F0 max and formants (F1, F2) as perceptual cues for naïve listeners’ prominence perception

Yoonsook Mo
Department of Linguistics
University of Illinois at Urbana-Champaign
Defining prosody

- Speech utterances are composed of hierarchically structured phonological phrases.
  - A prosodic **boundary** marks the phonological phrase juncture and serves to demarcate “chunks” of words.
- Within each utterance, some words or phrases are more **prominent** than others.
  - Prosodic prominence “highlights” a word or a phrase and conveys its status as focused or discourse-new.
- In this paper, in particular, prominence is of interest.
This talk focuses on the phonetic correlates of **prosodic prominence**, and is part of my larger study of phonetic correlates of prosodic structure in production and perception.
What is the phonetic expression of prominence?

- **Phonetic implementation**
  
  Speakers encode prosodic structure through the modulation of phonetic parameters.

- **Acoustic correlates of prominence**
  
  - **Fundamental frequency** (F0)
  - **Duration** (Fry, 1955 and 1958; Turk and Sawusch, 1996)
  - **Intensity** (Fry, 1955 and 1958; Kochanski, 2005)
  - **Sub-band intensities** (Sluijter and van Heuven, 1996; Heldner, 2001 and 2003)
  - **Formants**
  - **Spectral tilt** (Fant et al., 2000; Sluijter and van Heuven, 1996)
In my study

I investigate the phonetic encoding of prominence

- **14 vowels** in American English
- in everyday *conversational speech*
- from 38 *ordinary speakers* of American English
- by about 100 *untrained, ordinary listeners*

→ “Prominence” as judged by ordinary listeners, based only on auditory impression. No visual inspection of speech display.
In my study (continued)

- In other work I show duration, intensity and sub-band intensity measures to be important correlates of prominence. (Mo, 2008a and b)

- What effect, if any, does prominence have on F0 and on vowel formants?
  - Intonation
  - Hyper- vs. hypo- articulation
Previous studies

- Fundamental frequency (F0)
  - Height and shape of F0 contours are shown to be as a major correlates of prominence
    - Stressed vs. Unstressed (Lieberman, 1969; Cooper et al., 1985 among others)
    - Pitch accents (Gussenhoven et al., 1997; Hermes and Rump, 1994; Pirrehumbert, 1979; Terken, 1991 and 1994)

- Still controversial
  - Perception of focal status has not changed by gradual addition of F0 rise on non-focused words (Heldner and Strangert, 1997)
  - F0 plays a minor role in the automatic classification of pitch accent (Kochanski, 2005)
Previous studies (continued)

- Vowel quality
  - Acoustic studies (Sluijter and van Heuven, 1996; van Bergem, 1993)
  - Articulatory studies (Beckman et al., 1992; De Jong, 1995; Erickson, 2002; Cho, 2005)
Underlying mechanism

- **Sonority expansion** (Beckman et al., 1992)
  - Under accent, articulators move to increase “sonority”
  - More open vocal-tract

- **Hyperarticulation** (De Jong, 1995; Erickson, 2002)
  - Under accent, phonetic space of phonemic contrast expands
  - Feature distinctiveness is enhanced

- **Combination of sonority expansion and hyperarticulation** (Cho, 2005)
  - Under accent, more open
  - In front/ back dimension, more front or more back
Objectives of this study

- To investigate the phonetic properties that cue prominence in conversational speech by ordinary listeners
  - How does fundamental frequency vary?
  - How are formant structures modified?
- To evaluate which underlying mechanism better describes the phenomenon of prominence, as judged by listeners
A speaker marks a word as prosodically prominent in accordance with its pragmatic value (e.g., focused), position in the phrase, and other factors.

A speaker implements a prominent word with an F0 excursion, and with “enhanced” speech gestures that are longer, larger, or both. These effects are strongest on the lexically stressed syllable.

Listeners perceive a word as prominent based on acoustic evidence of the speaker’s “enhanced” speech gesture.

Therefore, words perceived as prominent will have stressed syllables that are acoustically “enriched”.

- Higher F0
- Higher F1 and more peripheral F2
Experimental Hypotheses

- **F0**
  - Vowels in words perceived as prominent will have higher F0 peaks.

- **Vowel quality**
  - Hyper-articulation: vowel formants will indicate more peripheral place of articulation, because prominence enhances phonemic contrast
    - High vowel: lower F1
    - Low vowel: higher F1
    - Front vowel: higher F2
    - Back vowel: lower F2
  
OR

  - Sonority Expansion: higher F1 regardless of vowel height
Methods
Materials and Participants

- **Materials**
  - 54 speech excerpts from 38 speakers in the Buckeye corpus of spontaneous speech of American English.
  - Sound files are equalized in their loudness level.
  - Length: 11 to 58 seconds.
  - Sound file presentation and its corresponding word transcripts

- **Participants**
  - 97 listeners from undergraduate Linguistics courses
  - Naïve in terms of phonetics and phonology of prosody transcription.
Transcription Task (continued)

- Simple definitions of prominence and boundary.
  - **Prominence** which “highlights” a word or a phrase and makes them “stand out” from other non-prominent words
  - **Boundary** which marks “a chunk of speech” and can help listeners interpret long stretches of continuous speech

- Playing sound files twice at their own pace.

- While listening, they marked “prominent” words and words at “juncture” using the following transcription marks:
  
  Prominence       Boundary
  word   word   word
  word | word   word
Results of transcription task

- Transcriptions pooled over listeners; each word is assigned a probabilistic P(rominence) and B(oundary) score ranged 0-1.

Speaker 26
### Reliability of transcription

- Fleiss' kappa inter-transcriber agreement scores and their corresponding z-scores

<table>
<thead>
<tr>
<th></th>
<th>Exp.1</th>
<th></th>
<th></th>
<th>Exp.2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Run 1</td>
<td>Run 2</td>
<td>Run 1</td>
<td></td>
<td>Run 2</td>
<td>Run 1</td>
</tr>
<tr>
<td></td>
<td>Grp 1</td>
<td>Grp 2</td>
<td>Grp 1</td>
<td>Grp 2</td>
<td>Grp 1</td>
<td>Grp 2</td>
</tr>
<tr>
<td>prominence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kappa</td>
<td>0.373</td>
<td>0.421</td>
<td>0.394</td>
<td>0.407</td>
<td>0.356</td>
<td>0.400</td>
</tr>
<tr>
<td>z</td>
<td>19.43</td>
<td>20.48</td>
<td>18.15</td>
<td>18.31</td>
<td>15.31</td>
<td>19.56</td>
</tr>
<tr>
<td>boundary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kappa</td>
<td>0.612</td>
<td>0.544</td>
<td>0.621</td>
<td>0.575</td>
<td>0.560</td>
<td>0.567</td>
</tr>
<tr>
<td>z</td>
<td>27.62</td>
<td>21.87</td>
<td>25.05</td>
<td>26.22</td>
<td>24.89</td>
<td>22.49</td>
</tr>
</tbody>
</table>

- Fleiss' statistic shows that transcribers agreement is significantly above chance levels at p<.001
- Untrained listeners’ transcription is reliable.
Acoustic measurements

- F0
  - Measured in 1ms interval
  - Smoothed by median-filtering with a 13 point window only at CV junctures
  - Interpolating F0 contours

- Formants
  - Steady state formants (F1 and F2) measured
    - Monophthong: at vowel midpoint
    - Diphthong: at 10% and 90% of the vowel
F0, F1 and F2 are extracted from the stressed vowels of each word in order to hold stress constant.

<table>
<thead>
<tr>
<th>Vowels</th>
<th>a</th>
<th>æ</th>
<th>õ</th>
<th>õ</th>
<th>õ</th>
<th>e</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>173</td>
<td>290</td>
<td>407</td>
<td>121</td>
<td>52</td>
<td>309</td>
<td>463</td>
</tr>
<tr>
<td>ι</td>
<td>121</td>
<td>214</td>
<td>475</td>
<td>306</td>
<td>211</td>
<td>72</td>
<td>183</td>
</tr>
<tr>
<td>Total</td>
<td>122</td>
<td>214</td>
<td>475</td>
<td>306</td>
<td>211</td>
<td>72</td>
<td>3398</td>
</tr>
</tbody>
</table>

Then the extracted acoustic measures are normalized.

\[ z = \frac{x - \bar{x}}{s} \]

- F0 with a 400ms analysis window
- Formants in the total phone space
Results & Discussion
Hypothesis: The more prominent a word is, the higher F0 max will be.

Pearson’s bivariate correlation analysis b/w F0 max and Pscores

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>ι</th>
<th>æ</th>
<th>ʌ</th>
<th>ɑ</th>
<th>α</th>
<th>ɪ</th>
<th>ʊ</th>
<th>ɪɛ</th>
<th>ɝ</th>
<th>e</th>
<th>i</th>
<th>ʊ</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>3398</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>
</tr>
<tr>
<td>F1</td>
<td>NA</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>F2</td>
<td>NA</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>F0 max</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>

The results support the hypothesis.
- Pscores are **positively correlated** with F0 max for the majority of vowels.
- Overall, words perceived as prominent have **higher F0 max**.
I like this slide! very clear!

Jennifer Cole , 2/8/2009
Vowel Quality

- Pearson’s bivariate correlation analysis b/w formants and Pscores

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>ɑ</th>
<th>æ</th>
<th>ʌ</th>
<th>ɔ</th>
<th>a</th>
<th>ɪ</th>
<th>ɛ</th>
<th>ɝ</th>
<th>e</th>
<th>ɪ</th>
<th>ʊ</th>
<th>a</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>3398</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>
</tr>
<tr>
<td>F1</td>
<td>NA</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>F2</td>
<td>NA</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>F0 max</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>

BECKMAN INSTITUTE
AT THE UNIVERSITY OF ILLINOIS

® 1867
Summary of the results

- **F1**
  - Pscores are positively correlated with F1 regardless of vowel height in all the monophthongs except the low back vowel, ə.
  - Pscores are negatively correlated with F1 of the glide part of two diphthongs, eɪ and aʊ.

- **F2**
  - Pscores are positively correlated with F2 of the front high vowel, i.
  - Pscores are negatively correlated with F2 of many central and back vowels and the nucleus part of two diphthongs, aɪ and oʊ.
I think you should just read out this summary while the audience views the table of results from the preceding slide. If you have a handout, you can include this slide on the handout, but you don’t have to show it. The next slide really delivers this information in a more digestible fashion!
Monophthongs in the vowel space
Diphthongs in the vowel space

Front

High

i

a

Low

Back

u

o
I changed this line segment to an arrow, showing the direction of movement of the diphthong. You should make the arrow head larger, and make the same change for the other diphthongs.

Jennifer Cole, 2/8/2009
Diphthongs in the vowel space

Front

High

Low

Back

I

u

a
Evaluation of hypotheses

- Hyperarticulation
  The stressed vowels perceived as prominent are peripheral in the vowel space.
  - Partially supported: front/ back dimension
  - The front vowel i, the nucleus of a₁, and the glide of e₁ are more front when perceived as prominent.
  - The vowels other than those listed above are more back when perceived as prominent.
Evaluation of hypotheses

- Sonority Expansion
  Regardless of vowel height, the stressed vowel in a prominent word is more open.
  - Supported
  - Vowels have more open vocal tract except the low vowel \( \alpha \) and diphthongs when perceived as prominent.
Evaluation of hypotheses

- The combination of Hypothesis 2 and 3 best account for the relation between formants and prosodic prominence.
  - In front/back dimension, peripheral vowel formants (F2) suggest that vowels are hyperarticulated under prominence.
  - In high/low dimension, higher vowel formants (F1) of non-low vowels suggest that sonority expands under prominence.
$R^2$ from stepwise regression analyses

$R^2$ (%)

Vowels
Results of regression models

- Regarding the results from stepwise regression analyses,
  - only a small portion of the variation in listeners’ response to prominence (ranged from 3.3% for /æ/ - 23.2% for /aɪ/) can be explained on the basis of those measures
  - Not a single acoustic measure is included in the regression model across all vowels
  - Not a unified regression pattern accounts for the variation of prominence
Conclusion

- In this study, prominence in conversational speech produced by ordinary speakers is judged by untrained ordinary listeners.
  - This transcription task approximates how listeners hear prosody in everyday conversation.
- Listeners’ perception of prominence is guided by the modulation of the patterns of F0, F1 and F2.
Conclusion

- No single acoustic measure and no single pattern of prominence marking across vowels
- Therefore, other acoustic measures as well as other factors that affect the acoustic properties of speech should also be examined.
  - Duration and intensities (Mo, 2008a and b)
  - Syntactic category information (Cole, Mo & Baek., 2008)
  - Word repetition and frequency (Cole, Mo & Hasegawa-Johnson, 2008)
Thank you very much!

**Acknowledgements**

This research is supported by NSF grants IIS 07-03624 and IIS 04-14117 to Jennifer Cole and Mark Hasegawa-Johnson.

Jennifer Cole, Linguistics, UIUC
Mark Hasegawa-Johnson, ECE, UIUC
Prosody-ASR group members
Two separate experiments are comprised of three runs.

Experiments are different in terms of the lengths of speech excerpts.
- Exp. 1: 11-22 sec
- Exp. 2: 31-58 sec

<table>
<thead>
<tr>
<th></th>
<th>Exp. 1</th>
<th></th>
<th>Exp. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Run 1</td>
<td>Run 2</td>
<td>Run 1</td>
</tr>
<tr>
<td>Grp 1</td>
<td>15</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>Grp 2</td>
<td>16</td>
<td>23</td>
<td>12</td>
</tr>
</tbody>
</table>

N of transcribers
Results of two-way ANOVA tests

- P-scores in two experiments are not statistically different. (F=3.028, p=.082)
- P-scores of 14 vowels are different from one another (F=7.509, p<.001)