Prominence perception in and out of context

Rory Turnbull, Adam J. Royer, Kiwako Ito, Shari R. Speer

Department of Linguistics, Ohio State University, Columbus, OH, USA

turnbull@ling.osu.edu

Abstract

The perception of prosodic prominence is known to be influenced by several distinct factors. In this study, we investigated the role of context, both global and local, in the prominence judgements of naïve listeners. Monolingual English listeners marked where they heard prominence on pairs of two-word phrases (e.g. *blue ball*, *green drum*). Stimuli varied in whether or not the first phrase implied a contrastive focus on the second phrase. We found clear evidence of a hierarchy of prominence across pitch accent types: L+H* > H* > !H* > unaccented. Additionally, we found that contrast status only affected prominence markings when the participants were made explicitly aware of the discourse context and were instructed to imagine themselves physically present to observe the conversation. This effect of global context suggests that information structure cannot be reliably interpreted in the absence of an established discourse context. Taken together, these results suggest that naïve listeners are sensitive to prominence differences at levels corresponding to categorical annotations. Perception of a word's relative prominence was consistently influenced by phonetic and phonological factors, while pragmatic factors (such as contrast-evoking context) required more elaborate plausibility manipulations in order to affect prominence perception.

Index Terms: prominence, perception, discourse, focus, contrast

1. Introduction

Intonational phonology assumes a strictly layered hierarchical prominence organization [1]. Words with pitch accent are more prominent than words without pitch accents, and the nuclear pitch accent—in American English, the final one in the utterance—is more prominent than other pitch accents [2]. For the most part, these assumptions have gone largely unchallenged.

Previous studies [3, 4, 5] have shown that the perception of prosodic prominence is based on the listeners' expectations in addition to properties of the signal. For instance, [3] found that information structural considerations, such as anticipating a contrastive focus, can influence perception of prominence such that a semantically or pragmatically salient word can be perceived as prominent even in the absence of acoustic or phonological salience. The scientific goal of this research project is to establish what factors underlie the perception of prominence—factors such as acoustics, phonology, lexico-syntactic phrasing, pragmatic context—and how these factors interact. Additionally, we seek to describe the constraints on the use of these factors predictively in speech comprehension [6].

Several investigations of prosodic prominence have obtained judgements from naïve, untrained listeners in tasks where they are asked to mark prominences on a transcript of aurally presented speech [4, 7, 8, 9, 10, 11]. For longer speech samples, such a task can involve considerable memory load. In the current study, we used a similar metalinguistic judgement task where native speakers of American English were asked to indicate which words in a short phrase sounded prominent. In particular, we examined the role of discourse-level factors and paradigmatic phonological structure in influencing listeners' judgements about which words are prominent.

2. Methods

2.1. Materials

Materials were selected from a ToBI-annotated corpus of spontaneous speech collected from naïve speakers instructing Christmas tree decoration [12]. Twenty-two utterances consisting of adjective-noun combinations, each denoting a particular tree ornament (e.g. red house), were extracted from one female speaker. Utterances either had the pitch accent tune [H* !H*] or [L+H* 0] on the adjective and noun respectively, where 0 represents an unaccented word. The first of these tunes can be considered a 'neutral' prosody, while the second is commonly associated with contrastive focus on the adjective. The stimulus phrases involved eight different adjectives (*beige*, *blue*, *brown*, *clear*, *gray*, *green*, *navy*, and *orange*) and six different nouns (*ball*, *bell*, *candy*, *drum*, *house*, and *onion*).

2.2. Procedure

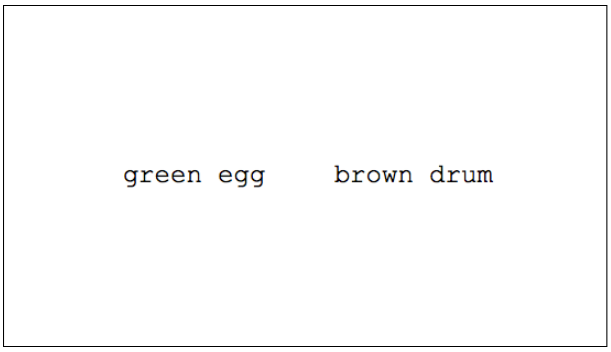
In each trial, a pair of two-word phrases were displayed on screen (see Figure 1), and after 250ms, the recordings were presented over headphones with 500ms of silence between the two. The participant's task was to highlight, using a button box, which words out of the four on screen sounded prominent. They could highlight as many or as few words as they liked, and there was no time constraint on their choices.[1]

Each pair of utterances was played twice, allowing the participants to double-check their marking before proceeding to the next trial. Each phrase pair either contained a lexically established contrast between first and second utterances (e.g. *blue ball*, *green ball*), and thus an implied contrastive adjective focus on the second adjective, or no contrast (e.g. *red house*, *green ball*). The pitch accent tune types in the first and second utterances were fully crossed, leading to 48 trials. Thus, this offered a total of 192 possible prominence markings, with an equal number of presentations of each pitch accent type; see Table 1.

Additionally, the pragmatic context of the phrases was manipulated between subjects. In the 'monologue' condition, the two phrases were presented one after another, with no intervening material during the 500ms interval. This way, the audio

[1] Participants also completed three other similar tasks; data from these other tasks are not analyzed in the current paper.

10.21437/SpeechProsody.2014-222



green egg   brown drum

Figure 1: *Example of visual presentation of orthographic representation of stimuli. Once a word was selected as prominent, it turned red.*

sounded like a monologue, a person saying a string of phrases. In the ‘plain dialogue’ condition, the two phrases were presented with a ‘connective’ utterance intervening between them. These connectives were extracted from the speech of the male Christmas tree decorator from the same corpus, and consisted of short utterances such as “okay, next”, “alrighty, next” and similar. In this condition, there was still a total of 500ms of silence between the first and second phrases.

Finally, the ‘narrative dialogue’ condition was exactly the same as the plain dialogue condition with the exception of the instructions to the participants. In this condition, participants were made fully aware of the provenance of the recordings and the purpose of the utterances (*viz.*, decorating a Christmas tree). They were presented with a short extract of the conversation between the instructor and the decorator, the visual materials used to elicit speech (see Figure 2), a diagram of the experimental apparatus and procedure, and the physical location the recordings were made. At the beginning of each block, they were instructed to imagine that they were physically present with both participants in the dialogue they heard, and to mark the words that sounded “important to the conversation”.

The experiment thereby constituted a 2×2×2×3 fully crossed design, with the relevant factors being pitch accent sequence of the first phrase ([H* !H*] vs [L+H* 0]), pitch accent sequence of the second phrase ([H* !H*] vs [L+H* 0]), contrast status of the pair of phrases (contrastive or non-contrastive), and pragmatic context of the utterances (monologue vs plain dialogue vs narrative dialogue). Pragmatic context was manipulated between subjects; all other factors were within subjects.

2.3. Participants

A total of 125 monolingual speakers of American English participated in the experiment for either partial course credit or a $10 payment. 43 participated in the monologue condition, 42 in the plain dialogue condition, and 40 in the narrative dialogue condition.

First phrase		Second phrase	
Adj	Noun	Adj	Noun
[H*	!H*]	[H*	!H*]
[H*	!H*]	[L+H*	0]
[L+H*	0]	[L+H*	0]
[L+H*	0]	[H*	!H*]

Table 1: *Summary of pitch accent sequences presented to participants.*

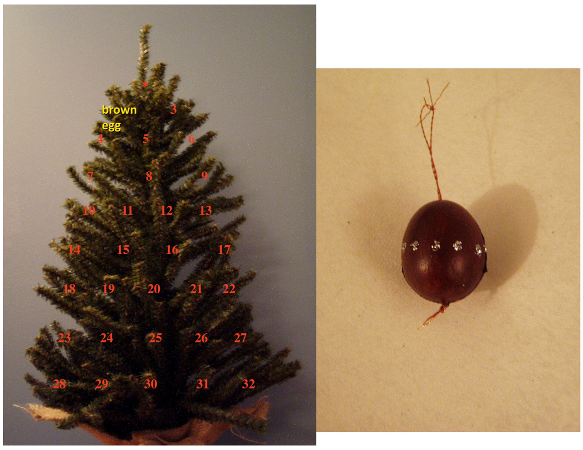


Figure 2: *Example visual materials from the original elicitation experiment [12] that was shown to participants in the narrative dialogue condition to demonstrate the illocutionary force of the extracts they will listen to. Depicted in this picture is a brown egg.*

2.4. Analysis

The results were analyzed using a mixed effect logistic regression model to predict whether or not the *second adjective* (the potentially putatively focused element) was marked as prominent. The model had fixed effects of the pitch accent sequence of the first phrase ([H* !H*] vs [L+H* 0]), pitch accent sequence of the second phrase ([H* !H*] vs [L+H* 0]), contrast status of the pair of phrases (contrastive or non-contrastive), and the condition (monologue vs plain dialogue vs narrative dialogue). All of the fixed effects were coded with sum contrasts, with the exception of condition which used treatment contrasts with the monologue condition as baseline. Additionally, since a number of participants reported during debriefing that they attended to the article *a* at the beginning of some phrases, the presence or absence of the article at the beginning of the first phrase and the second phrase were also added as fixed effects. Additionally, all possible three-way interactions between contrast status, 1st phrase pitch accent sequence, 2nd phrase pitch accent sequence, and condition were included, except for any interactions involving both of the pitch accent sequences.[2] Random intercepts of the second adjective lexical identity, second noun lexical identity, and subject identity were used, and a random slope for contrast status by subject.

In selecting the stimuli, we have relied upon ToBI transcriptions of the target phrases to classify them into groups for analysis. However, different coding systems exist, and it is theoretically possible to achieve different results from the use of different systems. In order to minimize this possibility, we sought independent phonetic evidence for classifying our stimuli, and so each stimulus phrase underwent acoustic analysis. For each phrase, measurements were made on both the adjective and the

[2] For clarity, the interaction term formula, in R-style syntax, was `contrast * condition * (1stpitchaccent + 2ndpitchaccent)`.

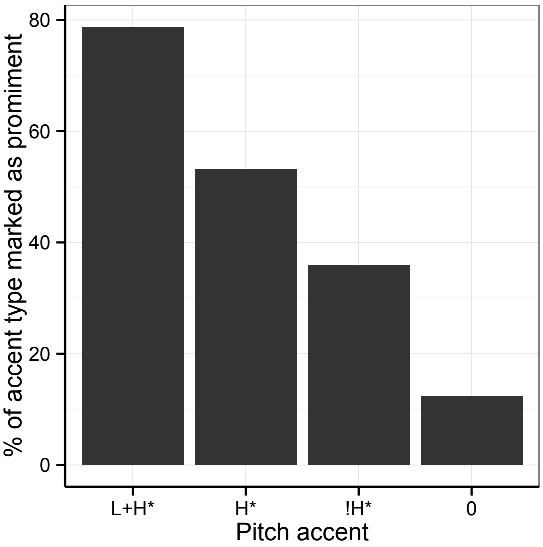


Figure 3: *Percentage of prominence endorsements on each type of pitch accent presented. 0 represents unaccented words.*

noun to extract the word duration, the vowel duration, the peak f0, the mean f0 within the vowel, and two measures of spectral tilt. The spectral tilt measures were the difference between the mean intensity of two different spectral bands, either 2kHz in bandwidth (i.e. 0-2kHz minus 2-4kHz) or 4kHz in bandwidth (i.e. 0-4kHz minus 4-8kHz). Additionally, two relative measures were taken which relate the adjective to the noun: the slope from the adjective peak f0 to the noun peak f0; and the slope from the adjective vowel mean f0 to the noun vowel mean f0. Each of these fourteen acoustic variables was entered as a predictor into a regression tree analysis [13] predicting the pitch accent sequence of the phrase (either [H* !H*] or [L+H* 0]). The resulting tree was pruned to minimize the cross-validation standard error, and of the predictor variables, only the f0 peak of the noun was found to act as a substantial cue to pitch accent. The tree correctly classified the phrases' pitch accent sequences 86.4% of the time (chance: 50%). This cue takes advantage of the relatively large f0 difference between !H* nouns and unaccented noun in our stimuli set. Therefore, with a small degree of error, it is possible to classify a phrase as [L+H* 0] if the noun peak f0 is low, and as [H* !H*] if the noun peak f0 is high.

A second mixed-effects logistic regression model was constructed, identical to the first except with the noun peak f0 measurements in place of the pitch accent transcriptions. This acoustically-based model allowed for a comparison of the phonological transcription with the acoustic details in their ability to account for the observed data. This comparison between the pitch-accent-based model and the acoustics-based model ensured that any observed effects were not simply artifacts of the transcription scheme or idiosyncrasies of the stimuli.

3. Results

Figure 3 depicts the overall prominence marking counts for each kind of pitch accent presented to the participants. This figure collapses over word position and condition; i.e., it depicts all of the prominence markings on all of the words in all of the conditions, split by pitch accent type. As can be seen, the endorsement rates depict a clear hierarchy of prominence, bolstering previous claims that the distinctions between these pitch accents in English (particularly between H* and L+H*) are mainly ones of prominence [14, 15, 16]. We now turn to the modeling results.

Effect	β	z	p
Intercept	0.922	2.228	0.026
PA2	1.676	10.462	< 0.001
Art1	-0.509	-6.220	< 0.001
Art2	0.358	2.232	0.026
Contrast × PA2	0.499	2.192	0.028
Contrast × PA2 × narrative	0.853	2.646	0.008

Table 2: *Summary of significant effects for the pitch-accent-based model. PA2 = pitch accent sequence of the second phrase; Art1 = presence of article in the 1st phrase; Art2 = presence of article in the 2nd phrase.*

The significant fixed effects of the pitch-accent-based model are summarized in Table 2. A significant effect of the pitch accent sequence of the second phrase was observed, such that adjectives with L+H* were more likely to be endorsed as prominent (78.9%) than those with H* (52.8%), as expected given previous research on accent type prominence (e.g. [14]). Two effects of article presence were observed: when the first phrase bore an article, the second adjective was significantly less likely to have a prominence marking (62.4%) than when the first phrase did not have an article (68.5%); similarly, when the second phrase bore an article, the second adjective was more likely to have a prominence marking (70.7%) than when there was no article (59.1%).

Additionally, an interaction between the pitch accent sequence of the 2nd phrase and contrast status was observed. This interaction is depicted in Table 3; when the sequence of phrases led to an implied contrastive focus on the adjective (e.g. *blue ball*, *green ball*), more prominence markings were observed on adjectives with a L+H* pitch accent. Interestingly, this effect of implied contrast was not observed on H* adjectives, suggesting that listeners did not interpret the [H* !H*] phrases with contrastive focus, despite the context.

Finally, and most crucially, a three-way interaction between contrast status, pitch accent, and pragmatic condition was observed. In contrastive contexts in the narrative dialogue condition, more endorsements were observed on L+H* adjectives when compared to the monologue condition. Essentially, the contrast effect for L+H* words (the two-way interaction mentioned in the preceding paragraph) is even stronger in the narrative dialogue condition. Recall that in the narrative dialogue condition, participants were encouraged to imagine themselves actually being present as the conversation took place. This effect is visualized in the right panel of Figure 4; note that in the other conditions, where the participants were not made aware

	H*	L+H*
Non-contrastive	54.1%	75.9%
Contrastive	51.5%	82.0%

Table 3: *Endorsement rates for second adjectives with different pitch accents in different contrast conditions.*

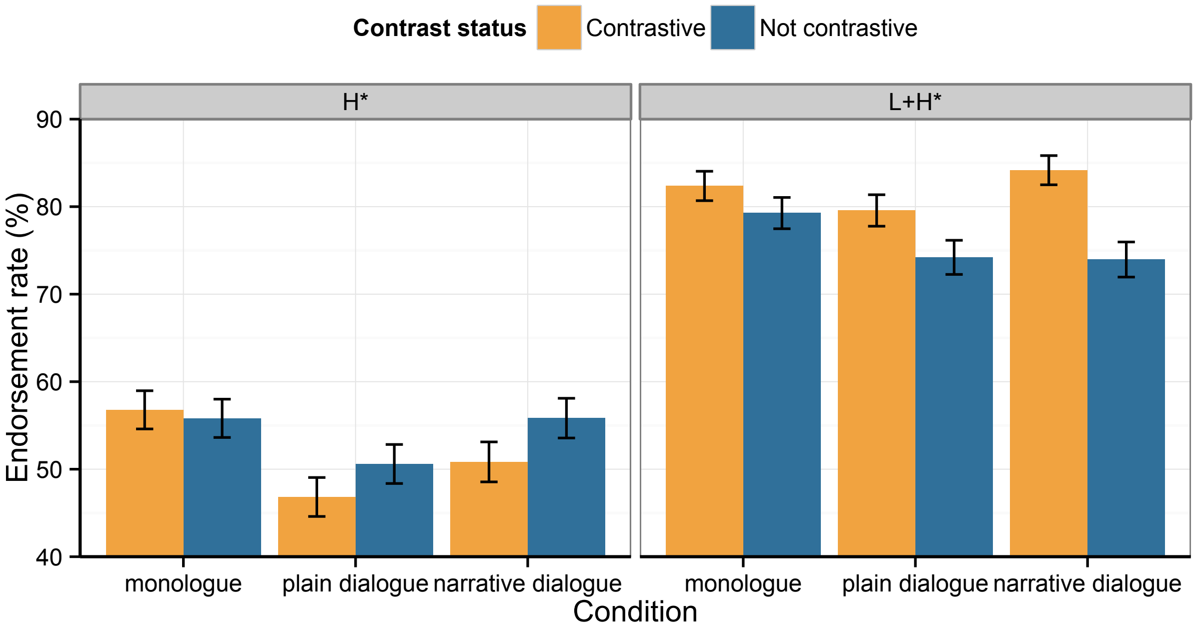


Figure 4: *Percentage of prominence markings on the second adjective of the phrase pair, broken up by pitch accent, pragmatic condition, and contrast status. Bars indicate standard error.*

of the discourse intent, the effect is much smaller.

The results of the acoustics-based model was comparable, in that the narrative dialogue condition sees contrast status effects that are not observed for the other conditions. See Table 4 for a summary of main effects. A likelihood ratio test comparing the two models confirmed that they did not differ in data likelihood ($p > 0.5$).

Effect	β	z	p
Intercept	9.721	8.835	< 0.001
Nf0	−0.040	−8.652	< 0.001
Art1	−0.440	−5.310	< 0.001
Contrast × Narrative	2.658	2.270	0.023
Contrast × Nf0 × Narrative	−0.012	−2.261	0.024

Table 4: *Summary of significant effects for the phonetic model. Nf0 = Noun peak f0; Art1 = presence of article in the 1st phrase.*

4. Discussion and conclusion

In addition to the expected effect where L+H* adjectives are marked as more prominent than H* adjectives, unexpected effects of article presence on the first and second phrases were observed. The trend in these effects appears to be that an article at the beginning of a phrase makes the adjective appear more prominent. During debriefing, some participants noted that they thought the article signaled “something really important coming up”, particularly when it was pronounced as [eɪ]. This intuition is supported by a model of the entire dataset, collapsing across conditions and phrase position, predicting prominence marking based on article presence and word class. The model revealed that in phrases that follow an article, adjectives are more likely to be marked as prominent ($\beta = 0.994$, $z = 16.884$, $p < 0.001$), while nouns are less likely to be marked ($\beta = -1.072$, $z = -13.859$, $p < 0.001$).

The pitch accent prominence hierarchy depicted in Figure 3 is particularly striking. However, care must be taken in its interpretation, since in this study all of the L+H*s and H*s were associated with phrase-initial adjectives, and all of the !H*s and unaccented words were phrase-final nouns. Nevertheless, our findings are consistent with the expected patterns of prominence in American English (e.g. [17]). The consistency of these findings also illustrate the fact that prominence is indicated in the same locations by the ToBI annotators of these stimuli, by our naïve participants, and by the results of the acoustic analysis.

Finally, our main result is an effect of discourse context evoked by elaborately illustrated task instructions. Only when participants were made fully aware of the intent of the discourse and instructed to imagine themselves as being physically present in the conversation was an effect of contrast status observed. This is to say, only when participants are able to construct a common ground with the interlocutors does their prominence perception reflect information-structural concerns. Prosodic prominence on its own can be perceived in extremely impoverished contexts (the monologue condition), but the information-structural notion of contrast requires an established discourse context before perception and interpretation is possible.

5. Acknowledgements

For comments: Speerlab discussion group. For partial funding: Buckeye Language Network Undergraduate Research Award.

6. References

[1] E. O. Selkirk, *Phonology and syntax: the relationship between sound and structure*. Cambridge, MA: MIT Press, 1986.

[2] D. R. Ladd, *Intonational Phonology*, 2nd ed. Cambridge: Cambridge University Press, 2008.

[3] J. Bishop, "Information structural expectations in the perception of prosodic prominence," in *Prosody and Meaning*, G. Elordieta and P. Prieto, Eds. Berlin: Mouton de Gruyter, 2012.

[4] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the percetion of prosodic prominence," *Laboratory Phonology*, vol. 1, pp. 425–452, 2010.

[5] C. Smith, "French listeners' perceptions of prominence and phrasing are differentially affected by instruction set," *Proceedings of Meetings on Acoustics*, vol. 19, p. 060191, 2013.

[6] K. Ito and S. R. Speer, "Anticipatory effects of intonation: Eye movements during instructed visual search," *Journal of Memory and Language*, vol. 58, no. 2, pp. 541–573, 2008.

[7] J. Buhmann, J. Caspers, V. J. van Heuven, H. Hoekstra, J.-P. Martens, and M. Swerts, "Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus," in *Proceedings of LREC*, 2002.

[8] J. Cole, Y. Mo, and S. Baek, "The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech," *Language and Cognitive Processes*, vol. 25, no. 7, pp. 1141–1177, 2010.

[9] K. Kohler, "The perception of prominence patterns," *Phonetica*, vol. 65, no. 4, pp. 257–269, 2008.

[10] B. M. Streefkerk, L. C. W. Pols, and L. F. M. ten Bosch, "Prominence in read aloud sentences, as marked by listeners and classified automatically," *Proceedings of the Institute of Phonetic Sciences*, vol. 21, pp. 101–116, 1997.

[11] M. Swerts, "Prosodic features at discourse boundaries of different strength," *Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 514–521, 1997.

[12] K. Ito and S. R. Speer, "Using interactive tasks to elicit natural dialogue," in *Methods in Empirical Prosody Research*, S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleink, N. Richter, and J. Schließer, Eds. Berlin: Walter de Gruyter, 2006, pp. 227–257.

[13] T. M. Therneau and E. J. Atkinson, "An introduction to recursive partitioning using the RPART routines," Section of Biostatistics, Mayo Clinic, Rochester, MN, Tech. Rep. 61, 1997.

[14] S. Calhoun, "The theme/rheme distinction: Accent type or relative prominence?" *Journal of Phonetics*, vol. 40, pp. 329–349, 2012.

[15] D. R. Ladd, "Metrical representation of pitch register," in *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, J. Kingston and M. E. Beckman, Eds. Cambridge: Cambridge University Press, 1990.

[16] D. R. Ladd and R. Morton, "The perception of intonational emphasis: continuous or categorical?" *Journal of Phonetics*, vol. 25, pp. 313–342, 1997.

[17] D. Büring, "On D-trees, beans and B-accents," *Linguistics and Philosophy*, vol. 26, pp. 511–545, 2003.