Uniformity and Variability among Speakers in the Acoustic Encoding of Prosody in Spontaneous Speech

Yoonsook Mo, Jennifer Cole
Beckman Institute, Department of Linguistics, University of Illinois at Urbana-Champaign

Given the multiplicity of acoustic cues to prosody,

Q1) What is the variation and uniformity across speakers in the phonetic implementation of prosody?

Q2) What are the underlying production mechanisms of prosody?

Q3) How are listeners affected by speaker variability in their interpretation of prosody?

Background

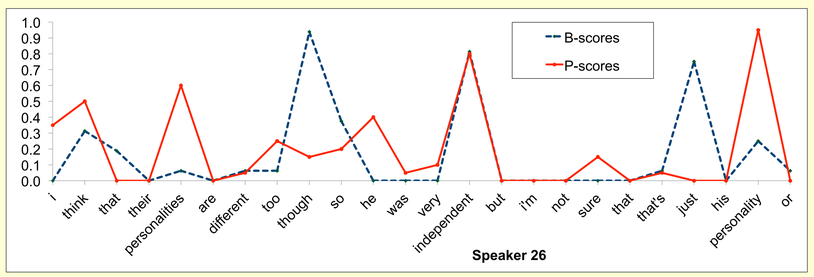
- Prior studies employing controlled "laboratory" speech (e.g. simple sentences, read speech) show acoustic effects of prosody in many languages, and also show that native listeners respond to acoustic prosodic cues in interpreting utterance meaning (Cho, 2005; Kochanski et al., 2005;Turk and Sawusch, 1996).
- Few studies attempt (1) to approximate prosody production and perception in everyday speech communication (Choi et al., 2005; Greenberg et al., 2003; Yoon et al., 2007) and (2) to directly examine speaker-dependent variability in the acoustic encoding of prosody (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.

Reliability Tests

Distribution of P- and B-scores



-RPT prosody annotation is reliable across transcribers by Fleiss' multi-rater's kappa agreement scores.

-Higher agreement for boundary than for prominence.

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

Acoustic measurements

- 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) *F2 measures are not reported here.*

Spearman's non-parametric correlation analysis

- **Prominence**

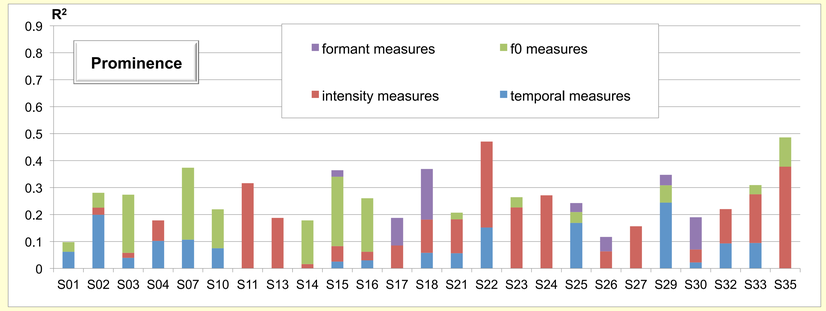
Prom	S01	S02	S03	S04	S07	S10	S11	S13	S14	S15	S16	S17	S18	S21	S22	S23	S24	S25	S26	S27	S29	S30	S32	S33	S35
dur	++	++	++	++	++	++		+	++	++	++		++	++	+			++	+	++	++	++	++	++	++
int		++	++	+	++		++	++	++	++	++	++	+	++	++	++	++	+	+	++	++	++	++	++	++
int_.5		++	+		+		++	++	+	++	++			++	++	++	++	++		++	+	++		++	++
int_1		++	+	+	++	+	++	++	++	++	++	++		++	++	++		++	+	++	++	++	++	++	++
int_2	+	++	++	+	++		+	++	++	++	++		+	+		++	+	++	+	++	+	++	+	++	++
int_4	+	++		++	+		++	++	++	++	++	+	++	+	++			++	+	+	+	++	+	++	+
f1		++			++	++		+	++	+	++	++	++	++		+			+	++	+	++		+	++
f0max	++	++	++		++	++	++	++	++	++	++	++		++	++		+	++		++	++	++		++	++
f0_R		++	++			++	+	++	+		++					++		++		+	++	+		+	

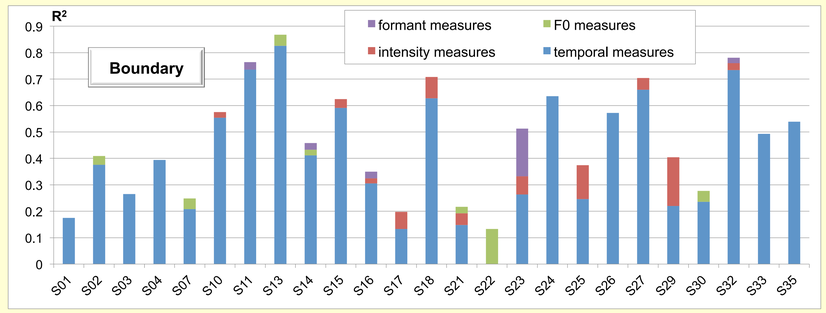
- **Boundary**

Bndry	S01	S02	S03	S04	S07	S10	S11	S13	S14	S15	S16	S17	S18	S21	S22	S23	S24	S25	S26	S27	S29	S30	S32	S33	S35
dur	++	++	++	++	++	++	++	++	++	++	++	++	++	++	+	++	++	++	+	++	++	++	++		
int					+				-	--			--	-		--		-		-			--	-	
int_.5						-			--	--	--		--	--		-				--	-		--	--	
int_1	+				++							+	--											-	+
int_2	+		-			-						+				--				+			--		+
int_4	++									-								-		--			--		
f1	++				+	+				+	++			+		+				++	++	++		+	++
f0max	+		++			+		+		++	++				++		+	++		+		++			
f0_R	-				--				-		-			--			+	-			+	--			
pause		++	++	++	++	++	++	++	++	++	++		++	++		++	+	++	++	++	++	++	++	++	++

- Across speakers, prosodic prominence is positively correlated with a large subset of the acoustic parameters and prosodic phrase boundary is positively correlated with temporal measures (V_duration and silent pause).
- Speaker variability:
  - Across speakers, there is no single acoustic parameter that is correlated with P-scores and the strength of correlation is not uniform.
  - Speakers vary in which acoustic parameters other than temporal parameters are correlated with B-scores and the direction and the strength of correlation is not uniform and varies across speakers.

Speaker-Dependent Regression Models of Prosody





- Speaker-dependent regression models account for 12 - 54% of variation in listeners' response to prominence and 18 - 87% of variation in boundary perception by speaker.
- 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. Yet, one speaker does not rely on temporal parameters at all and some speakers employ other acoustic parameters to cue for prosodic phrase boundary.

Discussion and Conclusion

1. Uniformity comes from the fact that
   - the phonetic characteristics are enhanced when a word is prominent and therefore, most or all acoustic parameters change in the direction of enhancing phonetic distinctiveness.
   - speech tempo reduces in the vicinity of a prosodic phrase boundary, which directly results in the elongation of the pre-boundary vowel and the following pause.
2. Variability comes from the fact that
   - speakers vary in their acoustic encoding of **prominence** in terms of the correlation strength
   - speakers vary in their acoustic encoding of **boundary** in terms of both the strength and direction
   - the contribution of each acoustic cue to listeners' prosody perception varies depending on speakers.
3. Prominence is manifested through the general acoustic enhancement of speech. Boundary is accomplished through slowing down speech.

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. 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.
3. Greenberg, S., Carvey, H., Hitchcock, L., and Chang, S., 2003. Temporal properties of spontaneous speech-a syllable-centric perspective, *Journal of Phonetics*, 31, 465-485.
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. Kochanski, G., Grabe, E., Coleman, J., and Rosner, B., 2005. Loudness predicts prominence: Fundamental frequency lends little , *Journal of the Acoustical Society of America*, 118 (2), 1038-1054.
6. 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).
7. Redi, L. and Shattuck-Hufnagel, S., 2001. Variation in the realization of glottalization in normal speakers, *Journal of Phonetics*, 29, 407-429.
8. Turk, A. E. and Sawusch, J. R., 1996. The processing of duration and intensity cues to prominence, *Journal of the Acoustical Society of America*, 99 (6), 3782-3790.
9. 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.