Autoencoder based Domain Adaptation for Speaker Recognition under Insufficient Channel Information

Suwon Shon, Seongkyu Mun*, Wooil Kim**, Hanseok Ko*

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
Korea University, Seoul, South Korea*
Incheon National University, South Korea**
Introduction

- Speaker recognition task

![Diagram of speaker recognition process]

- Input voice
- Speaker recognition system
- Enrolled speaker? True / False

Training Dataset
Introduction

• Channel domain mismatched condition

"Landline" Training Dataset
<out-of-domain>

"Cellular" input voice

Speaker recognition system

Enrolled speaker? True / False

Fail!
Introduction

• Domain adaptation challenge 2013 @ JHU workshop
  – SRE10 (evaluation) collected in 2010 (mostly cellular)
    * 7,169 target and 408,950 non-target trials
  – SWB collected from 1992-2000 (mostly landline), mismatched
  – SRE collected from 2004-2008 (mostly cellular), matched
    * Suppose we don’t have labels on SRE
Introduction

• Domain adaptation challenge 2013 @ JHU workshop
  – SRE10 (evaluation) collected in 2010 (mostly cellular)
    * 7,169 target and 408,950 non-target trials
  – SWB collected from 1992-2000 (mostly landline), mismatched
  – SRE collected from 2004-2008 (mostly cellular), matched
    * Suppose we don’t have labels on SRE

<table>
<thead>
<tr>
<th>System #</th>
<th>Unlabeled data</th>
<th>Labeled data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UBM, T</td>
<td>W,m</td>
</tr>
<tr>
<td>0*</td>
<td>SRE</td>
<td>SRE</td>
</tr>
<tr>
<td>1</td>
<td>SWB</td>
<td>SRE</td>
</tr>
<tr>
<td>2</td>
<td>SWB</td>
<td>SRE</td>
</tr>
<tr>
<td>3*</td>
<td>SWB</td>
<td>SWB</td>
</tr>
</tbody>
</table>

Table 2: SRE10 Test using DAC13 i-vector set.

Motivation

- Insufficient Channel Information

<table>
<thead>
<tr>
<th></th>
<th>SWB #spkrs</th>
<th>SRE #spkrs</th>
<th>SRE-1phn #spkrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>#spkrs</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>Avg. #calls/spkrs</td>
<td>10.6</td>
<td>9.6</td>
<td>6.77</td>
</tr>
<tr>
<td>Avg. #phone_num/spkr</td>
<td>3.8</td>
<td>2.8</td>
<td>1</td>
</tr>
</tbody>
</table>

<Statistics in DAC 13 i-vector Dataset>

<table>
<thead>
<tr>
<th>System #</th>
<th>UBM, T</th>
<th>W,m</th>
<th>WC,AC</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SWB</td>
<td>SRE</td>
<td>SRE</td>
<td>2.33</td>
</tr>
<tr>
<td>2</td>
<td>SWB</td>
<td>SRE</td>
<td>SWB</td>
<td>5.70</td>
</tr>
<tr>
<td>3</td>
<td>SWB</td>
<td>SRE-1phn</td>
<td>SRE-1phn</td>
<td>9.34</td>
</tr>
<tr>
<td>4</td>
<td>SWB</td>
<td>SRE-1phn</td>
<td>SWB</td>
<td>5.66</td>
</tr>
</tbody>
</table>

<SRE10 Test using DAC13 i-vector set>

Performance degraded by Insufficient channel information although it is matched domain dataset
Proposed Approach

- **Auto-encoder based Domain Adaptation (AEDA)**

![Diagram showing the proposed approach with symbols for out-of-domain and in-domain i-Vectors, adapted i-Vectors, and covariance sets.]

Useless because of insufficient channel information
Proposed Approach

- Auto-encoder based Domain Adaptation (AEDA)

Transferring labeled out-of-domain dataset to in-domain
Proposed Approach

• Autoencoder and Denoising Autoencoder

\[ \mathcal{L}(X, X') = ||X - X'||^2 \]

<Autoencoder>

\[ \mathcal{L}(X, X') = ||X - \tilde{X}'||^2 \]

<Denoising Autoencoder>
Proposed Approach

- Auto-encoder based Domain Adaptation (AEDA)
Proposed Approach

- Auto-encoder based Domain Adaptation (AEDA)

In-domain autoencoder (using unlabeled in-domain dataset)
Proposed Approach

- Auto-encoder based Domain Adaptation (AEDA)

\[ \Omega \times \tilde{X} = \times \tilde{X} \]

\[ X_{in} \approx \Omega_{in} \cdot A = \tilde{X}_{in} \]

domain transferring autoencoder (using labeled out-of-domain dataset)
Proposed Approach

• Sparse reconstruction

\[ \mathcal{L}(X, X') = ||X - \tilde{X}'||^2 \]

<Denoising Autoencoder>

\[ Y_{in} = X_{out}' \]

\[ X_{out} \approx \Omega_{in} \cdot A = \tilde{X}_{in} \]

Objective function: \( \min_{\alpha_j} \| \Omega_{in} \alpha_j - y_{in} \| ^2 + \gamma | \alpha_j | ^2 \)

\[ \mathcal{L}(X_{in}, Y_{in}) = ||X_{in} - Y_{in}||^2 \]

\[ = ||\tilde{X}_{in} - Y_{in}||^2 \]

<Out-of-domain transferring autoencoder>
Proposed Approach

- **Structure of Autoencoder which sharing hidden layer h**

- **Objective functions**
  - AE part
    \[
    \min_{f_{in}, g} \| X_{in} - Y_{in} \|^2 = \min_{f_{in}, g} \| X_{in} - g(f_{in}(X_{in})) \|^2
    \]
Proposed Approach

- **Structure of Autoencoder which sharing hidden layer h**

- **Objective functions**
  - AE part\(\min_{f_{in},g} \| X_{in} - Y_{in} \|^2 = \min_{f_{in},g} \| X_{in} - g(f_{in}(X_{in})) \|^2\)
  - Least angle regression for sparse reconstruction\(\min_{\alpha_j} \| \Omega_{in} \alpha_j - y_{in}^j \|^2 + \gamma | \alpha_j |^2\)
Proposed Approach

- **Structure of Autoencoder which sharing hidden layer h**

- **Objective functions**
  - **AE part** \( \min_{f_{in},g} \| X_{in} - Y_{in} \|^2 = \min_{f_{in},g} \| X_{in} - g(f_{in}(X_{in})) \|^2 \)
  - **DAE part** \( \min_{f_{out},g} \| X_{in} - Y_{in} \|^2 = \min_{f_{out},g} \| X_{in} - g(f_{out}(X_{out})) \|^2 \)
  - **AEDA** \( \min_{f_{in},f_{out},g} \| X_{in} - g(f_{in}(X_{in})) \|^2 + \| \tilde{X}_{in} - g(f_{out}(X_{out})) \|^2 \)

\[ X_{in} \approx \Omega_{in} \cdot A = \tilde{X}_{in} \]
Proposed Approach

• Structure of Autoencoder which sharing hidden layer h

\[ \| X_{in} - Y_{in} \|_2^2 \]

• AEDA
  – 600 dim i-vector with 1000 hidden node with learning rate 0.005

• Sparse reconstruction
  – Least Angle Regression(LARS)
  – Sparsity 0.01
  – Random 1500 spk i-vector for in-domain dictionary \( \Omega_{in} \)

• Performance
  – Using PLDA with 400 eigenvoice after 400 dim LDA transform
  – EER, DCF10, DCF08
## Experimental result

- **Auto-encoder based Domain Adaptation (AEDA)**

<table>
<thead>
<tr>
<th>#</th>
<th>Adaptation &amp; Compensation</th>
<th>WC,AC</th>
<th>EER</th>
<th>DCF10</th>
<th>DCF08</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>-</td>
<td>SRE-1phn</td>
<td>9.34</td>
<td>0.721</td>
<td>0.520</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>SWB</td>
<td>5.66</td>
<td>0.633</td>
<td>0.426</td>
</tr>
<tr>
<td>5</td>
<td>Interpolated [13]</td>
<td>SWB + SRE-1phn</td>
<td>6.55</td>
<td>0.652</td>
<td>0.454</td>
</tr>
<tr>
<td>6</td>
<td>IDV [15]</td>
<td>IDV-SWB</td>
<td>6.15</td>
<td>0.676</td>
<td>0.476</td>
</tr>
<tr>
<td>7</td>
<td>DICN [16]</td>
<td>DICN-SWB</td>
<td>4.99</td>
<td>0.623</td>
<td>0.416</td>
</tr>
<tr>
<td>8</td>
<td>DAE [23]</td>
<td>DAE-SWB</td>
<td>4.81</td>
<td>0.610</td>
<td>0.398</td>
</tr>
<tr>
<td>9</td>
<td>AEDA</td>
<td>AEDA-SWB</td>
<td>4.50</td>
<td>0.589</td>
<td>0.362</td>
</tr>
</tbody>
</table>

\(<\text{SRE10 evaluation result with DAC 13 Dataset when Unlabeled In-Domain Dataset is Available}>)\)
Conclusion

- Only small subset of unlabeled in-domain dataset is used for domain adaptation

- Insufficient channel information dataset is effectively used for transferring knowledge of in-domain

- Domain transferring autoencoder part of AEDA can be trained using sparse reconstruction without actual pair of in-domain and out-of-domain
Q & A

• Thanks!

• Domain related paper:
  Suwon Shon, Seongkyu Mun and Hanseok Ko,
  “Recursive whitening transformation for speaker recognition on Language Mismatched Condition”
  @ 4.9 Evaluation of Speaker and language identification systems session, Wednesday 10:00~12:00