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

Speaker Verification in Mismatched Conditions with Frustratingly Easy Domain Adaptation

Jahangir Alam, Gautam Bhattacharya, Patrick Kenny

Computer Research Institute of Montreal (CRIM)
Montreal (Quebec) Canada

jahangir.alam@crim.ca, gautam.bhattacharya@crim.ca

Abstract

The 2016 edition of the NIST speaker recognition evaluation tests the ability of speaker verification systems to deal with domain mismatch between development and test data. In order to adapt to new languages, a small amount of unlabeled, in-domain data was provided - warranting the need for an unsupervised approach to learn from this data. In this work we adapt a simple domain adaptation strategy to the speaker verification problem. We test our approach using two types of speaker embeddings - i-vectors and neural network based x-vectors. Despite the simplicity of our method, we show that it outperforms a competitive PLDA domain-adaptation approach in the i-vector domain (12.11% vs 12.68% EER), and works as well in the x-vector domain (8.93% vs 8.91% EER) when the dimension of x-vectors is reduced to 150 by using LDA. Finally, as our approach adapts the speaker embeddings, we combined our adapted embeddings with the PLDA adaptation approach. We achieved our best result (8.23% EER) using this strategy with x-vector embeddings.

1. Introduction

Speaker Verification (SV) involves the authentication of a person's identity given a small amount of audio from that speaker [1]. In the context of machine learning, SV represents a binary classification problem - Given the audio from a known speaker and an unknown test recording, the job of a SV system is to determine if the two recordings belong to the same speaker or not. Text-Independent SV represents the unconstrained form of the problem, wherein the phonetic content of the audio is irrelevant to the verification process. This is in contrast to text-dependent SV, which involves two levels of verification, both the phonetic and speaker level [2]. The National Institute of Standards and Technology (NIST) have held periodic speaker recognition evaluations (SRE), which have served as benchmark tasks in the speaker verification community. In recent years, the i-vector/PLDA speaker verification paradigm has established itself as the state-of-the-art in the field [3]. However, unlike previous SRE's, the evaluation in 2016 introduced new types of problems that were meant to better reflect the challenges of performing speaker verification in the real world.

Real world speaker verification systems are typically evaluated using speech that is described as out-of-domain, i.e. a mismatch exists between the speech used at test time and the speech the system was trained on. This issue is prevalent in several areas of machine learning including computer vision and natural language processing [4, 5, 6]. In the speaker verification research community the domain mismatch problem has received significant interest following the introduction of the domain adaptation challenge in 2013. Several successful domain adaptation strategies have been proposed based on the i-vector/PLDA speaker verification paradigm [7, 8, 9, 10].

The domain adaptation challenge explored the effects of channel mismatch between development and test datasets. For the challenge, the Switchboard dataset is used as background data, while the evaluation is performed on the SRE-2010 test set. The development audio consists of land-line calls while the test recordings are mainly cellular calls, a channel mismatch exists. We note that both datasets consist of English speakers. While the statistics of the SRE and Switchboard datasets are quite similar, speaker verification performance degrades significantly compared to a system trained on in-domain (SRE 2004-2008) data.
In 2016, NIST held a speaker recognition evaluation (SRE16). Unlike the domain adaptation challenge, the evaluation focused on language mismatched conditions. The development data consisted of a large amount of primarily English speakers in the form of the SRE 2004-2008 and Switchboard datasets. NIST also provided very small set of in-domain (non-English) unlabeled dataset while the evaluation dataset is spoken in Tagalog, Cantonese, Cebuano and Mandarin. Prior studies have considered multilingual dataset augmentation for language mismatched condition. However, application is possible only if sufficient in-domain dataset exists.

A common thread between several of the domain adaptation strategies explored for speaker verification is that they aim to adapt the system hyper-parameters, i.e. the Universal Background Model, Total Variability Matrix etc. In this work we make used of an adaptation approach that adapts the between class and within class matrices of the PLDA classifier [7], which also falls into this category.
Contrary to the techniques mentioned above, our proposed domain adaptation method works directly on the speaker embeddings. Consequently our approach is agnostic to both the classifier and the speaker embeddings. The approach used in this work was first presented for unsupervised domain adaptation in computer vision applications [11]. The authors showed that the technique worked on so-called shallow features, and further improvements could be obtained using features extracted from a deep neural network. From our experiments we show that the same trend is seen for speaker verification. In the i-vector domain (shallow features), we improve by 1.3% absolute compared to no domain adaptation, and for the x-vectors (deep features) the improvement is 1.8% absolute.

The remainder of the paper is organized as follows. We begin with a description of the domain adaptation strategy used in this work. This is followed by details of the different speaker embeddings we used in combination with domain adaptation. We

10.21437/Odyssey.2018-25

then proceed to our experiments and results. We finish with some concluding remarks and directions for future work.

2. Domain Adaptation for Speaker Verification

Domain adaptation problems are typically setup such that there is a large amount of out-of-domain data and a small amount of labeled or unlabeled in-domain data available to train machine learning systems. The NIST SRE-2016 presents exactly this type of problem. A vast majority of the training data consists of English speakers, while the test data consists of Asian language speakers. Apart from the difference in language, there is also significant mismatch in terms of channel conditions, given that the recordings are made in different parts of the world.

NIST also provided a small amount of unlabeled in-domain data, in order to adapt speaker verification systems trained on English speakers to the new domain. As mentioned in the previous section, a common aspect of many speaker verification domain adaptation approaches has been to adapt systems hyper-parameters. In this work we use the unsupervised PLDA adaptation approach [7], which adapts the within class and between class covariance matrices in PLDA to the in-domain data. To our knowledge, this approach in combination with the recently introduced x-vector speaker embeddings, represents the best published results on SRE-2016.

2.1. Domain adaptation using minimum divergence training

In our previous work [12], we proposed a novel unsupervised domain adaptation approach to compensate for the effects of mismatched domains introduced in the NIST speaker recognition evaluation (NIST SRE) 2016. It applies minimum divergence (MD) training to adapt a conventional i-vector extractor to the task domain. The main idea behind MD training is to transform standard normal priors to non-standard normal priors. Specifically, we take an out-of-domain trained i-vector extractor as an initialization and perform few iterations of minimum divergence training on the in-domain oriental data. In-domain oriental data comprised of unlabeled training data (2472) from SRE16 plus 2941 recordings from Mandarin, Chinese, and Tagalog from NIST 2004-2008 SREs. The out-of-domain set includes 53228 recordings from the Switchboard corpus and NIST 2004-2008 SREs excluding the Mandarin, Chinese, and Tagalog. Final i-vectors in [12] were obtained by projecting i-vectors with Nuisance Attribute Projection (NAP) and then applying length normalization.

2.2. Correlation Alignment for Unsupervised Domain Adaptation

In this section we present an overview of a recently introduced domain adaptation algorithm called Correlation Alignment (CORAL) [11], which works by aligning the distributions of out-of-domain and in-domain features in an unsupervised way. This is achieved by aligning second-order statistics, i.e. covariance.

To minimize the distance between the second-order statistics (covariance) of the out-of-domain and in-domain features, a linear transformation A to the original source features and the Frobenius norm is used as matrix distance metric:

$$\begin{aligned} min_A \|C_{\hat{S}} - C_T\|_F^2 \\ => min_A \|A^\top C_S A - C_T\|_F^2 \end{aligned} \tag{1}$$

It can be shown that the linear transformation A can be decomposed into two parts, the first can be interpreted as whitening the out-of-domain data, while the second part re-colors the whitened out-of-domain data with the in-domain covariance. We refer the reader to [11] for a detailed mathematical analysis of the algorithm. In practice the algorithm is implemented using classical whitening and coloring approaches. This makes it stable and highly efficient. The algorithm is presented below. We note that it can be implemented in 4 lines of MATLAB code.

```
Algorithm: CORAL for Unsupervised Domain Adaptation
Input: out-of-domain data $D_S$, in-domain data $D_T$
Output: Adjusted out-of-domain data $D_s^*$
$C_S$ = cov($D_S$)+eye(size($D_S$,2))
$C_T$ = cov($D_T$)+eye(size($D_T$,2))
$D_S = D_S * C_S^{-1/2}$                    %whitening
$D_S^* = D_S \; C_T^{1/2}$                     %re-coloring
```

The algorithm begins by computing the covariance matrices of the out-of-domain and in-domain datasets. Then the out-of-domain data is whitened using the out-of-domain data covariance and then re-colored using the in-domain data covariance.

The authors of CORAL showed experimentally that the approach worked on both shallow and deep features, with the best results being achieved using features extracted from deep convolutional networks. In this work we also test CORAL on shallow and deep speaker representations. For the shallow features we use i-vectors, while for deep speaker embeddings we use the recently proposed x-vectors [13]. As CORAL works by aligning data at the feature level, we also consider using it in combination with a PLDA adaptation algorithm. In the next section we elaborate on the speaker embeddings in more detail.

3. Speaker Embeddings

An important step in the speaker verification pipeline involves deriving a compact, low-dimensional representation from the audio of a speaker. We henceforth refer to such a representation as a speaker embedding. In this work we make use of two types of speaker embeddings, which are described in the following sections.

3.1. i-vectors

The conventional i-vector speaker embeddings, proposed in [3] represents the dominant approach in the speaker recognition field and is shown to provide excellent performance on text-independent task. In i-vector embeddings, the speaker and channel variability in speakers recording is compelled to lie in a low-dimensional space. Fundamentally, i-vector is the compact and fixed-length vector representation of a recording of arbitrary duration. This in turn allows for scoring of verification trials in a straightforward way; a simple cosine-scoring strategy provides impressive results [12, 14]. Further improvement can be achieved by using a more powerful classifier such as Probabilistic Linear Discriminant Analysis (PLDA) [14].

3.2. Neural Speaker Embeddings

With the advent of deep learning, there have been several efforts to learn deep speaker representations or embeddings using neural networks [15, 16, 17, 18, 19]. These networks are speaker-discriminative, and are trained by minimizing either the cross-entropy loss, or some form of contrastive loss. A particularity about the NIST-SRE data is that it consists of very long recordings. It has been observed that the embeddings obtained from a neural network are not able to outperform i-vectors on this timescale. On the other hand, neural speaker embeddings show a distinct advantage over i-vectors when we consider short recordings [18, 20].

Recently, the x-vector system was improved to give state-of-the-art performance on the SRE-2016 task [20, 21]. The model uses a time-delay structure with statistics pooling layers to learn speaker embeddings. The model is trained by minimizing the cross-entropy loss over speakers in the training set. It has been seen that verification performance can be improved through multi-condition training, i.e. by augmenting the training data with different kinds of noise such as simulated reverberation, babble noise, music and other additive noises to the clean training data [21].

4. Performance Evaluation

In order to evaluate the performances of CORAL-based unsupervised domain adaptation we carried out speaker verification experiments using two types of speaker representations, namely, i-vectors and deep neural networks-based x-vectors. Results are reported on the evaluation test set of NIST SRE 2016. Metric used for performance evaluation is equal error rates (EER). In this section we provide the details of our experimental setup as well as our speaker verification results.

4.1. Experimental setup

For speaker verification with i-vector speaker embeddings a 2048 Gaussians full covariance UBM is trained on the unlabeled major data. The i-vector extractor is trained on 90108 recordings taken from the Switchboard corpus and NIST 2004-2010 SREs. Extracted i-vectors are of 600 dimensional and LDA is used to reduce the dimension to 200. PLDA classifier is trained on the multi-condition training data generated by adding reverberation, music, babble noise and MUSAN noises to the NIST 2014-2010 SREs data. Unsupervised PLDA adaptation is done on the SRE 2016 unlabeled major data. For speaker verification with x-vector speaker embeddings a DNN is trained to discriminate between speakers on the multi-condition data generated by adding reverberation, music, babble noise and MUSAN noises to the Switchboard and NIST 2014-2010 SREs data. After training, 512-dimensional x-vectors are extracted from the affine component of 6th hidden layer. LDA is used to reduce the x-vector dimension of 150. PLDA classifier is trained on the multi-condition training (excluding the Switchboard portion) and adaptation of trained PLDA models are done in unsupervised fashion on the SRE 2016 unlabeled major data. In order to make comparison fair with i-vectors we also report results after reducing x-vectors dimension to 200 using LDA. In this work, for speaker verification with i-vector and x-vector speaker embeddings we used the recipe provided in [21].

4.2. Experimental results and discussion

We begin with the performance of our baseline i-vector/PLDA system and compare it to the PLDA domain adaptation algorithm as well as the approach proposed in this work. Table 1 shows the results on the NIST SRE 2016 (SRE16) evaluation test set. From the results we see that the both types of domain adaptation significantly improve the performance on the SRE16 test set. We also see that our proposed domain adaptation approach produces the best overall result, improving the pooled equal error rate to 12.11% from 12.68%. We obtained a relative improvement of 9.7% over the shallow i-vector/PLDA framework (without any domain adaptation) and over the PLDA adaptation approach a relative improvement of 4.4% is achieved by our adaptation approach. But we did not get any benefit by applying unsupervised PLDA adaptation over the adapted i-vectors. This is may be due to not tuning the parameters of PLDA adaptation method. We used the same parameters as used in the recipe provided in [21].

Table 1: Performances of speaker verification systems using i-vector speaker embeddings in terms of equal error rates (EER). The dimension of i-vectors after LDA is 200. The best results are highlighted in bold face. (A) indicates adaptation. The 3 error rates represent (cantonese/tagalog/pooled).

System	EER (%)
i-vector / PLDA	9.447 / 17.53 / 13.42
i-vector / PLDA (A)	8.172 / 17.25 / 12.68
i-vector (A) / PLDA	**7.55 / 16.71 / 12.11**
i-vector (A) / PLDA (A)	7.851 / 17.54 / 12.68

Next we turn our attention to x-vectors. These speaker embeddings have been shown to improve verification performance on the SRE16 task as compared to i-vectors. Furthermore, the performance of these embeddings in combination with adapted PLDA has led to state-of-the-art results. Both data augmentation and unsupervised PLDA adaptation play a crucial role to obtain better performance with x-vectors / PLDA paradigm. Tables 2 presents EERs achieved without and with different unsupervised adaptation approaches including the proposed approach when speaker verification is conducted using x-vectors. In this case the x-vectors are of 150-dimensional after applying LDA. From the results below it is clear that x-vectors perform better than i-vectors on this task. Furthermore, PLDA domain adaptation leads to larger improvement in terms of EER as compared to i-vectors. From lines 2 and 3 of table 2 we see that our proposed approach works almost as well as the PLDA adaptation strategy. In order to make a fair comparison with i-vectors, in table 3, we also report results with x-vectors when the dimension of x-vectors is reduced to 200 using LDA. It is clear from table 3 that the proposed approach outperforms the PLDA adaptation strategy for 200-dimensional x-vectors. We were able to achieve our best result by combining both methods, i.e. first adapting the x-vectors using our approach followed by adapting PLDA. Using this strategy and with 200-dimensional x-vectors (after applying LDA) we obtain the best performance (8.23% EER on pooled condition).

Table 4 demonstrates a comparison of performances of minimum divergence (MD) training-based unsupervised domain adaptation (3rd row of table 4) with other unsupervised domain adaptation techniques, such as domain adaptation by data augmentation (fast and last row of table 4), PLDA domain adaptation (2nd row of table 4), on NIST SRE 2016 task. The i-vector/PLDA system employ domain adaptation by data augmentation as the PLDA is trained with augmented or multi-condition data. Both the i-vector/adapted PLDA and x-vector/ PLDA paradigms use multi-condition training data either for training PLDA or for training x-vectors extractor as well as PLDA. Though multi-condition training is found helpful in the

Table 2: Performances of speaker verification systems using deep neural networks-based x-vector speaker embeddings in terms of equal error rates (EER). The dimension of x-vectors after LDA is 150. The best results are highlighted in bold face.(A) indicates adaptation. The 3 error rates represent (cantonese/tagalog/pooled).

System	EER (%)
x-vector / PLDA	7.369 / 16.29 / 11.74
x-vector / PLDA (A)	5.021 / **12.88** / 8.912
x-vector (A) / PLDA	5.052 / 12. 99 / 8.936
x-vector (A) / PLDA (A)	**4.695** / 12.94 / **8.753**

Table 3: Performances of speaker verification systems using deep neural networks-based x-vector speaker embeddings in terms of equal error rates (EER) when the dimension of x-vectors is reduced to 200 by applying LDA. The best results are highlighted in bold face. (A) indicates adaptation. The 3 error rates represent (cantonese/tagalog/pooled).

System	EER (%)
x-vector / PLDA	7.685 / 16.37 / 11.98
x-vector / PLDA (A)	5.006 / 12.36 / 8.688
x-vector (A) / PLDA	4.949 / 12.34 / 8.596
x-vector (A) / PLDA (A)	**4.46 / 12.12 / 8.229**

PLDA backend it is not beneficial in the i-vector extractor. On the otherhand, since deep neural networks are trained in supervised fashion use of data augmentation in deep learning is ubiquitous and it aids to generalize the network by creating diversity in the training data. Despite not taking any benefit by data augmentation MD training-based unsupervised domain adaptation provided comparable results to the x-vector/ PLDA paradigm and resulted in a relative improvements of 7.3% and 12.4% over the i-vector/ adapted PLDA and i-vector / PLDA frameworks, respectively.

Table 4: Comparison of Performances of minimum divergence training-based domain adaptation with other unsupervised domain adaptation techniques on NIST SRE 2016 task in terms of equal error rates (EER). Both i-vector and x-vector speaker representations are used. The lowest EERs are highlighted in bold face.

System	EER (%)
i-vector / PLDA	13.42
i-vector / adapted PLDA	12.68
adapted i-vector using MD training / PLDA [12]	**11.75**
x-vector / PLDA	**11.74**

5. Conclusion

In this work, we introduced a simple unsupervised domain adaptation approach to improve speaker verification performance under domain mismatch conditions. In this approach domain adaptation was performed by aligning the covariances of labeled out-of-domain and unlabeled in-domain data. Speaker verification experiments were conducted using i-vectors and x-vectors speaker embeddings on the SRE 2016 corpora and results were reported on the evaluation test set. Despite the simplicity this method was found effective to compensate for the mismatched domains. Proposed method provided reduced EER compare to a competitive PLDA domain-adaptation approach in the i-vector domain and worked as well in the x-vector domain. Finally, we achieved the lowest EER when our adapted deep speaker embeddings (x-vectors) were combined with the unsupervised PLDA adaptation approach. In our future work we would like to employ deep learning approaches for unsupervised domain adaptation problem for speaker verification.

6. References

[1] Douglas A Reynolds, "An overview of automatic speaker recognition technology," in *Acoustics, speech, and signal processing (ICASSP), 2002 IEEE international conference on*. IEEE, 2002, vol. 4, pp. IV–4072.

[2] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[3] Najim Dehak, Patrick J 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.

[4] Hal Daumé III, "Frustratingly easy domain adaptation," *arXiv preprint arXiv:0907.1815*, 2009.

[5] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015, pp. 1180–1189.

[6] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[7] Daniel Garcia-Romero, Alan McCree, Stephen Shum, Niko Brummer, and Carlos Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[8] Md Hafizur Rahman, Ahilan Kanagasundaram, David Dean, and Sridha Sridharan, "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, 2015, pp. 1017–1021.

[9] Stephen H Shum, Douglas A Reynolds, Daniel Garcia-Romero, and Alan McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," 2014.

[10] 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.

[11] Baochen Sun, Jiashi Feng, and Kate Saenko, "Return of frustratingly easy domain adaptation.," in *AAAI*, 2016, vol. 6, p. 8.

[12] Jahangir Alam, Patrick Kenny, Gautam Bhattacharya, and Marcel Kockmann, "Speaker verification under adverse conditions using i-vector adaptation and neural networks," *Proc. Interspeech 2017*, pp. 3732–3736, 2017.

[13] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 165–170.

[14] Themos Stafylakis, Patrick Kenny, Vishwa Gupta, Jahangir Alam, and Marcel Kockmann, "Compensation for phonetic nuisance variability in speaker recognition using dnns," in *Odyssey: The Speaker and Language Recognition Workshop*, 2016, pp. 340–345.

[15] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4052–4056.

[16] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[17] Chunlei Zhang and Kazuhito Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *Proc. of Interspeech*, 2017.

[18] Gautam Bhattacharya, Jahangir Alam, and Patrick Kenny, "Deep speaker embeddings for short-duration speaker verification," in *Proc. Interspeech 2017*, pp. 1517–1521. 2017.

[19] Lantian Li, Zhiyuan Tang, and Dong Wang, "Full-info training for deep speaker feature learning," *arXiv preprint arXiv:1711.00366*, 2017.

[20] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Proc. Interspeech 2017*, pp. 999–1003, 2017.

[21] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "X-vectors:robust dnn embeddings for speaker recognition," *Proc. ICASSP 2018*, 2018.