UNSUPERVISED DOMAIN ADAPTATION VIA DOMAIN ADVERSARIAL TRAINING FOR SPEAKER RECOGNITION

Qing Wang[1,2]*, Wei Rao*[2]*, Sining Sun*[1]*, Lei Xie*[1,*]*, Eng Siong Chng*[2,3]*, Haizhou Li*[2,4]

School of Computer Science, Northwestern Polytechnical University, Xi'an, China[1]
Temasek Laboratories, Nanyang Technological University, Singapore[2]
School of Computer Science and Engineering, Nanyang Technological University, Singapore[3]
Department of Electrical and Computer Engineering, National University of Singapore, Singapore[4]

ABSTRACT

The i-vector approach to speaker recognition has achieved good performance when the domain of the evaluation dataset is similar to that of the training dataset. However, in real-world applications, there is always a mismatch between the training and evaluation datasets, that leads to performance degradation. To address this problem, this paper proposes to learn the domain-invariant and speaker-discriminative speech representations via domain adversarial training. Specifically, with domain adversarial training method, we use a gradient reversal layer to remove the domain variation and project the different domain data into the same subspace. Moreover, we compare the proposed method with the other state-of-the-art unsupervised domain adaptation techniques for i-vector approach to speaker recognition (e.g. autoencoder based domain adaptation, inter dataset variability compensation, dataset-invariant covariance normalization, and so on). Experiments on 2013 domain adaptation challenge (DAC) dataset demonstrate that the proposed method is not only effective in solving the dataset mismatch problem, but also outperforms the compared unsupervised domain adaptation methods.

Index Terms— Domain Adversarial Training, Unsupervised Domain Adaptation, Speaker Recognition

1. INTRODUCTION

Conventional approaches of speaker recognition, such as i-vector [1] usually assume that training and evaluation data share the same probability distributions or the same feature space. Unfortunately, this assumption doesn't hold in many real-world applications because there is often domain mismatch between training and evaluation data. To alleviate the effect of domain mismatch, domain adaptation [2] is seen as a solution to mitigate the problem. The training and evaluation dataset are related to source domain and target domain, respectively, for speaker recognition domain adaptation.

According to the availability of labels for target domain, domain adaptation techniques for speaker recognition could be classified into two categories: supervised domain adaptation and unsupervised domain adaptation.

In supervised domain adaptation, we are given limited labeled data from the target domain. In [3], Garcia-Romero et al. treated within-speaker and between-speaker covariance as random variables and used maximum a posterior (MAP) to estimate these parameters conditioned on the target domain data. Unsupervised domain adaptation refers to the situation where some unlabelled data from the target domain are provided. It means that we will face some difficulties in performing channel compensation techniques (e.g., linear discriminative analysis (LDA), probabilistic LDA (PLDA) [4], [5]). To address this issue, three strategies are adopted: (1) the first method proposes to use clustering techniques to estimate the speaker labels for unlabelled target domain data firstly, such as [6], [7], and [8], (2) the second method selects the unlabelled target and source domain data to estimate the compensation model and compensate the domain mismatch, such as inter-dataset variability (IDV) [9], inter dataset variability compensation (IDVC) [10], and dataset-invariant covariance normalization (DICN) [11], and (3) the third method learns the domain-invariant space or maps the source domain data into target domain space and use the mapped source domain data with its speaker label to train LDA or PLDA. For example, Shon et al. [12] proposed the autoencoder based domain adaptation (AEDA), which combines an autoencoder with a denoising autoencoder to adapt resource-rich source domain data to target domain. Then, the transformed source domain data could be used for PLDA training.

This paper follows the third strategy on the unsupervised domain adaptation task and proposes to apply domain adversarial training (DAT) [13], [14] to address the domain mismatch problem. We apply DAT technique to alleviate the i-vectors mismatch across different domains. Under a multi-task learning framework [15], the approach jointly learns one feature extractor and two discriminative classifiers using one single DNN: the feature extractor is trained to extract domain-

*Lei Xie is the corresponding author.

invariant and speaker-discriminative features. As the main task, a speaker label predictor predicts speaker labels during training. As the second task, a domain classifier discriminates between the source and the target domains during training. With a gradient reversal layer that removes the domain variation, both domain i-vectors are projected into the same subspace. This approach only needs the labeled training data from source domain, and unlabeled data from target domain, so we call it unsupervised domain adaptation. Since the speaker labels of the source domain data are used to train the PLDA back-end, the training is carried out as in a mix of supervised (speaker) and unsupervised (domain) manner.

Besides exploring the effectiveness of DAT on speaker recognition, this paper also compares the performance of DAT with other state-of-the-art unsupervised domain adaptation methods. Experimental results on DAC 13 demonstrate that the proposed DAT method achieves the best performance. Moreover, compared with other unsupervised adaptation approaches, we can easily implement the proposed approach by simply augmenting a common feed-forward network with a few standard layers and a new gradient reversal layer.

1.1. Related Work

DAT has been applied in robust speech recognition in both supervised and unsupervised case. In noise robust speech recognition, they consider the clean speech as the source domain data while the noise speech as the target domain data. In [16], within the speech frame and its corresponding senone label of the labeled training data, DAT is used to learn the senone label classifier and domain classifier at the same time, using labeled source domain data and unlabeled target domain data. DAT was proposed to obtain adversarial senone-discriminative and domain-invariant representation. In [17], they applied DAT in a supervised way.

Both of them use DAT as the acoustic model in speech recognition, they extract the posterior probability from the senone label classifier for robust speech recognition decoding. But for our work, we use the DAT as the feature extractor to extract domain-invariant and speaker-discrimintive speech representations from the first hidden layer of feature extractor network part.

The rest of the paper is organized as follows. Section 2 details the DAT approach to speaker recognition. Section 3 introduces experimental setup. Section 4 presents and analyzes the experimental results. We conclude in Section5.

2. DOMAIN ADVERSARIAL TRAINING FOR SPEAKER RECOGNITION

2.1. Domain-Adversarial Training

We propose to project two different domains into a common subspace to eliminate the domain mismatch. This can be achieved by training a DAT [13] that learns a speaker-discriminative and domain-invariant feature representation, that are described as follows.

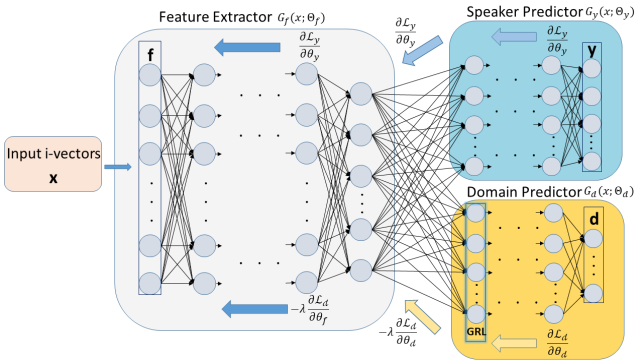

Fig. 1. Domain-Adversarial Training (DAT) framework include: feature extractor, speaker label predictor, domain predictor. A gradient reversal layer (GRL) is between feature extractor and domain predictor.

A conventional neural network associates input samples $\mathbf{x} \in X$ with data labels $\mathbf{y} \in Y$, where X and Y are input space and output space, respectively. Here in speaker recognition, $\mathbf{x}$ and $\mathbf{y}$ are i-vectors and speaker labels. However, the distribution $D(\mathbf{x}, \mathbf{y})$ may be different between training and evaluation dataset, which means domain mismatch exists. Assume there are two distributions $S(\mathbf{x}, \mathbf{y})$ and $T(\mathbf{x}, \mathbf{y})$ corresponding to source domain and target domain. Both of them are unknown. Due to the domain shift, S and T are similar but different.

The unsupervised domain adversarial training architecture is depicted in Figure 1. The architecture is based on a traditional feed-forward neural network. But different from a traditional network, it has two output layers, which are speaker label $\mathbf{y} \in Y$ and domain label $\mathbf{d} \in \{[0,1],[1,0]\}$. Denote with $\mathbf{d}_i$ ($[0,1]$ or $[1,0]$) for the i-th sample, which indicates whether $\mathbf{x}_i$ comes from the source domain ($\mathbf{x}_i \sim S(\mathbf{x})$ if $\mathbf{d}_i = [1,0]$) or from the target domain ($\mathbf{x}_i \sim T(\mathbf{x})$ if $\mathbf{d}_i = [0,1]$). Specifically, this model can be decomposed into three parts to perform different mappings: a feature extractor G_f, a speaker label predictor G_y and a domain predictor G_d. More formally, the mapping functions are:

$$\mathbf{f} = G_f(\mathbf{x}; \Theta_f); \quad (1)$$

$$\mathbf{y} = G_y(\mathbf{f}; \Theta_y); \quad (2)$$

$$\mathbf{d} = G_d(\mathbf{f}; \Theta_d); \quad (3)$$

where $\Theta_f, \Theta_y, \Theta_d$ are the parameters of the network (in Figure 1) and $\mathbf{f}$ is a D-dimension feature vector. From left to right in Figure 1, the features f are firstly extracted from the hidden layer. Our aim is to jointly train G_f, G_y and G_d. Specifically, we want to seek Θ_f to minimize the speaker label prediction loss and to maximize the domain classification loss at the same time, which can be done by a gradient reversal layer. Gradient reversal layer between the feature extractor and domain label predictor is introduced to search the saddle point

between speaker label classifier and domain classifier. This gradient reversal layer multiplies by a certain λ during the backpropagation. λ is a positive hyper parameter used to trade off two losses in practice. Gradient reversal layer ensures the feature distributions over the two domains are similar so that we can get domain-invariant and speaker-discriminative features.

Meanwhile, Θ_d is estimated to ensure that G_d will perform accurate domain classification. This is achieved by the loss function of this network:

$$E(\Theta_f, \Theta_y, \Theta_d) = \sum_{\substack{i=1,...N \\ \mathbf{d}_i=[1,0]}} L_y(G_y(G_f(\mathbf{x}_i; \Theta_f); \Theta_y), \mathbf{y}_i) - \lambda \sum_{i=1,...N} L_d(G_d(G_f(\mathbf{x}_i; \Theta_f); \Theta_d), \mathbf{d}_i) = \sum_{\substack{i=1,...N \\ \mathbf{d}_i=[1,0]}} L_y^i(\Theta_f, \Theta_y) - \lambda \sum_{i=1,...N} L_d^i(\Theta_f, \Theta_d) \quad (4)$$

where $L_y^i(.,.)$ and $L_d^i(.,.)$ are the loss of the i-th training sample for speaker label and domain predictors, respectively. We define a cross entropy function as the loss function.

According to the loss function, we can optimize the DAT network using stochastic gradient descent (SGD) [18] approach. We optimize the parameters so that:

$$(\hat{\Theta}_f, \hat{\Theta}_y) = \arg \min_{\Theta_f, \Theta_y} E(\Theta_f, \Theta_d, \Theta_y), \quad (5)$$

$$\hat{\Theta}_d = \arg \max_{\Theta_d} E(\Theta_f, \Theta_d, \Theta_y). \quad (6)$$

We noted that by maximizing Eq (4) for Θ_d we minimize the second item of Eq (4). In this way, Θ_d is optimaized for performance of domain predictor. We optimize Θ_f by minimizing the first item and maximizing the second item. With such an optimization strategy, we make sure the features extracted from the neural network are domain-invariant and classification-discriminative.

2.2. Extracting Speaker-Discriminative and Domain-Invariant Speech Representations

Fig. 2 shows how we use DAT strategy in speaker recognition. After training the domain adversarial neural network (DANN), we use enroll i-vector (i_e) and test i-vector (i_t) as the input to the DANN, and extract the new vectors $\hat{i_e}$, $\hat{i_t}$ from the hidden layer of feature extractor network of DANN. $\hat{i_e}$ and $\hat{i_t}$ are therefore expected to be domain-invariant and speaker-discriminative speech representation which stand in the same subspace. Then, we apply the pre-processing (whitening and length-norm [19]) to $\hat{i_e}$, $\hat{i_t}$. Finally, we use a scoring function to compute the scores between the speaker model and the test sample. In this paper, we adopt PLDA scoring.

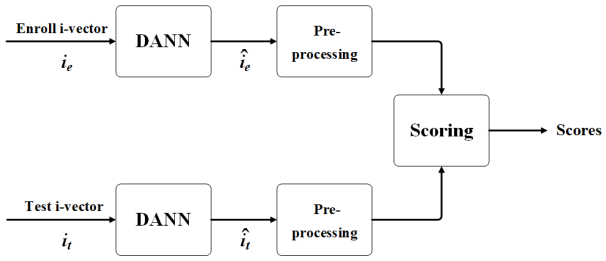

Fig. 2. Block diagram of DAT based speaker recognition. i_e and i_t represent the enroll and test i-vectors from the target domain, respectively. $\hat{i_e}$ and $\hat{i_t}$ indicate the extracted domain-invariant and speaker-discriminative speech representations.

3. EXPERIMENTAL SETUP

3.1. Evaluation Dataset

In this paper, we use 2013 domain adaptation challenge dataset (DAC 13) [20] as evaluation dataset. DAC 13 posed a task based on LDC telephone corpora which demonstrates the effect of dataset mismatch on hyper-parameters such as the latent speaker and channel factors for PLDA, and provided the audio lists and i-vectors of NIST SRE and SWB. In this paper, for fair comparison with other domain adaptation techniques, the DAC 13 i-vector dataset is used as the training and evaluation dataset, which contains i-vectors that were generated from UBM and total variability matrix (T-matrix) with 600-dim total factor space. Only SWB data were used to train the UBM and T-matrix. NIST SRE 2010, denoted as SRE10, telephone data is used as enroll and test sets. There are two dataset used for hyper-parameter training: the source domain SWB set consists of all telephone calls from all speakers taken from the Switchboard-I and Switchboard-II (all phases) corpora. The target domain SRE set consists of all the telephone calls without speaker labels taken from the NIST SRE 04, 05, 06, and 08 collections, while SRE-1phn is a reduced set of SRE with only the i-vectors from 1 telephone number per speaker, which makes it hard to estimate within-class variability because of the lack of speaker and channel information. This paper selected the more challenging SRE-1phn data for the domain adaptation task.

3.2. Domain-Adversarial Neural Network (DANN)

In the baseline experiment, we use SRE-1phn data to compute m and W and estimate the center mean and whitening matrix. Pre-processing by centering, whitening and length normalization is performed on all i-vectors. And we use SRE-1phn and SWB to estimate PLDA model, respectively, as the domain matched condition and domain mismatch condition baseline. The number of eigenvoices of PLDA is set to 400.

In the proposed DAT approach, i-vector pre-processing are done first. Training data of DANN consists of two parts: SWB i-vectors with speaker labels and SRE-1phn i-vectors without speaker label. SWB data are used to train the whole

Table 1. A comparative study between DAT and the state-of-the-art adaptation methods under DAC i-vector dataset

Systems#	Adaptation Methods	EER%	DCF10 [21]	DCF08
1	–	9.35	0.724	0.520
2	–	5.66	0.633	0.427
3	Interpolated [6] [12]	6.55	0.652	0.454
4	IDV [9] [12]	6.15	0.676	0.476
5	DICN [11] [12]	4.99	0.623	0.416
6	DAE [22] [12]	4.81	0.610	0.398
7	AEDA [12]	4.50	0.589	0.362
8	**DAT**	**3.73**	**0.541**	**0.335**

network while the SRE-1phn i-vectors are used to train the feature extractor and the domain classifier, because the data from target domain does not have speaker labels.

At the test stage, we use SWB, SRE-1phn, enroll and test data as the inputs to the network and extract the domain-invariant and speaker-discriminative speech representation from the first hidden layer of the feature extractor network. After that we do the PLDA back-end to obtain the scores.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1. DANN vs. State-of-the-art Unsupervised Domain Adaptation Methods

The experimental results are given in Table 1. System 1 is a domain matched condition, using known label target domain SRE-1phn data to estimate WC and AC. System 2 is the baseline of domain mismatched condition, using known label source domain data to estimate WC and AC. The EER of System 2 under domain mismatched condition is better than System 1 under domain matched condition. In [12], the authors believe that the performance degradation of System 1 is due to insufficient channel information. In a comparative study, we also report the results from other state-of-the-art studies in [12]. We observe that, by projecting the data to a common space with DAT approach, DAT (System 8) in Table 1 shows a 34% improvement over the System 2 baseline on EER.

4.2. The effect of λ in DANN

We also investigate the impact of the hyper-parameters λ, which used to trade off the two losses, on the performance of DANN. The impact of λ on EER, DCF10 and DCF08 are depicted in Fig. 3 and Fig. 4. When λ=0, the domain predictor is not trained. We can see the EER decreases as λ increases. We reach the lowest EER at λ=0.5. Similarly DCF10 and DCF08 decrease as λ increases. We reach the lowest DCF when λ=0.4.

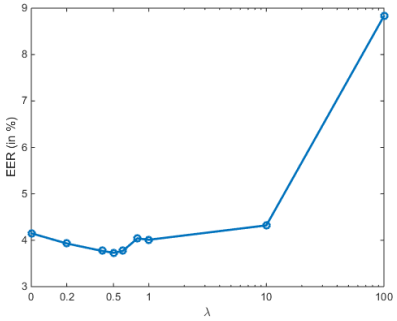

Fig. 3. EER of DANN as a function of λ in log scale.

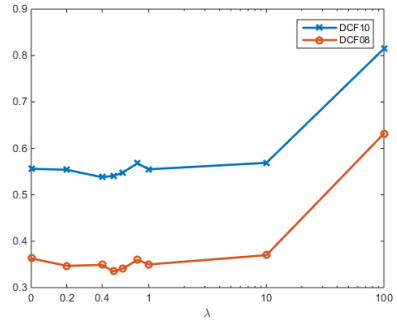

Fig. 4. DCF10 and DCF08 as a function of λ in log scale.

5. CONCLUSIONS

In this paper, we introduce an unsupervised domain adaptation approach-domain adversarial training for speaker recognition, which overcomes the domain mismatch problem in the speaker recognition by projecting the source domain and target domain data into the same subspace. As this approach doesn't require labeled data from the target domain, we call it unsupervised domain adaptation. But we should note that the training still requires the labeled training data from the source domain, therefore, the training is carried out as in a mix of supervised (speaker) and unsupervised (domain) manner. By DAT approach, we can obtain domain-invariant and speaker-discriminative speech representations. Experiments on DAC 13 i-vector dataset show that, the proposed approach improves the equal error rate from 5.66% to 3.73%, with 34% relative error reduction and outperforms the other compared domain adaptation techniques. In the future, we will explore the effectiveness of DAT on NIST SRE 16 database and compare the difference between DAT and the recently popular general adversarial network.

6. ACKNOWLEDGEMENT

The research work is supported by the National Key Research and Development Program of China (Grant No.2017YFB1002102) and the National Natural Science Foundation of China (Grant No.61571363).

7. REFERENCES

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio Speech Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[3] Daniel Garcia-Romero and Alan Mccree, "Supervised domain adaptation for i-vector based speaker recognition," in *ICASSP*, 2014, pp. 4047–4051.

[4] Simon J D Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," *Proceedings*, pp. 1–8, 2007.

[5] Pavel Matjka, Ondej Glembek, Fabio Castaldo, M. J. Alam, Oldich Plchot, Patrick Kenny, Luk Burget, and Jan ernocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," in *ICASSP*, 2011, pp. 4828–4831.

[6] 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.

[7] Daniel Garcia-Romero, Xiaohui Zhang, Alan Mccree, and Daniel Povey, "Improving speaker recognition performance in the domain adaptation challenge using deep neural networks," in *Spoken Language Technology Workshop*, 2015.

[8] Stephen H Shum, Douglas A Reynolds, Daniel Garcia-Romero, and Alan McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[9] Ahilan Kanagasundaram, David Dean, and Sridha Sridharan, "Improving out-domain plda speaker verification using unsupervised inter-dataset variability compensation approach," in *ICASSP*, 2015.

[10] Hagai Aronowitz, "Inter dataset variability compensation for speaker recognition," in *ICASSP*, 2014, pp. 4002–4006.

[11] Md Hafizur Rahman, Ahilan Kanagasundaram, David Dean, and Sridha Sridharan, "Dataset-invariant covariance normalization for out-domain plda speaker verification," in *INTERSPEECH*, 2015, pp. 1017–1021.

[12] Suwon Shon, Seongkyu Mun, Wooil Kim, and Hanseok Ko, "Autoencoder based domain adaptation for speaker recognition under insufficient channel information," in *INTERSPEECH*, 2017, pp. 1014–1018.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Franois Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2015.

[14] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.

[15] Jun Yu, Baopeng Zhang, Zhengzhong Kuang, Dan Lin, and Jianping Fan, "iprivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE Transactions on Information Forensics Security*, vol. 12, no. 5, pp. 1005–1016, 2017.

[16] Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, pp. 79–87, 2017.

[17] Yusuke Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," in *INTERSPEECH*, 2016, pp. 2369–2372.

[18] Léon Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

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

[20] "Jhu 2013 speaker recognition workshop," http://www.clsp.jhu.edu/wpcontent/uploads/sites/75/2015/10/WS13-Speaker-DAC.pdf.

[21] Alvin F Martin and Craig S Greenberg, "The nist 2010 speaker recognition evaluation," in *INTERSPEECH*, 2010.

[22] Oleg Kudashev, Sergey Novoselov, Konstantin Simonchik, and Alexandr Kozlov, "A speaker recognition system for the sitw challenge," in *INTERSPEECH*, 2016, pp. 833–837.