Modeling perceived prosody: Speaker-dependent vs. speaker-independent models
Yoonsook Mo, Jennifer Cole
Beckman Institute, Department of Linguistics, University of Illinois at Urbana-Champaign

Given the multiplicity of acoustic cues to prosody, Q1) In what ways do speakers vary in the phonetic implementation of prosody?

Q2) How are listeners affected by speaker variability in their interpretation of prosody?

Introduction
- In everyday conversation, speakers communicate pragmatic and discourse meaning through the prosodic form assigned to an utterance.
- Speakers modulate multiple acoustic parameters to signal prosodic structure.
- Listeners must attend to the acoustic variation in the speaker’s speech signal to fully recover the meaning intended by the speaker.

Goals of this study:
1. To gauge the extent of speaker variability in the acoustic implementation of prosodic structure and to identify common acoustic patterns that signal prosody across speakers in spontaneous conversational speech.
2. To identify the acoustic cues for prosody for each individual speaker, and their contribution to prosody perception.
3. To create speaker-dependent statistical models of the acoustic cues to prosody as perceived by “ordinary” listeners.

Prior studies employing controlled “laboratory” speech (e.g. simple sentences, read speech) show acoustic effects of prosody in many languages, and also show that native listeners respond to acoustic prosodic cues in interpreting utterance meaning (e.g., Kim et al., 2018).

Few studies directly examine speaker-dependent acoustic variation in prosody production (e.g. Knabe & Jung-Beeman, 2006).

Rapid Prosody Transcription (RPT)
1. 54 short excerpts (~11 – 56 sec) were selected from the Buckeye corpus of American English spontaneous speech (Piantadosi, 2017).
2. Orthographic transcripts were produced for each sound file, with no punctuation or capitalization.
3. 97 transcribers: UIUC undergraduates, untrained and unfamiliar with the phonetics and phonology of prosody.
4. After being given simple instructions and definitions of prominence and boundary, four groups of 12-20 subjects marked the locations of prosodic prominence and boundaries on the printed transcripts in a separate task in real time, based only on auditory impression (no visual speech display).
5. Collecting the transcription data from all subjects, each word in the set of excerpts was assigned probabilistic P prominences and B boundary scores depending on the number of transcribers who marked the word as prominent or as followed by a juncture.

Reliability tests
- Fleiss’ multi-rater’s kappa agreement scores are much above chance level: ordinary, untrained listeners’ perception of both prominence and boundary is systematic and reliable.
- Kappa scores for boundary are systematically higher than for prominence perception:
ordinary, untrained listeners’ perception of boundary is more consistent than their perception of prominence.

Regression models of perceived prosody as predicted by acoustic measures
- Acoustic measures are extracted from the lexically stressed vowels for prominence and the word-final lexically stressed vowels for boundary.
- Stress intensified according to the ISLE dictionary (Hasegawa-Johnson, 2009).
- Normalized acoustic measures:
  - temporal measures (Vdur and pause)
  - intensity measures (overall and subband intensities in 0-0.5, 0.5-1.0, 1.0-2.0, and 2.0-4.0 kHz)
  - F0 measures (local F0 maximum and F0 at the right edge)
  - formant measures (F1 and F2)

Speaker-dependent prosody models
- Speaker-independent models show duration and F0 as the strongest predictors of perceived prominence and 21 - 56% of variation in boundary perception.
- Speaker-dependent models of acoustic cues to prosody better account for variation in listeners’ response to prominence and 18 - 87% of variation in boundary perception.

Discussion and Conclusion
1. Speakers vary in their acoustic encoding of prosody: in terms of both the kinds of acoustic cues to prosody and their contribution to prosody perception.
2. Speaker-dependent models of acoustic cues to prosody better account for variation in listeners’ perception of prominence and boundary.

References

Acknowledgments
This research is supported by NSF grant 152-033-044 to Jennifer Cole and Mark Hasegawa-Johnson. Special thanks to the Prosody-ASR group members for their comments.

Variation by speakers
- Speaker-independent regression models explain only about 3 - 23% of variation for perceived prominence and 21 - 56% of variation in boundary perception.
- Speaker-dependent acoustic models account for 12 - 54% of variation in listeners’ response to prominence and 18 - 87% of variation in boundary perception.
- The speaker-dependent model shows duration and F0 as the strongest predictors of perceived prominence, with intensity and formants as weak predictors. But the speaker-dependent model shows that speakers vary in the sets of cues used to encode prominence. Some speakers rely more on F0 and duration as cues, but other speakers rely heavily or exclusively on intensity.
- The temporal measures of vowel duration and silent pause play a primary role in cueing boundaries across speakers.