A FRAME-SYNCHRONOUS PROSODIC DECODER FOR TEXT-INDEPENDENT DIALOG ACT RECOGNITION

Kornel Laskowski
Language Technologies Institute
Carnegie Mellon University, Pittsburgh PA, USA

Goal & Approach

Text-independent dialog act (DA) segmentation and classification in privacy-sensitive settings.

- Cannot compute ASR features → no words or word boundaries
  
- Anchor feature computation to unrecognized speech
- Construct an acoustic ASR-like decoder, whose states are:
  - Not phonemic sub-segments
  - But prosodic sub-phrases

Questions

1. Can the current 32 ms instantaneous prosodic feature vector and 8 ms frame step be extended to the decoder frame step of 100 ms, without negative impact on DA recognition performance?

2. Does feature-space combination with temporally adjacent features from the target participant improve DA recognition?

3. Does feature-space combination with temporally adjacent features from non-target participants improve DA recognition?

Techniques

Have:
- Hidden Markov model decoder
- 100 ms frame step, 100 ms frame size
- Specialized topology (split-and-merge talkspurts)
- 8 DA types
- 3 DA boundary states
- Instantaneous, frame-level prosodic features
- 8 ms frame step, 32 ms frame size
- Loudness (2): log-energy, delta-log-energy
- Speaking rate (2): cosine-Mel-energy, cosine-log-Mel-energy
- Voice quality (1): max-norm-autocorrelation
- Intonation (7): fundamental frequency variation (FFV) coefficients

Want to model prosodic context:
- Decoder frame step = feature computation frame step
- Consider adjacent target-speaker prosody
- Consider adjacent non-target-speaker prosody

Findings

1. A larger frame size (256 ms) improves DA classification only:
   - Mean DA \( F^\text{+} = +2.5\% \text{abs} \)
   - Boundary \( F^\text{−} = −0.2\% \text{abs} \)

2. Target-speaker prosodic context (1 s) improves DA recognition:
   - Mean DA \( F^\text{+} = +1.7\% \text{abs} \)
   - Boundary \( F^\text{−} = +4.0\% \text{abs} \)

3. Non-target-speaker prosodic context (1 s) improves DA segmentation only:
   - Mean DA \( F^\text{−} = −1.5\% \text{abs} \)
   - Boundary \( F^\text{+} = +1.8\% \text{abs} \)

Conclusions & Impact

1. Significant improvements in text-independent HMM-based DA recognition can be achieved with longer audio frames (256 ms vs 32 ms) and feature stacking (1 s).

II. Improvements observed despite the resulting much smaller amounts of training material.

III. Non-target-speaker prosody improves DA segmentation.

FUTURE WORK: Does non-target-speaker prosody improve over only non-target-speaker speech activity?

Experiments on ICSI Meeting Corpus, EvalSet (11 meetings)

<table>
<thead>
<tr>
<th>DA Types</th>
<th>Baseline (BL) 32 ms / 8 ms (12.5 frames)</th>
<th>Experiment 1 256 ms / 100 ms (1 frame)</th>
<th>Experiment 2 256 ms / 100 ms (11 frames)</th>
<th>Experiment 3 256 ms / 100 ms (w/ non-target)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>g-Opt c-Opt g-Opt c-Opt g-Opt c-Opt %rel, BL g-Opt c-Opt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean ( F )</td>
<td>31.5 33.7 35.5 36.2 36.6 37.9 42.9 35.8 36.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, floor holder</td>
<td>37.7 39.5 45.2 45.2 45.8 48.2 +22.0 **   45.5 47.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, hold</td>
<td>25.0 17.1 25.3 20.6 21.6 21.6 +26.1 *    16.5 12.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, floor grabber</td>
<td>7.2 7.2 8.2 8.2 10.1 10.1 +40.3 *       6.9 7.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, backchannel</td>
<td>48.0 64.6 57.5 64.4 59.9 64.2 −0.6     63.5 64.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, acknowledgment</td>
<td>19.0 20.9 25.2 25.2 25.3 25.3 +21.1 **  25.8 27.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, accept</td>
<td>9.5 8.9 17.5 17.5 21.9 22.5 +152.8 **  20.9 22.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, statement</td>
<td>85.8 91.8 88.8 91.9 88.5 92.0 +0.2 **  85.6 91.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, question</td>
<td>19.6 19.6 16.6 16.6 19.4 19.4 −1.0     18.5 18.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification error</td>
<td>25.9 16.6 21.6 15.9 21.7 15.8 −4.8     21.4 16.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DA Termination Types</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, completed</td>
<td>59.1 59.1 59.8 59.8 62.7 63.8 +8.0 **   61.0 63.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, interrupted</td>
<td>10.5 11.8 6.7 9.6 14.6 14.6 +23.7       28.4 28.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, abandoned</td>
<td>2.4 3.6 2.4 4.3 6.1 7.0 +94.4 **        4.6 5.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>any type, F</td>
<td>62.6 62.6 62.4 62.4 66.4 66.4 +6.1      66.9 68.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST error</td>
<td>66.5 63.0 66.5 63.0 66.1 58.5 −7.0      65.9 56.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>