Audiovisual Speech Activity Detection with Advanced Long Short-Term Memory

Fei Tao and Carlos Busso

Erik Jonsson School of Engineering & Computer Science at the University of Texas at Dallas, Richardson, Texas 75080, USA

Motivation

Background:
- Speech activity detection (SAD) is an important pre-processing step in speech-based interfaces
- Introducing visual information can improve performance and robustness
- Longer periodicity in the acoustic and visual features is crucial to distinguish speech activity
- Recurrent connections in LSTM only consider previous frame

Our Work:
- This study proposes to explore the advanced LSTM (A-LSTM) layers to improve the temporal dependency of our AV-SAD system

Experiment and Results

Speech Activity Detection:
- Baseline: BRNN proposed in [Tao & Busso, 2007] with LSTM
- Run experiment on Nvidia GTX 1070 (8GB)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Test Condition</th>
<th>Acc [%]</th>
<th>Pre [%]</th>
<th>Rec [%]</th>
<th>F1 [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRNN</td>
<td>HD</td>
<td>93.1</td>
<td>94.5</td>
<td>86.8</td>
<td>88.8</td>
</tr>
<tr>
<td>BRNN</td>
<td>TG</td>
<td>93.1</td>
<td>95.9</td>
<td>87.6</td>
<td>88.8</td>
</tr>
<tr>
<td>A-BRNN</td>
<td>HD</td>
<td>93.1</td>
<td>95.9</td>
<td>85.2</td>
<td>88.7</td>
</tr>
<tr>
<td>A-BRNN</td>
<td>TG</td>
<td>93.1</td>
<td>95.9</td>
<td>87.6</td>
<td>88.8</td>
</tr>
</tbody>
</table>

Non-speech Segments with Active Lip Motion:
- Evaluate the robustness to lip movements that are not associated with speech (smiles, lip-smack, deep breath)
- We manually identified 7,397 frames across speakers containing non-speech lip motion
- Report in F-score

<table>
<thead>
<tr>
<th>Approach</th>
<th>C</th>
<th>HD</th>
<th>N</th>
<th>C</th>
<th>TG</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRNN</td>
<td>94.2</td>
<td>93.1</td>
<td>94.7</td>
<td>89.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-BRNN</td>
<td>94.7</td>
<td>93.4</td>
<td>94.7</td>
<td>89.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SNR Analysis:
- HD channel is not affected much by noise (mic is close to the mouth)
- TG channel is affected since it is close to the noise source
- Clean session contains spontaneous speech, which is a harder task

Conclusions

- This study extended BRNN using A-LSTM for AV-SAD
- The proposed framework takes advantage of BRNN
- It has low latency and better time dependency modeling
- It is better in non-speech segments with active lip motion

Future Work
- The current implementation only uses A-LSTM in one layer, which is limited by the hardware requirement
- A-LSTM can be used in more layers.
- More frames in the past can be considered
- Learn facial and acoustic features with CNN
- Training the approach as an end-to-end system

This work was funded by NSF awards IIS-1718944 and IIS-1453781 (CAREER)