CNN with Phonetic Attention for Text-Independent Speaker Verification

Tianyan Zhou, Yong Zhao, Jinyu Li, Zhuo Chen, Jian Wu
{Tianyan.zhou; yonzhao; jinyli; zhuc; jianwu}@microsoft.com
CNN architecture

• CNN based system has become an effective solution for speaker verification due to the ability to capture local temporal and frequency patterns.

• Our network:
Attentive pooling

- Not all frames provide equal evidence to infer speaker identities.

- Instead of averaging, the attention mechanism provides a better alternate to actively select the hidden representations and emphasize speaker-discriminative information.

- We implement the shared-parameter non-linear multi-head attentive pooling (AP), similar to [1, 2].

\[
U = \text{softmax}(V^T \tanh(W^T H + b))
\]

\[
z = UH^T
\]

U: attention weights
H: segment-level embedding
z: utterance-level embedding

Phonetic attention[3]

• Given the local connectivity and spatial contiguity in convolutional operations, we cannot directly feed the phonetic features into the CNN network.

• Transformation network is jointly trained with the following CNN.

Visualization

• non-speech areas are mostly assigned with smaller weights than voiced areas.

• phonetic attention can effectively emphasize or deemphasize the importance of frames depending on the speech content.
BLSTM-CNN Network

Input features $x_1, x_2, \ldots, x_T$

Speaker embedding $z$

Attentive pooling

CNN (2-D)

Projection

BLSTM

Augmented feature maps
System combination for Open Training

- EER reported on VoxCeleb1-test.

<table>
<thead>
<tr>
<th>System</th>
<th>Training set</th>
<th>Inputs</th>
<th>Network</th>
<th>Aggregation</th>
<th>EER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VoxCeleb1&amp;2</td>
<td>MFCC</td>
<td>xvectors</td>
<td>-</td>
<td>3.10</td>
</tr>
<tr>
<td>2</td>
<td>VoxCeleb2</td>
<td>LFB</td>
<td>ResNet</td>
<td>AP</td>
<td>1.91</td>
</tr>
<tr>
<td>3</td>
<td>VoxCeleb2</td>
<td>LFB</td>
<td>BLSTM+ResNet</td>
<td>AP</td>
<td>1.66</td>
</tr>
<tr>
<td>4</td>
<td>VoxCeleb1&amp;2</td>
<td>LFB</td>
<td>ResNet</td>
<td>TAP</td>
<td>2.44</td>
</tr>
<tr>
<td>5</td>
<td>VoxCeleb1&amp;2(short)</td>
<td>LFB</td>
<td>ResNet</td>
<td>TAP</td>
<td>3.52</td>
</tr>
<tr>
<td>6</td>
<td>VoxCeleb1&amp;2</td>
<td>LFB</td>
<td>ResNet</td>
<td>AP</td>
<td>2.04</td>
</tr>
<tr>
<td>7</td>
<td>VoxCeleb1&amp;2</td>
<td>LFB+PBN</td>
<td>ResNet</td>
<td>AP</td>
<td>1.53</td>
</tr>
<tr>
<td>8</td>
<td>VoxCeleb1&amp;2(w/sim)</td>
<td>LFB+PBN</td>
<td>ResNet</td>
<td>AP</td>
<td>1.61</td>
</tr>
<tr>
<td>fusion</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.34</td>
</tr>
</tbody>
</table>
System combination for Fixed Training

- EER reported on VoxCeleb1-test.

<table>
<thead>
<tr>
<th>System</th>
<th>EER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet</td>
<td>1.91</td>
</tr>
<tr>
<td>BLSTM+ResNet</td>
<td>1.66</td>
</tr>
<tr>
<td>Fusion</td>
<td>1.55</td>
</tr>
</tbody>
</table>
Contacting jianwu@Microsoft.com or 1-425-7654885 during interspeech if interested