Naïve listeners' prominence and boundary perception

Yoonsook Mo, Jennifer Cole, Eun-Kyung Lee

Department of Linguistics
University of Illinois at Urbana-Champaign
{ymo; jscole; eklee}@uiuc.edu

Abstract

This paper examines how ordinary listeners, naïve with respect to the phonetics and phonology of prosody, perceive the location of prosodic boundaries that demarcate speech "chunks" and prominences that serve a "highlighting" function, in spontaneous speech (Buckeye corpus). Over 70 naïve listeners marked the locations of prominences and boundaries in a real-time transcription task. Fleiss' multi-transcribers' reliability tests show that naïve transcribers are consistent in their perception of prosodic boundaries and prominences. Specifically, we observe higher multi-transcriber agreement scores for boundary marking than for prominence marking. Variation between transcriptions of the same speech excerpt produced by different listeners reveals individual differences in the perception of prominences and boundaries. Variation in Fleiss' multi-transcribers' agreement scores for excerpts from different speakers suggests that speakers vary in how they structure an utterance prosodically and/or in how effectively they cue prosodic structure. We also find that nuclear prominences are more consistently perceived by naïve listeners than prenuclear prominences. The finding that naïve listeners agree well above chance on the location of prosodic events indicates that naïve transcription is a valid method for prosody analysis which can augment analysis based solely on expert labeling.

1. Introduction

Speech utterances are more than the linear concatenation of words. Prosodic structures determine the hierarchical organization of words and phrases and the prominence relation among them. The prosodic boundaries encoded in speech mark the hierarchical prosodic phrase structure and serve to demarcate "chunks" of words, while prosodic prominences serve to "highlight" a word or a phrase and convey their status as focused or discourse-new. Therefore, understanding prosodic structures plays an important role in the listener's comprehension of speech utterances.

Previous studies have examined the perception of prosodic boundaries and prominences [1, 5, 6, 7, 8]. Streefkerk et al. [5, 6], for instance, showed reliable inter-labeler agreement on the perception of prominence with read-aloud Dutch sentences. Buhmann et al. [1] also showed that non-expert transcriber labeling can obtain considerably high inter-transcriber agreement rates for both prominence and boundary on materials from the spoken Dutch corpus. Moreover, Buhmann et al. [1] found that boundary annotations reach considerably higher kappa scores (the statistical measure of rater agreement) than prominence annotations. More recently, Yoon et al. [8] examined inter-transcriber agreement with two phonetically trained labelers. In that study, transcribers labeled prosodic events on a subset of the Switchboard corpus, an American English corpus of spontaneous, casual telephone conversations, using a simplified ToBI labeling system. The results showed consistently high rates of inter-transcriber agreement for labels marking the presence of pitch accents, phrasal accents and boundary tones. Moreoever, the agreement scores on the presence and the choice of boundary tones are higher than those of the presence of prominence.

These studies collectively show some degree of consistency among listeners in their perception of prosody; however, we note methodological shortcomings in several instances. Two of the studies demonstrated results based on the readings of simple sentences, not spontaneous speech [5, 6]. One other study arrived at results based on transcriptions from only a few prosodically trained lablers [8]. In the Buhmann et al. [1] and Yoon et al. [8] studies, labeling was aided with visual inspection of the speech display and not done in real-time. The time requirements of visually-guided, expert transcription raise questions about the usefulness of this method for analysis of large corpora.

We address the challenge of efficient and reliable prosody transcription for spontaneous, conversational speech with a transcription method that (i) optimizes efficiency and (ii) delivers probabilistic prosody transcriptions from untrained listeners, naïve with respect to prosodic analysis. Using this method we investigate how naïve listeners perceive prosody in conversational American English, in a real-time listening task. We hypothesize that the naïve listeners in our study will perceive prominences and boundaries consistently, but with higher agreement rates on boundary perception than on prominence perception in accordance with previous studies [1, 7]. We also look for differences in the perception of accent as a function of location in the prosodic phrase, to see if labelers show more consistency in the perception of nuclear accents compared to pre-nuclear accents. In addition, we examine patterns of variation across transcribers, and differences in agreement rate that are dependent on the speaker.

2. Methodology and analysis

2.1. Materials

A total of 36 sound files are created from excerpts of interviews from the Buckeye corpus of spontaneous American English [3], comprising two excerpts of about 20 sec. from each of 18 speakers. We divide the excerpts into two sets so that each set includes one excerpt from each speaker. The 18 excerpts in a set are then divided into two separate blocks, one for prominence transcription and the other for boundary transcription. Within each block, the sound files are randomized for each subject. A printed transcript of the word content in each excerpt is provided for each subject, with excerpts ordered to match the ordering of the sound files they will hear. Words are separated by a space with no punctuation

or capitalization. Speech errors and disfluencies are included in the word transcripts.

2.2. Transcription procedure

74 students were recruited from undergraduate Linguistics courses at the University of Illnois to participate in one of two runs of our transcription experiment (hereafter Experiments 1 and 2). In each experiment, participants are divided into two groups. In one group, transcribers do prominence transcription in the first block, and boundary transcription in the second. The other group has the reverse order.

The experiment runs were held in a computer classroom with each participant seated at a separate computer, equipped with headphones and a printed transcript of the excerpts they will hear. In a brief introduction, participants are told the goal of the study (to find out how ordinary listeners perceive the prosody of conversational speech) and are administered informed consent. Participants are then instructed to mark up their transcript by underlining words they hear as "prominent" and by marking a vertical bar between words that belong to different "chunks" of the utterance, while listening to speech excerpts played in real time. A prominent word is defined as a word that is "highlighted for the listener, and stands out from other non-prominent words", while a chunk is defined as a grouping of words "that helps the listener interpret the utterance", and that chunking is "especially important when the speaker produces long stretches of continuous speech". Participants play each sound file twice at their own pace, marking their transcripts as they listen. Changes to the transcription can be made on the second play of the sound file. It is important to note that participants do not view any graphical display of the speech signal; the transcription is made solely on the basis of auditory impression in concert with the printed transcript.

Each excerpt receives prominence and boundary markings by different transcriber groups. Data from three transcribers are excluded due to failure to follow the transcription guidelines. Transcriptions are coded for speaker and transcriber, and each word is coded for the number of transcribers who marked a prominence and the number of transcribers who marked a boundary. This number ranges from 0 (no transcriber marks prominence or boundary) to a maximum number that corresponds to the total number of transcribers who marked that excerpt (between 15 - 22). This coding yields a graded prominence and boundary score for each word.

In addition to the naïve transcriptions, we also produced an "expert" transcription for eight excerpts to compare the perception of nuclear vs. prenuclear accents. The expert transcription is the consensus labeling from three transcribers, trained in phonetics and experienced in the ToBI system for Mainstream American English.

2.3. Analyses

The transcribed data are subject to four analyses. The first analysis evaluates the reliability of prominence and boundary labels based on tests of inter-transcriber reliability for all transcriptions pooled over speakers and transcribers. The second analysis looks at variation in speakers' expression of prosody as a function of differences in inter-transcriber agreement rates across speakers. The third analysis evaluates variation in transcribers' sensitivity to prosody cues by comparing the frequency of prosody labels across transcribers. The fourth and final analysis looks at transcription reliability of accent as a function of accent type—nuclear vs. pre-nuclear—through a comparison of the naïve transcription data with the expert transcription, where nuclear and pre-nuclear pitch accents can be identified.

3. Results

3.1. Multi- and Inter-transcribers' reliability test

The reliability of the naïve prosody transcriptions pooled over all 74 transcribers is evaluated using Fleiss' kappa coefficient [2] and their z-normalized scores, as shown in Table 1. We choose Fleiss' kappa coefficient because it provides a single coefficient as a measure of agreement between all pairs of transcribers, while Cohen's kappa calculates agreement only between a single pair of transcribers, and multi-transcriber agreement is approximated using mean kappa scores [1, 8].

Table 1: *Fleiss' kappa scores and z-normalized socres*

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**

The z-scores in Table 1 are all greatly above the significance level with a 99% confidence level and show that agreement among naïve listeners on the perception of prominence and boundary is much above chance. They also show that the agreement for boundary perception is higher than those for prominence perception.

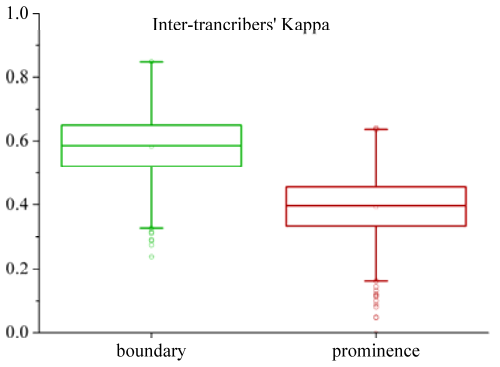

Figure 1: *Distributions of Cohen's kappa coefficients as a measure of inter-transcriber agreement on prominence and boundary labels.*

Variation in agreement for prominence and boundary labels between individual pairs of transcribers is evaluated using Cohen's kappa coefficients. Figure 1 shows the distribution of Cohen's kappa coefficients over all pairs of transcribers.

Kappa scores range from -0.003 to .644 for prominence labels and from 0.240 to 0.850 for boundary labels.

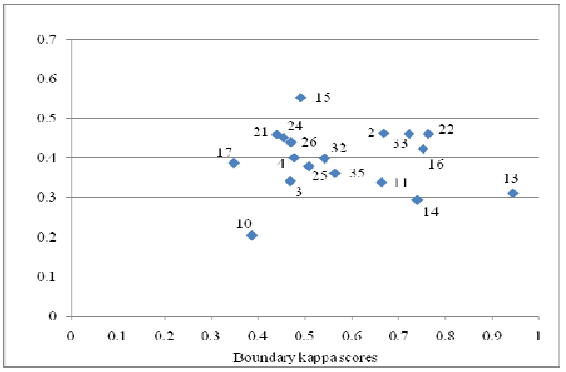

Figure 3: *The distribution of Fleiss' multi-transcribers' kappa coefficients by speaker in Experiment 1 and 2 pooled.*

Figure 3 shows the distribution of Fleiss' kappa coefficients (based on all transcriber pairs) for each of the 32 speakers. Z-scores range from 3.25 - 13.77, and all 32 scores are significant at $p < .001$.

3.2. Speaker and listener dependent variation

Figure 4 plots the mean interval (counted in words) between accents and between boundaries for each speaker, pooling transcriptions over all transcribers who labeled that speaker. Variation between speakers may reflect variation in the frequency of accent and boundary, and/or it may reflect variation in the salience of cues to accent and boundary produced by different speakers.

Figure 4 shows that for many speakers the mean interval between boundary labels (3.8 - 10.1 words) is slightly longer than the mean interval between prominence labels (4.1 - 8.1). However, the grand mean of the boundary interval values, calculated across all speakers (7.1 words) is close in value to the grand mean of the prominence interval values (6.4 words). Some speakers exhibit longer mean intervals between prominences than between boundaries, which means that for those speakers at least some transcribers marked prosodic phrase intervals that lack even a single prominent word.

The mean intervals between prominence and boundary labels was also calculated for each transcriber, this time pooling data across all the speakers transcribed by that transcriber. Figure 5 shows the distribution of the mean interval values. The mean prominence interval ranges from 3.7 - 18.3 words, with a grand mean interval length of 7.5 words. The mean boundary interval ranges from 4.6 - 15.3 words with a grand mean interval length of 7.4 words. This means that listener dependent variation is larger than speaker dependent variation.

3.3. Perception of nuclear vs. prenuclear prominences

The naïve transcriptions are compared to expert transcriptions to see if there is any difference in agreement rates for nuclear accents (i.e., accents that are final in the prosodic phrase) and pre-nuclear accents, as shown in Table 2. For example, naïve transcribers agree with experts in categorizing 5289 words as non-prominent, while they categorize as non-prominent 185 words recognized by expert transcribers as prominent.

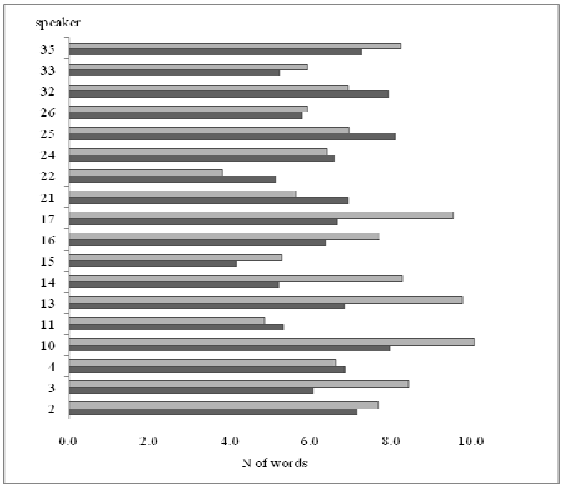

Figure 4: *Mean intervals between prominence (black bars) and boundary (grey bars) labels by speaker. Each speaker's data is pooled across transcribers.*

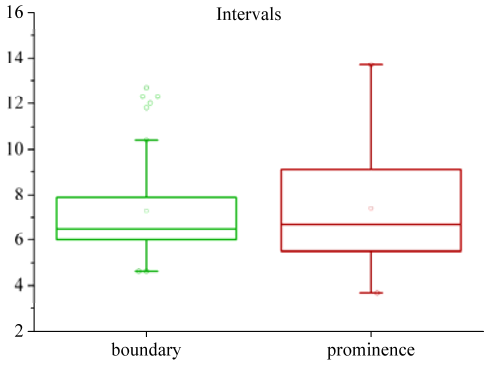

Figure 5: *Distributions of mean intervals between prominence and boundary labels by transcriber.*

Table 2: *Confusion matrix of nuclear and pre-nuclear prominences comparing naïve and expert ("trained") labelers.*

naïve listeners / trained listeners	No prominence	Prominence	Total
No prominence	5289	185	5474
Pre-nuclear Prom.	1647	556	2230
Nuclear Prominence	731	507	1238
Total	7694	1248	8942

Results from chi-square tests show that there is a statistically significant difference in naïve listeners' perception of prominence ($\chi^2 = 1485.0 > 13.82$, $\alpha = 0.001$) depending the locations of prominence within a prosodic phrase. Furthermore, repeated chi-tests show that there are significant differences between the perception of (i) pre-nuclear

prominence and no prominence (χ^2 = 846.7 > 10.85, α = 0.001), (ii) nuclear prominence and no prominence (χ^2 = 1541.5 > 10.85, α = 0.001), and (iii) nuclear prominence and pre-nuclear prominence (χ^2 = 96.1 > 10.847, α = 0.001).

4. Discussion

This study looks at variability in the prosodic features naïve listeners perceive in conversational speech. Results from inter-transcriber agreement tests show that naïve listeners are consistent in their labeling of prominent words and prosodic boundaries, indicating regular patterns in listeners' perceptions. A further finding is that boundary perception is more consistent across listeners than prominence perception, based on higher Fleiss' agreement scores for boundary agreement than for prominence, in all analyses.

Upon closer examination, however, we also find great speaker- and listener-dependent variation in the perception of prosodic prominence and phrase boundary. Naïve listeners' pairwise agreement scores on the perception of boundary range from 0.24 (fair) to 0.85 (very good) with a mean agreement score of 0.582, while the pairwise agreement scores on the perception of prominence range from - 0.003 (almost perfect disagreement and not larger than agreement scores by chance) to 0.64 (good), with a mean agreement score of 0.39. These results indicate that not all pairs of transcribers agree similarly in their judgment of prominence or boundary, i.e., there is significant variation in perception across listeners. In addition, about 96.3% of pairwise Cohen's kappa coefficients for boundary agreement are over 0.4 (generally considered the low end of the "good" agreement range), while only 48.9% of prominence agreement scores are over 0.4. This tells us that not only there is an overall higher agreement rate for boundary perception, but listener-dependent variation is less for boundary perception than prominence perception.

To examine speaker-dependent variation, we compare z-normalized statistics of multi-transcribers' Fleiss' kappa coefficients across speakers. The kappa coefficients range from 0.35 - 0.95 (mean = 0.58) for boundary and from 0.20 - 0.59 (mean, 0.39) for prominence. These results show that although agreement scores for both prominence and boundary perception are significantly high, listeners' agreement scores vary greatly depending on the speaker.

Measures of the mean number of words between prominence or boundary labels in a given excerpt offer another index of speaker- and listener-dependent variation. The mean interval length for both prominence and boundary interval varies across the group of transcribers who labeled that excerpt. This pattern of variation may reflect variation in speakers' production of prosody cues and/or variation in listeners' sensitivity to those cues.

Lastly, this study shows different patterns of perception for different kinds of prominences. Comparing naïve listeners' transcripts with the transcripts from three prosodically trained labelers, we find that naïve listeners are more accurate in perceiving nuclear prominences (those that are marked as final in the prosodic phrase by expert transcribers) than they are for pre-nuclear prominences. This finding suggests that the cues to prominence may be more salient with nuclear prominences than with pre-nuclear prominences.

5. Conclusion

This study shows that ordinary listeners perceive prosodic prominences and boundaries with consistency that is well above chance, and that consistency among listeners is greater for prosodic boundary perception than prominence perception. This study also shows that listeners are more consistent and accurate in the perception of nuclear prominence than in the perception of pre-nuclear prominence. The method of naïve, "real-time" prosody transcription developed in this work is an efficient means to obtain prosody labels that are gradient (or probabilistic), reflecting patterns of speaker and listener variation in the production and perception of prosody in spontaneous speech.

6. 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 Prosody & ASR research group for their comments.

7. References

[1] 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.

[2] Fleiss, J. L. 1971. Measuring nomial scale agreement among many raters. Psychological Bulletin. 76, 3780382.

[3] Ladd, R. D., 1998. *Intonational phonology*. Cambridge University Press.

[4] 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).

[5] Streefkert, 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.

[6] Streefkert, B. M.; Pols, L. C. W.; Bosch, L. F. M. ten, 1998. Automatic detection of prominence (as defined by listeners' judgements) in read aloud Dutch sentences. *Proceedings of ICSLP 1998* (Sydney). 3, 683-686.

[7] Wagner, P., 2005. Great expectations-Introspective vs. perceptual prominence ratings and their acoustic correlates. *Proceedings of Interspeech 2005* (Lisbon). 2381-2384.

[8] 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.