Unsupervised Domain Adaptation via Domain Adversarial Training for Speaker Recognition

Qing Wang\textsuperscript{1,2}, Wei Rao\textsuperscript{2}, Sining Sun\textsuperscript{1}, Lei Xie\textsuperscript{1}, Eng Siong Chng\textsuperscript{2,3}, Haizhou Li\textsuperscript{4}

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

18 April, 2018
Outline

- Introduction
- Proposed Method
- Experimental Setup and Result
- Conclusions
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,
Domain adaptation for speaker recognition

- Training dataset \(\rightarrow\) Source domain
- Evaluation dataset \(\rightarrow\) 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.
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.

- **Domain Adversarial Training (DAT)**

![Diagram of DAT in Speaker Recognition](image)
Method

- 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

\[
E(\Theta_f, \Theta_y, \Theta_d) = \sum_{i=1,...,N} L_y(G_y(G_f(x_i; \Theta_f); \Theta_y), y_i) - \lambda \sum_{i=1,...,N} L_d(G_d(G_f(x_i; \Theta_f); \Theta_d), d_i)
\]

\[
= \sum_{i=1,...,N} L^i_y(\Theta_f, \Theta_y) - \lambda \sum_{i=1,...,N} L^i_d(\Theta_f, \Theta_d)
\]
Method

- 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 $\hat{i}_e$, $\hat{i}_t$ 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

<table>
<thead>
<tr>
<th></th>
<th>SWB</th>
<th>SRE</th>
<th>SRE-1phn</th>
</tr>
</thead>
<tbody>
<tr>
<td>#spks</td>
<td>3114</td>
<td>3790</td>
<td>3787</td>
</tr>
<tr>
<td>#calls</td>
<td>33039</td>
<td>36470</td>
<td>25640</td>
</tr>
<tr>
<td>#calls/spkrs</td>
<td>10.6</td>
<td>9.6</td>
<td>6.77</td>
</tr>
<tr>
<td>#phone_num/spkr</td>
<td>3.8</td>
<td>2.8</td>
<td>1.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>System#</th>
<th>Pre-processing</th>
<th>PLDA</th>
<th>EER%</th>
<th>DCF10</th>
<th>DCF08</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SRE</td>
<td>SRE</td>
<td>2.33</td>
<td>0.402</td>
<td>0.235</td>
</tr>
<tr>
<td>2</td>
<td>SRE</td>
<td>SWB</td>
<td>5.65</td>
<td>0.632</td>
<td>0.427</td>
</tr>
<tr>
<td>3</td>
<td>SRE-1phn</td>
<td>SRE-1phn</td>
<td>9.35</td>
<td>0.724</td>
<td>0.520</td>
</tr>
<tr>
<td>4</td>
<td>SRE-1phn</td>
<td>SWB</td>
<td>5.66</td>
<td>0.633</td>
<td>0.427</td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>System#</th>
<th>Adaptation Methods</th>
<th>EER%</th>
<th>DCF10</th>
<th>DCF08</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>-</td>
<td>5.66</td>
<td>0.633</td>
<td>0.427</td>
</tr>
<tr>
<td>5</td>
<td>DAT</td>
<td>3.73</td>
<td>0.541</td>
<td>0.335</td>
</tr>
</tbody>
</table>
Experimental Setup and Result

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

### EER%

<table>
<thead>
<tr>
<th>Method</th>
<th>Interpolated</th>
<th>IDV</th>
<th>DICN</th>
<th>DAE</th>
<th>AEDA</th>
<th>DAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpolated</td>
<td>6.55</td>
<td>6.15</td>
<td>4.99</td>
<td>4.81</td>
<td>4.5</td>
<td>3.73</td>
</tr>
</tbody>
</table>

### DCF10%

<table>
<thead>
<tr>
<th>Method</th>
<th>Interpolated</th>
<th>IDV</th>
<th>DICN</th>
<th>DAE</th>
<th>AEDA</th>
<th>DAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpolated</td>
<td>0.652</td>
<td>0.676</td>
<td>0.623</td>
<td>0.61</td>
<td>0.589</td>
<td>0.541</td>
</tr>
</tbody>
</table>

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

Thank you!