Listening for sound, listening for meaning: Task effects on prosodic transcription

Jennifer Cole [1], *Timothy Mahrt* [1], *José I. Hualde* [1,2]

[1] Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[2] Department of Spanish, Italian and Portuguese , University of Illinois at Urbana-Champaign, Urbana, IL, USA

jscole@illinois.edu, tmahrt2@illinois.edu, jihualde@illinois.edu

Abstract

The perception of prosodic structure (phrasal prominences and boundaries) may depend in part on acoustic cues in the speech signal and in part on utterance meaning as related to syntactic structure and discourse context. In this study we ask if listeners are able to differentially weigh acoustic and meaning-based cues to prosody. We test naïve subjects' transcription of prominences and boundaries in spontaneous American English under three different conditions, all of which involve listening to audio recordings and marking prominences and boundaries on a transcript. The three conditions differ in the instructions given to transcribers. In one condition, subjects were instructed to transcribe prominence and boundaries based on meaning criteria, in a second condition they were told to transcribe based on criteria of acoustic salience, and a third condition had less specific instructions, without explicit reference to either meaning-based or acoustic cues. Our results show that subjects perform differently when focusing on meaning than when focusing on acoustics, especially for prominence marking, where partially different sets of words are selected as prominent under the two tasks. Boundary marking is more similar under the two instructions, with acoustic criteria resulting in more listeners marking a given word as pre-boundary, but with boundaries marked largely on the same words in both tasks. With non-specific instructions, performance was similar to that obtained under acoustic-based instructions. We report on agreement rates within and across conditions. This study has implications for models of prosody perception and the methodology of prosodic transcription.

Index Terms: prosody, prominence, boundaries, prosodic transcription

1. Introduction

Prosodic prominences and boundaries are assigned to utterances based on many factors related to syntactic structure and discourse context. Because of these dependencies, the acoustic cues that signal prosody also serve as cues to the linguistic context of the prosodically marked word and the utterance to which it belongs. For instance, listeners interpret syntactic structure based in part on acoustic cues that signal prosodic boundaries [1,2,3,4], and the interpretation of the focus and information status of a word is influenced by acoustic cues to prominence [5,6,7,8,9]. The influence of prosodic cues on discourse processing is such that a mismatch between the discourse context and a word's prosodic form can disrupt processing, as shown by evidence from eye-tracking [6,8,10] and ERP studies [11].

The studies cited above, and many others, demonstrate the role of prosody in communicating meaning related to the syntactic, semantic and discourse context of an utterance. While the evidence shows that listeners attend to the prosodic cues present in the acoustic speech signal, it's possible that listeners' perception of prosody is also driven by expectations about the prosodic form of an utterance given its syntactic properties and its semantic and discourse context, in the same way that expectations play a crucial role in word recognition (and in the visual domain). In this paper we examine prosody perception due to acoustic cues and due to expectations from factors related to syntactic, semantic and discourse context (hereafter *meaning-based cues*), and ask whether listeners can focus differentially on acoustic and meaning-based cues in identifying prosodic prominence and phrase boundaries in spontaneous speech. We examine listeners' perception of prosody using the method of Rapid Prosody Transcription (RPT) developed by one of the authors (JC) for investigating prosody through the analysis of judgments made by naïve native speakers of English [12]. This method and prior findings are introduced in the next section.

1.1. Rapid Prosody Transcription

Under the RPT methodology multiple listeners (between 10-20 in prior experiments) make auditory judgments about the location of prosodic phrase boundaries and prominences in an audio speech recording, based only on (the individual listener's) auditory impression and with no visual inspection of the graphical speech display. For each word of transcribed speech two continuous-valued prosody features are calculated, representing the proportion of transcribers who perceived the word as prominent (the p-score), and the proportion who perceived the word as final in a prosodic phrase (the b-score). A p-score of 0 shows agreement among all transcribers that the word is not prominent while a p-score of 1 shows agreement that the word is prominent. Values in between 0 and 1 reflect disagreement among transcribers. The prosody scores can be viewed as a measure of the probability that a random listener (from the same speech community) will perceive a given word as prominent, or as preceding a prosodic phrase boundary. Fig. 4 below shows an example of the p- and b-scores for each word in a fragment of a speech sample from this study, based on the aggregated transcriptions of 16 listeners.

Cole and her colleagues conducted two studies of prosody perception with American English spontaneous speech using RPT, testing the relative contribution of signal-based processing (from acoustic cues) and expectation-based processing (from syntactic and information-based cues) in non-expert listeners' judgments of prosodic prominence and phrase boundaries. Cole et al. [13] investigated p-scores and their relationship to various acoustic cues previously found to correlate with prosodic prominence such as increased duration, increased intensity, and the presence of a pitch accent. They found a positive correlation between these acoustic measures

10.21437/SpeechProsody.2014-160

and p-scores: higher values of the acoustic measures (e.g., longer duration, higher intensity) predict higher p-scores (indicating higher agreement among listeners that a word is prominent) for a given word. However, p-scores were also correlated with measures of word surprisal—non-acoustic cues such as word frequency and number previous mentions of a word in the discourse. Similarly, Cole et al. [14] found that syntactic context predicts the perception of boundaries in spontaneous speech, in addition to and partly independent of acoustic cues.

These two prior studies using RPT provide evidence that in perceiving prosodic prominences and phrase boundaries, listeners are influenced both by acoustic cues and by cues related to the syntactic role of a word and its meaning in relation to discourse context, and that these cues function at least partly independently of one another. However, these studies do not fully indicate to what extent acoustic and meaning-based cues are different and whether listeners can attune their attention to either acoustic or meaning-based cues, diminishing the other. These issues are the point of departure for the present study.

2. Methodology

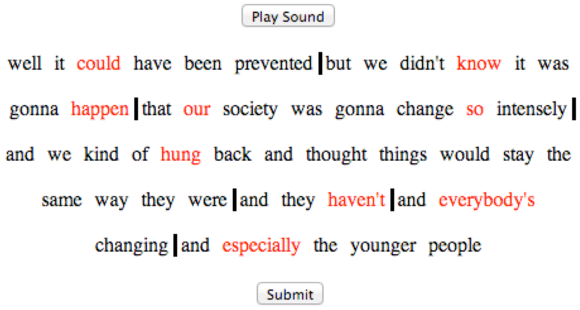
2.1. Subjects and Materials

This experiment uses the Rapid Prosody Transcription method [12] to obtain prominence and boundary judgments from 15 naïve native speakers of English, all students at the University of Illinois. Sixteen short excerpts (~18 s each) from sixteen different speakers in the Buckeye corpus of spontaneous English speech [15] were used in this study. The total number of words summed over all excerpts was 925. This dataset is a subset of the dataset used in the study by [12] and comparisons with the findings of that study are included in Section 3. The transcription experiment was conducted in a quiet, computer-equipped room. Subjects proceeded through the experiment at their own pace using LMEDS, a customized software application developed by the authors.

2.2. LMEDS

A customized web interface, LMEDS, the Language Markup and Experimental Design Software, was developed to administer the experiment materials electronically. LMEDS is a generic toolset that simplifies the creation of custom experimental setups, such as those needed for an RPT experiment, as well as the aggregation of data collected during the experiment.

In this LMEDS experiment each excerpted speech sample was presented on its own page. Each of these pages presents a button for playing the audio file, a transcript where each word is clickable, and a button to progress to the next phase (Fig. 1). Each participant (hereafter, transcriber) first listened to the audio twice while clicking on individual words to mark a perceived boundary after the word. The transcribers had no training in phonetics and were not shown any visual display of the speech waveform, spectrogram or pitch track. After two passes through the file, listening and marking boundaries, transcribers then listened to the audio passage two more times, clicking on words perceived as prominent. The interface displayed the location of a selected boundary with a thick vertical line and indicated a word marked as prominent by changing the font color of the word to red. While marking prominences, transcribers were able to see, but not modify, the boundaries they had just placed. After annotating a transcript for both boundaries and prominences, subjects would progress to the next page in the experiment.



well it could have been prevented | but we didn't know it was gonna happen | that our society was gonna change so intensely | and we kind of hung back and thought things would stay the same way they were | and they haven't | and everybody's changing | and especially the younger people

Figure 1. Example transcript and action buttons for a speech excerpt displayed in LMEDS, with prominences and boundaries as marked by an individual transcriber.

2.3. Task instructions

The experiment was divided into two blocks. Each block specified the criteria transcribers were to use in making judgments about the location of prominences and boundaries. In the *acoustics* block, subjects were asked to mark a boundary where they heard a ‘break, discontinuity or disconnection in the speech stream, strong or subtle’ and were asked to mark a prominence where they heard a word stand out by ‘being louder, longer, more extreme in pitch, or more crisply articulated.’ In the *meaning* block, subjects were asked to mark boundaries where the audio could be ‘segmented with minimal disruption of the meaning of the speech’ and were asked to mark the words that ‘convey the main points of information as you think the speaker intended.’

3. Results

In our analysis we consider the RPT task under three different instruction sets: the explicit acoustic-based and meaning-based instruction sets that were run for the present study and the less explicit instructions used in [12,13,14].

Our first question is whether or not subjects performed differently across these three tasks. One way to assess differences due to task instructions is through inter-transcriber agreement, as reflected in p-scores and b-scores. Greater agreement would indicate stronger and more consistent cues to prosody under the stated criteria (acoustic or meaning-based). Fig. 2 shows that the distribution of b-scores and p-scores across all of the recordings are largely similar across the three transcription tasks. Most words receive a b-score of zero, indicating that no transcribers marked a boundary following the word, with a nearly flat distribution of b-scores greater than zero. The distribution of p-scores is similar, with the majority of words receiving a p-score of zero (again, indicating that no transcribers marked the word as prominent). However, there are also a significant number of words with low, non-zero p-scores (i.e, words that very few transcribers marked as prominent). Although there is some variation across tasks, the overall trend in transcriber agreement for p-scores or b-scores is the same.

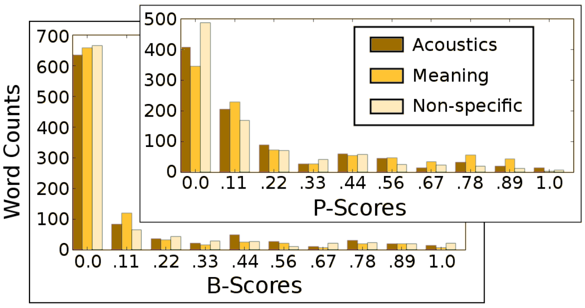


Figure 2. Distribution of b- and p-scores across the three tasks differing in transcription instruction.

The distributions of b-scores and p-scores also show that across tasks transcribers are marking a similar number of words as prominent or as preceding a boundary. For instance, looking at the histogram of b-scores we see that the number of words with a b-score of zero (indicating that no transcriber marked the word for a boundary) is very similar across the three tasks. We also observe that boundary labeling is more conservative than prominence labeling: in all tasks there are more words with a b-score of zero than there are words with a p-score of zero, which means that there are relatively fewer words being marked by transcribers as pre-boundary compared to the number of words marked as prominent. One exception to the overall similarity in prosody scores across tasks, as shown in Fig. 2, is the lower number of words with a p-score of zero under the meaning-based and acoustic criteria compared to the non-specific criteria. This finding indicates that transcribers are more likely to judge a word as prominent when attending to specific cues than when judging prominence in a non-specific way .

The distributions of p-score and b-score values for all tasks show a high agreement among transcribers on words that are not prominent (p-score=0), and on words that are not preceding a boundary (b-score=0), but it reveals little about the patterns of agreement on individual words as marked under different transcription instructions. We are interested to know if the transcribers weigh acoustic and meaning-based criteria differently on the basis of the task instructions, marking different words as prominent/important or as pre-boundary, across tasks.

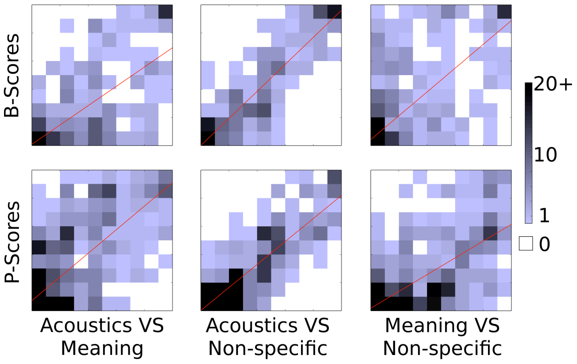


Figure 3. Linear correlations on top of density plots comparing b-scores (top) or p-scores (bottom) in one task against the corresponding score for the same word in a different task. The plots represent data density (# of datapoints in the same region of the plot) with values on the color scale as shown in the legend.

We examine task effects on transcription by comparing b-scores and p-scores for each word from one task with the scores for the same word from a different task, using correlation and linear regression analysis (Fig. 3). All correlations are significant, and r^2 values are above 0.5 for all comparisons (Table 1), indicating that the selection of words that are prominent and in pre-boundary position is similar across tasks, though not identical. The transcriptions that are most correlated (with the highest r^2) are those based on acoustic and non-specific criteria. The meaning-based transcription is less correlated to transcription under either acoustic or non-specific criteria. These findings show that transcribers are able to weigh acoustic and meaning-based cues differently when specifically instructed, but that in the absence of instructions calling for attention to meaning-based criteria, transcribers rely more on acoustic cues in marking prominences and boundaries. Thus, providing different instructions to subjects can indeed cause them to attune their attention to different types of information in speech.

	Acoustics VS Meaning	Acoustics VS Non-specific	Meaning VS Non-specific
B-scores	0.616	0.892	0.528
P-Scores	0.575	0.828	0.540

Table 1. Linear regression coefficients (r^2) for data plotted in Figure 3. (p < 0.01 for all reported values)

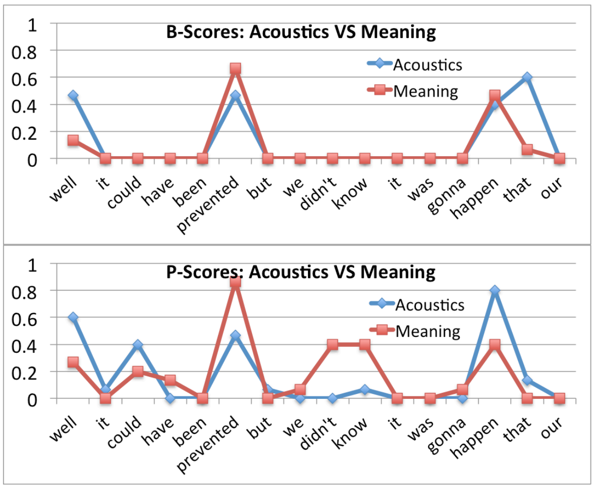


Figure 4. B-scores and p-scores for words from a fragment of a single excerpt, comparing the acoustics-based task with the meaning-based task.

The relatively lower correlation between acoustic and meaning-based prosody scores tells us that there are frequent mismatches in transcribers' marking of words in these two tasks. There are two scenarios that can explain such differences. One possibility is that transcribers in one task are marking a subset of the words that are marked in the other task. Taking p-scores as an example, it could be words that are selected as prominent under meaning-based criteria are also selected under acoustic criteria, but not vice versa. Under this scenario, (nearly) all words with p-scores greater than zero in the meaning-based transcription would also have non-zero p-scores in the acoustic transcription, but in addition, some words with non-zero p-scores in the acoustic transcription would have a p-score of zero in the meaning-based task. The second scenario is where *different* words are marked as prominent under the two transcription criteria. This pattern of mismatch would result in words that have p-scores of zero in one task and non-zero p-scores in the other task, and vice-versa.

Fig. 4 plots b-scores and p-scores for individual words for a fragment of one speech excerpt, comparing acoustic-based with meaning-based transcription. Recall that the regression analysis (Table 1) shows substantial disparity between tasks in both b-scores and p-scores. The example in Fig. 4 suggests that task-related differences pattern differently for p-scores compared to b-scores. The b-scores under the two tasks (top panel) are similar in the selection of words that are marked as pre-boundary by any transcriber—the lines graphing b-scores nearly lie on top of one another—with b-scores in the two tasks differing mostly in the *number* of transcribers who mark a word as pre-boundary. On the other hand, the p-scores under the two tasks (bottom panel) differ substantially in which words are marked as prominent—the lines graphing the p-scores do not appear so nearly as lying on top of one another.

To further investigate the nature of the disparity between acoustic and meaning-based transcription, we compare prosody scores between the tasks after binning all scores into two values: 0 and 1. All b-score and p-score values of zero remain at zero, and values greater than zero are set to 1. This amounts to labeling a word as prominent if one or more transcriber marks it so, and otherwise labeling it as not-prominent. B-scores are similarly transformed to the labels boundary and not-boundary. Using these binned results, we can now ask how often two tasks align in their scores, which reveals the extent to which the same words are selected as prominent or as pre-boundary in both tasks, disregarding differences in the number of transcribers (>0) who agree in marking the word. In this analysis we ask whether a word that is labeled as prominent by acoustic criteria is also labeled as prominent under meaning-based criteria, and similarly for boundary labeling.

Fig. 5 shows the patterns of agreement between acoustic and meaning-based transcription for prominence and boundary labels (P/B) from the binned p-score and b-score data. Words counted in the Agree groups have the same P (or B) label in both tasks (acoustic, meaning), where 1 marks P (or B) and 0 marks not-P (or not-B). Words counted as Disagree are labeled differently in the two tasks. We observe several interesting findings. First, the vast majority of the words in the speech samples (89%) have the same boundary label across tasks (words in the 'Agree' groups), with most words assigned the not-boundary label. This meets our expectation, given that most prosodic phrases contain more than one word. However, we also note that there is high agreement across tasks for words with the boundary label, and only 11% of words are assigned different boundary labels in the two tasks. Turning to the prominence labels, we again note that overall more words are marked as prominent than are marked as (pre-)boundary, but there is also a lower overall level of agreement between tasks in prominence labeling, with only 76% of words assigned the same prominence label (the Agree groups). In other words, 24% of the words in this sample are marked as prominent in one task but not in the other. Words with disagreeing labels include those marked as prominent under acoustic criteria but not under meaning-based criteria, and vice-versa. These findings confirm the patterns shown in Fig. 4, that differences in p-scores between the tasks reflect the selection of different words marked as prominent, where differences in b-scores tended to reflect differences in the number of transcribers marking a word as pre-boundary more than differences in which words are marked.

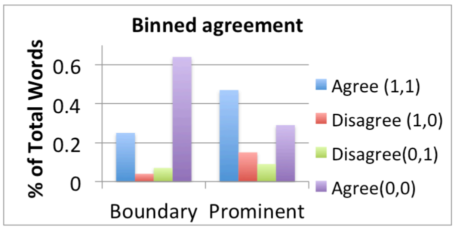


Figure 5. Number of words (% of total) that agree or disagree in prominence and boundary labels across acoustic and meaning-based tasks.

In both the histograms (Fig. 2) and the agreement ratings (Fig. 5) we see that there are more words have a b-score of 0 than there are words having a p-score of zero. For the p-scores, that mass is mostly redistributed to the low, non-zero p-scores (between 0.1-0.3). One reason for this discrepancy between prominence and boundary marking might be the number and variety of factors that condition the placement of prosodic prominence vs. boundaries. Thus, a boundary may be placed at a major syntactic juncture and also preceding disfluency. Prominence, on the other hand, seems to be conditioned by a greater variety of factors. Importantly, words that are acoustically salient do not necessarily convey new or pragmatically important information [16]. There may be differences among transcribers and/or across tasks in the weighting of these factors in the perception of prominence, resulting in greater variability in prominence marking. Another consideration is that the acoustic correlates of prominence seem to be more variable across speakers and utterances compared to the acoustic correlates of boundaries [17], and it's possible that transcribers vary in their sensitivity to individual cues.

4. Conclusions

This study compared prosody transcription under task conditions that focus listeners' attention on acoustic vs. meaning-based criteria. Transcription of prominence and boundaries in spontaneous American English was conducted by non-expert listeners. The findings show similar frequency of boundary and prominence marking across tasks, a lower frequency of boundaries than prominences, and higher agreement among transcribers in the location of boundaries. Task-related differences were also observed: more frequent prominence marking under meaning-based criteria, and a greater disparity between tasks in the individual words that are marked as prominent than there is for words marked as preceding a boundary. Overall there is more uniformity across transcribers and across tasks in boundary marking, parallel to results on inter-transcriber reliability for the ToBI prosodic transcription system [18,19]. This finding calls for future work on the status of prominence in speech production and perception, and on the criteria for prominence transcription. Our ongoing work investigates acoustic cues and also compares p-scores and b-scores gathered in a text-only condition with those obtained under the conditions described in this paper.

5. Acknowledgements

This study is supported by NSF BCS 12-51343.

6. References

[1] Schafer, A., Speer, S., Warren, P., & White, S. D. Intonational disambiguation in sentence production and comprehension. Journal of Psycholinguistic Research, 29(2):169-182, 2000.

[2] Carlson, K., Clifton, C., Jr., & Frazier, L. Prosodic boundaries in adjunct attachment. Journal of Memory and Language, 45:58–81, 2001.

[3] Clifton, C., Carlson, K., & Frazier, L. Informative prosodic boundaries. Language and Speech, 45:87–114, 2002.

[4] Snedeker, J. and Truesell, J. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. Journal of Memory and Language, 48:103-130, 2003.

[5] Birch, S., & Clifton, C. Jr. Focus, accent, and argument structure: Effects on language comprehension. Language and Speech, 38(4):365-392, 1995.

[6] Dahan, D., Tanenhaus, M. K., & Chambers, C. G. Accent and reference resolution in spoken-language comprehension. Journal of Memory and Language, 47(2):292-314, 2002.

[7] Chen, A., den Os, E. & de Ruiter, J.P. Pitch accent type matters for online processing of information status: Evidence from natural and synthetic speech. The Linguistic Review, 24, 317–344, 2007.

[8] Ito, K., & Speer, S. R. Anticipatory effects of intonation: Eye movements during instructed visual search. Journal of Memory and Language, 85(2):541-573, 2008.

[9] Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. Interpreting pitch accents in on-line comprehension: H* vs. L_H*. Cognitive Science, 32, 1232-1244, 2008.

[10] Arnold, J. E. (2008). *THE BACON* not *the bacon*: How children and adults understand accented and unaccented noun phrases. Cognition, 108(1), 69-99.

[11] Magne, C., Astésano, C., Lacheret-Dujour, A., Morel, M., Alter, K., & Besson, M. On-line processing of “pop-out” words in spoken French dialogues. Journal of cognitive neuroscience, 17(5):740- 756, 2005.

[12] Mo, Y., Cole, J., Lee, E. Naïve listeners’ prominence and boundary perception. In Proceedings of the 4th Speech Prosody, Campinas, Brazil, 2008.

[13] Cole, J., Mo, Y., & Hasegawa-Johnson, M. Signal-based and expectation-based factors in the perception of prosodic prominence. Laboratory Phonology, 1:425–452, 2010.

[14] Cole, J., Mo, Y., & Baek, S. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. Language and Cognitive Processes, 25(7):1141–1177, 2010.

[15] Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., et al. Buckeye corpus of conversational speech (2nd release). Columbus, OH: Department of Psychology, Ohio State University, 2007. Retrieved March 15, 2006, from www.buckeyecorpus.osu.edu.

[16] Calhoun, S. The centrality of metrical structure in signaling information structure: A probabilistic perspective. Language, 86:1-42, 2010.

[17] Mo, Y. Prosody production and perception with conversational speech. PhD Thesis, University of Illinois, 2011.

[18] Pitrelli, J.F., Beckman, M.E., & Hirschberg, J. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, 123-126, 1994.

[19] Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In Proceedings of the International Conference on Speech and Language Processing, Jeju, Korea, 2729-2732, 2004.