# Prosody perception by naïve listeners: Evidence from a large multi-transcriber reliability study

Yoonsook Mo, Jennifer Cole, Eun-Kyung Lee

The Beckman Institute and the Department of Linguistics
University of Illinois at Urbana-Champaign

**Q1) How do ordinary listeners perceive the location of prosodic boundaries and prominences?**

**Q2) How can we measure the reliability of prosodic transcription across multiple listeners?**

Prosodic features play an important role in communication.

1. A prosodic boundary demarcates a speech chunk, which is usually a semantically coherent unit of speech.
2. A prosodic prominence highlights a word or phrase in speech, signaling its information contribution.

## Goals of this study:

1. to find an appropriate measure of transcription reliability between naïve transcribers in a task of real-time prosody transcription
2. to look at variation across listeners and across speakers

**Prior studies** test the consistency of prosody perception across transcribers, but with certain limitations:

1. Materials: single, simple sentences or read speech (Streefkerk et al. 1997, 1998)
2. Transcribers: few, trained in prosody and phonetics (Yoon et al. 2004)
3. Procedure (Buhmann et al. 2002; Yoon et al. 2004)
   - Aided by visual inspection
   - Complex annotation scheme
   - Replays allowed as many times as listeners want
4. Analysis: simple % agreement scores or Cohen's pairwise agreement statistic.

# Methodology

## Materials

1. Selected from the Buckeye corpus of American English spontaneous speech (Pitt et al. 2007)
2. 38 short excerpts (about 20 sec): 19 speakers * 2 excerpts each
3. Sound files are blocked into two groups, one for prominence labeling and the other for boundary labeling, and then randomized within a group.
4. Materials are prosodically transcribed by 74 subjects, undergraduates at UIUC.

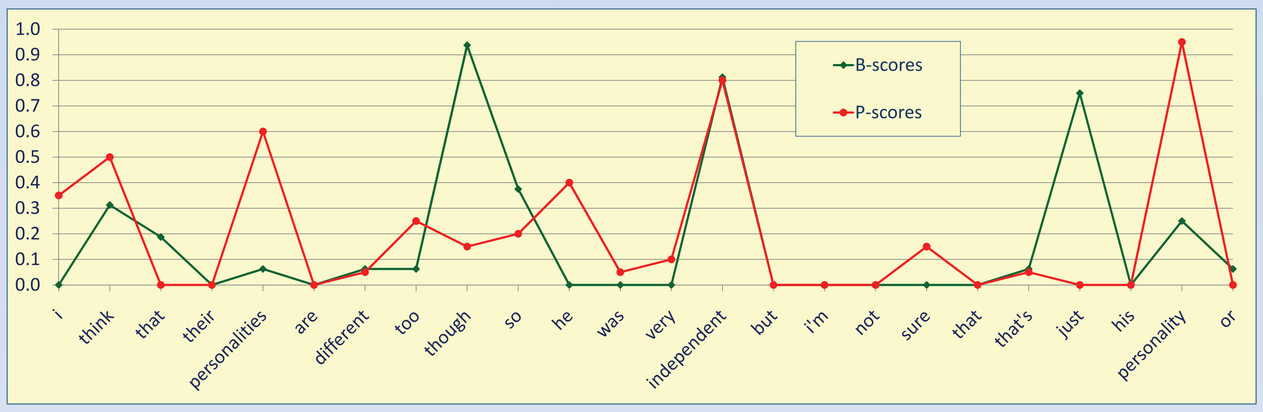
| # of transcribers | Exp. 1 | | Exp. 2 | |
|---|---|---|---|---|
| | Grp. 1 | Grp. 2 | Grp. 3 | Grp. 4 |
| prominence | 16 | 20 | 16 | 20 |
| boundary | 16 | 20 | 15 | 22 |

## Procedures

1. Simple definitions of prominence and boundary and short instructions are provided
2. Subjects see a transcript of each excerpt without any punctuation or capitalization
3. Subjects listened twice to sound files in real time, through headphones, with no visual speech display.
4. Half of the listeners listened for prominence first and the other half listened for boundary first.
5. Listeners marked up a transcript of each excerpt
   - word word word word
   - word | word word | word

# Results

Probabilistic Prominence and Boundary scores




## Fleiss' multi-rater Kappa agreement scores and z-test results

$$K = \frac{P(A) - P(E)}{1 - P(E)},$$

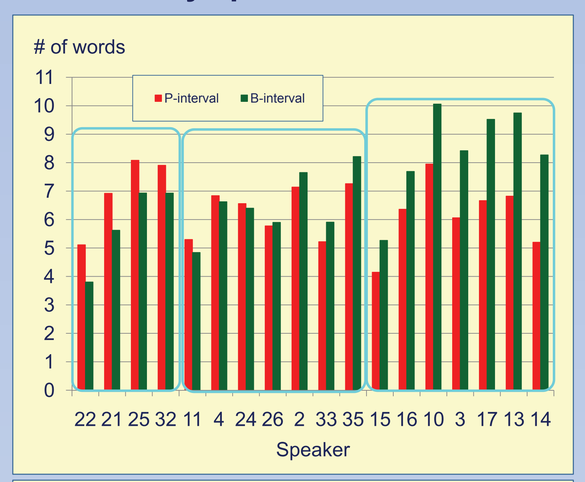
$$P(A) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n} n_{ij}^2 - N \cdot n}{N \cdot n(n+1)}, \quad P(E) = \sum_{j=1}^{k} p_j^2$$

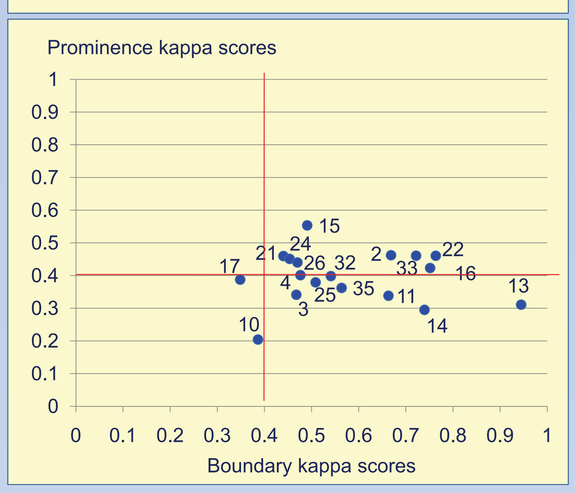
P(A): proportions of times that raters actually agree
P(E): proportions of times that raters would agree by chance
*N*: number of subjects
*n*: number of ratings per subject
*k*: number of categories in which assignments are made
$p_j$: proportion of all assignments which are to the j-th category

| z=2.32, α=0.01 | | Exp.1 | | Exp. 2 | |
|---|---|---|---|---|---|
| | | Grp.1 | Grp.2 | Grp.3 | Grp.4 |
| prominence | Kappa | 0.373 | 0.421 | 0.394 | 0.407 |
| | z | **19.43** | **20.48** | **18.15** | **18.31** |
| boundary | Kappa | 0.612 | 0.544 | 0.621 | 0.575 |
| | z | **27.62** | **21.87** | **25.05** | **26.22** |

**- All agreement scores are statistically significant.**
**- Boundary agreement is consistently higher than prominence agreement.**

## Variation by speakers

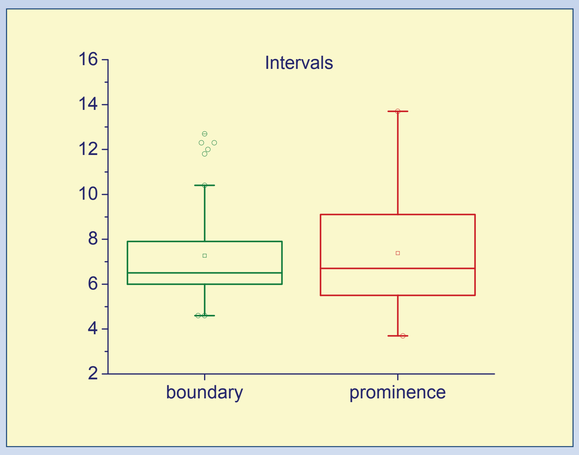


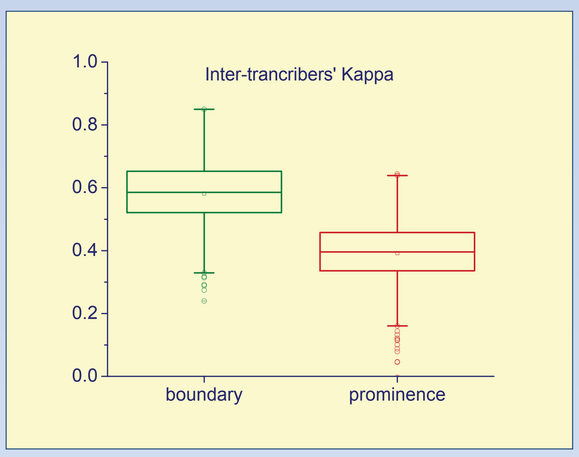

- Average intervals between prominences: 6.4 words (range: 4.1 ~ 8.1 words)
- Average intervals between boundaries: 7.1 words (range: 3.8 ~ 10.1 words)
- There is variation across speakers in the length of the interval between prominences, and between boundaries, as judged by listeners.
- There is variation across speakers in the degree of cross-transcriber agreement for prominence and boundary, based on Fleiss' kappa scores.

**** Reference lines are placed to mark levels of agreement judged to be "fair" for pairwise agreement; this corresponds to even better agreement comparing multiple transcribers (N>2).*

## Variation by listeners (*N* = 74)




Distribution of boundary and prominence intervals across listeners.




Distribution of kappa scores for boundary and prominence agreement, as calculated for each pair of listeners.

# Discussion and conclusion

1. Statistically significant Fleiss' multi-rater agreement scores confirm that listeners' responses are systematic and relate to linguistic information in the utterance.
2. Higher agreement scores for boundary than for prominence show that listeners are more consistent in their perception of boundary location than for prominence location.
3. Speaker-dependent variation in the perception of prosody indicates that speakers vary in how they structure an utterance prosodically and/or in how effectively they cue prosodic structures in the speech signal.
4. Listener-dependent variation indicates that listeners differ in their sensitivity to the locations of prosodic prominence and boundary, for the same set of speakers.

**References**

- Buhmann, J.; Caspers, J.; Heuven, V. J. van; Hoekstra, H.; Martens, J-P.; Swerts, M., 2002. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. *Proceedings of LREC 2002* (Las Palmas). 779-785.
- 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 Psychology, Ohio State University (Distributor).
- Streefkerk, B. M.; Pols, L. C. W.; Bosch, L. F. M. ten, 1997. Prominence in read aloud sentences as marked by listeners and classified automatically. *Proceedings of IFA* (Amsterdam). 101-116.
- Streefkerk, B. M.; Pols, L. C. W.; Bosch, L. F. M. ten, 1998. Automatic detection of prominence (as defined by listeners' judgments) in read aloud Dutch sentences. *Proceedings of ICSLP 1998* (Sydney). 3, 683-686.
- Yoon, T-J.; Chavarría, S.; Cole, J.; Hasegawa-Johnson, M., 2004. Intertranscriber Reliability of Prosodic Labeling on Telephone Conversation using ToBI. *Proceedings of Interspeech 2004* (Jeju). 2722-2732.

**Acknowledgements**
This study is supported by NSF IIS-0703624 and ISF-0414117. We would like to thank Steve Winters, Zak Hulstrom, our participants and the members of the Beckman Institute Prosody & ASR research group for their comments.