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[4]

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

18 April, 2018

西北工業大學 Northwestern Polytechnical University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NUS National University of Singapore

Outline

- Introduction
- Proposed Method
- Experimental Setup and Result
- Conclusions



Introduction | Domain Mismatch in Speaker Recognition

- Conventional approaches of speaker recognition usually assume that training and evaluation data share the same probability distributions or the same feature space.

- However, in the real-world application, there is always a mismatch between the training and evaluation datasets, which leads to the **domain mismatch** in speaker recognition.

- **Domain adaptation** is seen as a solution to alleviate the domain mismatch,

Introduction | Previous work

- Domain adaptation for speaker recognition
  - Training dataset ⟶ Source domain
  - Evaluation dataset ⟶ Target domain

- According to the availability of labels in target domain:
  - Supervised domain adaptation
  - Unsupervised domain adaptation
    - Use clustering techniques to estimate speaker label of unlabeled target domain data.
    - Select the unlabeled target and source domain data to estimate a compensation model to compensate the domain mismatch.
    - Learn the domain-invariant space or map the source domain data into target domain space and use the mapped source domain data with its speaker label to train LDA or PLDA.
      - Autoencoder based Domain Adaptation (AEDA): adapt source domain data to target domain.



Introduction | Contribution

- Apply Domain Adversarial Training (DAT) [2] to solve the domain mismatch problem in speaker recognition.
- Project the source domain data and target domain data into the common domain.
- Learn the domain-invariant and speaker-discriminative speech representations.

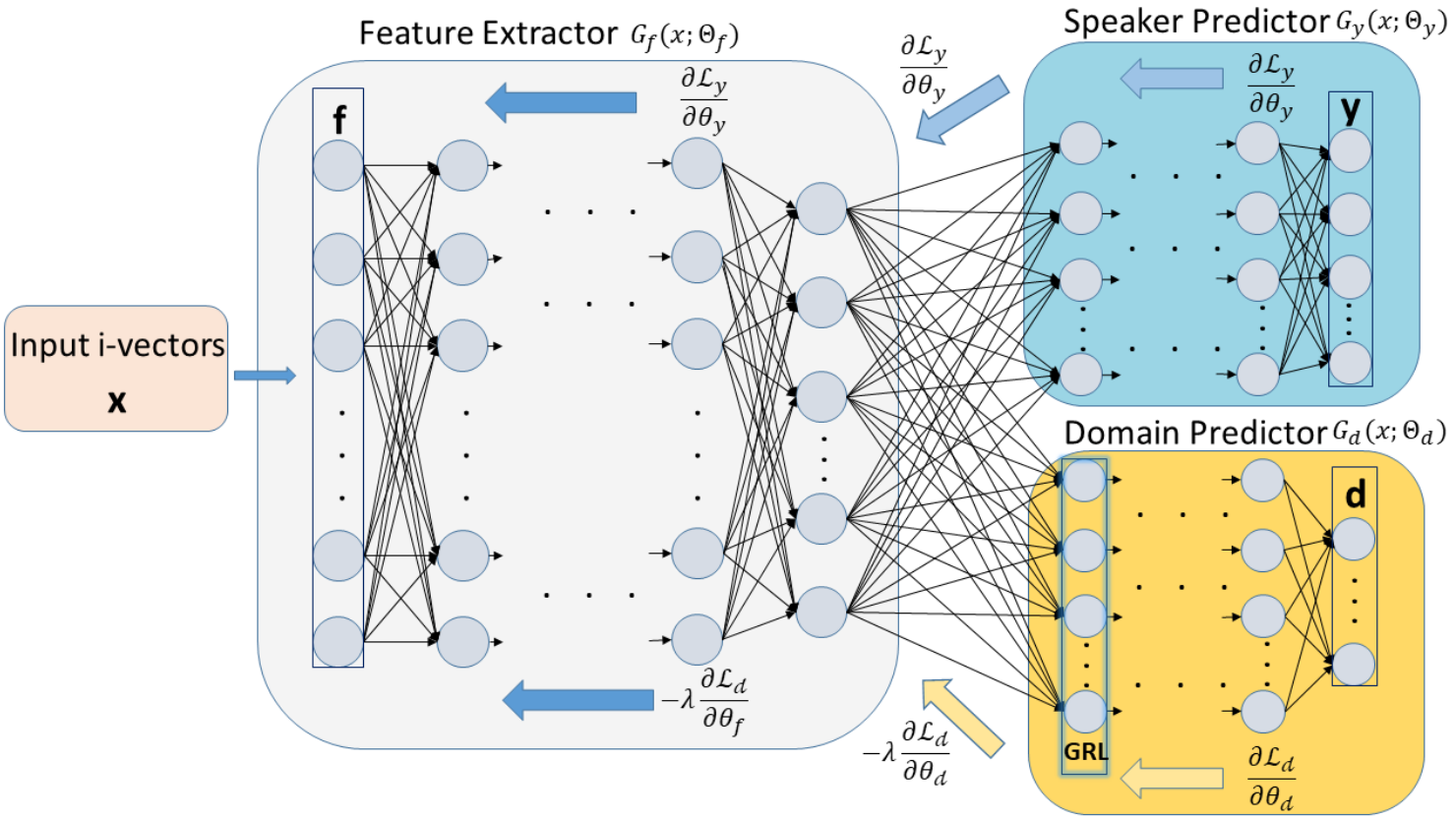
[1] Yaroslav Ganin et al. “Domain-Adversarial Training of Neural Networks”. In: Journal of Machine Learning Research 17.1 (2015), pp. 2096–2030.
[2] Sining Sun et al. “An unsupervised deep domain adaptation approach for robust speech recognition,” Neurocomputing, pp. 79– 87, 2017.



Method

DAT in Speaker Recognition

- Domain Adversarial Training (DAT)






西北工業大學 Northwestern Polytechnical University
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
NUS National University of Singapore

Method

DAT in Speaker Recognition

- Gradient Reversal Layer (GRL)
  - ensures the feature distributions over the two domains are similar so that we can get domain-invariant and speaker-discriminative features.
  - multiplies by a certain **negative** hyper parameter during the backpropagation, used to trade off two losses.
- Loss Function

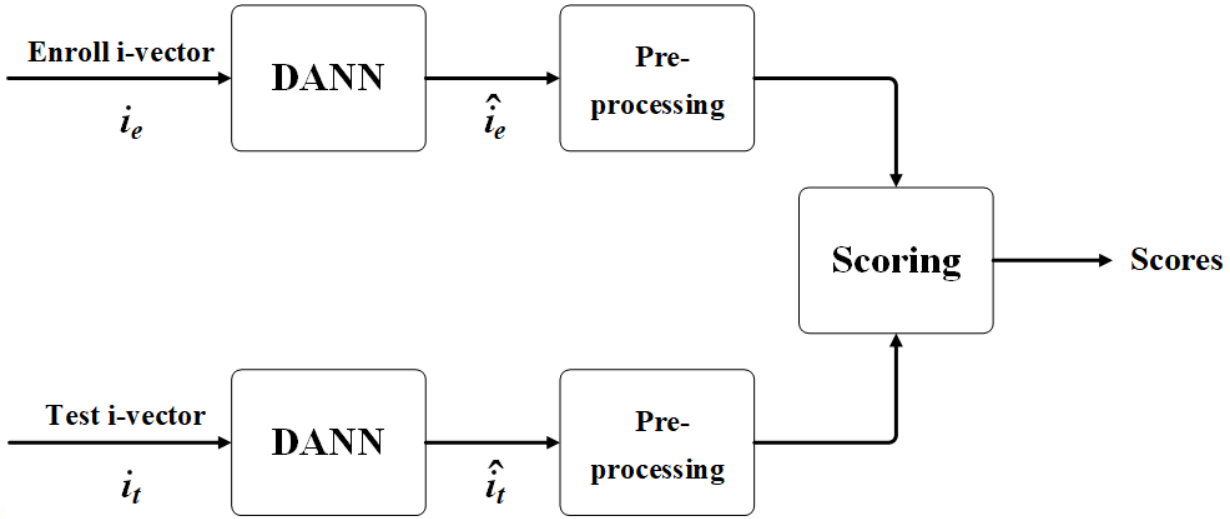
$$E(\Theta_f, \Theta_y, \Theta_d) = \sum_{\substack{i=1,\dots N \\ \mathbf{d}_i=[1,0]}} L_y(G_y(G_f(\mathbf{x}_i;\Theta_f);\Theta_y), \mathbf{y}_i) - \lambda \sum_{i=1,\dots N} L_d(G_d(G_f(\mathbf{x}_i;\Theta_f);\Theta_d), \mathbf{d}_i)$$

$$= \sum_{\substack{i=1,\dots N \\ \mathbf{d}_i=[1,0]}} L_y^i(\Theta_f, \Theta_y) - \lambda \sum_{i=1,\dots N} L_d^i(\Theta_f, \Theta_d)$$



Method | i-vector Scoring

- Domain Adversarial Neural Network (DANN): we call the model trained by DAT method as DANN
  - Input: enroll data i-vector (i_e) and test i-vector (i_t)
  - Extract new vectors $\widehat{i_e}$, $\widehat{i_t}$ from the hidden layer of the feature extractor sub-network from DANN
  - Domain-invariant and speaker-discriminative speech representations



Experimental Setup and Result

- Dataset:
  - 2013 domain adaptation challenge dataset (DAC 13) i-vector Dataset
    - Source domain data: SWB
    - Target domain data: SRE, SRE-1phn
    - Test data: SRE10 telephone data

	SWB	SRE	SRE-1phn
#spks	3114	3790	3787
#calls	33039	36470	25640
#calls/spkrs	10.6	9.6	6.77
#phone_num/spkr	3.8	2.8	1.0

i-vector Statistic in DAC 13 i-vector Dataset



Experimental Setup and Result

- Baseline Experiments:
  - System1: domain match
  - System2: domain mismatch
  - System3: domain match & insufficient channel information
  - System4: domain mismatch

System#	Pre-processing	PLDA	EER%	DCF10	DCF08
1	SRE	SRE	2.33	0.402	0.235
2	SRE	SWB	5.65	0.632	0.427
3	SRE-1phn	SRE-1phn	9.35	0.724	0.520
4	SRE-1phn	SWB	5.66	0.633	0.427



Experimental Setup and Result

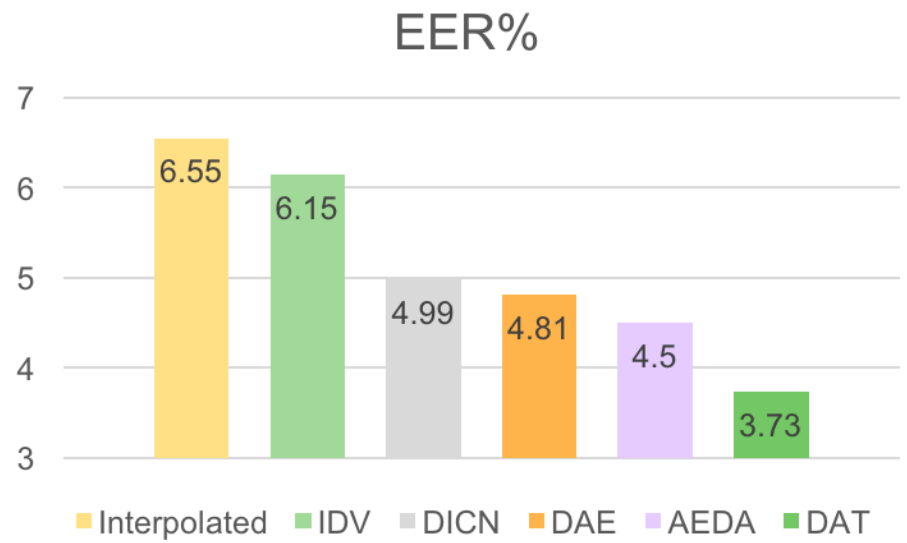
- DAT method Experiments:
  - Training data of DANN:
    - SWB i-vectors with speaker labels (used to train the whole network)
    - SRE-1phn i-vectors without speaker label (used to train the feature extractor and domain classifier)
  - Baseline systems:
    - System 4: PLDA ⟶ SWB (Source domain data)

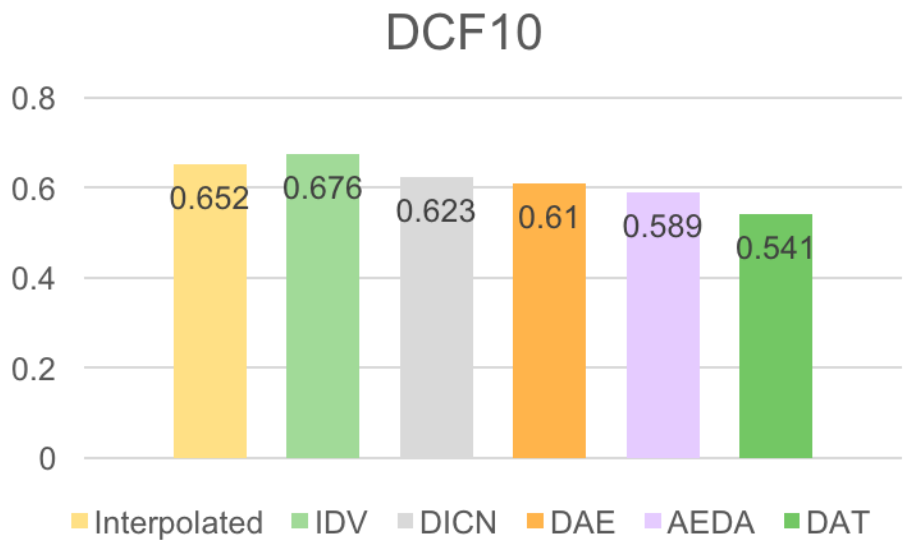
System#	Adaptation Methods	EER%	DCF10	DCF08
4	-	5.66	0.633	0.427
5	**DAT**	**3.73**	**0.541**	**0.335**



Experimental Setup and Result

- DAT vs. state-of-the-art unsupervised domain adaptation methods





Interpolated
IDV: Inter-dataset variability compensation
DICN: Dataset-Invariant Covariance Normalization
DAE: Denoising Autoencoder
AEDA: Autoencoder based Domain Adaptation



Conclusions

- We have proposed to perform domain adversarial training for speaker recognition.
- DAT overcomes the domain mismatch problem by projecting the source domain and target domain data into the same subspace.
- By DAT approach, we can obtain domain-invariant and speaker-discriminative speech representations.
- In future work, we will explore the effectiveness of DAT on NIST SRE 16 database and compare the difference between DAT and the generative adversarial network.



Reference

- [3] Suwon Shon et al. “Autoencoder based Domain Adaptation for Speaker Recognition under Insufficient Channel Information”. In: INTERSPEECH. 2017, pp. 1014–1018.
- [4] Daniel Garcia-Romero et al. “Unsupervised domain adaptation for i-vector speaker recognition”. In: Proceedings of Odyssey: The Speaker and Language Recognition Workshop. 2014.
- [5] Ahilan Kanagasundaram et al. “Improving out-domain PLDA speaker verification using unsupervised inter-dataset variability compensation approach”. In: ICASSP. 2015.
- [6] Md Hafizur Rahman et al. “Dataset-Invariant Covariance Normalization for Out-domain PLDA Speaker Verification”. In: INTERSPEECH. 2015, pp. 1017–1021.
- [7] Oleg Kudashev et al. “A speaker recognition system for the sitw challenge,” in INTERSPEECH, 2016, pp. 833–837.



Thank you!




西北工業大學
Northwestern Polytechnical University
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
NUS National University of Singapore