3.67%	0.0166	2.62%	0.0142
0.7	4.04%	0.0184	5.16%	0.0246	3.72%	0.0173	2.72%	0.0148
0.6	4.18%	0.0190	5.34%	0.0263	4.14%	0.0187	2.80%	0.0151
0.5	4.43%	0.0203	5.81%	0.0283	4.55%	0.0208	2.98%	0.0154

(b) NIST 10 core-core condition								
Weight (α)	Interview-interview		Interview-telephone		Interview-microphone		Telephone-telephone	
	EER	DCF_{old}	EER	DCF_{old}	EER	DCF_{old}	EER	DCF_{old}
Baseline system (Pooled approach)								
-	6.76%	0.0292	4.41%	0.0220	4.10%	0.0196	3.41%	0.0152
New approach (Linear weighted approach)								
1.0	6.56%	0.0281	4.48%	0.0203	3.81%	0.0194	3.42%	0.0148
0.9	6.40%	0.0274	4.04%	0.0200	3.80%	0.0190	3.53%	0.0156
0.8	6.36%	0.0276	4.04%	0.0201	3.80%	0.0194	3.67%	0.0160
0.7	6.41%	0.0284	4.10%	0.0209	3.94%	0.0198	3.67%	0.0164
0.6	6.54%	0.0294	4.27%	0.0224	4.27%	0.0203	3.67%	0.0165
0.5	6.78%	0.0301	4.35%	0.0237	4.23%	0.0211	3.81%	0.0179

conditions. This is because the microphone-estimated GPLDA parameters provide a poor estimate of the true parameters due to the scarcity of microphone data. Based upon this outcome, we believe that if the amount of microphone is further increased, speaker verification system could achieve further improvement in microphone conditions when the α_{tel} is less than 0.8.

At the optimal α_{tel} of 0.8, it can be clearly seen that the linear weighted GPLDA approach shows over 9% relative improvement in DCF for NIST 2008 *interview-interview* and mismatched conditions, and over 5% relative improvement in EER for NIST 10 *interview-interview* and mismatched conditions.

7. Conclusion

In this paper, we introduced a new approach to Gaussian probabilistic linear discriminant analysis (GPLDA) to estimate reliable model parameters as a linear weighted model taking more input from the large volume of available telephone data and smaller proportional input from limited microphone data. When compared to a traditional pooled GPLDA approach, this technique showed 9% relative improvement in DCF for NIST 2008 *interview-interview* and mismatched conditions, and over 5% relative improvement in EER for NIST 2010 *interview-interview* and mismatched conditions. Based upon these results, we believe that linear-weighted GPLDA will provide a better approach than pooled GPLDA, allowing for the further improvement of GPLDA speaker verification in conditions with limited development data.

8. Acknowledgements

This project was supported by the European Commission Marie Curie ITN "Bayesian Biometrics for Forensics" (BBfor2) net-

work and the Spanish Ministerio de Economia y Competitividad under the project TEC2012-37585-C02-01.

9. References

- [1] J. Gonzalez-Dominguez, B. Baker, R. Vogt, R. Gonzalez-Rodriguez, and S. Sridharan, "On the use of factor analysis with restricted target data in speaker verification," pp. 103–108, 2010.
- [2] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Proc. Odyssey Speaker and Language Recognition Workshop*, pp. 28–33, 2010.
- [3] M. Senoussaoui, P. Kenny, P. Dumouchel, and F. Castaldo, "Well-calibrated heavy tailed Bayesian speaker verification for microphone speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 4824–4827, IEEE, 2011.
- [4] 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.
- [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, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic*, 2010.
- [7] 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.
- [8] 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.
 - [9] 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.
 - [10] A. Kanagasundaram, D. Dean, S. Sridharan, and R. Vogt, “PLDA based speaker verification with weighted LDA techniques,” in *Proc. Odyssey Workshop*, no. 34-38, 2012.
 - [11] NIST, “The NIST year 2008 speaker recognition evaluation plan,” tech. rep., NIST, 2008.
 - [12] NIST, “The NIST year 2010 speaker recognition evaluation plan,” tech. rep., NIST, 2010.
 - [13] S. Shum, N. Dehak, R. Dehak, and J. Glass, “Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification,” *Proc. Odyssey*, no. 76-82, 2010.