Incorporating Tone-related MLP Posteriors in the Feature Representation for Mandarin ASR

Xin Lei, Mei-Yuh Hwang, and Mari Ostendorf
SSL-LAB, Department of Electrical Engineering, University of Washington, Seattle, WA 98195

Overview

- **Motivation**
  - Tone has a crucial role in Mandarin speech
  - Tone depends on the F0 contour of a time span much longer than a frame; frame-level F0 and its derivatives only capture part of the tonal patterns

- **Our approach**
  - Use MLP with longer input window to extract tone-related posteriors as a better model of tone
  - Incorporate tone-related posteriors in featural vector

Tone-related MLP Posteriors in the Feature Representation for Mandarin ASR

- **Overall configuration**
- **Tone-related Feature Extraction**
  - Overall configuration
  - Consider two types of features:
    - Tone-predictive feature
    - Tone posterior feature
  - Tone posterior feature
    - MLP: Sparse in time and more impact
    - MLP: Short 3-frame sections of [PREV-F0] features
    - Tone posterior feature
    - MLP: Spreads. All posteriors plus one silence, one for laughter, one for all other information
    - PCS: Stops the cut-off point for

Speech Recognition Results

<table>
<thead>
<tr>
<th>Feature</th>
<th>Dimension</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP+F0</td>
<td>42</td>
<td>35.7%</td>
</tr>
<tr>
<td>PLP+SVM</td>
<td>42</td>
<td>36.8%</td>
</tr>
<tr>
<td>PLP+F0+(tone posterior)</td>
<td>45</td>
<td>36.3%</td>
</tr>
<tr>
<td>PLP+SVM+(tone posterior)</td>
<td>45</td>
<td>38.0%</td>
</tr>
<tr>
<td>PLP+PCA(tone posterior)</td>
<td>48</td>
<td>35.6%</td>
</tr>
</tbody>
</table>

Experiment Setup

- **CCTA**
  - Mandarin conversational telephone speech data collected by HKUST
  - Training set: train04, 251 phone calls, 57.7 hours of speech
  - Testing set: dev04, 24 calls totaling 2 hours, manually segmented

- **Cross-word system results**
  - Tone posterior features by themselves outperform F0
  - Tone posteriors and the plain F0 features complement each other, though with some overlap
  - More improvement is achieved by using tone-related posteriors than tone

Future Work

- Context-dependent tone ("Shi"-time) identification
- SVM boosting

References

- H. Hermansky et al., "Tandem connectionist feature extraction for conventional HMM systems", ICASSP 2000
- N. Morgan et al., "TRAPping conversational speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition", ICASSP 2004