Deep Learning for Prominence Detection in Children’s Read Speech

Mithilesh Vaidya, Kamini Sabu, Preeti Rao
Department of Electrical Engineering, Indian Institute of Technology Bombay, India

Motivation
- Humans convey rich non-phonetic information during speech delivery
- Prosody – suprasegmental attributes of speech which convey both linguistic and para-linguistic information
- Expressiveness - important prosodic component of spoken language fluency which conveys linguistic information
- Proficient readers are good at:
  - Emphasis - Stress on the right words to convey novel info.
  - Phrasing – group words together to convey syntactic info.
- Phrasing: Spoken language: ; Punctuation: Text
- Pitch, intensity, duration – acoustic correlates of prosody
- Typical changes: word lengthening, within-word pitch variation

Waveform-based feature extraction
- Words sleeping, came and tail made prominent by increasing duration (waveform) and varying pitch (blue contour)

Dataset
- Goal: Predict oral reading fluency of school children
- 41,286 words across 790 utterances of read stories (~52 words/utterance)
- 4 hours 20 minutes of speech at 16 KHz sampling rate
- 35 (middle-school) speakers
- Each word rated for presence/absence of prominence and boundary by 7 naive listeners using RPT methodology (Cole et al., 2017)
- Votes scaled to 0-1; goal is to predict degree of prominence for each word
- Dataset split into 3 speaker non-overlapping folds
- Evaluation: Pearson correlation coefficient

Results

<table>
<thead>
<tr>
<th>No.</th>
<th>Input</th>
<th>Acoustic Model</th>
<th>Layer 1 (type, width, stride)</th>
<th>Pearson correl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>A34</td>
<td>BGRU</td>
<td>-</td>
<td>0.726</td>
</tr>
<tr>
<td>2.</td>
<td>Wav</td>
<td>CRNN</td>
<td>Standard, 51, 1</td>
<td>0.692</td>
</tr>
<tr>
<td>3.</td>
<td>Wav</td>
<td>CRNN</td>
<td>Sinc, 51, 1</td>
<td>0.712</td>
</tr>
<tr>
<td>4.</td>
<td>Wav</td>
<td>CRNN</td>
<td>Sinc, 31, 2</td>
<td>0.721</td>
</tr>
<tr>
<td>5.</td>
<td>A34 + Wav</td>
<td>CRNN</td>
<td>Sinc, 31, 2</td>
<td>0.735</td>
</tr>
</tbody>
</table>

- Both capture peaks around: 200 Hz (Pitch) and 1100 Hz (First formant)
- Standard conv response is noisy as compared to Sinc -> sign of overfitting
- Sinc does a better job at capturing overall spectral envelope

Constrained (Sinc) filters better than unconditioned kernels which overfit on our task and dataset
- Conditioning on phrase boundary in the presence of a shared Sinc layer boosts performance
- Significant complementary information in lexical features such as word embeddings and POS tags

References: