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 (Cho, 2005).
- Few studies directly examine speaker-dependent acoustic variation in prosody production (Redi & Shattuck-Hufnagel , 2001).

Rapid Prosody Transcription (RPT)

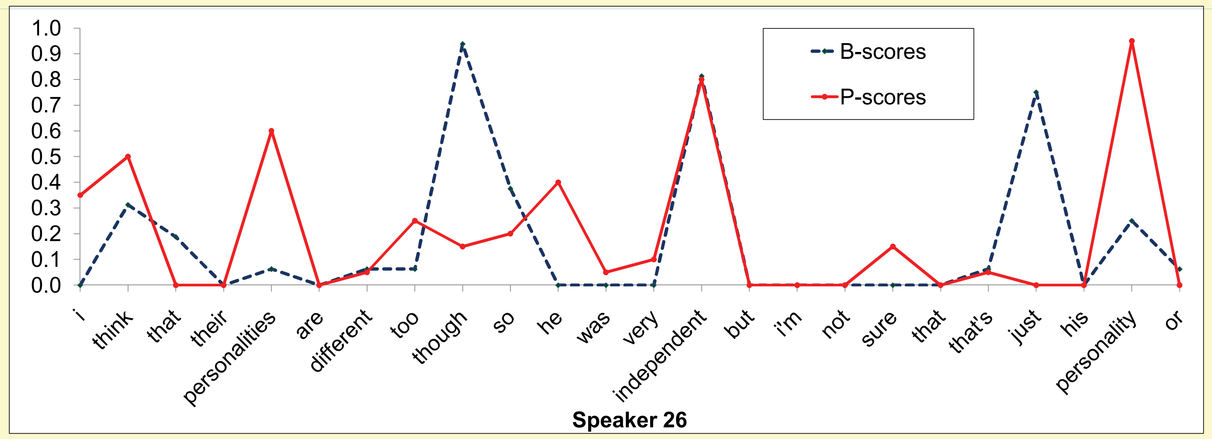
1. 54 short excerpts (~11 – 58 sec) were selected from the Buckeye corpus of American English spontaneous speech (Pitt et al., 2007).
2. Orthographic transcripts were produced for each sound file, with no punctuation or capitalization.
3. 97 transcribers: UIUC undergraduates, untrained and unfamiliar 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(rominence)- and B(oundary)-scores depending on the number of transcribers who marked the word as prominent or as followed by a juncture.

References
1. Choi, J.-Y., Hasegawa-Johnson, M., and Cole, J., 2005. Finding intonational boundaries using acoustic cues related to the voice source, *Journal of the Acoustical Society of America*, 118 (4), 1-9.
2. Cole, J., Kim, H., Choi, H., and Hasegawa-Johnson, M. 2007. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News, Ohio State speech, Journal of Phonetics, 35, 180-209
3. Cho, T., 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a, i/ in English. *Journal of the Acoustical Society of America*. 117 (6). 3867-3878.
4. Hasegawa-Johnson, M. and Fleck, M., 2007. ISLE Dictionary version 0.2.0, downloaded Oct. 19, 2007 from http://www.isle.uiuc.edu/dict/index.html
5. Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E., 2007. *Buckeye Corpus of Conversational Speech* (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of PsychologyUniversity (Distributor).
6. Redi, L. and Shattuck-Hufnagel, S., 2001. Variation in the realization of glottalization in normal speakers, *Journal of Phonetics*, 29, 407-429.
7. Yoon, T-J., Cole, J, and Hasegawa-Johnson, M., 2007. On the edge: Acoustic cues to layered prosodic domains. In proceedings of ICPhS (Saarbruken, Germany), 1017-1020.

Acknowledgements
This research is supported by NSF grant IIS 07-03624 to Jennifer Cole and Mark Hasegawa-Johnson. Special thanks to the Prosody-ASR group members for their comments

Results

Distribution of P- and B-scores



Reliability tests

z=2.33, α=0.01		Exp.1		Exp. 2		Exp. 3
		Grp 1	Grp 2	Grp 1	Grp 2	
prominence	Kappa	0.377	0.399	0.346	0.448	0.377
	z	**25.2**	**22.7**	**21.5**	**33.4**	**32.8**
boundary	Kappa	0.601	0.587	0.532	0.640	0.580
	z	**33.0**	**31.4**	**26.8**	**37.3**	**44.3**

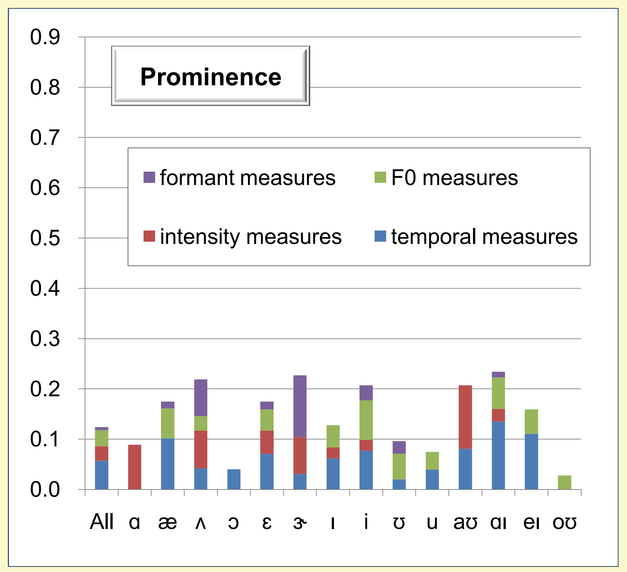
-Fleiss' multi-rater's kappa agreement scores are much above chance level: *ordinary, untrained listeners' perception of both prominence and boundary is* ***systematic and reliable****.*

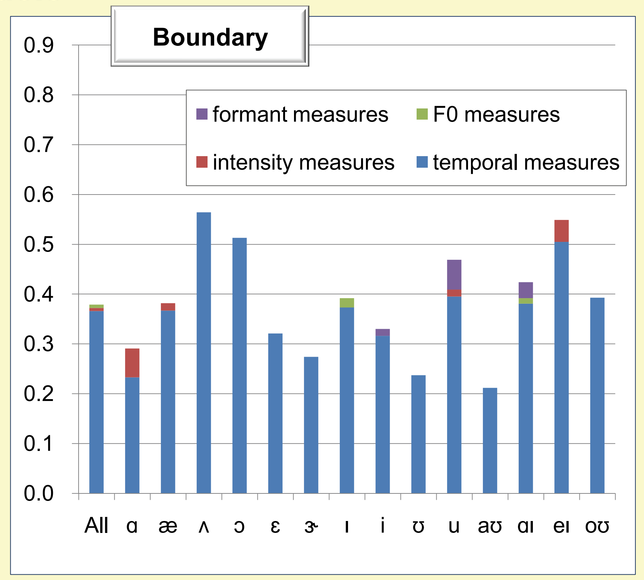
-Kappa scores for boundary are systematically higher than for prominence perception: *ordinary, untrained listeners' perception of boundary is more consistent than their perception of prominence.*

Regression models of perceived prosody as predicted by acoustic measures

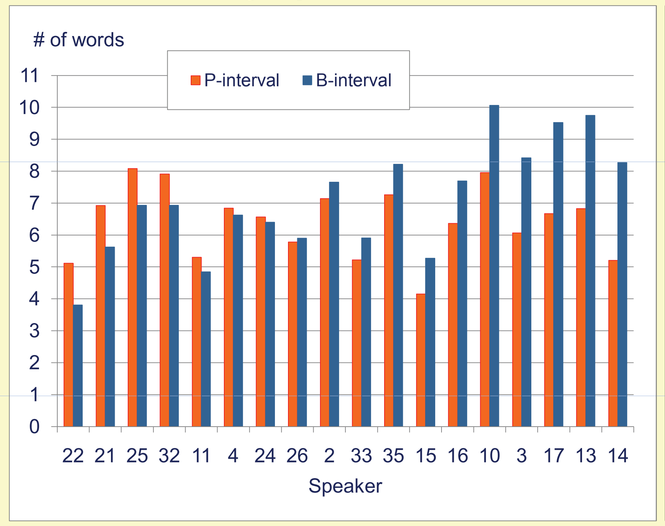
- Acoustic measures are extracted from the lexically stressed vowels for prominence and the word final lexically stressed vowels for boundary.
- Stress indentified according to the ISLE dictionary (Hasegawa-Johnson and Fleck, 2007)
- Normalized acoustic measures
  - temporal measures (Vdur and pause)
  - 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)

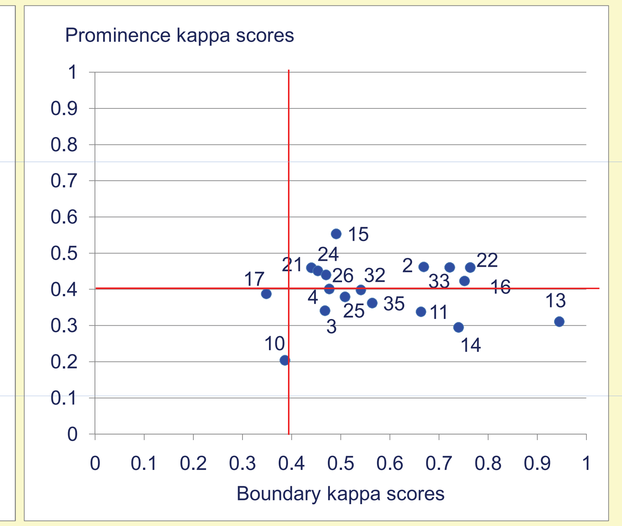
Speaker-independent prosody models



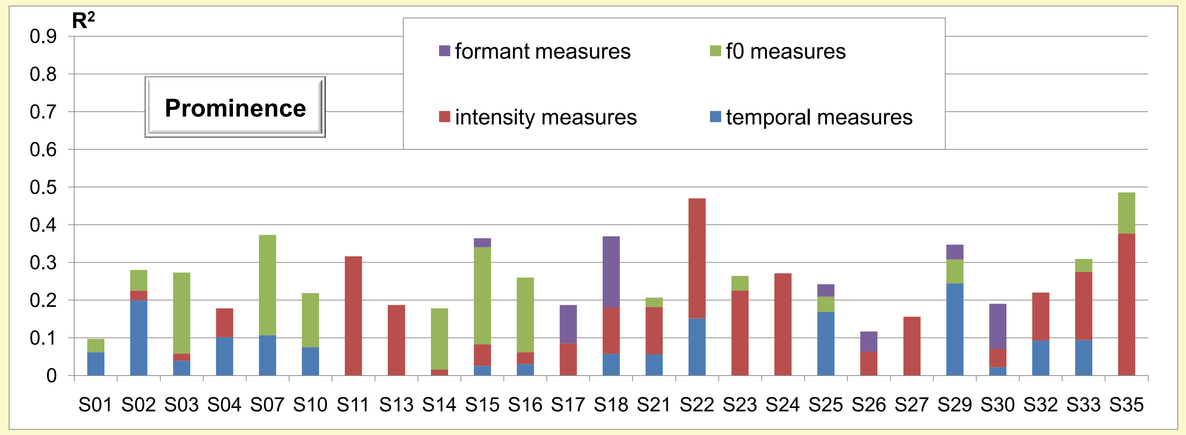


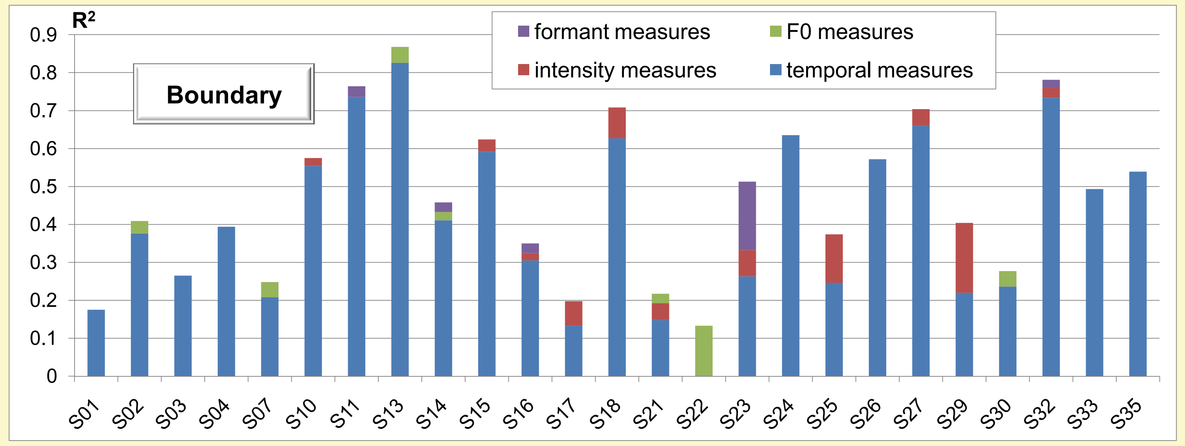
Variation by speakers





Speaker-dependent prosody models





- Speaker-independent regression models explain only about 3 - 23% of variation for perceived prominence and 21 - 56% of variation in boundary perception.
- Speaker-dependent regression models account for 12 - 54% of variation in listeners' response to prominence and 18 - 87% of variation in boundary perception.
- The speaker-independent 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 set 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.

Discussion and Conclusion

1. Speakers vary in their acoustic encoding of prosody: in terms of both the kinds of acoustic cues and the contribution of each acoustic cue to listeners' perception.
2. Speaker-dependent models of acoustic cues to prosody better account for variation in perceived prosody

Viewed from the perspective of the listener…

Speakers vary in their assignment of prosodic features to a spoken utterance and there is not a unique phonetic expression of prosodic structure across speakers. In particular, prominence is not uniquely cued by F0 or any other single acoustic measure. This finding points to the importance of considering a range of acoustic measures in investigating prosody in spontaneous speech.