**Perceived Prosody: Phonetic bases of Prominence and Boundaries**

Jennifer Cole       University of Illinois  
Louis Goldstein     USC and Haskins Laboratories  
Argyro Katsika      Yale and Haskins Laboratories  
Yoonsook Mo         University of Illinois  
Emily Nava          USC and Haskins Laboratories  
Mark Tiede          Haskins Laboratories

Supported by NSF IIS 07-03624
What are the phonetic bases of perceived prosody?

- Do untrained listeners agree on the location of perceived prosodic prominences and boundaries for the same utterance?
- What are the correlates of perceived prosody?
  - Acoustic
  - Articulatory
Naïve prosody transcription

Adapting a method from Buhmann et al. 2002

• The transcribers: many (74; 38) listeners with no training.
• The transcriptions: locate prominent words and boundaries, ignoring differences in tune and strength
• The analysis: evaluates variation in prosodic transcription across listeners and assigns probabilistic prosody labels
• Speed: Real-time comprehension, to diminish strategic analysis
• Reliability: measured using Fleiss’ Kappa statistic to calculate agreement rates for multiple (> 2) transcribers.
Method: Naïve prosody transcription

- Listeners hear speech excerpts from one of two corpora:
  - **Expt. 1: Buckeye** (Pitt et al. 2007)
    - spontaneous speech
    - 38 short excerpts
  - **Expt. 2: Xray µbeam** (Westbury, 1994)
    - 74 sentences produced by 6 talkers + null talker (reading only)

Listeners annotate a transcript of each excerpt that has no capitalization or punctuation:

- **Prominence**: word  word  word ....
- **Boundary**:  word | word  word ...

- Transcriptions pooled over listeners to obtain two population-wise prosody scores for each word:
  - **P-score & B-score**
Expt. 1: Acoustic correlates of perceived prosody in BUCKEYE

- **Materials**
  - 38 short excerpts (19 speakers x 2 excerpts)
  - 11-25 sec. each

- **Participants**
  - 37 in each of 2 experiments (replications)
  - Each participant performs prominence task for half the materials and boundary task for half the materials.
Results: Prosody Transcriptions

Probabilistic prosody scores by word

Prosody score

Prosody score

\[ p(P) \]

\[ p(B) \]
Fleiss’ multi-rater Kappa coefficient & z-statistic were used to assess agreement

Agreement between transcribers is highly significant for labeling of prominent words and boundaries

Agreement is higher for boundaries than for prominent words.
Results: Assessing inter-transcriber agreement

Pairwise transcriber agreement by Cohen’s Kappa statistic

Mean Kappa for Boundary = 0.582
Mean Kappa for Prominence = 0.392

Two transcribers listening to the same speaker perceive prosody differently!
Summary of transcription findings

- Naïve transcribers agree on the location of prosodic boundaries and prominent words in spontaneous speech at levels well above chance.

- Agreement is highest for prosodic phrase boundaries.

- Transcriptions reveal variation across listeners and across speakers.

- Variation across listeners yields two probabilistic prosody scores for each word.
All stressed vowels and vowels from word-final syllables were extracted using time-aligned phone and word transcriptions.

Acoustic measures:

- **prominence** stressed vowels
- **boundaries** word-final stressed vowels
- Duration: from the onset to the offset of the stressed vowel
- Overall intensity: mean RMS across vowel interval
- Spectral emphasis: bandpass filtered intensity in sub-bands: 0-500, 500-1000, 1000-2000, and 2000-4000 Hz
- F1 and F2 (prominence data only)

The extracted acoustic measures were z-normalized within each vowel phoneme, across speakers.
Acoustic correlates of perceived prosody: vowels (Mo, 2008)

Correlation of P-scores and B-scores with acoustic measures
Pearson’s $r$ for significant correlations in shaded cells; $p<0.05$; 1-tail

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>æ</th>
<th>ʌ</th>
<th>oʊ</th>
<th>aɪ</th>
<th>e</th>
<th>ɛ</th>
<th>ɪ</th>
<th>i</th>
<th>oʊ</th>
<th>u</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong> (total =1617)</td>
<td>81</td>
<td>129</td>
<td>211</td>
<td>58</td>
<td>28</td>
<td>140</td>
<td>187</td>
<td>66</td>
<td>114</td>
<td>209</td>
<td>156</td>
<td>103</td>
</tr>
<tr>
<td>Duration</td>
<td>0.3</td>
<td>0.2</td>
<td>0.22</td>
<td>0.49</td>
<td>0.42</td>
<td>0.24</td>
<td>0.3</td>
<td>0.24</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall INT</td>
<td>0.31</td>
<td>0.14</td>
<td>0.2</td>
<td>0.22</td>
<td>0.28</td>
<td>0.21</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SB (0-.5kHz)</td>
<td>0.16</td>
<td>0.13</td>
<td>0.24</td>
<td>0.19</td>
<td>0.14</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SB (.5-1)</td>
<td>0.35</td>
<td>0.16</td>
<td>0.25</td>
<td>0.22</td>
<td>0.29</td>
<td>0.34</td>
<td>0.211</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SB (1-2)</td>
<td>0.34</td>
<td>0.21</td>
<td>0.26</td>
<td>0.23</td>
<td>0.27</td>
<td>0.34</td>
<td>0.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SB (2-4)</td>
<td>0.23</td>
<td>0.15</td>
<td>0.15</td>
<td>0.17</td>
<td>0.19</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>æ</th>
<th>ʌ</th>
<th>oʊ</th>
<th>aɪ</th>
<th>e</th>
<th>ɛ</th>
<th>ɪ</th>
<th>i</th>
<th>oʊ</th>
<th>u</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>0.3</td>
<td>0.51</td>
<td>0.63</td>
<td>0.45</td>
<td>0.8</td>
<td>0.52</td>
<td>0.62</td>
<td>0.54</td>
<td>0.61</td>
<td>0.48</td>
<td>0.43</td>
<td>0.66</td>
</tr>
<tr>
<td>Overall INT</td>
<td>-0.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Correlation of P-scores and B-scores with acoustic measures (Mo, 2008)
Acoustic correlates of perceived prosody: vowels

<table>
<thead>
<tr>
<th></th>
<th>α</th>
<th>æ</th>
<th>ʌ</th>
<th>ɔ</th>
<th>̄a</th>
<th>ə</th>
<th>ɪ</th>
<th>ɛ</th>
<th>ɝ</th>
<th>ɪ</th>
<th>ɪ</th>
<th>ʊ</th>
<th>ʊ</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>173</td>
<td>290</td>
<td>407</td>
<td>121</td>
<td>52</td>
<td>309</td>
<td>463</td>
<td>122</td>
<td>214</td>
<td>475</td>
<td>306</td>
<td>211</td>
<td>72</td>
<td>183</td>
</tr>
<tr>
<td><strong>F1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>end</td>
<td>0.18</td>
<td>0.23</td>
<td>0.208</td>
<td>0.19</td>
<td>0.34</td>
<td>0.13</td>
<td>0.14</td>
<td>0.29</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beg.</td>
<td>-0.19</td>
<td>-0.16</td>
<td>beg.</td>
<td>-0.34</td>
<td>end</td>
<td>-0.13</td>
<td>-0.10</td>
<td>end</td>
<td>0.25</td>
<td>beg.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>end</td>
<td>-0.19</td>
<td>-0.16</td>
<td>beg.</td>
<td>-0.34</td>
<td>end</td>
<td>-0.13</td>
<td>-0.10</td>
<td>end</td>
<td>0.25</td>
<td>beg.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pearson’s $r$ for significant correlations in shaded cells; $p<0.05$; 1-tail. For diphthongs formants are extracted from beginning (10%) and end (90%) of vowel interval.

Note: data set is larger than that used for energy and duration analysis by including words from a second transcription study (Mo, in prep.)
Summary: Acoustic correlates

- **Prominence:** there is no single acoustic correlate of perceived prominence for all 14 vowels

  +++ **Duration** is the strongest correlate, and is significant for 9/14 vowels

  ++ **Formants** (F1, F2) are weaker correlates for 9/14 vowels, though not always the same vowels that show duration correlate

  + **Spectral emphasis** in the .5-1 kHz band is a weak correlate for 8/14 vowels.

- **Boundary:**

  ++++ **Duration** is a reliable correlate of perceived boundary for all vowels.

  A longer vowel in the word-final syllable predicts a greater likelihood of boundary perception.
Expt. 2: Gestural Kinematics of Perceived Prosody in µBEAM

• In each trial listeners hear and see 5 sentences from one µ-beam talker at a time, strung together with no punctuation.

• label either boundaries (‘chunking’) by clicking between words, or prominent words by clicking on the prominent word.

• 38 listeners made both boundary and prominence judgements on 6 talkers plus a reading-only (null talker) condition for a total of 74 sentences.

• Boundary and prominence data collected during separate sessions.
Results: Assessing inter-transcriber agreement

- Agreement among all transcribers is highly reliable.
  - **Prominence**: Fleiss’ Kappa = 0.23, p<.0001
  - **Boundary**: Fleiss’ Kappa = 0.31, p<.000
- As with Buckeye study, agreement for boundaries is higher than prominences
Selection of prosodic events for kinematic analysis

- Listeners transcribed the same utterances produced by 6 different talkers and from transcript for null talker.

- 30 sentences exhibiting prosodic events of interest were selected, words for which talkers differed in either P-scores or B-scores.

- Differences in transcription must be due to material in the signal.
Talkers

Boundary

Prominence

Task 78 Sentence 3 (Index 50)
Selection of consonant gestures for kinematic analysis

- **Time span:** from the stressed syllable preceding the prosodic event of interest to the stressed syllable following the event.

- In time span of interest, constriction maxima for each of the **consonant** gestures were identified (algorithmically but hand-guided). Maxima were identified in the appropriate markers: (Labials--LL, Coronals--T1, Dorsals--T4).

- Closing (onset) and Release (offset) movements of gestures associated with the labeled constriction maxima during time span of interest were obtained algorithmically using velocity criteria.
Group Labeling Task:

Example gesture, showing labeling criteria

JW24_TP012
/m/ in “walk more”

1) Gestural Onset
2) Peak Velocity (onset)
3) Nuclear Onset
4) Maximum Constriction
5) Nuclear Offset
6) Peak Velocity (offset)
7) Gestural Offset

Onsets delimited using 20% of local min:max range criterion
Kinematic Measures

- Kinematic measures (following Perkell al. 2002) computed over the gestural CLO and REL.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUR</td>
<td>movement duration (<em>gestural onset : offset; ms</em>)</td>
</tr>
<tr>
<td>DIST</td>
<td>distance along the x-y path (<em>path integral; mm</em>)</td>
</tr>
<tr>
<td>PEAK</td>
<td>peak speed (<em>maximum speed reached during mvt; cm/sec</em>)</td>
</tr>
<tr>
<td>STIFF</td>
<td>peak speed / distance (<em>relative index of movement “stiffness”</em>)</td>
</tr>
<tr>
<td>C</td>
<td>peak speed / average speed (<em>average speed = DIST / DUR</em>)</td>
</tr>
<tr>
<td>NPEAKS</td>
<td>number of acceleration peaks (<em>index of movement smoothness</em>)</td>
</tr>
<tr>
<td>SYM</td>
<td>speed symmetry (<em>ratio of acceleration duration / DUR</em>)</td>
</tr>
<tr>
<td>CURV</td>
<td>movement curvature (<em>ratio of DIST / straight-line distance</em>)</td>
</tr>
</tbody>
</table>
Relation of consonant gesture

Kinematics and prosody

- Compare kinematic measures for consonant gestures at prosodic events of interest:
  - following boundary vs. no following boundary
  - prominent vs. not prominent
- 25% of listener judgements was used as the threshold for defining a word as a boundary or prominence location.
Example sentence: Chunking

![Graph showing task 78 sentence 3 (Index 50)]

### Percent of M11 : KV1Y duration

<table>
<thead>
<tr>
<th>LK05 (T4)</th>
<th>JW12 (VC/CV)</th>
<th>JW14 (VC/CV)</th>
<th>JW15 (VC/CV)</th>
<th>JW16 (VC/CV)</th>
<th>JW21 (VC/CV)</th>
<th>JW24 (VC/CV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUR</td>
<td>178.5</td>
<td>164.8</td>
<td>171.7</td>
<td>185.4</td>
<td>171.6</td>
<td>212.8</td>
</tr>
<tr>
<td>DIST</td>
<td>12.5</td>
<td>14.3</td>
<td>14.6</td>
<td>10.8</td>
<td>19.1</td>
<td>17.7</td>
</tr>
<tr>
<td>STIFF</td>
<td>0.916</td>
<td>0.902</td>
<td>1.018</td>
<td>1.382</td>
<td>1.023</td>
<td>0.676</td>
</tr>
<tr>
<td>C</td>
<td>1.636</td>
<td>1.486</td>
<td>1.747</td>
<td>2.563</td>
<td>1.756</td>
<td>1.439</td>
</tr>
<tr>
<td>NPEAKS</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>SYM</td>
<td>0.192</td>
<td>0.750</td>
<td>0.320</td>
<td>0.852</td>
<td>0.360</td>
<td>0.710</td>
</tr>
<tr>
<td>CURV</td>
<td>1.254</td>
<td>1.248</td>
<td>1.096</td>
<td>1.096</td>
<td>1.070</td>
<td>2.125</td>
</tr>
</tbody>
</table>

no boundary

boundary
Each bar represents token from one talker
Boundary: Means across talkers

- **DUR**
  - CLO
  - REL

- **PEAK**
  - CLO
  - REL

- **STIFF**
  - CLO
  - REL

- **NPEAKS**
  - CLO
  - REL
Boundary: Means across talkers
Summary: Boundary

- Lengthening/slowing of C gesture (e.g., Byrd, 2000) immediately preceding perceived boundary is robust enough that it can be observed across subjects with no normalization.
  - Too few tokens for statistical analysis.
- Lengthening/slowing is observable mainly in gesture release in DUR, STIFF, PEAK.
- No spatial effect (DIST) is observed.
- No systematic temporal effects are observed in the initial C of a CVC word that immediately precedes a perceived boundary.
Prominence: Means across talkers

**CLO**

- **DUR**
  - Prominence
  - No prominence

- **STIFF**
  - NPEAKS

**REL**

- **PEAK**
  - CLO
  - REL

- **DUR**
  - CLO
  - REL

- **NPEAKS**
  - CLO
  - REL

Legend:
- **Red** prominence
- **Blue** no prominence
Prominence: Means across talkers

- **DIST**: Comparison of prominence and no prominence across different distances.
- **SYM**: Similar comparison for symmetry.
- **CURV**: Comparison for curvature.
- **C**: Comparison for another measure.

Key:
- Red: Prominence
- Blue: No prominence

Notes:
- The graphs illustrate the means across talkers for each category.
Summary: Prominence

- Temporal effects are observable on prominent word: DUR, STIFF, SYM.
- Locus of effect is in constriction formation (CLO), not release.
- Temporal effects do not include Peak Velocity.
Overall Summary: Naïve prosodic transcription

- efficient
- reliable
- valid: correlated with acoustic (vowels) and articulatory (consonant) variables


