Odyssey 2018 The Speaker and Language Recognition Workshop
26-29 June 2018, Les Sables d'Olonne, France

Domain-invariant I-vector Feature Extraction for PLDA Speaker Verification

Md Hafizur Rahman[1], *Ivan Himawan*[2], *David Dean*[3], *Clinton Fookes*[4], *Sridha Sridharan*[5]

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
{m20.rahman[1], i.himawan[2], c.fookes[4], s.sridharan[5]}@qut.edu.au, ddean@ieee.org[3]

Abstract

The performance of the current state-of-the-art i-vector based probabilistic linear discriminant analysis (PLDA) speaker verification depends on large volumes of training data, ideally in the target domain. However, in real-world applications, it is often difficult to collect sufficient amount of target domain data for successful PLDA training. Thus, an adequate amount of domain mismatch compensated out-domain data must be used as the basis of PLDA training. In this paper, we introduce a domain-invariant i-vector extraction (DI-IVEC) approach to extract domain mismatch compensated out-domain i-vectors using limited in-domain (target) data for adaptation. In this method, in-domain prior information is utilised to remove the domain mismatch during the i-vector extraction stage. The proposed method provides at least 17.3% improvement in EER over an out-domain-only trained baseline when speaker labels are absent and a 27.2% improvement in EER when speaker labels are known. A further improvement is obtained when DI-IVEC approach is used in combination with a domain-invariant covariance normalization (DICN) approach. This combined approach is found to work well with reduced in-domain adaptation data, where only 1000 unlabelled i-vectors are required to perform better than a baseline in-domain PLDA approach.

1. Introduction

In recent times, speaker verification has become popular research field, which aims to verify authenticity of a claimed identity using information extracted from its acoustic speech signal. Over the last few years, the state-of-the-art text independent speaker verification has been greatly influenced by i-vector based probabilistic linear discriminant analysis (PLDA) [1], which resulted in an excellent performance on recent speaker recognition evaluations (SREs). However, this success largely depends on the volume of the training speech data from the target domain. It was found that the performance of a PLDA system trained only on SWB and evaluated on NIST is worse compared to a PLDA system trained on NIST data, even though both Switchboard (SWB) and Mixer-NIST consist of telephone data. This problem was characterized as domain mismatch and introduced at the John Hopkins University (JHU) Summer workshop in 2013 [2].

Recently, researchers have proposed several techniques to improve the performance of speaker verification system, when PLDA models are initially trained on out-domain data. These can be broadly categorized into supervised and unsupervised techniques. Garcia-Romero *et al.* [3] found that there is no major detrimental effect of training the UBM and total variability matrix with out-domain data. Accordingly, they left these hyper-parameters in the out-domain and proposed four similar performing supervised adaptation techniques for PLDA parameters including fully Bayesian adaptation, approximate MAP adaptation, weighted likelihood and PLDA parameter interpolation. Villalba *et al.* [4] proposed another Bayesian PLDA parameters adaptation technique for limited supervised target domain data. They used a fully Bayesian approach and a variational approximation to compute the intractable posterior using conjugate priors, and gained relatively good performance than [3]. Wang *et al.* [5] introduced two transfer learning method to transfer the target domain features to the out-domain. They used maximum likelihood linear transformation (MLLT) technique to estimate the transfer parameters and expectation maximization (EM) to get the adapted PLDA parameters. These techniques performed better than interpolation of PLDA parameters [3]. Hong *et al.* [6] proposed another transfer learning method using Bayesian joint probabilities, where Kullback-Leibler (KL) divergence is used to maximize the optimization function to share the target knowledge between the two domains. Aronowitz [7] proposed an inter-dataset variability compensation (IDVC) technique based on nuisance attribute projection (NAP) to minimize the domain mismatch in the i-vector subspace. He partitioned the out-domain training dataset into 12 small subsets, and trained an IDV subspace spanned by the 12 centers of those subsets. This i-vector subspace domain-mismatch compensation has proven to be very efficient than other PLDA parameter adaptation techniques. Singer *et al.* [8] proposed a library whitening technique which can manually adjust the whitening scheme. This approach adapts the specific whitener data automatically to compensate the domain mismatch from out-domain data.

For unsupervised adaptation, Villalba *et al.* [9] proposed PLDA parameter adaptation technique, where unknown labels are modeled as latent variables and the variational Bayes approach is used to predict their posterior distributions. This technique improved the out-domain speaker recognition performance with only 200 in-domain speakers for adaptation data. Garcia-Romero *et al.* [10] used agglomerative hierarchical clustering (AHC) to label the unsupervised in-domain adaptation data. They also investigated the same AHC approach with a DNN/i-vector system and achieved the best performance using a DNN to collect the sufficient statistics for speaker modeling [11]. Glembek *et al.* [12] introduced a within-speaker covariance correction (WCC) approach, where they separated the between dataset covariance matrix from within-class covariance matrix for unsupervised adaptation of the LDA subspace. This approach has found to very effective than other unsupervised PLDA parameter adaptation techniques. Kanagasundaram *et al.* [13] proposed an unsupervised IDVC approach, in contrast to Aronowitz's approach [7], they shifted the out-domain i-vectors by capturing domain variability from the in-domain mean i-vector. In [14], Rahman *et al.* introduced unsupervised dataset invariant covariance normalization (DICN) , where they

10.21437/Odyssey.2018-22

trained the DICN matrix to capture the domain variability from the global mean i-vector and compensated this variability in the out-domain i-vectors for LDA and PLDA training. Rahman *et al.* also proposed domain-invariant mismatch modelling (DMM) [15] technique to model the domain mismatch part of the out-domain training i-vectors. In this technique the domain mismatch is first modelled from all out-domain i-vectors using maximum a posteriori (MAP) estimation. The new mismatch compensated clean out-domain i-vectors are then computed by the difference between the original and domain mismatch part of the i-vector. Also, combining the DICN approach with the DMM approach has been proven to be very efficient in domain mismatch compensation while having only a small amount of unlabelled in-domain data for adaptation.

In this paper, we present a domain-invariant i-vector extraction (DI-IVEC) technique using prior in-domain information to reduce the domain mismatch in the i-vector subspace. This approach is motivated by the introduction of a source-specific estimation of informative priors for i-vector extraction [16]. After extracting the compensated i-vectors, we also combine DI-IVEC with DICN approach to further compensate the mismatch in the i-vector subspace.

The rest of this paper is organized as follows: Section 2 details the DI-IVEC feature extraction technique. Section 3 describes details of the DICN approach. Section 4 describes the length-normalized GPLDA system. The experimental protocol and corresponding results are described in Section 5 and Section 6. Finally, Section 7 concludes the paper.

2. Domain-invariant I-vector Feature Extraction

In recent times, different unsupervised domain adaptation techniques have been proposed to compensate domain variability from the training i-vectors prior to PLDA modelling [7, 13]. However, none of these methods use in-domain prior information during the i-vector extraction stage to produce domain mismatch-compensated clean i-vectors. This section introduces a domain-invariant i-vector extraction (DI-IVEC) technique utilising prior in-domain information to reduce domain mismatch from the training data in the i-vector space. This approach is motivated by the introduction of a source-specific estimation of informative priors for i-vector extraction [16].

In i-vector representation [17], the speaker and channel-dependent GMM super-vector $\mathbf{m}$ can be represented via a single total variability subspace as follows,

$$\mathbf{m} = \mathbf{m}_0 + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{m}_0$ is a speaker- and session-independent UBM super-vector, $\mathbf{T}$ is a low-rank total-variability matrix and $\mathbf{w}$ is total variability factor which is assumed to be normally distributed $\mathcal{N}(0, \mathbf{I})$. For any given observation vector $\mathbf{X}$, the aim here is to determine the posterior distribution of $\mathbf{w}$ as follows,

$$p(\mathbf{w} \mid \mathbf{X}) = \mathcal{N}\left(\boldsymbol{\phi}, \mathbf{L}^{-1}\right), \tag{2}$$

where $\boldsymbol{\phi}$ is the desired i-vector and $\mathbf{L}$ is the precision matrix.

The i-vector extraction is based on the zero-order ($\mathbf{N}$) and centralised first-order ($\tilde{\mathbf{F}}$) Baum-Welch statistics defined as,

$$\mathbf{N}(c) = \sum_t \gamma_t(c), \tag{3}$$

$$\tilde{\mathbf{F}}(c) = \sum_t \gamma_t(c)(\mathbf{x}_t - \mathbf{m}_c), \tag{4}$$

where, c is the Gaussian component, $\mathbf{x}_t$ is the feature frame at time t, $\gamma_t(c)$ is the occupancy of the frame $\mathbf{x}_t$ to Gaussian component c and $\mathbf{m}_c$ is the mean vector of the c-th component of UBM super vector, $\mathbf{m}_0$.

With a standard normal prior $\mathcal{N}(0, \mathbf{I})$, the i-vector $\boldsymbol{\phi}$ can be extracted as follows,

$$\begin{aligned} \mathbf{L}_1 &= \mathbf{I} + \mathbf{N}\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\mathbf{T}, & (5)\\ \boldsymbol{W}_1 &= \mathbf{T}^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{F}}, & (6)\\ \boldsymbol{\phi} &= \mathbf{L}_1^{-1}\boldsymbol{W}_1, & (7) \end{aligned}$$

where $\mathbf{N}$ is a diagonal matrix whose diagonal blocks are $\mathbf{N}_c\mathbf{I}$ for $c = 1, ..., C$ and $\tilde{\mathbf{F}}$ is formed through the concatenation of the centralised first-order statistics $\tilde{\mathbf{F}}_c$ for any given observation vector $\mathbf{X}$. The covariance matrix $\boldsymbol{\Sigma}$ represents the residual variability not captured by $\mathbf{T}$.

In order to extract the domain-mismatch compensated i-vectors, a total variability matrix $\mathbf{T}$ already trained on out-domain data is used to extract the i-vectors. Now, instead of assuming the standard normally distributed priors with mean 0 and covariance $\mathbf{I}$, out-domain i-vectors are assumed to have new distribution with mean $\hat{\boldsymbol{\mu}}$ and covariance $\hat{\boldsymbol{\Sigma}}$ [18], i.e $\boldsymbol{\phi}_{out} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$.
where,

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\phi}_{in}^i, \tag{8}$$

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{\phi}_{in}^i - \hat{\boldsymbol{\mu}})(\boldsymbol{\phi}_{in}^i - \hat{\boldsymbol{\mu}})^T \\ + \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{I} + \mathbf{N}_{in}^i\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)^{-1}, \end{aligned} \tag{9}$$

where N is the number of total in-domain i-vectors. The closed form solution of this problem is found in [19]. The aim here is to find the posterior distribution that best matches this prior by minimising the Kullback-Leibler (KL) divergence of the desired prior distribution.

In the next step, out-domain first-order statistics $\tilde{\mathbf{F}}_{out}$ are re-centred again by removing $\mathbf{T}\boldsymbol{\mu}_d$ projection from the global mean $\boldsymbol{\mu}_c$ as follows,

$$\begin{aligned} \hat{\mathbf{F}}_{out} &= \sum_t \gamma_t(c)(\mathbf{x}_t - \boldsymbol{\mu}_c - \mathbf{T}\boldsymbol{\mu}_d), & (10)\\ &= \tilde{\mathbf{F}}_{out} - \mathbf{N}_{out}\mathbf{T}\boldsymbol{\mu}_d, \end{aligned}$$

where $\boldsymbol{\mu}_d$ is the mean of the domain mismatch part of the i-vectors and determined by,

$$\boldsymbol{\mu}_d = \frac{1}{M}\sum_{i=1}^{M}\left(\boldsymbol{\phi}_{old}^i - \hat{\boldsymbol{\mu}}\right), \tag{11}$$

where M is the total number of old out-domain i-vectors and $\boldsymbol{\phi}_{old}$ is the old out-domain i-vectors extracted using total-variability matrix $\mathbf{T}$.

Finally, the domain mismatch-compensated out-domain i-vectors are extracted as follows,

$$\begin{aligned} \mathbf{L}_2 &= \hat{\boldsymbol{\Sigma}}^{-1} + \mathbf{T}^T\boldsymbol{\Sigma}^{-1}\mathbf{N}_{out}\mathbf{T}, & (12)\\ \boldsymbol{W}_2 &= \mathbf{T}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{F}}_{out}, & (13)\\ &= \mathbf{T}^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{F}}_{out} - \mathbf{T}^T\boldsymbol{\Sigma}^{-1}\mathbf{N}_{out}\mathbf{T}\boldsymbol{\mu}_d, \\ \boldsymbol{\phi}_{out} &= \mathbf{L}_2^{-1}\boldsymbol{W}_2. & (14) \end{aligned}$$

Algorithm 1: EM Algorithm for DI-IVEC training

Input : $\boldsymbol{\phi}_{in} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_N\}$
$\boldsymbol{\phi}_{old} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_M\}$
Output: $\boldsymbol{\phi}_{out}$
Initialization: $\mathbf{N}_{in}$,$\mathbf{N}_{out}$,$\tilde{\mathbf{F}}_{out}$,$\mathbf{T}$,$\boldsymbol{\Sigma}$
$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}_{in}^{i}$
$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{\phi}_{in}^{i} - \hat{\boldsymbol{\mu}})(\boldsymbol{\phi}_{in}^{i} - \hat{\boldsymbol{\mu}})^T +$
$\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{I} + \mathbf{N}_{in}^{i}\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)^{-1}$
$\boldsymbol{\mu}_d = \frac{1}{M}\sum_{i=1}^{M}\left(\boldsymbol{\phi}_{old}^{i} - \hat{\boldsymbol{\mu}}\right)$
begin
E-Step:
$\boldsymbol{\phi}_{out} = (\hat{\boldsymbol{\Sigma}}^{-1} +$
$\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\mathbf{N}_{out}\mathbf{T})^{-1}\left[\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{F}}_{out} - \mathbf{T}^T\boldsymbol{\Sigma}^{-1}\mathbf{N}_{out}\mathbf{T}\boldsymbol{\mu}_d\right]$
$\boldsymbol{\phi}_d = \boldsymbol{\phi}_{out} - \hat{\boldsymbol{\mu}}$
$E\{\ln p(\phi_d \mid \boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)\} = \sum_{j=1}^{M} \ln p(\phi_d^j \mid \boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$
M-Step:
$\boldsymbol{\mu}_d = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{\phi}_d^i$
$\boldsymbol{\Sigma}_d = \frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{\phi}_d^i - \boldsymbol{\mu}_d)(\boldsymbol{\phi}_d^i - \boldsymbol{\mu}_d)^T$
until convergence

After one successful estimation of the domain mismatch-compensated out-domain i-vectors, in the next stages $\boldsymbol{\mu}_d$ and $\boldsymbol{\Sigma}_d$ are estimated as follows,

$$\boldsymbol{\mu}_d = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{\phi}_d^i, \tag{15}$$

$$\boldsymbol{\Sigma}_d = \frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{\phi}_d^i - \boldsymbol{\mu}_d)(\boldsymbol{\phi}_d^i - \boldsymbol{\mu}_d)^T. \tag{16}$$

where $\boldsymbol{\mu}_d$ and $\boldsymbol{\Sigma}_d$ represent the mean and co-variance of the domain mismatch part of the i-vectors.

Algorithm 1 describes the E and M steps of an EM algorithm used to estimate the domain mismatch-compensated out-domain training i-vectors.

3. DICN approach

In our previous work [14], we proposed DICN approach to compensate the domain mismatch in the i-vector subspace. In this approach, the global mean i-vector is used to capture the inter-domain variability using the outer product of the difference between all i-vectors (in-domain and out-domain) and global mean i-vector. After extracting the domain-invariant i-vectors as described in Section 2, we applied the DICN approach to compensate the domain variability further in the i-vector subspace. The domain mismatch using DICN approach can be captured as follows,

$$\boldsymbol{\Sigma}_{DICN} = \frac{1}{Q}\sum_{q=1}^{Q}(\boldsymbol{\phi}_q - \boldsymbol{\mu}_g)(\boldsymbol{\phi}_q - \boldsymbol{\mu}_g)^T \tag{17}$$

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

$$\boldsymbol{\mu}_g = \frac{1}{Q}\sum_{q=1}^{Q} \boldsymbol{\phi}_q \tag{18}$$

The scaling matrix $\mathbf{A}$ is calculated using the Cholesky decomposition of $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}_{DICN}^{-1}$. Later, the DICN compensated out-domain i-vectors are extracted as follows,

$$\boldsymbol{\phi}_{DICN} = \mathbf{A}^T\boldsymbol{\phi}_{out} \tag{19}$$

4. Length-normalized GPLDA

In this paper, we used length-normalized GPLDA approach introduced by Garcia-Romero *et al.* [20] for speaker and session modeling, which is computationally more efficient than heavy-tailed (HTPLDA) [1]. The benefit of the length-normalization approach is that it transforms the non-Gaussian behaviour of i- vectors into Gaussian, which involves three steps: centering, whitening and length normalization. In GPLDA modeling, the length-normalized i-vector can be decomposed into speaker and channel dependent part as follows,

$$\boldsymbol{\phi}_r = \bar{\boldsymbol{\phi}} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2r} + \boldsymbol{\epsilon}_r \tag{20}$$

where for given speaker recordings $r = 1, ...R$; $\mathbf{U}_1$ is the eigenvoice and $\mathbf{U}_2$ is the eigenchannel matrices; $\bar{\boldsymbol{\phi}} + \mathbf{U}_1\mathbf{x}_1$ is the speaker dependent part with covariance matrix of $\mathbf{U}_1\mathbf{U}_1^T$ and $\mathbf{U}_2\mathbf{x}_{2r} + \boldsymbol{\epsilon}_r$ is the channel dependent part with covariance matrix of $\mathbf{U}_2\mathbf{U}_2^T + \Lambda^{-1}$.

The GPLDA scoring is computed using the batch likelihood ratio between a target and test i-vector [1]. For given target i-vector $\boldsymbol{\phi}_{target}$ and test i-vector $\boldsymbol{\phi}_{test}$, the batch likelihood ratio can be calculated as follows,

$$\text{Score} = \ln \frac{P(\boldsymbol{\phi}_{target}, \boldsymbol{\phi}_{test} \mid \mathcal{H}_1)}{P(\boldsymbol{\phi}_{target} \mid \mathcal{H}_0)P(\boldsymbol{\phi}_{test} \mid \mathcal{H}_0)} \tag{21}$$

where $\mathcal{H}_1$: The speakers are same, $\mathcal{H}_0$: The speaker are different.

5. Experimental setup

The development dataset is derived from both in-domain (Mixer-NIST) and out-domain (SWB) datasets as described in DAC [2]. The in-domain dataset comprises 2,675 female and 1,115 male speakers, for a total of 36,470 sessions collected from NIST-2004, 2005, 2006 and 2008 SRE datasets, and the out-domain dataset contains 1,653 female and 1,462 male speakers, for a total of 33,039 sessions telephone data accumulated from Switchboard I, II phase I, II, III corpora as described in [2]. A subset of in-domain data containing different number of speakers (50, 100, 300, 500, 700 speakers), sessions/speakers (2, 4, 6, 8 sessions/speaker) and up to maximum 90 seconds of active speech length per utterance (15, 30, 60, 90 seconds) was collected for limited data investigations. For score normalisation, a subset of an in-domain dataset containing 1,526 female speakers, for a total of 1,972 sessions, and 1,115 male speakers, for a a total of 1,436 sessions is adopted as an in-domain score normalisation dataset. Similarly, the out-domain score normalization dataset is collected from a subset of the original out-domain dataset, containing 1,125 female speakers with 1,872 sessions and 1,125 male speakers with 1905 sessions.

For speaker modelling, 13-dimensional feature-warped MFCCs with Δ and $\Delta\Delta$ coefficients are extracted from raw speech signal using 25 ms frames with 10 ms frame shift. An energy-based VAD removes the silence frames from the feature stream while using an energy threshold of 5.5 across the zero coefficient of extracted MFCC features to perform VAD. Two gender dependent 512-mixture UBMs are trained and used for Baum-Welch (BW) statistics calculation for total-variability space training and i-vector extraction. Later, 500-dimensional i-vector extractor reduces the dimension of the GMM supervector

Table 1: Performance comparison of PLDA speaker verification on the common set of the NIST-2010 core-core evaluation condition, where PLDA and score normalisation are trained on both in-domain and out-domain data.

PLDA training	Score normalisation	UBM/i-vector EER	DNN/i-vector EER
Out-domain	In-domain	5.36%	4.05%
	Out-domain	5.62%	4.96%
	None	6.01%	5.03%
In-domain	In-domain	4.03%	3.15%
	Out-domain	4.32%	3.85%
	None	4.52%	3.98%

into a low dimensional subspace defined by the matrix $\mathbf{T}$. Prior to GPLDA modelling, LDA subspace reduces the dimension of the i-vectors, where LDA subspace is trained by selecting most significant 150 eigenvectors from 500 eigenvectors based on highest eigenvalues. In order to convert the heavy-tailed behaviour of i-vectors into Gaussian, i-vectors are whitened and length-normalized before GPLDA modeling. For GPLDA modeling, the best 120 eigenvoices (N_1) are selected by sorting the eigenvectors according to decreasing eigenvalues for speaker subspace training. Finally, the S-normalisation is applied on raw scores to reduce the score variability before evaluation. The

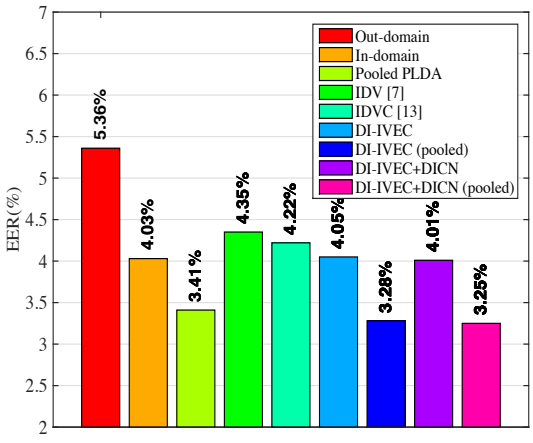
In this paper, all of the experiments are evaluated on core-core telephone-telephone conditions of NIST 2010 evaluation plan and performances are measured using EER.

For DNN/i-vector framework, the Kaldi toolkit [21] is used for ASR DNN training. A feed-forward network with back propagation estimation is used for DNN training with five hidden layer configurations using about 300 hours of out-domain (SWB) data. The hidden layers use a p-norm activation function (where p=2). The input layer takes 40-dimensional MFCC features with five-frame temporal context, and cepstral mean subtraction (CMS) performed over a window of six seconds. Each hidden layer has 350 nodes, the output dimension is 3500, and a softmax output layer computes posteriors for 5,346 senone targets. The force-alignment is applied between state-level transcripts and corresponding speech signals to generate HMM state-alignment labels for DNN training.

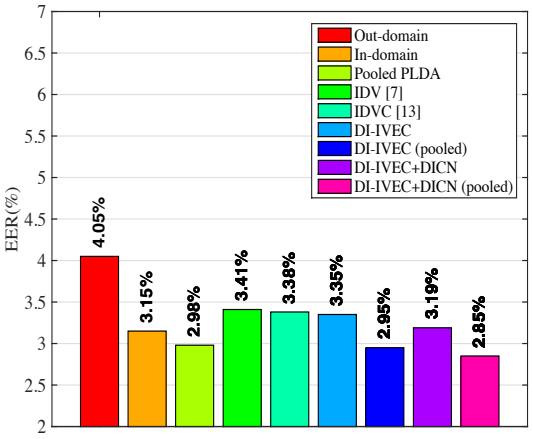
6. Experimental results

6.1. Baseline performance

Table 1 presents the LDA-projected out-domain and in-domain baseline PLDA speaker-verification performance, evaluated on core-core *telephone-telephone* condition. Although both in-domain and out-domain datasets consist of telephone data with almost similar statistics and distribution, experimental results show the performance difference between out-domain and in-domain systems due to inherent domain mismatch. Also, score normalisation plays a vital role in overall system performance. Without any score normalisation, UBM/i-vector system performance gets worse by 36.8% employing out-domain data rather than in-domain for PLDA training. For a DNN/i-vector system,



(a) UBM/i-vector based PLDA system.



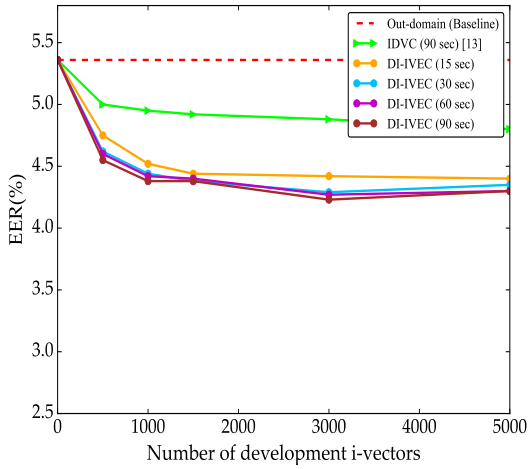
(b) DNN/i-vector based PLDA system.

Figure 1: Performance comparison of different domain adaptation techniques with DI-IVEC approach (a) UBM/i-vector-based PLDA system (b) DNN/i-vector-based PLDA system.

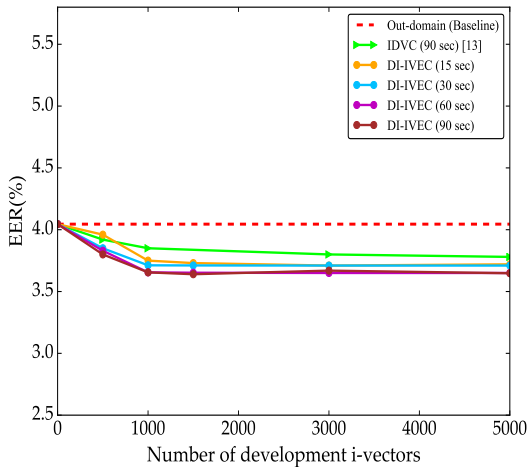
this performance gap is 26.4%. Also, the best result is achieved while training PLDA and score normalisation on in-domain data. These results clearly suggest that the score normalization data should always match with the target domain data to deliver best possible performance. Since the score normalisation statistics trained on in-domain data perform better than out-domain score normalisation, the rest of the experiments in this paper are presented using in-domain data for score normalisation.

6.2. Domain adaptation performance

Figure 1 presents the performance of the proposed DI-IVEC approach compared to other unsupervised domain adaptation techniques (like IDV [7], IDVC [13]) for improving out-domain PLDA system performance while using a full in-domain dataset for domain adaptation training. Experimental results show that compensating domain mismatch during the i-vector extraction stage results in an entirely considerable amount of performance improvement and yields 24.4% and 17.3% out-domain performance improvement for UBM/i-vector and DNN/i-vector systems, respectively. Also, combining a DICN approach with a

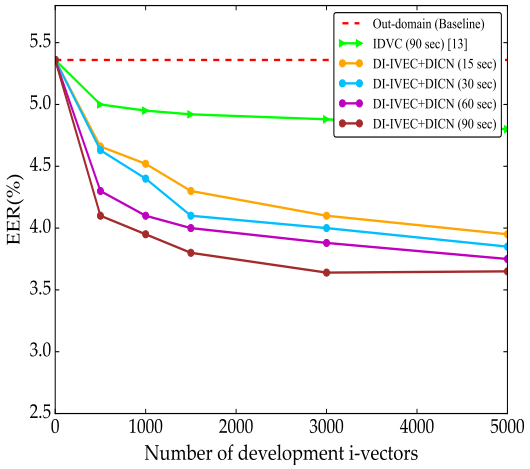


(a) UBM/i-vector based PLDA system.

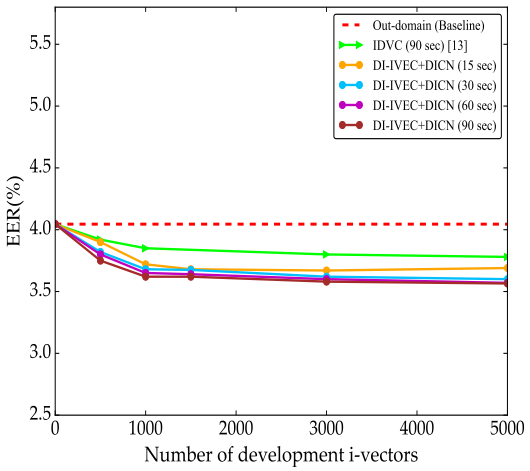


(b) DNN/i-vector based PLDA system.

Figure 2: Speaker-verification performance using DI-IVEC domain adaptation with limited in-domain i-vectors. (a) UBM/i-vector-based PLDA system (b) DNN/i-vector-based PLDA system.



(a) UBM/i-vector based PLDA system.



(b) DNN/i-vector based PLDA system.

Figure 3: Speaker verification performance using DI-IVEC+DICN domain adaptation with limited in-domain i-vectors. (a) UBM/i-vector based PLDA system (b) DNN/i-vector based PLDA system.

DI-IVEC approach (DI-IVEC+DICN) gives an additional improvement over the out-domain baseline, suggesting that DICN subspace transformation is very useful in compensating unwanted domain variability that cannot be compensated during the i-vector extraction stage. Furthermore, in the presence of labelled data, out-domain performance can be improved further by using labelled data pooled with out-domain data for PLDA training. This DI-IVEC (pooled) approach produces at least 27.2% EER improvement over the out-domain baseline, 6.3% EER improvement over in-domain baseline systems and 1% EER improvement over pooped PLDA systems, for both UBM/i-vector and DNN/i-vector systems. Subsequently, the combined DI-IVEC+DICN (pooled) approach shows the best performance compared to the other approaches presented in this section.

6.3. Limited data experiments

The previous section demonstrated the performance of the proposed DI-IVEC approach while using a full in-domain dataset for adaptation. However, it is also essential to learn the effectiveness of this technique under restricted-data conditions as well. These data-scarce conditions are created by reducing the amount of training data (500, 1000, 1500, 3000, 5000 i-vectors) as well as reducing the session length (15, 30, 60, 90 seconds) of each i-vector for domain adaptation training. These limited-data domain adaptation performances are compared with IDVC[13] approach, trained with 90-second sessions.

6.3.1. DI-IVEC approach

Figure 2 shows the performance of the DI-IVEC approach using limited unlabelled in-domain data for adaptation. Regardless of the length of the sessions, this approach shows higher performance accuracy over IDVC [13] and obtains most of the performance improvement with only 1000 in-domain i-vectors. Although increasing the adaptation data further does not show substantial performance variation, using a longer session length shows a direct influence on the overall system performance. For both UBM/i-vector and DNN/i-vector systems, 150-second

utterances produce the best speaker-verification performance. However, with only 15 seconds long 1000 in domain i-vectors, DI-IVEC approach yields 16% and 7.6% out-domain performance improvement for UBM/i-vector and DNN/i-vector system, respectively.

6.3.2. DI-IVEC+DICN approach

In Section 6.2, it was shown that the best domain adaptation performance could be achieved by combining a DICN approach with a DI-IVEC approach. This section extends these investigations to analyse the performance of this combined approach under limited-data conditions. As expected, Figure 3 illustrates the superior performance of this combined DI-IVEC+DICN approach over the IDVC and DI-IVEC approach alone (as presented in Figure 2). Like other limited-data experiments, this approach also shows notable out-domain performance improvement with only 1000 in-domain i-vectors. However, while using 15-second sessions for domain adaptation, this approach finds at least 7.8% improvement over the out-domain baseline for both UBM/i-vector and DNN/i-vector systems.

7. Conclusion

This paper introduced the DI-IVEC approach to extract the domain-mismatch compensated out-domain i-vectors to improve the GPLDA speaker verification system. Experimental results suggested that DI-IVEC approach can improve the speaker verification performance significantly by using only small amount of adaptation data. Also, an additional domain mismatch can be compensated in the i-vector subspace when DI-IVEC is used in combination with DICN, thus improving the overall speaker verification performance. A significant improvement in EER of 27.2% is achieved over the out-domain baseline when speaker labels are available. In our experiments with limited data, we required only 15 sec long 1000 unlabeled i-vectors for the DI-IVEC approach to performing similarly well as the in-domain baseline.

8. Acknowledgements

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

9. References

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

[2] Domain Adaptation Challenge, "Speaker and language recognition summer workshop, Tech. Rep., 2013," in *2013 speaker recognition workshop. Available online: http://www. clsp. jhu. edu/workshops/archive/ws13-summerworkshop/groups/spk-13*, 2013.

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

[4] Jesús Villalba and Eduardo Lleida, "Bayesian adaptation of PLDA based speaker recognition to domains with scarce development data," in *The Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2012.

[5] Qiongqiong Wang, Hitoshi Yamamoto, and Takafumi Koshinaka, "Domain adaptation using maximum likelihood linear transformation for PLDA-based speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 5110–5114.

[6] Qingyang Hong, Jun Zhang, Lin Li, Lihong Wan, and Feng Tong, "A transfer learning method for PLDA-based speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 5455–5459.

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

[8] Elliot Singer and Douglas A Reynolds, "Domain mismatch compensation for speaker recognition using a library of whiteners," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 2000–2003, 2015.

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

[10] Daniel Garcia-Romero, Alan McCree, Stephen Shum, Niko Brümmer, and Carlos Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *The Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2014, pp. 260–264.

[11] Daniel Garcia-Romero, Xiaohui Zhang, Alan McCree, and Daniel Povey, "Improving speaker recognition performance in the domain adaptation challenge using deep neural networks," in *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 378–383.

[12] Ondrej Glembek, Jeff Ma, Pavel Matejka, Bing Zhang, Oldrich Plchot, Lukas Burget, and Spyros Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 4032–4036.

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

[14] Md Hafizur Rahman, Ahilan Kanagasundaram, David Dean, and Sridha Sridharan, "Dataset-invariant covariance normalization for out-domain PLDA speaker verification," in *Proceedings of Interspeech*, September 2015, pp. 1017–1021.

[15] Md Hafizur Rahman, Ivan Himawan, David Dean, and Sridha Sridharan, "Domain mismatch modeling of out-domain i-vectors for PLDA speaker verification," in *Proceedings of Interspeech*, 2017, pp. 1581–1585.

[16] Sven Ewan Shepstone, Kong Aik Lee, Haizhou Li, Zheng-Hua Tan, and Søren Holdt Jensen, "Total variability modeling using source-specific priors," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 504–517, 2016.

[17] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[18] Niko Brummer, "EM for probabilistic LDA," *Agnitio Research, Cape Town, Tech. Rep*, 2010.

[19] Liping Chen, Kong Aik Lee, Bin Ma, Wu Guo, Haizhou Li, and Li Rong Dai, "Minimum divergence estimation of speaker prior in multi-session PLDA scoring," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 4007–4011.

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

[21] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.