IMPROVING OUT-DOMAIN PLDA SPEAKER VERIFICATION USING UNSUPERVISED INTER-DATASET VARIABILITY COMPENSATION APPROACH

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

Experimental studies have found that when the state-of-the-art probabilistic linear discriminant analysis (PLDA) speaker verification systems are trained using out-domain data, it significantly affects speaker verification performance due to the mismatch between development data and evaluation data. To overcome this problem we propose a novel unsupervised inter dataset variability (IDV) compensation approach to compensate the dataset mismatch. IDV-compensated PLDA system achieves over 10% relative improvement in EER values over out-domain PLDA system by effectively compensating the mismatch between in-domain and out-domain data.

Index Terms— speaker verification, PLDA, domain adaptation, inter-dataset variability

1. INTRODUCTION

A significant amount of development data, especially in the presence of large intersession variability, is required to develop a speaker verification system. Recent studies have found that when speaker verification is developed in one domain data and evaluated in another domain data, the dataset mismatch significantly affects the speaker verification performance [1, 2, 3]. Therefore significant amount of target domain data is required to develop speaker verification system in order to achieve state-of-the-art performance. However, it is hard to collect adequate amount of target domain data, specially speaker labelled data in real world environments. In recent times, researchers have been proposing several approaches to achieve state-of-the-art speaker verification performance if significant amount of out-domain data and limited in-domain unlabelled is available. This problem is defined as domain adaptation.

Recently, Garcia-Romero *et at.* [1] have found that the adaptation of the PLDA parameters produces the largest gains, and universal background model (UBM) and total-variability matrix would not be required to estimate on in-domain data. They have studied several supervised approaches, including fully Bayesian adaptation, approximate *maximum a posteriori* (MAP) adaptation, weighted likelihood [1]. Aronowitz [2] introduced inter dataset variability compensation (IDVC) to explicitly compensate for dataset shift in the i-vector space, which is based on nuisance attribute projection (NAP) method. For IDVC estimation, out-domain Switchboard dataset is partitioned into several sub datasets. Recently, Garcia-Romero *et at.* [4] have also introduced agglomerative hierarchical clustering (AHC) based unsupervised approach for domain adaptation.

In this paper, a novel unsupervised inter-dataset variability (IDV) is introduced in order to compensate the mismatch between out-domain data and in-domain data. Our approach is similar to the IDVC approach proposed by Aronowitz in [2] but in contrast, out-domain Switchboard dataset is not required to be partition into several subsets to estimate the inter dataset variability compensation. Recently, we have proposed short utterance variance (SUV) approach to capture the utterance variation for short utterance PLDA speaker verification system [5, 6]. In this paper, similar idea is taken to capture the variation between in-domain and out-domain data. The variation between in-domain and out-domain data is defined as the outer product of difference between out-domain i-vectors and average of in-domain i-vectors. The IDV compensation matrix is estimated using Cholesky decomposition of inverse of variation matrix. We analyse how the limited in-domain unlabelled data that is available for IDV compensation estimation, affects the speaker verification performance.

This paper is structured as follows: Section 2 details the i-vector feature extraction techniques. Section 3 details the inter dataset variability compensation approach. Section 4 explains the Gaussian PLDA (GPLDA) based speaker verification system. 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 Gaussian mixture model (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,

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

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(0,1)*. Extracting an i-vector from the total-variability subspace is essentially a *maximum a-posteriori adaptation* (MAP) of $\mathbf{w}$ in the subspace defined by $\mathbf{T}$. An efficient procedure for the optimization of the total-variability subspace $\mathbf{T}$ and subsequent extraction of i-vectors is described Dehak *et al.* [8, 9]. 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. IDV COMPENSATION APPROACH

When PLDA speaker verification is trained using out-domain data, it significantly affects the speaker verification performance due to mismatch between development data and evaluation data. Inter dataset compensation techniques are required to compensate this mismatch. Aronowitz [2] introduced inter dataset variability compensation (IDVC) to explicitly compensate for dataset shift in the i-vector space, which is based on NAP method. Out-domain Switchboard dataset is required to partition into sub datasets in order to estimate the IDVC approach.

In this section, we introduce a different IDV approach to that proposed in [2] to compensate the mismatch between in-domain and out-domain data. For this estimation, out-domain Switchboard dataset is not required to partition into sub datasets. Recently, we have proposed SUV estimation approach for short utterance speaker verification system [5]. Similarly to the SUV estimation [5], the mismatch between in-domain and out-domain is captured using the outer product of the difference between the out-domain i-vectors and average of speaker unlabelled in-domain i-vectors. The dataset mismatch variation, $\mathbf{S}_{IDV}$, can be calculated as follows,

$$\mathbf{S}_{IDV} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}_n^{OD} - \mathbf{w}_{avg}^{ID})(\mathbf{w}_n^{OD} - \mathbf{w}_{avg}^{ID})^T \quad (2)$$

where $\mathbf{w}_n^{OD}$ is out-domain i-vectors, and $\mathbf{w}_{avg}^{ID}$ is average of in-domain unlabelled i-vectors. The IDV decorrelated matrix, $\mathbf{D}$, is calculated using the Cholesky decomposition of $\mathbf{DD}^T = \frac{1}{\mathbf{S}_{IDV}}$. After the IDV decorrelated matrix, $\mathbf{D}$, is estimated, inter dataset variability compensated out-domain i-vectors are extracted as follows,

$$\hat{\mathbf{w}}_{IDV} = \mathbf{D}^T\mathbf{w} \quad (3)$$

Once inter-dataset variability compensated i-vectors, LDA projection is applied to compensate the additional session variation prior to the PLDA modelling and reduce the dimensionality [10], which is explained in following in Section 3.1.

3.1. LDA approach

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

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

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

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 $\mathbf{S}_b\mathbf{v} = \lambda\mathbf{S}_w\mathbf{v}$. The IDV-compensated LDA-projected i-vector can be calculated as follows,

$$\hat{\mathbf{w}}_{IDV\text{-}LDA} = \mathbf{A}^T\mathbf{w} \quad (8)$$

After LDA-projection, length-normalized GPLDA model parameters are estimated in as described in Section 4.

4. LENGTH-NORMALIZED GPLDA SYSTEM

4.1. PLDA modelling

In this paper, we have chosen the length-normalized GPLDA, as it is also a simplified and computationally efficient approach [11]. The length-normalization approach is detailed by Garcia-Romero *et al.* [11], and this approach is applied on development and evaluation data prior to GPLDA modelling. A speaker and channel dependent length-normalized i-vector, $\hat{\mathbf{w}}_r$ can be defined as,

$$\hat{\mathbf{w}}_r = \bar{\hat{\mathbf{w}}} + \mathbf{U}_1\mathbf{x}_1 + \boldsymbol{\varepsilon}_r \quad (9)$$

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 $\mathbf{\Lambda}^{-1}$. We assume that precision matrix ($\mathbf{\Lambda}$) is full rank. Prior to GPLDA modelling, standard LDA approach is applied to compensate the additional channel variations as well as reduce the computational time [12].

4.2. GPLDA scoring

Scoring in GPLDA speaker verification systems is conducted using the batch likelihood ratio between a target and test i-vector [13]. Given two i-vectors, $\mathbf{w}_{target}$ and $\mathbf{w}_{test}$, the batch likelihood ratio can be calculated as follows,

$$\ln \frac{P(\mathbf{w}_{target}, \mathbf{w}_{test} \mid H_1)}{P(\mathbf{w}_{target} \mid H_0) P(\mathbf{w}_{test} \mid H_0)} \tag{10}$$

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

5. EXPERIMENTAL METHODOLOGY

The proposed methods were evaluated using the 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 $P_{target} = 0.01$ [14]. Outer-domain data is defined as Switchboard I, II phase I, II, III corpora, and in-domain data is defined as NIST 2004, 2005 and 2006 SRE corpora.

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 Switchboard I, II phase I, II, III 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 was trained using Switchboard I, II phase I, II, III corpora. For out-domain PLDA speaker verification system, the GPLDA parameters were trained using Switchboard I, II phase I, II, III corpora. We empirically selected the number of eigenvoices (N_1) equal to 120 as best value according to speaker verification performance over an evaluation set. 150 eigenvectors were selected for LDA estimation. S-normalisation was applied for experiments. The randomly selected telephone and microphone utterances from NIST 2004, 2005 and 2006 were pooled to form the NIST S-normalisation dataset, and randomly selected utterances from Switchboard I, II phase I, II, III were pooled to form the Switchboard S-normalisation dataset [15].

Table 1. *Performance comparison of LDA-projected GPLDA systems on common condition of NIST 2008 short2-short3 evaluation condition when GPLDA and score normalization is trained using out-domain and in-domain data.*

GPLDA training	Score normalization			
	Out-domain data		In-domain data	
	EER	DCF	EER	DCF
Out-domain	4.69%	0.0232	3.87%	0.0169
In-domain	3.62%	0.0177	**3.38%**	**0.0160**

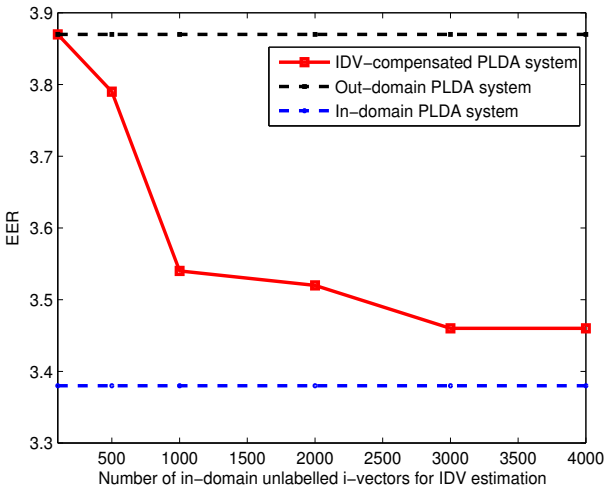

Fig. 1. *Comparison of IDV-compensated PLDA system against in-domain and out-domain PLDA system.*

6. RESULTS AND DISCUSSIONS

6.1. Out-domain PLDA speaker verification system

In this section, the performance of LDA-projected PLDA speaker verification system was compared on NIST 2008 short2-short3 condition when GPLDA and score normalization were respectively trained using in-domain and out-domain data. Table 1 compares the performance of in-domain and out-domain PLDA speaker verification system. It can be observed from Table 1 that though GPLDA is trained using out-domain data, if in-domain data is used for score normalization, it significantly improves the speaker verification performance as score normalization data behaviour matches with evaluation data. Further, it was also found that when GPLDA and score-normalization are trained using in-domain data, the system achieves the best performance.

6.2. IDV-compensated PLDA speaker verification system

In previous section, it was found that if GPLDA and score normalization are trained using in-domain data, PLDA speaker

verification achieves the best performance. However, in real world scenario, it is hard to collect labelled in-domain data, whereas unlabelled data can be collected. It was also found that if speaker verification is developed using out-domain data, system achieves poor performance due to mismatch between development data (out-domain) and evaluation data (in-domain).

Though unlabelled in-domain data can be collected, it is hard to collect huge amount of data. In Figure 1 we present the results for our IDV-compensated PLDA speaker verification system. We have experimented with our approach when limited amount of unlabelled in-domain data is used to estimate the IDV compensation matrix.

Figure 1 compares the EER values of IDV-compensated PLDA system against in-domain and out-domain PLDA systems when IDV compensation matrix is trained using different amount of unlabelled in-domain data. IDV-compensated PLDA system achieves over 10% relative improvement in EER values over out-domain PLDA system showing that our IDV approach effectively compensates the mismatch between in-domain and out-domain data.

7. CONCLUSION

The novel unsupervised inter dataset variability (IDV) compensation approach was proposed in this paper to improve the out-domain PLDA speaker verification systems. It is well known that when PLDA is trained using out-domain data, it significantly affects speaker verification performance due to the mismatch between development data and evaluation data. We introduced a novel unsupervised IDV compensation approach that compensates for the dataset mismatch. Our IDV-compensated PLDA system achieved over 10% relative improvement in EER values over the out-domain PLDA system showing that the IDV approach compensates the mismatch between in-domain and out-domain data.

8. ACKNOWLEDGEMENTS

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

9. REFERENCES

[1] Daniel Garcia-Romero and Alan McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4047–4051.

[2] Hagai Aronowitz, "Inter dataset variability compensation for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4002–4006.

[3] Ondrej Glembek, Jeff Ma, Pavel Matejka, Bing Zhang, Oldrich Plchot, Lukáš Burget, and Spyros Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," .

[4] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsuprvised domain adaptation for i-vector speaker recognition," in *Proc. Odyssey Workshop*, 2014.

[5] 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.

[6] A. Kanagasundaram, D.B. Dean, and S. Sridharan, "Short utterance PLDA speaker verification using SNWLDA and variance modelling techniques," in *Proceedings of the 15th Australian International Conference on Speech Science and Technology*. The Australian Speech Science & Technology Association, 2014.

[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*, 2009, p. 1559 1562.

[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. PP, no. 99, pp. 1 –1, 2010.

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

[10] 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.

[11] 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.

[12] 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.

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

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

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