Queensland University of Technology
Brisbane Australia

This is the author's version of a work that was submitted/accepted for publication in the following source:

Rahman, Md Hafizur, Kanagasundaram, Ahilan, Dean, David, & Sridharan, Sridha
(2015)
Dataset-invariant covariance normalization for out-domain PLDA speaker verification. In
Proceedings of the 16th Annual Conference of the International Speech Communication Association, Interspeech 2015, International Speech Communication Association, Maritim International Congress Center, Dresden, Germany, pp. 1017-1021.

This file was downloaded from: http://eprints.qut.edu.au/85083/

© Copyright 2015 [Please consult the author]

Notice: *Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:*

Dataset-Invariant Covariance Normalization for Out-domain PLDA Speaker Verification

Md Hafizur Rahman[1], *Ahilan Kanagasundaram*[2], *David Dean*[3], *Sridha Sridharan*[4]

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
{m20.rahman[1], a.kanagasundaram[2], s.sridharan[4]}@qut.edu.au, ddean@ieee.org[3]

Abstract

In this paper we introduce a novel domain-invariant covariance normalization (DICN) technique to relocate both in-domain and out-domain i-vectors into a third dataset-invariant space, providing an improvement for out-domain PLDA speaker verification with a very small number of unlabelled in-domain adaptation i-vectors. By capturing the dataset variance from a global mean using both development out-domain i-vectors and limited unlabelled in-domain i-vectors, we could obtain domain-invariant representations of PLDA training data. The DICN-compensated out-domain PLDA system is shown to perform as well as in-domain PLDA training with as few as 500 unlabelled in-domain i-vectors for NIST-2010 SRE and 2000 unlabelled in-domain i-vectors for NIST-2008 SRE, and considerable relative improvement over both out-domain and in-domain PLDA development if more are available.

Index Terms: speaker verification, PLDA, DICN, domain adaptation

1. Introduction

In the past few years extensive research has been conducted in the field of speaker verification. Numerous methods have been proposed, like joint factor analysis (JFA) [1] and i-vector [2] based subspace modelling techniques, that have resulted in excellent speaker verification performance. But most techniques have only been investigated in relatively clean environments, with huge amounts of 'in-domain' development data. However, if we use this clean data for development of real world applications, it would produce poor performance, because of many factors that are not always considered in clean development data. One of the key reason is the mismatch between development and evaluation dataset.

The performance variation due to cross-domain speaker verification development and evaluation was first addressed at the Summer Workshop at Johns Hopkins University (JHU) held in 2013 [3]. Results presented in that workshop clearly showed the performance gap between in-domain and out-domain development for speaker verification. This task was deemed the 'Domain Adaptation Challenge' (DAC) at that workshop.

In response to this poor cross-domain speaker verification performance, Garcia-Romero *et al.* [4] found that training UBMs and total-variability matrices on in-domain or out-domain data have very limited effect on overall performance, but the effect on PLDA parameters were more pronounced. They investigated four adaptation techniques for supervised domain adaptation of PLDA parameters: fully Bayesian adaptation, approximate MAP, weighted likelihood and SPLDA parameter interpolation. Each of these techniques performed very similarly. Villalba *et al.* [5] introduced a variational Bayesian technique for adapting PLDA models from labeled out-domain to unlabeled in-domain data. Recently, Garcia-Romero *et al.* also introduced an agglomerative hierarchical clustering (AHC) method to cluster unlabeleld in-domain data for domain adaptation. To compensate the dataset shift in i-vector space Aronowitz [6] introduced an inter-dataset variability compensation (IDVC) technique based on nuisance attribute projection (NAP). Glembek *et al.* [7] proposed within-speaker covariance correction (WCC) and extended unsupervised adaptation of the LDA matrix to compensate the mismatch between training and testing datasets. Recently, Kanagasundaram *et al.* [8] introduced an improved IDVC technique, where dataset variability is captured using difference between out-domain i-vectors and average of in-domain i-vectors.

In this paper, a novel dataset invariant covariance normalization (DICN) approach is introduced to compensate the mismatch between in-domain and out-domain dataset in the i-vector space. Instead of capturing the mismatch directly between out-domain and in-domain data [8], we captured the mismatch as compared to the global mean i-vector. In this approach we used a set of unlabelled in-domain i-vectors and captured the mismatch using the difference between all i-vectors (in-domain and out-domain) and the global mean i-vector.

The rest of the paper is structured as follows: Section 2 details the i-vector feature extraction techniques. Section 3 details the DICN approach. Section 4 explains the linear discriminant analysis (LDA), and Section 5 presents GPLDA based speaker verification system. The experimental setup and corresponding results are given in Section 6 and Section 7. Finally, Section 8 concludes the paper.

2. I-vector based speaker verification

Single subspace-based i-vector speaker verification was first proposed by Dehak *et al.* [2]. This approach was inspired by his previous work, finding speaker discriminant information was lost in the discarded channel space of the earlier joint factor analysis (JFA) technique [9]. Unlike the JFA approach, in i-vector feature extraction the GMM super-vectors are represented in a single subspace called total-variability subspace. Both speaker and channel dependent GMM super-vector in i-vector can be represented by,

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

where $\boldsymbol{m}$ is the speaker and session independent UBM super-vector, $\boldsymbol{T}$ is a low rank matrix. $\boldsymbol{w}$ is total variability factor

which is normally distributed. We used only out-domain (SWB) data for total-variabilty subspace ($R_w = 500$) training. A detail procedure of total-variability subspace, $\boldsymbol{T}$, training and i-vector extraction is described in [2, 10].

3. DICN approach

One of the prime reason of poor out-domain PLDA speaker verification performance is the mismatch between in-domain and out-domain data in the i-vector subspace. Aronowitz proposed the IDVC [6, 11] approach to compensate dataset shift in i-vector space based on nuisance attribute projection (NAP). In this approach, the out-domain dataset is first partitioned into 12 separate subsets and the centers of these subsets are used to find a subspace (the inter-dataset variability subspace) spanned by the 12 centers using principal component analysis (PCA). Later, Kangasundaram *et al.* [8] proposed another IDVC approach where dataset variability is directly captured using average of unlabelled in-domain data.

In this section we introduce our novel DICN approach to capture the dataset variability more efficiently in the i-vector space. In this approach a global mean i-vector is determined using all in-domain and out-domain i-vectors and domain variability is captured using the outer product of the difference between all i-vectors (in-domain and out-domain) and the global mean i-vector. The dataset mismatch using DICN can be captured as follows,

$$\boldsymbol{S}_{DICN} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{w}_n - \bar{\boldsymbol{w}})(\boldsymbol{w}_n - \bar{\boldsymbol{w}})^{'} \tag{2}$$

where N is the total number of i-vectors (in-domain and out-domain) and $\bar{\boldsymbol{w}}$ is the global mean, which can be calculated as follows,

$$\bar{\boldsymbol{w}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{w}_i \tag{3}$$

The matrix $\boldsymbol{A}$ is used to scale the subspace, where $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{S}_{DICN}^{-1}$. Later, DICN compensated out-domain i-vectors are extracted as follows,

$$\boldsymbol{w}_{DICN} = \boldsymbol{A}^T\boldsymbol{w}_{out} \tag{4}$$

4. Linear discriminant analysis (LDA)

LDA is a channel compensation method [2, 12], which attempts to define new spatial axes that minimize the intra-class variance caused by channel effects and maximize the variance between speakers through the eigenvalue decomposition of,

$$\boldsymbol{S}_b\boldsymbol{v} = \tau\boldsymbol{S}_w\boldsymbol{v} \tag{5}$$

where τ is the eigenvalues, $\boldsymbol{v}$ is the eigenvector, $\boldsymbol{S}_w$ is within class matrix and $\boldsymbol{S}_b$ is between class matrix.

We compute the within class and between class matrix as follows,

$$\boldsymbol{S}_b = \sum_{s=1}^{S} n_s(\bar{\boldsymbol{w}}_s - \bar{\boldsymbol{w}})(\bar{\boldsymbol{w}}_s - \bar{\boldsymbol{w}})' \tag{6}$$

$$\boldsymbol{S}_w = \sum_{s=1}^{S}\sum_{1=1}^{n_s}(\boldsymbol{w}_i^s - \bar{\boldsymbol{w}}_s)(\boldsymbol{w}_i^s - \bar{\boldsymbol{w}}_s)' \tag{7}$$

where S is the total number of out-domain speakers, n_s is the number of sessions of speaker s. $\bar{\boldsymbol{w}}_s$ is the mean i-vector for each speaker and $\bar{\boldsymbol{w}}$ is the mean of all speakers which are defined by,

$$\bar{\boldsymbol{w}}_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\boldsymbol{w}_i^s \tag{8}$$

$$\bar{\boldsymbol{w}} = \frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\boldsymbol{w}_i^s \tag{9}$$

where N is the total number of sessions.

In the low dimensional space resulting from the linear transformation $\boldsymbol{G}$, the within class and between class matrices become $\boldsymbol{S}_w = \boldsymbol{G}^T\boldsymbol{S}_b\boldsymbol{G}$. An optimal transformation $\boldsymbol{G}$ would maximize trace ($\boldsymbol{S}_b$) and minimize ($\boldsymbol{S}_w$),

$$\max_{G}\{\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\} \tag{10}$$

Finally, the DICN compensated LDA projected i-vector can be calculated as follows,

$$\boldsymbol{w}_{DICN-LDA} = \boldsymbol{G}^T\boldsymbol{w}_{DICN} \tag{11}$$

5. Length-normalized GPLDA system

Recently, Garcia-Romero *et al.* [13] have introduced the length-normalized GPLDA approach, which is comparable to, but computationally more efficient than heavy-tailed (HTPLDA). This approach is used to transform the non-Gaussian i-vector feature behaviour into Gaussian i-vector feature behaviour. This technique consists of two steps: (1) linear whitening and (2) length-normalization. A linear-whitened i-vector $\boldsymbol{w}_{DICN-LDA-wht}$ can be estimated as follows,

$$\boldsymbol{w}_{DICN-LDA-wht} = \boldsymbol{d}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{w}_{DICN-LDA} \tag{12}$$

where $\boldsymbol{U}$ is an orthonormal matrix containing the eigenvectors and $\boldsymbol{d}$ is a diagonal matrix containing the corresponding eigenvalues. A length-normalized i-vector $\boldsymbol{w}^{norm}$ can be found as follows,

$$\boldsymbol{w}_{DICN-LDA}^{norm} = \frac{\boldsymbol{w}_{DICN-LDA-wht}}{\|\boldsymbol{w}_{DICN-LDA-wht}\|} \tag{13}$$

A speaker and channel dependent length-normalized i-vector can be defined as,

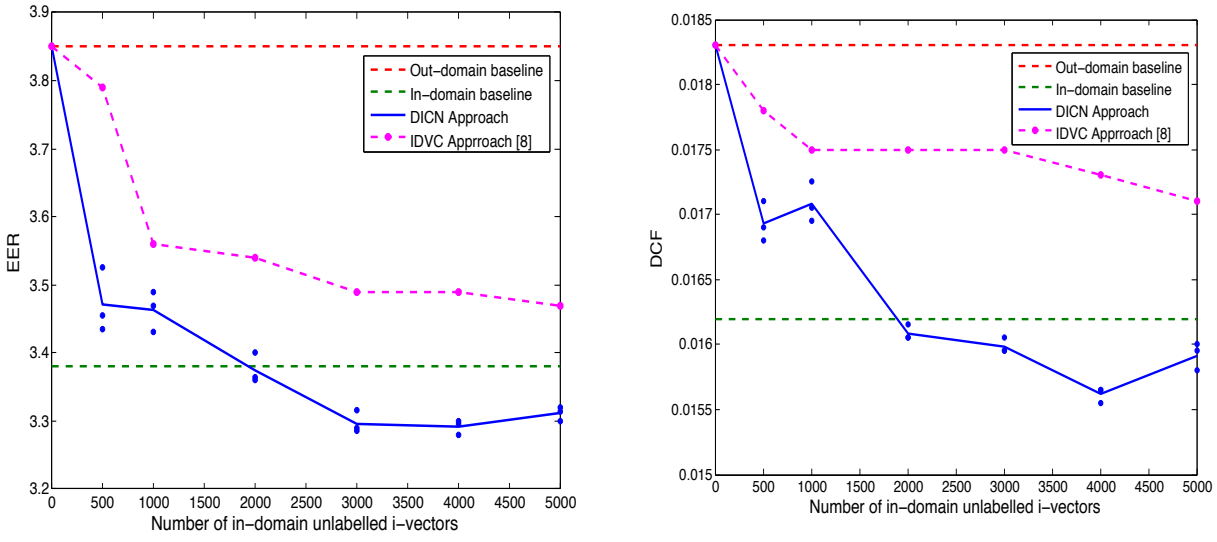
$$\boldsymbol{w}_{DICN-LDA-r}^{norm} = \boldsymbol{w}_{DICN-LDA}^{norm} + \boldsymbol{U}_1\boldsymbol{x}_1 + \boldsymbol{U}_2\boldsymbol{x}_{2r} + \boldsymbol{\epsilon}_r, \tag{14}$$

where for given speaker recordings $r = 1, 2, ...R$; $\boldsymbol{w}_{DICN-LDA}^{norm} + \boldsymbol{U}_1\boldsymbol{x}_1$ is the speaker specific part and $\boldsymbol{U}_2\boldsymbol{x}_{2r} + \boldsymbol{\epsilon}_r$ is the channel specific part; The covariance matrix of the speaker part is $\boldsymbol{U}_1\boldsymbol{U}_1^T$ and the covariance matrix of the channel part is $\boldsymbol{U}_2\boldsymbol{U}_2^T + \boldsymbol{\Lambda}^{-1}$. Training of the eigenmatrices $\boldsymbol{U}_1$and $\boldsymbol{U}_2$ are also same as learning the eigenvoice matrix $\boldsymbol{V}$ in JFA.

GPLDA scoring is calculated using the batch likelihood ratio [14]. Given target i-vectors $\boldsymbol{w}_{DICN-LDA-target}$ and test i-vectors $\boldsymbol{w}_{DICN-LDA-test}$, batch likelihood ratio can be calculated as follows,

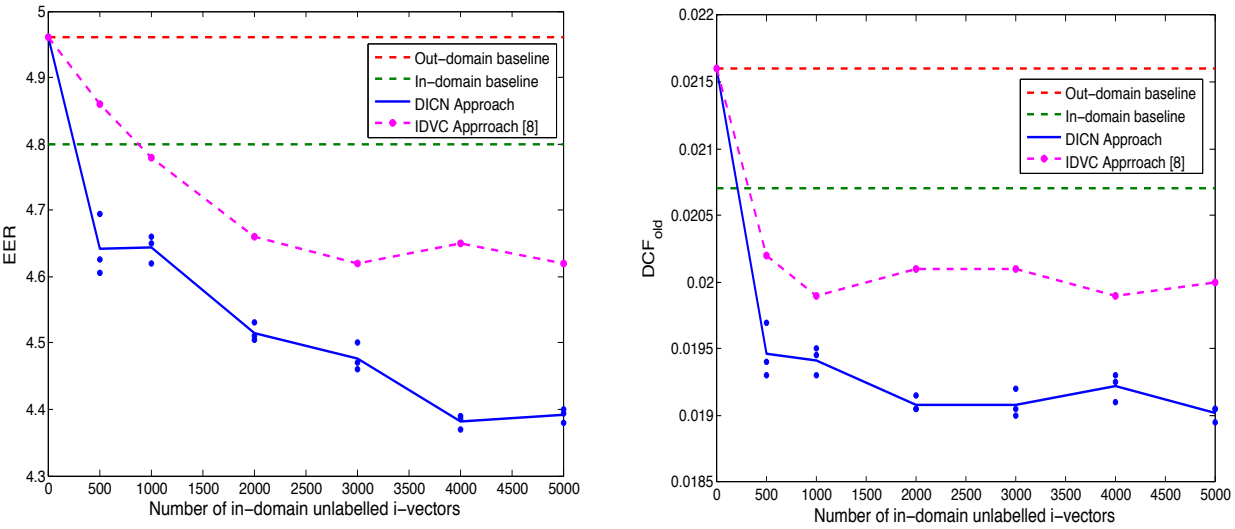
$$\ln\frac{P(\boldsymbol{w}_{DICN-LDA-target}, \boldsymbol{w}_{DICN-LDA-test} \mid H_1)}{P(\boldsymbol{w}_{DICN-LDA-target} \mid H_0)P(\boldsymbol{w}_{DICN-LDA-test} \mid H_0)} \tag{15}$$

where H_1: The speakers are same, H_0: The speaker are different



(a) EER (b) DCF

Figure 1: *The performance comparison of DICN compensated out-domain PLDA system against in-domain baseline, out-domain baseline and IDV compensated [8] out-domain PLDA system on the common set of the 2008 NIST SRE short2-short3 condition.* (a) *EER,* (b) *DCF.*



(a) EER (b) DCF_{old}

Figure 2: *The performance comparison of DICN compensated out-domain PLDA system against in-domain baseline, out-domain baseline and IDV compensated [8] out-domain PLDA system on the common set of the 2010 NIST SRE core-core condition.* (a) *EER,* (b) DCF_{old}.

6. Experimental setup

In our experimental setup, the out-domain dataset is defined as the Switchboard I, II phase I, II, III corpora which consists of 1115 male speakers with 13380 total sessions and 1231 female speakers with 14772 total sessions. The in-domain data is defined as NIST 2004, 2005 and 2006 SRE corpora. We adopted 13 dimensional feature-warped MFCCs with appended delta coefficients. Two gender dependent universal background models (UBM) with 512 Gaussian mixtures were trained on the out-domain (SWB) data. Baum-Welch statistics were calculated using those UBMs before training gender dependent total-variability sub-space with dimension of $R_w = 500$. Total-variability sub-spaces were also trained on the out-domain data. Prior to PLDA training, the out-domain i-vectors were projected using DICN and a 150-dimensional LDA space. We used length-normalized out-domain i-vectors for GPLDA parameter training. We empirically selected 120 best eigenvoices (dimension of $\boldsymbol{U}_1$) according to speaker verification performance. A full precision matrix was used for $\boldsymbol{\Lambda}$, rather than the diagonal. S-normalisation was used in the experiments as defined

Table 1: *Comparison of PLDA speaker verification on the common set of the NIST-2008 short2-short3 and NIST-2010 core-core evaluation conditions. GPLDA and score normalization are trained using both in-domain and out-domain data.*

GPLDA training	Score normalization	NIST-2008 EER	NIST-2008 DCF	NIST-2010 EER	NIST-2010 DCF_{old}
In-domain	In-domain	**3.38%**	**0.0162**	**4.80%**	**0.0207**
	Out-domain	3.62%	0.0177	5.08%	0.0217
Out-domain	In-domain	3.85%	0.0183	4.96%	0.0216
	Out-domain	4.70%	0.0230	5.68%	0.0268

Table 2: *Comparison of DICN compensated out-domain PLDA speaker verification with out-domain baseline and IDVC approach*

GPLDA training	Approach	NIST-2008 EER	NIST-2008 DCF	NIST-2010 EER	NIST-2010 DCF_{old}
Out-domain	–	3.85%	0.0183	4.96%	0.0216
	IDVC [8]	3.47%	0.0171	4.62%	0.0200
	DICN	**3.29%**	**0.0154**	**4.10%**	**0.0172**

in [14]. We randomly selected telephone and microphone sessions from NIST 2004, 2005 and 2006 datasets and combined them to form the NIST S-normalisation dataset, and randomly selected sessions from Switchboard I, II phase I, II, III datasets and combined them to form the SWB S-normalisation dataset.

Experiments were evaluated using the NIST 2008 and NIST 2010 Speaker Recognition Evaluation (SRE) corpora. For NIST 2008 evaluation the equal error rate (EER) and the minimum decision cost function (DCF) were used, calculated using $C_{miss} = 10, C_{FA} = 1$, and $P_{target} = 0.01$. For NIST 2010 evaluation the EER and the old minimum decision cost function (DCF_{old}) were used, calculated using $C_{miss} = 10, C_{FA} = 1$, and $P_{target} = 0.01$.

7. Results and discussion

This section presents the speaker verification performance with GPLDA paremeters trained on in-domain (NIST) and out-domain (SWB) data. Evaluation was performed on NIST 2008 short2-short3 condition and NIST 2010 core-core condition. Table 1 represents the in-domain and out-domain PLDA speaker verification performances. The results clearly show the performance gap between in-domain and out-domain GPLDA speaker verification performance. Accordingly, the best performance was achieved by training both the GPLDA parameters and score normalization statistics using only the in-domain data.

The out-domain PLDA speaker verification results with DICN trained on all in-domain data are shown in Table 2. Results clearly indicate that DICN with all in-domain data yields a 14% relative improvement in EER and a 15% relative improvement in DCF over the out-domain baseline for NIST-2008 evaluation and a 20% relative improvement in both EER and DCF_{old} over the out-domain baseline for NIST-2010 evaluation.

We also trained the DICN matrix with limited, unlabelled, in-domain data to replicate a data scarce condition. We selected unlabelled data in a random manner with three different seeds to show the performance variation of the system with different data subsets. Figure 1 represents the performance of NIST-2008 evaluation on short2-short3 condition and Figure 2 represents the performance of NIST-2010 evaluation on core-core condition with DICN compensated PLDA system against in-domain baseline, out-domain baseline and IDV compensated [8] out-domain PLDA systems. The average performance over the three trials is indicated by the solid line between the actual trials. For NIST-2008 evaluation there is a 14% relative performance improvement in EER and a 13% relative performance improvement in DCF with respect to the out-domain baseline. Also, for NIST-2010 evaluation there is an 11% relative performance improvement in EER and also a 12% relative performance improvement in DCF_{old} with respect to the out-domain baseline performance. Results also indicate that we require only 500 in-domain i-vectors for NIST-2010 evaluation and 2000 in-domain i-vectors for NIST-2008 evaluation to perform better than in-domain baseline PLDA. The performance variation between the three trials is mostly due to small data mismatch between NIST-2004, 2005 and 2006 datasets.

8. Conclusions

One of the main reason for poor out-domain PLDA speaker verification performance is the dataset mismatch between in-domain data and out-domain data. In this paper we introduced a dataset-invariant covariance normalisation (DICN) approach to compensate this dataset variability. We carried out the evaluation on both NIST-2008 and NIST-2010 SRE corpora, and demonstrated the performance of the out-domain PLDA system with DICN, without DICN and IDVC approach. For DICN training we presented the data scarce condition to replicate the real world scenario. Results clearly indicate that with DICN compensation PLDA speaker verification performs well enough to beat the out-domain as well as in-domain baseline performance.

9. Acknowledgements

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

10. References

[1] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[2] 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, 2011.

[3] Domain Adaptation Challenge 2013, Johns Hopkins University, 2013 speaker recognition workshop. Available online: http://www.clsp.jhu.edu/workshops/archive/ws13-summer-workshop/groups/spk-13/.

[4] D. Garcia-Romero and A. 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.

[5] J. Villalba and E. Lleida, "Unsupervised adaptation of PLDA by using variational bayes methods," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 744–748.

[6] H. 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.

[7] O. Glembek, J. Ma, P. Matejka, B. Zhang, O. Plchot, L. Burget, and S. Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4032–4036.

[8] A. Kanagasundaram, D. Dean, and S. Sridharan, "Improving out-domain PLDA speaker verification using unsupervised inter-dataset variability compensation approach," *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015.

[9] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification." in *INTERSPEECH*, vol. 9, 2009, pp. 1559–1562.

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

[11] H. Aronowitz, "Compensating inter-dataset variability in PLDA hyper-parameters for robust speaker recognition," *Speaker and Language Recognition Workshop (IEEE Odyssey)*, pp. 282–286, 2014.

[12] 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 Recogntion Workshop*, 2010, pp. 28–33.

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

[14] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.