

JOURNAL ARTICLE

Individual differences and patterns of convergence in prosody perception

Joseph Roy¹, Jennifer Cole^{1,2} and Timothy Mahrt³

¹ University of Illinois, US

² Northwestern University, US

³ Aix-Marseille Université, FR

Corresponding author: Joseph Roy (jroy042@illinois.edu)

The challenge of prosodic annotation is reflected in commonly reported variability among trained annotators in the assignment of prosodic labels. The present study examines individual differences in the perception of prosody through the lens of prosodic annotation. First, Generalized Additive Mixed Models (GAMMs) reveal the non-linear pattern of some acoustic cues on the perception of prosodic features. Second, these same models reveal that while some of the untrained annotators are using these cues to determine prosodic features, the magnitude of effect differs quite dramatically across the annotators. Finally, the trained annotators follow the same cues as subsets of the untrained annotators, but present a much stronger effect for many of the cues. The findings show that while prosody perception is systemically related to acoustic and contextual cues, there are also individual differences that are limited to the selection and magnitude of the factors that influence prosodic rating, and the relative weighting among those factors.

Keywords: prosody; annotation; individual differences; generalized additive mixed models; inter-rater reliability; speech transcription

1. Introduction

Through the modulation of pitch, tempo, loudness, voice quality, and other properties of speech, prosody serves many functions in spoken language. There is striking variation in prosodic patterning across languages and dialects, reflecting grammatical differences in the prosodic structures that locate prominences and boundaries, in their tonal specification, and in their phonetic realization. Languages also vary in the linguistic function of prosody. For instance, many if not all languages are reported to use prosodic phrasing to structure speech into units that are morpho-syntactically or semantically delimited. Yet languages may differ in the alignment of prosodic phrases to morpho-syntactic units, in the size of prosodic phrases and the number of phrase level distinctions, or in the phonological features used to demarcate the edges of prosodic phrases (Jun, 2005, 2014; Selkirk, 2011). Similar variation is observed in the role of prosodic prominence in encoding focus. Some languages mark focus with prominence, while other languages do not, or use prominence in conjunction with syntactic devices for the expression of focus (Féry, 2013).

Understanding the characteristics and linguistic functions of prosody in a given language requires an analysis that relates prosody in production and perception. An understanding of prosody in production involves knowing how the acoustic correlates of prosody (e.g., measures of F0, intensity, duration, and spectral balance) pattern in relation to morpho-syntactic structure, semantic focus, and information status. Additionally, it is necessary to know how listeners perceive prosody as a function of the acoustic cues, and

how perceived prosody influences the comprehension of word, sentence, and discourse meaning. Prosodic annotation is one of the tools of prosodic analysis that is useful in this regard. A prosodic annotation represents a subjective interpretation of the prosodic substance of an utterance, typically that of a trained annotator, that transduces the acoustic speech signal onto a symbolic representation in terms of a small set of prosodic features. While prosodic data from other experimental methods are also used in prosody research, a distinct advantage of prosodic annotation is that it can in principle be used with any kind of speech data, to assess the full spectrum of prosody expressed in speech as it is produced in natural contexts of linguistic communication.

Prosody researchers have long relied on prosodic annotation to represent the cognitive, phonological specification that mediates the phonetic realization of prosody (e.g., in time-varying patterns of F₀, intensity, duration, spectral balance, and other acoustic quantities) and the linguistic meaning associated with that phonetic expression (see reviews in Wagner & Watson, 2010; Cole, 2015). Approaches to prosodic annotation vary, and the methods and best practices are still very much in development (D'Imperio et al., 2016), but all annotation systems include a set of features that mark prosodic prominence and a set of features marking prosodic phrase boundaries—properties of hierarchical phonological phrase structure (see Ladd, 2008 for details). The prosodic features that mark these structural properties are typically specified as tones. Prosodic annotation involves assigning a tonally specified boundary feature to a word that is initial or final in the prosodic phrase, and a tonal prominence feature ('pitch accent') to the word that is the head of a prosodic phrase, with optional prominence features assigned to one or more words preceding the head (details vary across languages—see Jun, 2005, 2014; Büring, 2016). Heads may be located adjacent to a prosodic phrase edge, or in languages like English and German, in other positions due to constraints relating prominence and semantic focus (Büring, 2016; Chafe, 1987; Féry, 2011, 2013; Ladd, 2008; Selkirk, 1995; *inter alia*). Despite the fact that prominence and boundary features are typically specified as tonal features, it is important to note their phonetic realization may involve acoustic correlates of stress other than pitch (see Cole, 2015 for overview).

Grammatical constraints govern the prosodic structures that underlie the assignment of prosodic prominence and boundary features, but depending on the language, these constraints may be probabilistic, giving speakers considerable flexibility in the specification of prosodic structure. For example, in English a speaker has the option of marking only the head of the prosodic phrase as prominent (e.g., *Georgia sent Randolph a letter*), or marking additional prominences on words preceding the nuclear prominence (e.g., *Georgia sent Randolph a letter*). Similarly, an English speaker can choose to parse a sentence as a single prosodic phrase with a boundary at the beginning and end of the sentence (e.g., [She met Randolph at the library]), or as a series of smaller phrases contained within a larger phrase, in which case there will be additional prosodic phrase boundaries internal to the sentence (e.g., [[She met Randolph] [at the library]]). Evidence that speakers exercise these options is found in several studies showing speaker variation in the prosodic structures and tonal features used in a given syntactic or discourse context and in the acoustic expression of prosodic features (e.g., Peppé et al., 2000; Grabe, 2004; Cole et al., 2007; Yoon, 2010; Cangemi et al., 2015).

Variation in prosodic patterning within and across speakers means that utterances with similar syntactic, semantic, and pragmatic properties may be realized with a variety of prosodic patterns, which can make the task of prosodic annotation difficult. The annotator must decide which prosodic structures and tonal features best characterize the prosody of an utterance on the basis of sometimes faint or ambiguous cues in the speech signal. Even trained annotators, working independently, often disagree on the prosodic label assigned

to individual words (Pitrelli et al., 1994; Syrdal & McGory, 2000; Yoon et al., 2004; Breen et al., 2012). The inherent subjectivity of prosodic annotation raises a concern about the reliability and the validity of any particular set of prosodic annotations.

Inter-annotator agreement is of key importance for establishing the reliability of an annotation system, and also for any analysis that makes use of annotated features. Yet agreement is a criterion that is costly to measure and to achieve. Annotators must undergo substantial training, typically over a period of many weeks, after which their annotations may be compared against other annotators. Prosodic annotation can be a very slow process, in large part due to the challenges mentioned above. For instance, annotation using the ToBI system (Beckman & Ayers, 1997; Jun & Fletcher, 2014) is reported to take from 100–200 times the duration of the speech audio file to complete (Syrdal et al., 2001), a number which is consistent with our experience as well. Reliability testing requires that a minimum of two (and ideally, more) annotators work independently to annotate the same speech materials. Thus, many dozens of hours of annotation effort are required to establish a minimum measure of reliability for an annotation system, before the system can be employed for annotation of a specific dataset. Even after an initial demonstration that an annotation scheme meets an established reliability criterion, the researcher must demonstrate that the annotation scheme is consistently and reliably employed in the annotation of their dataset. One way to do this is to have two or more annotators work together to assign prosodic features to some or all of the data, resolving disagreements through consensus or arbitration. Any such approach represents a substantial commitment of time and money on the part of the researcher.

The focus on resolving inter-annotator disagreement in prosodic annotation belies the important question of why annotators disagree, and whether there is any pattern to their disagreement. Inter-annotator disagreement may arise due to differences in the mapping from acoustic cues to prosodic features, reflecting individual differences in the perceptual sensitivity to a cue, or in cue weighting. Yet another possibility is that annotators differ in their sensitivity to contextual factors that predict the occurrence of a prosodic feature. Such differences would be unsurprising in light of the noted variation in the prosodic patterns produced by speakers, and the possibility that listeners may differ in their prior exposure to those variable patterns. An annotation practice that resolves annotator differences fails to acknowledge the possibility that inter-annotator disagreement may reflect genuine differences among listeners in the prosodic patterns they perceive for a given utterance—differences that could have consequences for comprehension of sentence and discourse meaning.

The present study examines individual differences in the perception of prosody through the lens of prosodic annotation. Our primary goal is to explore the patterning of individual differences in prosodic annotation in relation to the presence of acoustic cues and other properties of the linguistic context. A related goal is to explore annotation reliability among a larger group of annotators than is typically employed for research purposes. We are further interested in discovering how prosodic annotation differs as a function of annotator training (comparing trained vs. untrained annotators), time allowed for annotation, and depending on whether annotation is informed by visual inspection of the speech waveform, spectrogram, and pitch track. For this we compare ‘expert’ ToBI annotations, which are commonly used in prosody research, with annotations from ordinary listeners using the Rapid Prosody Transcription (RPT) system, as described further below.

This study contributes to a growing interest in individual differences in the study of prosody (e.g., Dillely & Heffner, 2013; Ladd et al., 2013; Cangemi et al., 2015; Bishop, 2017). We believe that knowledge about individual differences in prosody perception is important for the use of prosodic annotation in linguistic research, since individual

differences related to perception may be a source of inter-annotator disagreement. A further motivation for this work is to understand the nature and extent of individual differences in the perception of prosody in spontaneous, conversational speech of the sort we experience in everyday life. Ultimately, an understanding of individual differences in the perception of prosody may aid in understanding differences in language processing, comprehension, and communication style.

1.1 Exploring individual differences in prosody

A typical experimental setup is meant to mitigate or control for individual differences to examine a treatment effect or group effect. Traditional statistical models treat participant effects as ancillary to the fixed effects component of the model. The study presented below inverts this traditional focus to explore patterns of individual differences in the set of acoustic and contextual variables that inform prosodic ratings of words in conversational speech. We examine prosodic ratings from a prosodic annotation task with untrained listeners, which are also compared with a consensus labeling from a pair of trained annotators using the ToBI system. The relationship between these prosodic ratings and predictor variables is analyzed using regression models to explore patterns of inter-annotator agreement, and the relationships that hold between prosodic ratings and cues that are present in the acoustic signal and in the broader linguistic context of the utterance.

This exploratory analysis is organized around four questions. The first three seek to establish that there are systematic patterns of agreement and disagreement among annotators (trained and untrained) in the prosodic rating of words in our speech sample. Questions 1–3 have been explored in our prior work (Cole et al., 2010a, b), but are examined again here with an expanded focus. Moreover, we must establish systematicities in patterns of inter-annotator agreement for our present dataset to merit the further exploration of individual differences in those data, which is the focus of Question 4.

- [Q1] Are untrained annotators, lacking experience with the metalinguistic task of annotation, able to rate the prosodic prominence and boundary status of a word in a systematic fashion, as reflected in above-chance measures of inter-annotator agreement? An extension of this question is whether prominences and boundaries are rated with similar reliability, assessed again through inter-annotator agreement. This question is motivated by findings from our prior RPT study (Cole et al., 2010a, b) showing higher agreement for boundary ratings compared with prominence, which however is at odds with findings of a roughly equal level of agreement for boundary and prominence ratings in prior studies with ToBI annotation (Breen et al., 2012). We also explore how the pattern of inter-annotator agreement varies according to the number of annotators who are compared.
- [Q2] Are there more words for which annotators agree on the presence vs. absence of prominence (or boundary) compared to words where they disagree, or alternatively, are there roughly equal numbers of words at each possible level of inter-annotator agreement? This amounts to a question about the distribution of words in our dataset over the range of possible agreement values. Our interest here is in probing the salience of a binary distinction for prominence, and similarly for boundary. If annotators perceive a clearly marked, binary distinction for the presence/absence of prominence (or boundary) in the speech signal, then we expect a majority of words to have high inter-annotator agreement. On the other hand, if prominence (or boundary) distinctions are not clearly perceived due to ambiguous or faint cues in the speech signal, we expect a lower overall rate of agreement and many words with intermediate agreement values.

- [Q3] What is the relationship between prosodic rating and the presence of acoustic and contextual cues for the prosodic ratings from our pooled annotators? We want to know if the same set of acoustic and contextual factors shown to cue the perception of prosodic features in prior studies are also serving as prosodic cues for the untrained annotators in our study, as representatives of a larger population of listeners. We investigate the influence of acoustic and contextual cues taken together, and their individual influence on prosodic rating.
- [Q4] Are there individual differences among annotators in which acoustic or contextual factors serve as cues for prosodic rating, or in the weighting of those cues? Here we investigate individual differences in perceptual processing as a possible basis for inter-annotator disagreement in prosodic rating. If disagreement arises from differences in cue selection or cue weighting, we predict systematic patterns in individual differences. Specifically, we predict that individual listeners may show a null effect for a given predictor, or a difference in the magnitude of the effect, compared to the overall pattern over the group of annotators, but qualitative differences in the effect pattern of individual annotators are not *expected*. We reason that individual differences should not fundamentally restructure the mapping from a phonological prosodic feature to its phonetic expression, or between a prosodic feature and a discourse meaning. An extension of this question explores similarities among annotators in the patterning of individual differences in cue selection and weighting.

Following the presentation of our Methods in Section 2, we present results from a prosodic annotation task in Section 3, which are discussed in Section 4 in relation to our four exploratory questions. The paper ends with a brief conclusion and prospectus for future work in Section 5.

2. Methods

2.1 Participants and materials

Perceptual ratings of prosodic boundaries and prominences were obtained from 32 students at the University of Illinois who self-reported as monolingual speakers of American English with no deficits in hearing or reading.¹ These participants² produced prosodic ratings of words in a sample of American English from the Buckeye Corpus (Pitt et al., 2007). The speech sample consists of conversational speech excerpts from the narrative (monologue) speech of 16 different speakers, with each speaker's excerpt presented in a separate audio file (duration range 13–24s; average 18s), for a total of 932 words over all excerpts (290s total; average 18.13s per speaker).

2.2 Prosodic transcription with rapid prosodic transcription and ToBI

The method of Rapid Prosodic Transcription (RPT) was used to elicit prosodic ratings (i.e., annotations) of each word (Cole et al., 2010a; Cole et al., 2010b). Transcription tasks were administered using a custom web-based presentation and annotation tool (LMEDS: Mahrt, 2016). Participants listened to audio files presented through headphones while seated at a computer monitor. The text for each audio file was modified to remove all

¹ The prosodic annotation data reported in this paper are part of a larger study, in which prosodic annotations were additionally collected from remote annotators using the crowd-sourcing platform of Mechanical Turk. See Cole et al. (2017) for a comparison of the annotations from the 'lab' study reported in this paper and crowd-sourced annotations from annotators in the US and India.

² The 'participants' in this study are our prosodic annotators. We favor the term 'participant' in the context of a statistical model, and 'annotator' when matters related to prosody perception and annotation behavior.

punctuation and capitalization as potential orthographic cues to prosodic phrasing. The text was presented in the LMEDS browser window in 12-point black font, such that the complete text transcript of the audio file could be viewed in one screen view. Participants were not shown any graphical display of the speech signal (F0 trace, waveform, or spectrogram), nor did they receive any training or coaching on the acoustic encoding of prosody. Participants rated prosodic boundaries and prominences in real time while listening to the audio file, in two separate transcription tasks with the same audio files. Participants were given the very brief instructions shown in (1) and (2), and were asked to use the mouse to click words in the speech excerpt that they heard as preceding a boundary (task 1) or as prominent (task 2). We refer to the selection of a word via mouse click as ‘marking,’ a term that is used synonymously with prosodic ‘rating’ in this paper.

- 1) Instructions for boundary marking: “Speakers break up utterances into *chunks* that group words in a way that helps the listener interpret the utterance. You will mark locations where you hear a boundary between two chunks of speech. Note that chunks can vary in size, and boundaries do not necessarily correspond to locations where you would place a comma, period, or other punctuation mark, so you must really listen and mark the boundary where you hear a juncture between chunks.”
- 2) Instructions for prominence marking: “In normal speech, speakers pronounce some word or words in a sentence with more *prominence* than others. The prominent words are in a sense highlighted for the listener, and stand out from other non-prominent words. Your task is to mark words that you hear as prominent in this way.”

A vertical bar marking the prosodic juncture appeared following each word selected in the first task as preceding a boundary, and these junctures remained visible throughout the second task (**Figure 1**). Words selected as prominent in the second task immediately appeared in red font.³ Participants were required to listen to an excerpt twice in each task, and they could modify their selection of a word for boundary or prominence marking by repeat selection of the word with the mouse click (where each selection toggled the prosodic

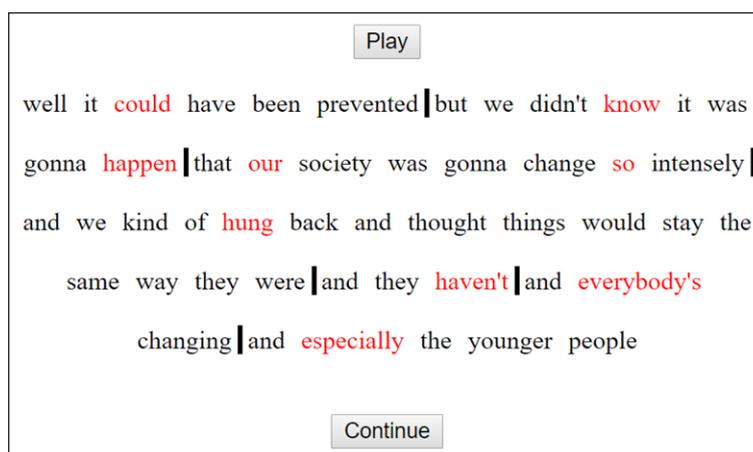


Figure 1: Example of LMEDS screen view in the prominence annotation task that followed boundary annotation of the same excerpt. See text for details.

³ The sequencing of boundary marking followed by prominence marking was adopted on the basis of trial runs showing that initial chunking of the speech helps constrain the listener’s focus for the more challenging task of prominence marking. Comparison of RPT performed with both sequences shows no significant differences in the overall patterning of results.

marking on or off), as many times as they wanted, but without being able to stop or restart playback of the audio file. Participants were given one practice item, and no feedback on their annotations.

On completion of the annotation task every word in the speech sample was coded with a “0” or “1” value for prominence, and with a “0” or “1” value for the presence of a following boundary (i.e., coding a word as final or non-final in a perceived chunk). This coding was performed for each individual participant, resulting in a total of 32 prominence ratings and 32 boundary ratings for each word in the dataset.

An independent consensus prosodic annotation was carried out by 2 trained ToBI labelers (including the second author) using the full inventory of ToBI tone labels for pitch accents, phrase accents, boundary tones, and break indices (Veilleux et al., 2006). The ToBI annotations were performed in the conventional fashion, based on auditory impression together with visual inspection of the speech display (in Praat), and with annotators exercising full control over audio playback and display settings such as zoom, and formant and intensity contour overlays.

2.3 Predictors

2.3.1 Selection of factors analyzed and predicted effects on prosodic ratings

We examine a set of factors that predict variation in prosodic annotation for the content word subset of the annotated dataset. Closed-class function words (e.g., determiners, prepositions, conjunctions) were almost never rated as prominent nor as followed by a prosodic boundary in this corpus, so the part of speech labels alone predict the prosodic status of function words with high accuracy.⁴ Removing function words from the corpus leaves a subset of 477 content words, which as a class are more variable in their prosodic status. The content word subset is used for the regression analyses presented below.

Seven factors were examined as predictors of prosodic rating, as listed in **Table 1**. Four ‘acoustic factors’ were chosen based on prior studies showing that phone duration (as just one among several temporal measures of local speech rate), between-word pause duration, intensity, and F0 are signal correlates of prosodic prominence and prosodic phrase boundaries in American English.⁵ Predicted effects of acoustic factors are as follows. Prominence marking is predicted for words that exhibit a local peak in rate-adjusted phone duration, intensity, and F0. Boundary marking is predicted for words with lengthened phone duration, words followed by longer pauses, and words with lower intensity.⁶ Measurement methods for all predictors are discussed below.

Part of speech and Word Frequency were chosen as ‘contextual (i.e., top-down) factors’ on the basis of which a listener may expect to encounter a prosodic boundary or a prominent word due to grammatical rules or constraints that govern prosodic phrase structure and the distribution of prosodic features associated with those structures. Prominence marking is expected for words that introduce new information to the discourse (Halliday, 1967; Ladd, 1980, 2008; see Baumann & Reister, 2012 for an overview), so we predict that less predictable words, e.g., words with low unigram frequency, are more likely

⁴ Function words are not considered in the present analysis, though we acknowledge that the assignment of corrective or contrastive focus to a function word (an infrequent event) may result in the association of the function word with prominence.

⁵ For acoustic correlates of boundaries see Lehiste (1973); Wightman et al. (1992); Kraljic and Brennan (2005); Kim et al. (2006). Acoustic correlates of prominence are reported in Sluijter and van Heuven (1996); Kochanski et al. (2005); Wang and Narayanan (2007); and acoustic correlates of information status or focus (understood here as determinants of prominence) are reported in Xu and Xu (2005); Breen et al. (2010); and Katz and Selkirk (2011). For recent overviews see Wagner and Watson (2010); Cole (2015).

⁶ F0 correlates of boundary are also expected, but may include both rising, falling, and plateau contours, which combined would cancel one another in statistical analysis. Further analysis of F0 correlates of both prominence and boundary is planned using contour stylization and clustering of contour shapes, following Reichel (2014).

Table 1: Predictors of prosody ratings used in regression models. All predictors are measured at the word level, for each content word in the speech corpus.

Predictor	How measured	Expected effects on prosodic rating (details in Section 2.3.1)
Local intensity (RMS)	RMS intensity of primary stressed vowel, z-transformed in window of 5 primary stressed syllables centered on target syllable.	High intensity expected to increase prominence rating; low intensity expected to increase boundary rating.
Local Max F0 (Hz)	The maximum of the log-F0 in the primary stressed vowel, measured using autocorrelation with floor of 70 Hz and ceiling of 250 Hz, z-transformed in window of 5 primary stressed syllables centered on target syllable.	High local max F0 expected to increase prominence rating; no uniform effect of F0 on boundary rating is expected.
Word Frequency	Log (unigram) frequency of word based on its occurrence in the Switchboard corpus.	Low word frequency expected to increase prominence rating; no predicted effect on boundary rating.
Word Phone-rate (phones/second)	A local rate-dependent measure of word duration that takes into account the number of phones in the word. Low values indicate slowed local tempo.	Low phone-rate expected to increase prominence and boundary rating.
Post-word Pause Duration (s)	An interval of (near-)zero energy following the word. Not measured for turn-final words.	Longer pause predicts increased boundary rating; no expected effect on prominence.
Part of Speech (POS)	Manually annotated for 15 POS labels; results reported for N, Adj, Adv, and Vb.	No predicted effect on prominence; increased boundary rating weakly expected for verbs, predicate adjectives, nouns and adverbs.
Boundary Marked (yes/no)	Codes an annotator's rating of boundary in relation to his/her rating of prominence for the target word. Only used in prominence model.	A word that is rated as pre-boundary is more likely to be rated as prominent than a word that is not pre-boundary.

to be marked as prominent than words that are more predictable, e.g., those with high unigram frequency (Pan & McKeown, 1999; Calhoun, 2010). Furthermore, words that are typically used to introduce new referents into the discourse (e.g., adjectives and nouns) are more likely to be marked as prominent than words that are less likely to introduce new referents, such as verbs and function words (Hirschberg, 1993; Chen & Hasegawa-Johnson, 2004). Certain lexical items frequently occur with prominence, e.g., adverbs used as intensifiers (*extremely, incredibly*), or adjectives used in emphatic descriptions (*ecstatic, astonishing*), and the frequent occurrence of such words in the corpus could boost the prominence profile of the entire adverb or adjective category. Effects on prominence rating due to intensifiers or emphatic adjectives may be reflected in the part of speech effects reported here, but they are not further investigated. Boundary marking is expected to be only weakly influenced by Part of Speech. Syntactic phrase structure rules determine which part of speech classes can occur in the initial or final position of syntactic constituents, so to the extent that prosodic phrases align with major syntactic constituents, we can derive predictions about the position of words in prosodic phrases. The strongest

prediction concerning Part of Speech is that function words are less likely than content words to be in the final position in a prosodic phrase, but among content words, which are our sole focus of analysis, we have no strong prediction. In phrase-final position we expect to find verbs (for intransitive sentences), predicate adjectives, nouns (for transitive sentences), and post-verbal adverbs. All of these categories are also licensed in a variety of non-final positions, weakening the predictive power of Part of Speech for boundary marking. We are not aware of any basis for predicting that Word Frequency plays a role in boundary marking.

For prominence marking we also consider boundary marking of the same word as a predictor. We expect a relationship between prominence and boundary rating due to the patterning of nuclear prominence in English. In English utterances with no marked focus (i.e., in sentences with broad focus), the rightmost content word in a prosodic phrase is marked with an obligatory prominence-lending pitch accent.⁷ In many sentences the word with nuclear prominence is also the final word in the prosodic phrase, resulting in the expected co-occurrence of prominence and boundary marking on the same word.

It's important to note that we do not maintain that the seven predictors of prominence and boundary rating examined here are the only factors to influence prosody perception in American English, nor that they are a sufficient subset to model listener response patterns. A comprehensive study of prosody perception would need to examine other acoustic measures, such as sub-band frequency (Sluijter & van Heuven, 1996), measures of F0 contours (Möhler & Conkie, 1998; Reichel, 2014), and measures of irregular pitch periods (Redi & Shattuck-Hufnagel, 2001), along with features related to syntactic phrase structure (Speer et al., 2011), 'accentability' (Calhoun, 2010), and the referential status of expressions (Ito & Speer, 2008; Baumann & Reister, 2012). Nonetheless, though the predictors examined here are clearly not exhaustive, they do define an informative subspace to begin exploring individual differences in the perception of prosody.

2.3.2 Measurement and labeling methods for selected predictors

Acoustic measures of duration, F0, and intensity were extracted in regions defined by the phone and word alignments distributed with the Buckeye corpus, as follows. The effects of prosody on measures of acoustic duration are captured here using Pfitzinger's (1998) measure of local phone rate, which tracks local changes in tempo that result in lengthening or shortening of a phone.⁸ Unlike an absolute measure of word duration, the phone-rate measure is robust to differences in speaking rate across speakers and utterances, and to the phonological length of the word. We have also confirmed that the local phone-rate measure is independent of Word Frequency, which is not true for an absolute measure of word duration. Although the phone-rate measure can be used to calculate a local-rate-dependent measure of word duration, we use the simpler phone-rate measure as the 'Duration' predictor in the regression analyses presented below. The reader should bear in mind that the expected effect of prosody on local phone-rate is opposite of its effect on absolute word duration: An increase in local phone-rate yields a decrease in word duration (faster tempo leads to shorter words), while a decrease in phone-rate yields an increase in word duration (slower tempo leads to longer words).

⁷ More specifically, nuclear prominence is assigned to the rightmost content word in the rightmost argument following the verb in transitive sentences. In intransitive sentences the nuclear prominence does not as consistently occur in final position, with the nucleus assigned either to the subject or the verb, depending on semantic conditions (Gussenhoven, 1984; Ladd, 2008, pp. 223–251).

⁸ Pfitzinger's (1998) algorithm creates a phonerate curve over the entire utterance (in 500 ms windows with time step of 10 ms.). The mean and variance of these values across the utterance are used to normalize the measure of word phonerate, calculated by dividing the number of phones in a word by the word's duration and then z-transforming that value using the phonerate mean and *SD*.

The duration of silent pause following the target word was measured for all words except the final word in a speaker's turn, based on a criterion of energy at or near zero. This pause duration measure is labeled 'Post-pause Duration.' We note that the pause duration measure may include the silent closure interval of a voiceless stop at the end of the target word, or a similar voiceless closure interval at the beginning of the following word. It is not possible to distinguish silent stop-closure intervals from silent pause (with no corresponding articulatory closure) based only on acoustic evidence.

Intensity was calculated as the RMS intensity of the vowel that bears primary lexical stress (based on dictionary specification). F0 was extracted from Praat's autocorrelation method with floor and ceiling set to 70 Hz and 250 Hz, respectively, for speech samples from male speakers, and 100 Hz and 300 Hz, respectively, for female speakers, following Vogel et al. (2009). Manual testing of parameter settings for a subset of our speech sample confirmed that these floor and ceiling values minimized halving and doubling errors in the F0 estimation, while also capturing most if not all of the F0 maxima and minima in vowel regions in the speech sample. Max F0 was calculated from the extracted F0 values as the maximum of the log-transformed F0 in the interval of the vowel bearing the primary lexical stress. The intensity and F0 values were locally normalized using a z transform in a window of five stressed syllables centered on the target syllable, which gives measures of the difference in F0 peak and overall intensity of the target word's stressed syllable relative to the stressed syllables in its nearby left and right context.

Words were manually annotated with their part of speech from a set of 15 part of speech labels. Results are reported here only for the content words, comprising nouns (132 word tokens in the corpus), adjectives (59 words), adverbs (92 words), and verbs (194 words). Function word categories and their counts are presented in Appendix B.

Word Frequency was calculated as the log frequency of each word as it occurs in the Switchboard corpus of telephone conversation speech (Godfrey et al., 1992). The Switchboard corpus is similar to the Buckeye corpus in speech genre (spontaneous, conversational speech between unfamiliar partners), but Switchboard is a much larger corpus with approximately 240 hours of speech from over 500 speakers, and therefore provides a more robust estimate of Word Frequency in the spoken language.⁹

The seven predictor variables listed in **Table 1** were used in each of the statistical analyses reported below.

2.4 Statistical Methods

As a preliminary to the analysis of individual differences in prosody perception, we establish the degree of inter-annotator agreement in the prosodic rating of words presented in the RPT tasks. The overall level of agreement among the 32 untrained annotators across all annotated materials is calculated using Fleiss's kappa from the irr library in R (Gamer et al., 2014). Fleiss's kappa statistic measures the observed agreement in relation to the expected level of agreement by controlling for the overall frequency of each label (here, 0 or 1 for each of prominence and boundary). A Fleiss' kappa statistic value of 1 indicates perfect agreement and values at or below 0 indicate chance agreement or disagreement.

Our primary interest in this study is stated in research question 4 (Q4, from Section 1), concerning differences among individual listeners in the factors that influence their

⁹ Measures of word frequency may vary somewhat across corpora, reflecting different patterns of language usage in relation to dialect, genre and social factors. For words from the Buckeye subset studied here, we have compared word frequency values based on the entire Buckeye corpus (approximately 20 hours of speech), and based on the Switchboard corpus (approximately 240 hours). Spearman's correlation coefficient for these two frequency measures of our data is $R = 0.89$, which supports the use of Switchboard as the reference dataset for our frequency analysis.

perception of prosodic boundaries and prominences. For example, do some listeners attend more to F0 cues, while others attend more to temporal cues, or intensity? Or, do listeners vary in the relative weighting of acoustic cues vs. contextual factors? These questions are explored through regression analyses testing the contribution of each predictor variable to prosodic marking. All continuous predictors are normed by `scale()` in R, transforming them to z-scores. There are two motivations for doing this: First, it helps the computational algorithms converge if the data is scaled; second, and more importantly, in logistic regression (and GAMM extensions) it allows for direct comparison of all effects in terms of a standard deviation change in the predictor and that predictor's estimated effect in the likelihood of the response (here, prosodic marking). We use a generalized additive mixed effects model (GAMM) (Wood, 2006, 2011) in order to assess these questions, with a random intercept for (lexical) word and random smooths for each predictor, by participant.

GAMMs allow us to simultaneously account for non-linearity in the predictor's effect on the dependent variable (here, binary prosodic marking) and account for the non-linear deviations of participants from that overall effect. Random intercepts and slopes are fairly common in regression methods used in linguistics (Baayen et al., 2008; Gries, 2015 *inter alia*),¹⁰ but random smooths are not as common. The motivation for a random smooth is similar to random intercepts and slopes, except that with a random smooth, a researcher is accounting for the difference in each participant's effect for a fixed effect in a non-linear manner. A further benefit of the GAMM is that we can test for the significance of both our main effects and the random smooths (or the estimated differences for each participant). A GAMM also provides the percentage of deviance explained, which is a generalized measure of R^2 for non-normal response data (Wood, 2006, p. 84), and a useful measure for determining how much the prosodic ratings are explained by the information in the predictors.¹¹

We think GAMMs are well suited for modeling prosodic annotation due to likely sources of nonlinearities in the association between predictor variables and prosodic feature assignment. First, there may be gradation in the degree of prominence for a given word that is related to the specific type of information status or focus of the word, or similarly, gradations in the degree of prosodic juncture for a word related to its specific syntactic context. Such distinctions provide a source of possible non-linearities in the association between

¹⁰ In fact, our original analysis of this data included a generalized linear mixed effects model: However, a model with participant slopes failed to converge for the data forcing us to consider non-linearity in the data (via GAMMs) as a cause.

¹¹ Alongside the benefit of modeling non-linear random effects of individual participants, we must acknowledge some potential drawbacks of GAMMs. First is the difficulty in the inclusion of interactions of smoothed predictors and their interpretation in a non-linear space. Further, the analogue of collinearity in these models, concurvity, is one that is difficult to describe, let alone assess. As Ramsay et al. (2000) note, the mathematical formalism underlying concurvity is not intuitive for researchers. A good intuitive definition comes from Wood's function help page in `mgcv` library for the concurvity function: "Concurvity occurs when some smooth term in a model could be approximated by one or more of the other smooth terms in the model." Concurvity can be measured between pairs of predictors by the `mgcv` library and the measure it produces varies between 0, no concurvity, to 1, high concurvity. Another potential drawback of GAMMs is more general and related to the use of statistical models in a Null Hypothesis Significance Testing (NHST) framework where the purpose of the statistical analysis is to assess the likelihood of a non-zero (in most cases) or more extreme effect estimate under the assumption there is no effect. With GAMMs, the alternative hypothesis (i.e., the hypothesis being tested) is not just an increase or decrease of our dependent variable with respect to a variable of interest, but a wide range of possible functions. It is difficult to frame a specific hypothesis for a smooth function: The null hypothesis (that there is no change in y with respect to x) is held against the alternative hypothesis that there is a change in y with respect to x , while allowing that the change may or may not flip direction and magnitude and there may even be statistical overlap with the null effect across the full range of x . While confronting non-linear effects in the data is an important feature of GAMMs, the problems of overfitting the data, concurvity between predictors and not being able to assess confirmatory hypotheses in the same way should be taken into account when building a model by building higher order interaction terms into the model if possible.

acoustic cues and prominence or boundary label for a given word. Second, a continuous predictor variable (including all the acoustic measures) may have a multi-modal distribution over its range, which could be another source of non-linearity in the association of the predictor with prosodic marking. For example, we examine a word’s maximum F0 value as a predictor of its perceived prominence (through pitch accenting) and also as a predictor of boundary marking tones. But the maximum F0 of a word is itself dependent on many factors, such as the position of the word in the prosodic phrase (determining the effect of declination), the loudness of the phrase, and the height of the vowel in which F0 is measured. With these known influences on F0, it is difficult to a priori predict linearity in the effect of F0 on prosodic marking. Finally, even if we set aside the question of whether the predictors have an overall linear relationship with the probability of prosodic marking, there is no evidence from prior research that individual differences in the effects of a predictor (relative to the overall effect) are linear, nor is there any basis for predicting linear effects of individual differences over the range of each predictor (i.e., the ability to be captured in a random slope rather than a random smooth, as discussed below).

Figure 2 represents the visualization of simulated data fit with a linear mixed effects model with a random slope and intercept for a predictor, x1 (in R syntax: `lme1 = lmer(y ~ x1 + (1 + x1|subject), data = newDat)`). This model allows for subject variation with respect to the intercept and slope for a variable as seen in the fitted line (in blue) for each subject.

In **Figure 3**, we have an example of random smooths by a predictor, x1, for subject (in R syntax: `gam1 = gam(y ~ s(x1) + s(Subject, x1, bs = “fs”, m = 1), data = newDat)`). A regular generalized additive model allows for a non-linear relationship between y and x1. A random smooth is an extension of random slopes and intercepts seen in **Figure 2** and is meant to account for non-linear deviations from the overall fit by subject.

In the two GAMMs presented below, a random smooth for participant of each fixed effect is included in the model. These random smooths will allow us to explore both overall cues used in the selection of prosodic marking and individual differences in the relationship between the cues and prosodic marking. GAMM results below are displayed with visualizations produced using the visreg library in R (Breheny & Burchett, 2017).

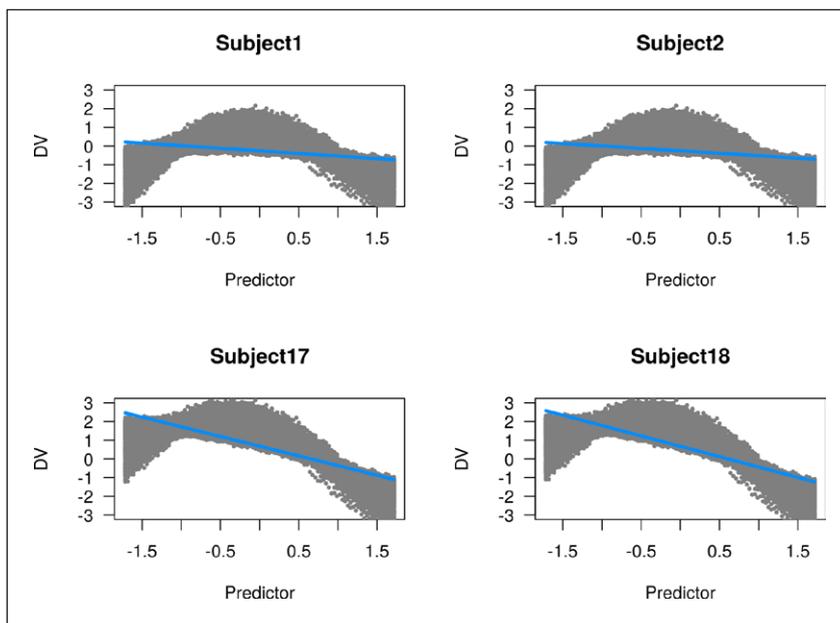


Figure 2: Example of random slopes and intercepts for subjects.

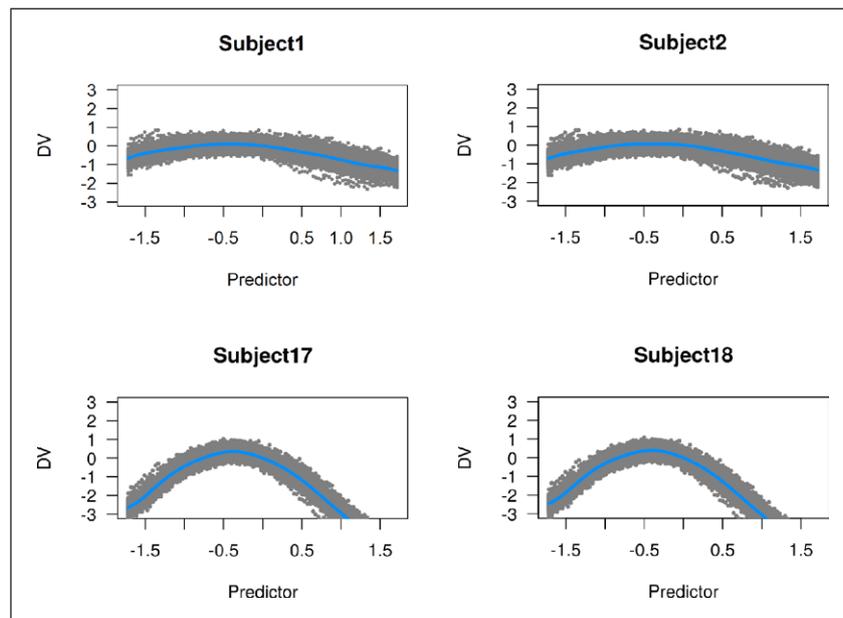


Figure 3: Example of random smooths for subjects.

3. Results

Below we present results from the analysis of boundary and prominence ratings from 32 annotators and 1 ToBI annotation. The ToBI annotation counts 276 accented words from the total speech sample of 931 words (including function words, this is 42.1% of the total) and 179 words marked for boundary features (19.4% of total). This yields an average ratio of 1.5 prominences for every boundary rating. In the ratings from 32 untrained annotators over just the 477 content words, there are a maximum of 15264 possible boundary marks and the same number of possible prominence marks (with every word marked by every annotator). Against this maximum, the observed number of boundary marks is 2198 (14.4%) and the observed number of prominence marks is 3979 (26.1%). This gives an average ratio of approximately 2 prominences per interval demarcated by prosodic boundaries.

3.1 Inter-annotator agreement

In order to assess overall agreement of the 32 untrained annotators across the annotations, we calculated multi-rater agreement rates using Fleiss's kappa from the irr library in R (Gamer et al., 2012). Fleiss's kappa statistic takes values that range from 0 to 1, with zero for the minimal level of agreement and 1 for perfect agreement. The annotators have a kappa of .52 and .28 for boundary and prominence marking, respectively. This indicates a high-moderate and low-moderate level of agreement between annotators for boundary and prominence marking, respectively, but whether or not the annotators are using the same cues even when they agree on an annotation has to be further explored.

3.2 Distribution of prosodic scores by predictors

This section reports on the relationship between the distributions of acoustic and contextual cues, and prosodic ratings. **Figures 4** and **5** display plots showing the relationship between the distribution of predictor variables and RPT ratings for boundary and prominence, presented as the average rating over the group of 32 annotators. Specifically, for every content word in the database, the b-score (boundary) and p-score (prominence) are calculated by the proportion of annotators that identified that word as preceding a boundary or as prominent. In **Figure 4**, the acoustic predictors are plotted by the b-scores

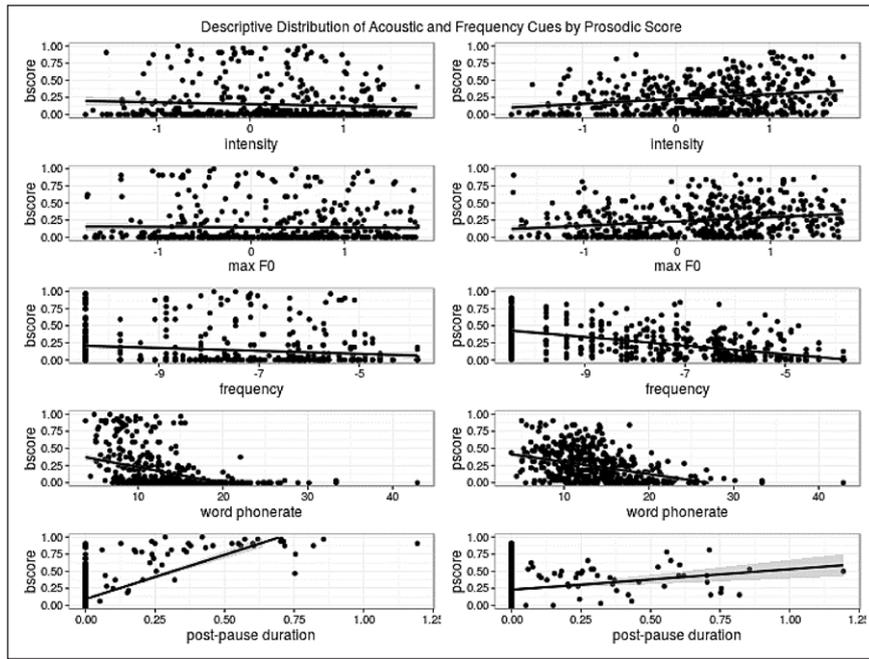


Figure 4: Distributions of average boundary (left) and prominence (right) ratings across range of each continuous-valued predictor variable. Average boundary and prominence ratings are plotted on the y-axes, with range [0, 1], based on the ratings of all annotators for all words. The x-axes show the range of observed values (z-normalized for Intensity and max F0) for each continuous predictor.

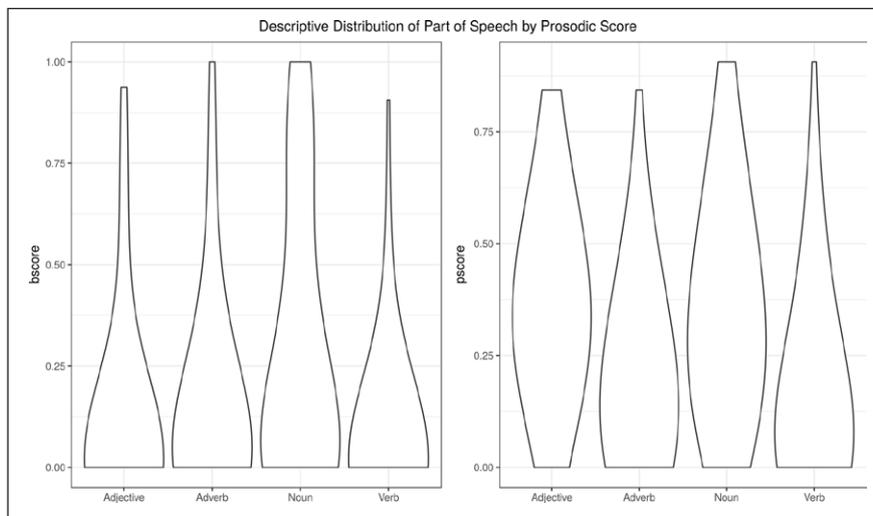


Figure 5: Violin plots for prosodic scores and Part of Speech. This shows the density of the data in each part-of-speech category across the average boundary (left) and prominence (right) ratings (the b-score and p-score).

and p-scores with a trend line (in black). **Figure 5** presents violin plots showing the distribution of p-scores and b-scores for each POS category. These plots show several properties of the data. First, the plots reveal the b-score and p-score distributions as skewed towards zero values, indicating that many words are rated consistently among listeners as *not* being prominent, and as *not* being followed by a prosodic boundary. Conversely, there are relatively few words that listeners agree are prominent or preceding a boundary—data are sparser at the high end of the y-axis on these plots. Second, with the exception of Post-pause Duration and b-scores, the predictor variables each show substantial spread over a

wide range of p-scores and b-scores. This is clearly not the kind of bimodal distribution we would expect if, for example, the prominent/non-prominent distinction was conveyed through the patterning of individual cues (and similarly for the boundary/non-boundary distinction). A more strongly bimodal distribution of p-scores and b-scores across the range of values for an individual predictor cue would be consistent with higher levels of inter-annotator agreement in prominence and boundary rating, but as shown in section 3.1, inter-annotator agreement is moderate for our data. Indeed, based on these plots we do not expect any single factor to emerge as a sufficient cue to prosody on its own, and we further expect that individual annotators may differ in where they locate a category split between boundary/non-boundary (or prominent/non-prominent) along the continua defined by these predictors.

3.3 Annotator group size

The GAMM results reveal a lot of noise in the relationship between each predictor variable and the prosodic ratings. This raises the question of how many annotations need to be collected for the signal to overcome the noise. This is an important question for the researcher looking to use RPT as a means to obtain prosodic annotations for a speech database. It's also a question of interest for the understanding of the extent of individual differences in the overall population. We chose the technique of bootstrapping (resampling) the different size groups of annotators (from 2 to 31) to test the sampling error around our kappa (Fleiss's kappa) for the marking of boundary and prominence. Bootstrapping resamples the data and a statistic of interest in order to estimate the sampling error, the amount of noise due to random sampling, in a statistic of interest for a given data set. Chihara and Hesterberg (2012) provide an introduction to such techniques for statistical analysis. We bootstrap (see Appendix C for R-code) 10,000 samples in order to assess how much sampling error¹² there is in our data depending on how many annotators were included in the analysis. This provides an initial analysis of how many annotators we would need to include in the analysis to reasonably expect results within .10 of the true population kappa if we ran another experiment on the same population. Because we are bootstrapping the kappa statistic for 10,000 simulations of $n = 2$ annotators, $n = 3$ annotators, etc., the average kappa for each n is going to be the overall kappa of the data set. Bootstrapping does not provide a better estimate of kappa than calculating kappa for the overall population. It does provide a way to access the sampling error across different n s for the data that can be generalizable to other experimental designs of prosodic marking.

In **Figure 6**, the estimate of the sampling error for each group size of annotators indicates that somewhere around 7 annotators yields an estimated sampling error of .05 and somewhere around 10 annotators would be needed to get below an estimated sampling error of .05. The .05 sampling error used is where a researcher could expect for the data sampled and the kappa statistic of interest, that 96% of the new samples would be within .10 (plus or minus two times the sample error) of the kappa from the sample size of 32. One may wonder why more annotators are necessary to reach the stated criterion for boundary marking than for prominence marking. This demonstrates a benefit of bootstrapping: It indicates that there is more noise in the kappa statistic for boundary marking than prominence marking. While the overall agreement is higher for boundary marking than prominence marking, the kappa statistic is noisier for boundary marking in the data.

¹² Assessing questions of fixed sample size and kappa sampling error require bootstrapping. The central limit theorem cannot be applied here since we are looking at fixed n (and not as n goes to infinity). Further, it is not known if the kappa statistic for what we are measuring is actually normally distributed or converges much more slowly to normality (see Hesterberg, 2008 for the importance of this distinction).

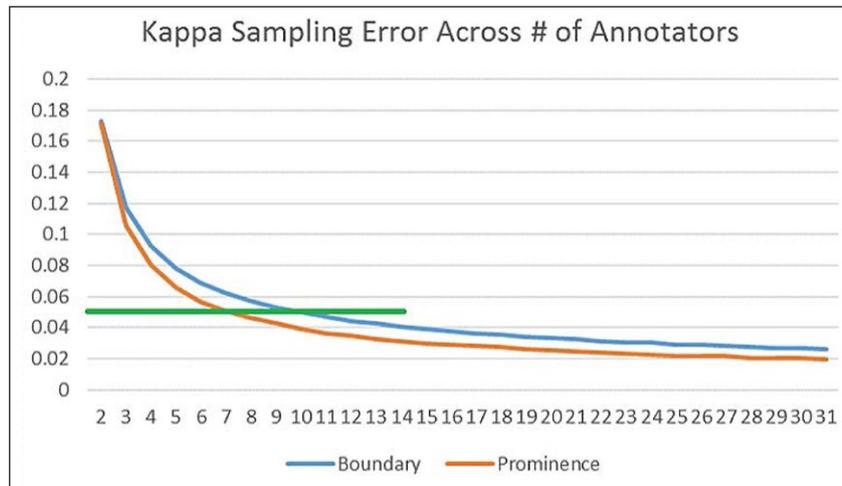


Figure 6: The estimate of sampling error for each # of annotators.

3.4 Generalized additive mixed models (GAMMs): Overall effects

GAMMs are used to extract from the data information about each of the cues and the differences in the contribution of each cue as a predictor of individual annotators' marking of prosodic features. Results for the full model are presented first, combining all predictors with subject and word as random factors, with data from all participants. Following that we present visualizations of the results from the same model showing the effects of individual predictors, and finally, visualizations of results for individual participants, again from the same full model.

3.4.1 Results from full model

Table 2 and **Table 3** display the parametric and non-parametric tests for each of our predictors in the model (Wood, 2006, 2011)¹³ for boundary and prominence marking. For boundary marking, the predictors and random smooths account for 64% of the deviance in the model. For prominence marking, the percentage of deviance explained is substantially less, at 37%. The only component that fails to obtain statistical significance is the random smooth for Part of Speech by subject, indicating that there is a shared effect of Part of Speech on boundary marking that does not differ by subject.¹⁴

3.4.2 Effects of individual predictors over all participants

The GAMM results indicate that the predictors have a non-null contribution to prosodic marking considering data from all annotators. In order to understand the nature of each effect, visualizations are presented below of model estimates for the probability of prosodic marking across the range of values for each predictor, as reported in **Table 2**. In each of these visualizations of GAMM estimates, the y-axis represents the estimated probability of the prosodic feature and the x-axis represents the range of values for the

¹³ For categorical predictors (i.e., Part of Speech and Boundary Marking), parametric tests of significance can be used. For the smooths produced by the GAMM, only non-parametric tests are available. The results are separated into two tables for presentation purposes only—they come from the single GAMM of each type of prosodic marking represented in Appendix A.

¹⁴ When we examine the estimated concavity for the models, all pairwise results for the main effects are less than .5. There were a few instances of the possibility of higher order interaction with estimated concavity between .7–.8, but only between different subject smooths. In order to remedy this, several higher order subject smooths were fit to interactions of predictors as well as interactions of predictors without subject terms, but none converged. All subject smooth terms were concave with their associated main predictor, but this is to be expected as the overall main effect and subject specific curves are related and is why $m = 1$ was chosen for the smooths to down-weight individual differences in favor of the overall effect.

Table 2: Non-parametric test for smooth terms (fixed and random effects).

Term	Boundary (Deviance Explained = 63%)				Prominence (Deviance Explained = 37%)			
	Effective	Residual	Chi.sq	p-value	Edf	Ref DF	Chi.sq	p-value
	DF ¹⁵	Effect DF						
s(word.phonerate)	2.96	3.60	175.57	<.001	5.13	6.09	78.52	<.001
s(log.f0.sv.max.norm)	8.12	8.66	38.19	<.001	7.35	8.12	128.90	<.001
s(int.sv.norm)	7.65	8.32	61.13	<.001	6.86	7.71	22.73	.003
s(log.wordfreq.switchboard)	1.00	1.00	5.06	0.02	1.00	1.00	40.88	<.001
s(durpostpause)	4.26	4.87	249.54	<.001	5.74	6.45	16.39	.02
s(word)	178.33	270	1253.24	<.001	225.75	270	1458.47	<.001
s(subject,word.phonerate)	34.42	287	55.34	<.001	63.439	287	168.76	<.001
s(subject,log.f0.sv.max.norm)	2.26	287	2.45	<.001	28.80	287	42.09	<.001
s(subject,int.sv.norm)	2.25	287	2.45	<.001	40.74	287	71.35	<.001
s(subject,log.wordfreq.switchboard)	16.54	287	21.63	0.004	75.90	287	234.19	<.001
s(subject,durpostpause)	57.39	287	184.71	<.001	2.54	287	2.81	<.001
s(subject,pos)	0.01	124	0.01	0.39	28.37	124	41.95	.001
s(subject,boundarym)					22.08	62.00	64.66	<.001

Table 3: Parametric tests for categorical predictors.

Term	Boundary				Prominence			
	Estimate	Std. Error	z	p-value	Estimate	Std. Error	z	p-value
(Intercept)	-4.5389	0.3756	-12.085	<.0001	-0.56	0.21	-2.62	.01
POS 1 [Adjective vs Noun]	1.2896	0.3539	3.644	0.0003	-1.28	0.21	-6.18	<.0001
POS 2 [Adverb vs Noun]	1.3413	0.3526	3.804	0.0001	-1.01	0.18	-5.63	<.0001
POS 3 [Verb vs Noun]	0.2479	0.3839	0.646	0.5184	-1.23	0.19	-6.18	<.0001
Boundary Marked					0.47	0.12	3.92	.0001

given predictor in our data. The black line is the actual estimated probability of prosodic marking, while the grey area represents the standard error around the estimate. Along the x-axis, there are short black bars that represent the distribution of our data across each predictor with the thickness representing the amount of data and white space where there is no data.

In all the visualizations, participant numbers are consistent, e.g., participant #21 in one graph is the same as participant #21 in all other graphs. The ToBI annotation is shown in all graphs as participant #33 and is included to allow comparison of trained and untrained annotators in the same statistical model. Supplementing the visualizations shown below, Appendix A presents visualizations for the complete set of predictors for both boundary and prominence marking, by participant, and a visualization with estimates for all participants overlaid (in different colors). Appendix A also presents the R code (R Core Team, 2017) for our models and the visualizations.

Figure 7 presents the estimated probability of marking across the range of Intensity values. The probability of boundary marking is greatest for the lowest Intensity values,

¹⁵ Effective Degrees of Freedom are the model df that are used for the smoothing functions. They represent a function of non-linearity as 1.0 indicates a linear relationship between the DV and s(IV) and further than 1.0 indicates non-linearity. This measure is described in Wood (2006, pp. 170–171). The residual effective DF are those degrees of freedom not used by the smooth function.

with a general decrease in the likelihood of boundary marking as Intensity increases. The effect of Intensity on prominence marking is more muted with some non-zero fluctuation in the estimated probability of marking, but has no general increase or decrease.

The estimated effect of the locally normalized measure of Max F0 on prosodic marking is shown in **Figure 8**. The probability of boundary marking (left panel) shows only slight variation in relation to normalized Max F0, with a very small peak at high values. The

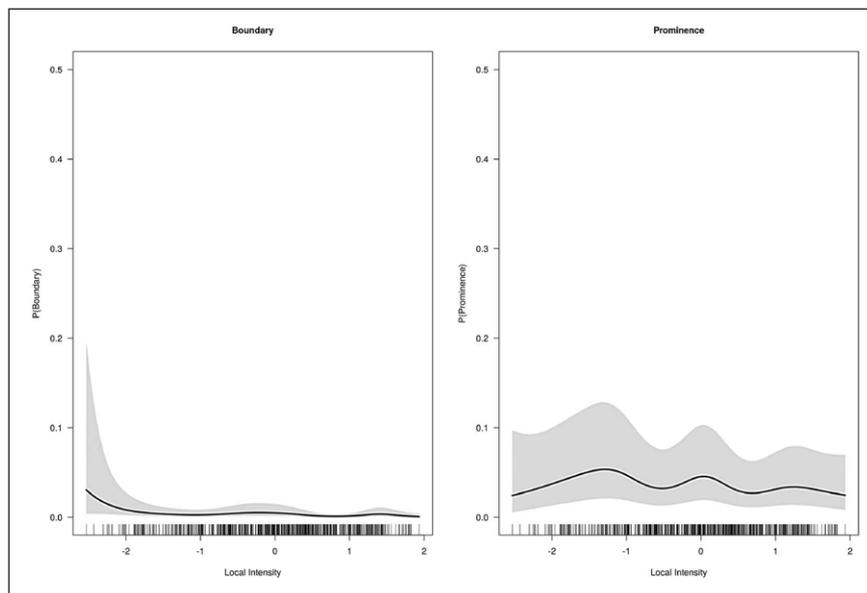


Figure 7: Estimated probability of prosodic marking (y-axis) across the range of normalized local Intensity values (x-axis), for boundary marking (left panel) and prominence marking (right panel). The black line shows the model estimated probability and the grey band shows the confidence interval around the estimate. The distribution of Intensity values in the data is shown by the thickness of the short black bars along the x-axis, with white intervals at values where there are no data. Intensity is measured in the stressed vowel as described in Section 2.3.

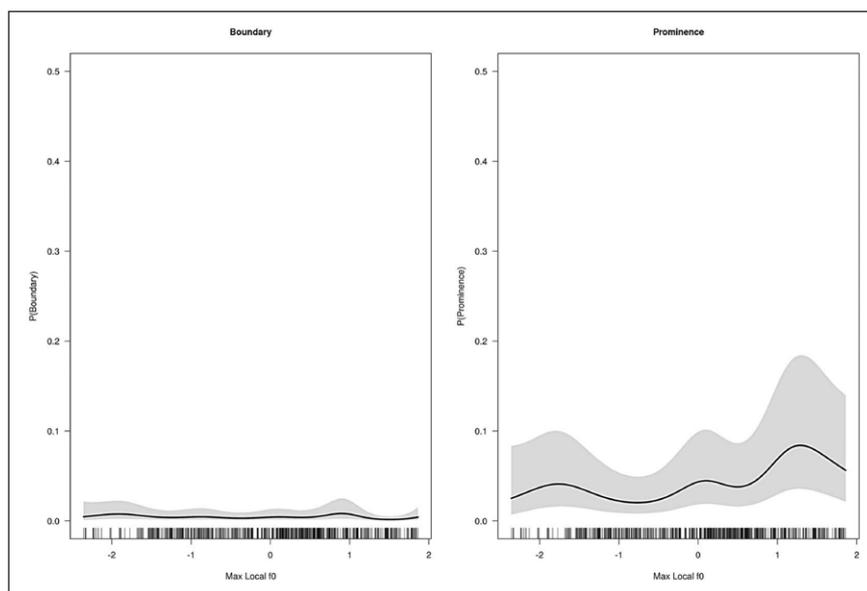


Figure 8: Estimated probability of prosodic marking (y-axis) across the range of normalized values of Max F0 (x-axis), for boundary marking (left panel) and prominence marking (right panel). Plot details as in Figure 6. Max F0 measured in the stressed vowel as described in Section 2.3.

probability of prominence marking (right panel) has a slightly stronger peak at a high value of normalized Max F0. For both boundary and prominence, the effect of max F0 is mildly non-linear.

The estimated effect of Word Frequency on the likelihood of prosodic marking is shown in **Figure 9**. This effect, though significant, is very small for boundary marking (left panel): Words with very low log frequency have a slightly greater probability of being marked as preceding a boundary. The effect of Word Frequency for prominence marking (right panel), however, is much greater, with a substantial increase in the likelihood of prominence marking for low frequency words, and a general decrease in likelihood of marking as Word Frequency increases.

Figure 10 shows the estimated effect of Word Phone-rate on prosodic marking. There is a very dramatic decrease in the likelihood of boundary marking (left panel) as Word Phone-rate increases—i.e., as speech tempo increases, producing shortened phone and word durations. This effect is repeated, but much shallower, with prominence marking (right panel).

Figure 11 shows the effects of the following pause duration (Post-pause Duration) on prosodic marking of a word. As Post-pause Duration increases beyond a value of about 50 ms, there is a very large increase in the likelihood of boundary marking (left panel) that peaks at a Post-pause Duration of about 600 ms. Very long pauses that exceed 600 ms are less frequent in our data, and show an unexpected decrease in the likelihood of boundary marking in comparison to words with the peak pause duration of about 600 ms. Data sparsity for high Post-pause duration makes it difficult to interpret the source of this effect, but it's possible that some very long pauses reflect hesitation disfluencies that the listener does not interpret as a prosodic boundary. The right panel of **Figure 10** shows that the effect of Post-pause Duration on prominence marking, though significant, is much smaller, with a slight peak in the likelihood of prominence marking for words that are followed, again, by a pause of about 600 ms.

The estimated effects of Part of Speech category on prosodic marking is shown in **Figure 12**. Part of Speech has a minimal effect on the probability of boundary marking

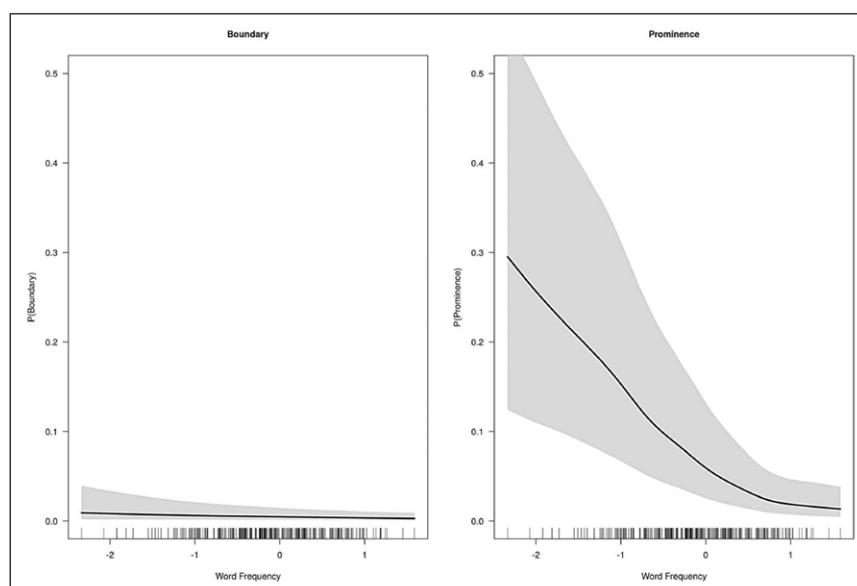


Figure 9: Estimated probability of prosodic marking (y-axis) across the range of log Word Frequency values (x-axis), for boundary marking (left panel) and prominence marking (right panel). Plot details as in Figure 6. Word Frequency calculated as described in Section 2.3.

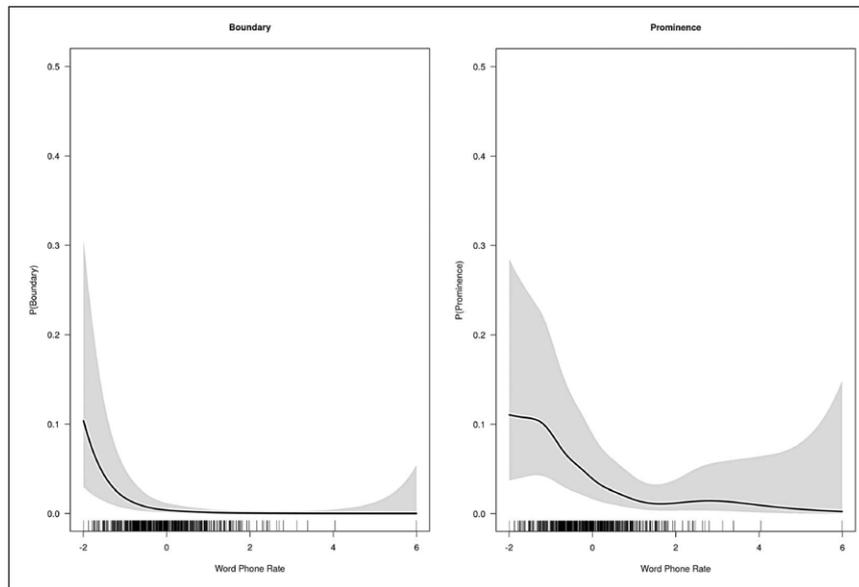


Figure 10: Estimated probability of prosodic marking (y-axis) across the range of Word Phone-rate values (x-axis), for boundary marking (left panel) and prominence marking (right panel). Plot details as in Figure 6. Word phone-rate calculated as described in Section 2.3.

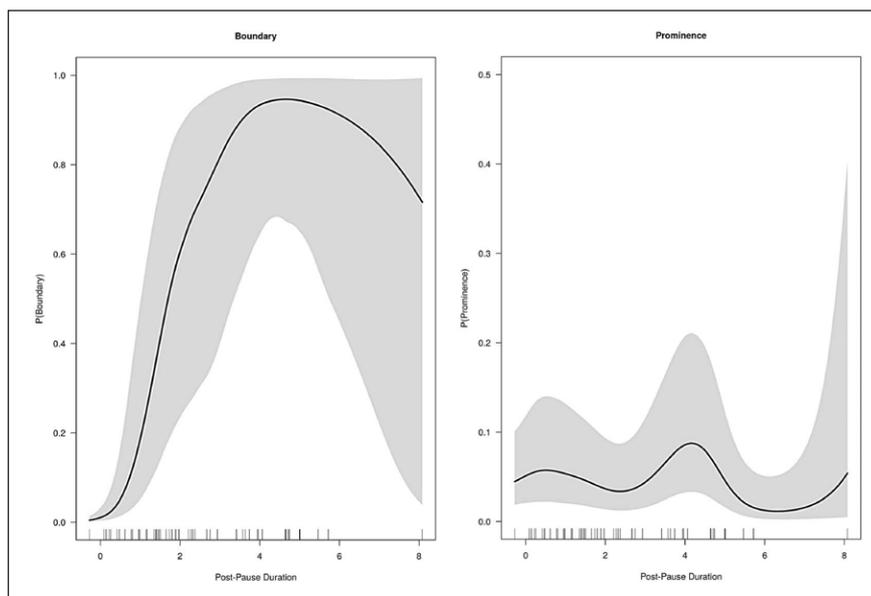


Figure 11: Estimated probability of prosodic marking (y-axis) across the range of Post-pause duration values (x-axis), for boundary marking (left panel) and prominence marking (right panel). Plot details as in Figure 6. Post-pause duration measures the duration of a silent pause following the target word, as described in Section 2.3.

(left panel); however, for prominence marking (right panel), there is a much greater differentiation among Part of Speech categories, with nouns exhibiting a substantially greater likelihood of prominence marking.

The estimated probability of prominence marking for tokens marked as boundaries (= 1) is presented in **Figure 13**. Words that an annotator marks as preceding a boundary have a slightly greater likelihood of being marked by the same annotator as prominent.

3.4.3 Interim Summary

The overall effects of each predictor on the probabilities of boundary and prominence marking are summarized in **Table 4**. Effects are described in terms of the values of the

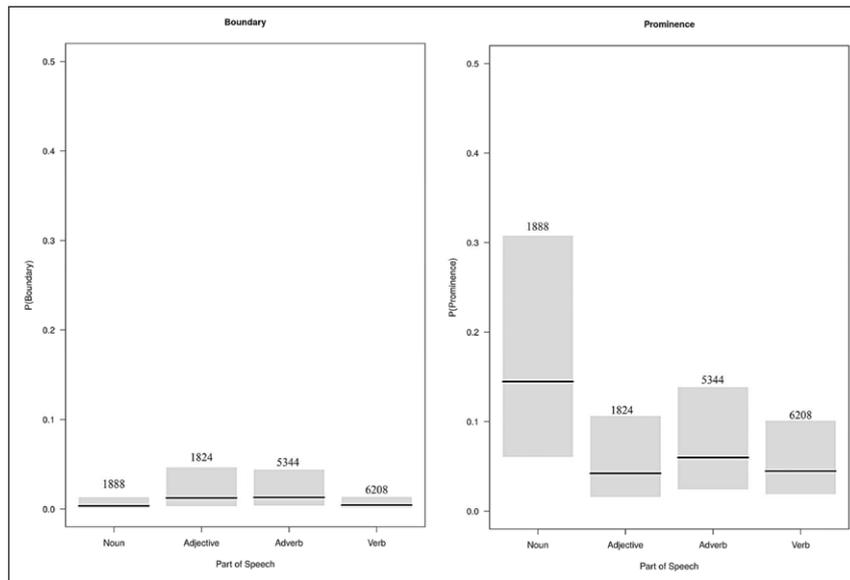


Figure 12: Estimated probability of prosodic marking (y-axis) by Part of Speech category (x-axis), for boundary marking (left panel) and prominence marking (right panel).

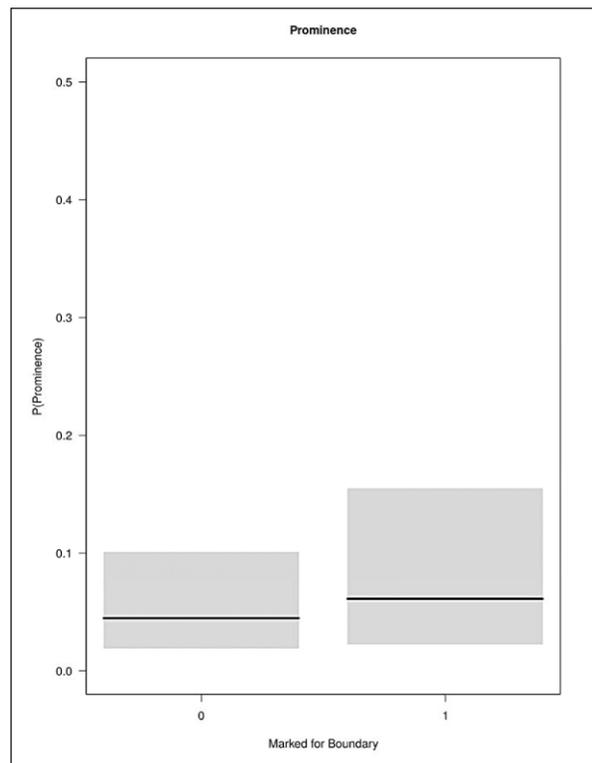


Figure 13: Estimated probability of prominence marking for words not marked for boundary (left bar) and those marked for boundary (right bar).

predictor variable that define a peak in the probability of boundary (or prominence) marking—these are the values for which the predictor is potentially most effective as a perceptual cue for boundary (or prominence). Effects are also described in terms of the pattern of co-variation: Does the probability of boundary or prominence marking vary proportionally, either positively or negatively, with changes in the predictor variable? If so, over what range of predictor values? Effects that are unpredicted or that run counter to prediction are in grey. A more detailed discussion of these results in light of the predicted effects is in Section 4.3.

Table 4: Summary of Interim Results.

Predictor variable	Increase in Probability of Boundary	Increase in Probability of Prominence
Local Intensity	Peaks at lowest Intensity, and steeply diminishes as Intensity increases to mid-low values	Small peaks at mid-low and average Intensity
Local Max F0	Two small peaks, at very low Max F0 and very high Max F0	Peaks at near-highest Max F0, with mild growth as Max F0 increases from average values ($z = 0$); secondary peak at very low Max F0
Word Frequency	Peaks at lowest Word Frequency, and mildly diminishes as Word Frequency increases	Peaks at lowest Word Frequency, and very steeply diminishes as Word Frequency increases
Word Phone-rate	Peak at lowest Phone-rate, and steeply diminishes as Phone-rate increases up to approximately median value	Peaks at lowest Phone-rate, and moderately diminishes as Phone-rate increases over most of its range
Post-pause Duration	Peak at near-highest Post-pause duration, with strong growth trend over almost entire range of Post-pause Duration	Very mild peak for values in small interval at the high end of predictor range
Part of Speech	Very minimal effects, with slight increase for Adjectives, Adverbs and Verbs	Substantial increase for Nouns, and slight increase for Adverbs

Considering the numeric effect of each predictor in increasing the probability of boundary and prominence marking, we can identify the strongest predictors for each prosodic feature marked by our participants. The predictors with the strongest effect on boundary marking are Post-pause Duration, and Word Phone-rate, in that order. The predictor with the strongest effect on prominence marking is Word Frequency, followed by Part of Speech (Noun), and Word Phone-rate, in that order. The effects of the strongest predictors for boundary (Post-pause Duration and Word Phone-rate) are substantially greater than effects of the strongest predictor of prominence (Word Frequency), and dwarf the rest of the small though significant effects.

3.5 Individual differences

In this section, we explore the patterns by which each participant deviates from the overall pattern of effects presented in the previous section. The individual differences analysis is performed based on the GAMM visualizations for the random effects of individual participants. These graphs are structured identically as the GAMM visualization graphs presented above. Critically, the effects for all predictor variables are displayed on the same scale on the y-axis (probability of boundary/prominence marking), to facilitate easy visual comparison across graphs. We illustrate the analysis with a complete set of plots for the main effects of Intensity by participant. The full set of graphs for all predictor variables is presented in Appendix A, and is summarized in Section 4.4 in an analysis of how annotators can be grouped on the basis of their individual pattern of effects.

In **Figures 14** and **15**, each participant's estimated effect across Intensity is presented for prosodic marking. For the marking of boundary (**Figure 14**), many of the participants show a flat, near-zero effect of Intensity. Only a subset of participants seems sensitive to Intensity as a cue for boundary marking, with an effect pattern that resembles the overall effect of Intensity shown above in **Figure 6** (left panel), where boundary marking has a mild peak for words with the lowest Intensity values, and falls sharply as Intensity rises to mid-low values. Among the participants who show the greatest sensitivity to Intensity

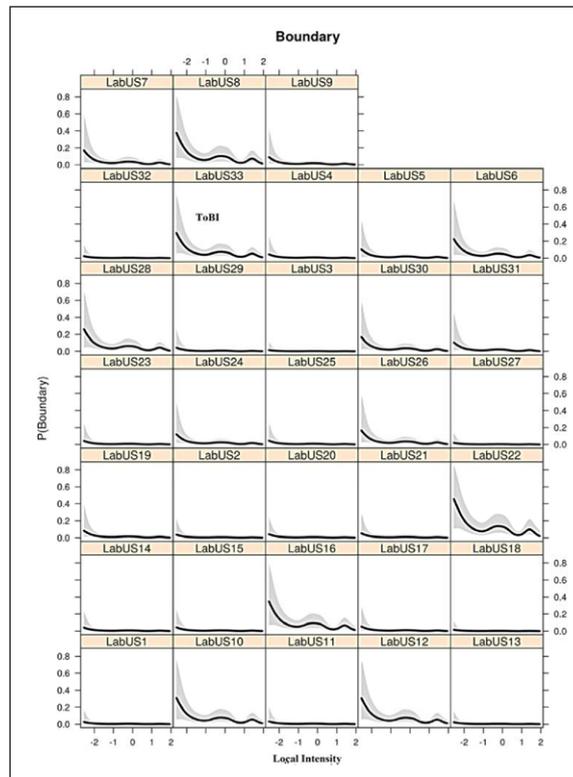


Figure 14: The estimated effect of Intensity by Participant for Boundary Selection Estimated probability of boundary marking (y-axis) across the range of Intensity values (x-axis), for each participant. Plot details as in Figure 6. Participant ID is shown above each plot. Participants 1–32 are untrained annotators, while participant 33 represents the consensus ToBI labeling.

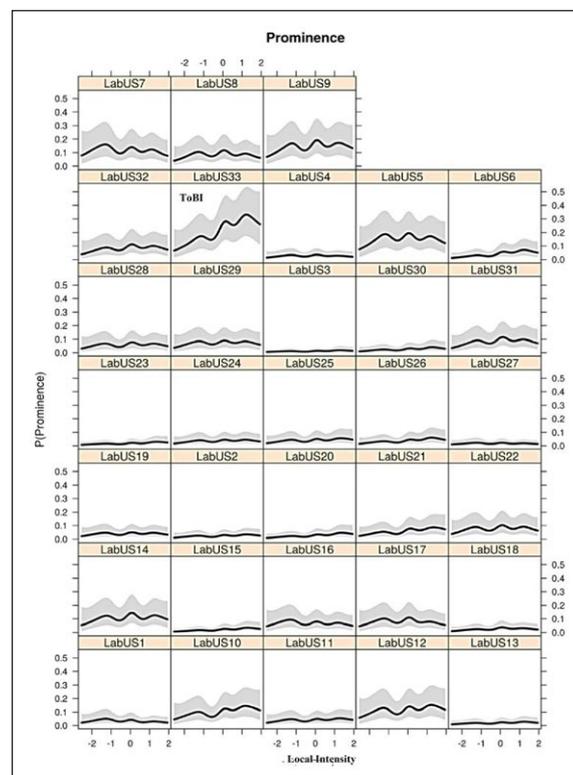


Figure 15: Estimated probability of prominence marking (y-axis) across the range of Intensity values (x-axis), for each participant. Plot details as in Figure 6. Participant ID is shown above each plot. Participants 1–32 are untrained annotators, while participant 33 represents the consensus ToBI labeling.

as a cue to boundary marking is participant #33, representing the expert annotation from the trained ToBI annotators. This finding, where the ToBI annotation is among the few participants showing the strongest effect of a predictor, is repeated in many of the visualizations in Appendix A as well, and is discussed further in Section 4.4.

A similar differentiation of individual sensitivities to this cue is seen in the marking of prosodic prominence across the participants (**Figure 15**). While many participants show a flat, near-zero effect, some show an effect pattern that resembles the overall effect in **Figure 6** (right panel), with a few mild peaks in the probability of prominence marking for Intensity values at mid-low (near $z = -1$) and average (near $z = 0$) values. Some participants show an additional mild peak in the likelihood of prominence for high Intensity values (near $z = 1$), which is less evident in the overall results shown in **Figure 6** (right panel). Notably, the consensus ToBI labeling shows the same pattern of effects of Intensity on prominence marking as other participants who respond to this cue, but with much greater sensitivity.

Considering the individual participant graphs produced for all predictors (shown in Appendix A), we find a similar global pattern of results. There is a lot of variability across participants in the effect of each predictor on the probability of prosodic marking. Yet remarkably, for participants who show a non-null effect (i.e., the predictor has a non-zero effect on the probability of prosodic marking, for at least some values in its range) the general pattern of co-variation between the predictor variable and the probability of prosodic marking—as illustrated by the plotted estimate (the black line in the graphs)—does not change, though the magnitude of the effect may differ across participants. These findings of individual differences—and similarities—are discussed further in Section 4.4.

4. Discussion

We have investigated the prosodic marking of a word (for boundary and prominence) in relation to the several measures of acoustic correlates of prosody (duration, pause, intensity, and F0), and in relation to several contextual measures (Word Frequency, Part of Speech, Boundary Marking) to test the central question (Q4) that individual listeners may vary in a constrained manner from the overall pattern of effects these predictors have on prosodic ratings from the sample population as a whole. The analysis includes prosodic ratings from 32 untrained listeners and one consensus ToBI annotation performed by a pair of trained annotators. Before assessing the results for Question 4, we review the findings as they relate to the three preliminary questions addressing the overall pattern of inter-annotator agreement and the main effects of individual predictors as prosodic cues against the data from the larger sample population.

4.1 Inter-annotator agreement (Q1, Q2)

We observe moderate inter-annotator agreement for the boundary and prominence marking in our data, based on the binary 0/1 coding of each annotator's rating of content words in our corpus, with kappa scores in the low to middle range. This finding addresses our Question 1, indicating substantial differences in the prosodic ratings assigned to words in this sample by our participants. At the same time, the agreement rate is well sufficiently high to indicate that annotators are behaving somewhat systematically. This overall level of agreement affords the possibility that rating disagreements arise due to individual differences in the selection or weighting of various cues to prosody, including the predictors tested here.

The results also address Question 2, demonstrating higher agreement for boundary marking compared to prominence marking. Comparing the current findings with our prior results using RPT (e.g., Cole et al., 2010a, 2010b), we observe similar agreement rates for

boundary rating, and somewhat lower agreement for prominence rating. In that earlier study, Fleiss' kappa scores were separately calculated for four cohorts of 15–22 annotators each, with kappa scores for boundary ratings between .54–.62, and kappa scores for prominence ratings between .37–.42.¹⁶

4.2 Average prosodic ratings differentiate presence vs. absence of prosodic features (Q2)

We address Question 2 by examining the average prosodic ratings for each word (i.e., the b-scores and p-scores). The prediction is that there will be two peaks in the distribution of b-scores and p-scores: One peak near zero representing words that no or almost no annotators marked for boundary or prominence, and a second peak near 1 representing words that all or almost all annotators marked for boundary or prominence. These high levels of annotator agreement would arise for words that a speaker produces with salient acoustic cues that distinguish the presence vs. absence of a prosodic boundary, or salient cues distinguishing prominence from non-prominence. The prediction is that annotators will readily perceive a boundary or prominence in the presence of a perceptually salient cue, resulting in high agreement. Similarly, high agreement is predicted in the presence of cues that unambiguously signal the *absence* of a boundary or a non-prominent word. Another possibility is that the linguistic context strongly predicts the presence or absence of a prosodic boundary, or prominence. If predictions from contextual factors are a strong influence on perceived prosody, we may find high annotator agreement in strongly predictive contexts. When contextual factors and acoustic cues are both salient and congruent in cueing the presence vs. absence of boundaries and prominences for a given word, we expect the highest level of annotator agreement. If there is a prevalence of words with salient and congruent cues, it will be reflected in the distribution of average prosodic ratings with peaks near zero and one.

The average prosodic ratings illustrated in **Figures 7–12**, which are even more clearly represented in the histograms in Appendix C, provide only partial evidence for the predicted distribution, and thus only partial evidence of salient and congruent cueing of a binary boundary and prominence distinction. We see that the b-score and p-score distributions are both strongly skewed by a high number of zero values: Annotators agree that most words are *not* preceding a prosodic boundary, and *not* prominent. On the other hand, the b-score and p-score distributions are quite sparse at higher values, indicating that there are very few words that all, or even many, annotators perceive as being followed by a boundary, or as prominent. The finding that there are many words for which annotators do not agree in their rating of boundary or prominence suggests that salient acoustic cues to boundary and prominence are very often not present in the speech signal. Alternatively, it may be the case that there are many words where the acoustic cues for boundary or prominence are incongruent with the linguistic context. For instance, an incongruence may arise when acoustic cues for boundary or prominence are present for a word in a syntactic or discourse context where a boundary or prominence is not predicted. Or, in an opposite scenario, an incongruence may arise when acoustic cues fail to signal the presence of a boundary or prominence in a context where it is predicted. The lack of agreement in marking the presence of a prosodic feature is consistent with the hypothesis

¹⁶ Note that because we have only a single ToBI annotation in this study, we cannot assess annotator agreement for the ToBI annotation. Inter-annotator agreement reported for prior studies of ToBI annotation reliability (Yoon et al., 2004; Breen et al., 2012) indicate higher levels of agreement than we found for RPT annotations, with kappa scores for prominence ranging from .75–.78, and kappa scores for boundaries ranging from .58–.77. But direct comparison with our RPT kappa scores is not meaningful because the ToBI reliability studies used very few annotators (2 in Yoon et al. study; 4 in the Breen et al. study), and computed agreement using Cohen's pairwise kappa.

of individual differences, but before we turn to those results we first consider the findings for individual predictors against the full group of listeners.

4.3 Cue-based prosodic ratings (Q3)

Question 3 asks whether prosodic ratings from the full group of listeners are predicted by the collective set of cues available to the listener. The GAMM results in **Tables 2 and 3** provide confirmation of this hypothesis. Despite the small number of cues we included in the full GAMM, the results indicate that the model explains 63% of the deviance in boundary ratings, and 37% of the deviance in prominence ratings. Furthermore, all predictors and random effects are significant in the model, with only a single non-significant interaction term (the smooth for Part of Speech by participant). The findings also underscore an asymmetry between boundary and prominence ratings, with boundary ratings more strongly explained in terms of the cues selected for the model. This finding is almost certainly related to the higher kappa scores for inter-annotator agreement for boundary marking (Section 4.1).

The model estimated effects of predictors on prosodic marking are for the most part as predicted in Section 2.3.1. Here we briefly review the GAMM estimated effects of individual predictors, as visualized in **Figures 7–12** (and summarized following **Figure 12**) with an eye to identifying the manner in which the predictor may function as a cue for boundary or prominence rating over the entire group of listeners in our sample. In briefest summary, the model estimated effects for individual predictors do in all but three instances conform to predictions about how the predictors function as cues; unexpected or null findings were obtained for Intensity and Post-pause Duration effects on prominence marking, and Part of Speech effects on boundary marking. Details as follows.

- **Intensity:** The predicted effect on boundary marking is observed only for below average Intensity values ($z < 0$). The downward trend diminishes, with minor fluctuations for above average Intensity values ($z > 1$). This pattern indicates that strong local dips in Intensity serve as a cue for boundary marking, while local upward spikes in Intensity are seemingly uninformative about boundaries. Predicted effects on prominence marking are not clearly confirmed—there is no consistent trend for increased likelihood of prominence marking with higher values of Intensity (cf., Kochanski et al., 2005). There is a weak trend in the predicted direction over the very lowest range of Intensity values, but the trend reverses well before the middle of the Intensity distribution, suggesting that very low Intensity is incommensurate with prominence marking. Intensity values that are moderately low or higher are relatively uninformative about prominence.
- **F0:** For both boundary and prominence marking, there is a weak bimodal pattern in the effect of Max F0, with small increases in the likelihood for both boundary and prominence marking for very low values of Max F0, and for very high values. Only the effect at high Max F0 was predicted, but the estimated bimodal pattern of the F0 effects are consistent with the view of contrastive High and Low tone prosodic features. In addition to these local effects of low and high Max F0, there is a broad trend of gradient increase in prominence marking over F0 values ranging from average ($z = 0$) to high ($z > 0$). This trend is in the predicted direction and confirms high F0 as a cue to prominence marking.
- **Word Frequency:** The predicted effect on prominence marking is confirmed, with low frequency words much more likely to be marked as prominent. A weak but unpredicted finding is the very small effect observed for boundary marking,

where words with the lowest frequency are slightly more likely to be perceived as preceding a boundary. There is no strong basis for predicting low frequency words to be positioned at the end of a prosodic phrase, though it's possible that this weak effect in our data is related to the preference in English for topics or themes (typically given information) to be positioned at the beginning of a sentence (Steedman, 2008).

- **Word Phone-rate:** The very strong effect on boundary marking confirms our prediction, but the effect is apparent only for words with Word Phone-rate values in the lower quartile or slightly above. In other words, only words that exhibit an extremely slower tempo than the immediately surrounding context show an increased likelihood of boundary marking. The effect on boundary marking is near zero for words with more modestly slowed Phone-rate. In contrast, the predicted effect of Word Phone-rate on prominence marking is only weakly confirmed. The model estimated effect is more uniformly gradual across a broader range of Word Phone-rate values, extending approximately up to the upper quartile of values, with slower rates (i.e., longer phone durations) associated with an increased likelihood of prominence marking.
- **Post-pause Duration:** There is a very dramatic, nearly linear effect of Post-pause Duration on boundary marking, which goes in the predicted direction, up to almost the end of the range of pause duration values. For words with the very longest Post-pause Duration, the trend appears to reverse, but data is very sparse in this region of the predictor range. The model estimates only a weak effect of Post-pause Duration on prominence marking, which is limited to duration values in the vicinity of 600 ms. This inflection point is close to the peak in the boundary marking effect, suggesting that Prominence may be marked on the basis of an apparent Boundary cue, but perhaps for words that lacked other indicators of boundary status. This question is not explored further here.
- **Part of Speech:** Although there were no strong predictions for Part of Speech effects on boundary marking for the content words included in our analysis, we nonetheless find a minimal, though significant decrease in the likelihood of boundary marking for nouns relative to the other three categories. This effect, though small, is unexpected given that nouns that are final in the grammatical subject constituent are also expected to be final in a prosodic phrase under some accounts of the correspondence between syntactic and prosodic phrase structure, as are object nouns that are final in the verb phrase (Büring, 2016). On the other hand, the slightly greater likelihood of boundary ratings for verbs, predicate adjectives and adverbs, relative to nouns, could be related to the prevalence of those part of speech categories at the end of a sentence, also final in the prosodic phrase. Further explanation for this small effect is not explored here. The model estimated effects of Part of Speech on prominence marking partially confirm predictions, with only nouns but not adjectives showing an increased likelihood of prominence marking. The more moderate effect for adverbs is most likely due to emphasis on intensifiers, though we leave this question for future analysis.
- **Boundary marking:** The predicted effect of increased prominence marking for words marked by the same annotator as preceding a Boundary is weakly confirmed. The effect is small in magnitude, though significant, suggesting either that nuclear prominence on the phrase-final word is not so common in our materials, and/or that a final nuclear prominence does not have increased perceptual salience relative to prominences in other phrasal positions.

The strongest evidence for the predictive value of a cue for prosodic rating would be a finding that each predictor effectively separates the data into two subsets for each prosodic feature: Boundary/no-boundary, and prominent/not-prominent. The visualizations of estimated effects of individual predictors shown in **Figures 7–12** provide only partial confirmation of this pattern. Word Phone rate and Post-pause Duration appear to effectively distinguish words with a very high likelihood of boundary marking from other words, while the noun category in Part of Speech appears to effectively distinguish words with higher likelihood of prominence marking. Word Frequency also has a strong effect on prominence marking, but the effect is uniformly gradient across the entire range of frequency values, making it difficult to see where a threshold marking a binary prominence distinction might be located.

4.4 Patterns of individual participant differences

The GAMM results reported above for both prominence and boundary have included all of the data from the 32 untrained annotators plus the ToBI annotation. In this section, we discuss the GAMM results for the random smooth of each predictor by participant, addressing Question 4, about how individual listeners deviate from the pattern of overall effects of predictors on prosodic marking. As discussed in Section 1, we predict that individual differences may occur in the selection or non-selection of a predictor as a cue for prosodic marking, or in cue weighting, but not in ways that would change the overall pattern that relates variation in the cue to variation in the likelihood of prosodic marking. The patterning of individual deviations from the overall effect of a predictor on prosodic rating can be assessed qualitatively by visual inspection of the GAMM visualizations in Appendix A, and confirm our prediction: If a cue is a significant predictor of prosodic rating for any number of annotators, then it has the same pattern of influence for all of those annotators. The effect of a significant cue is always in the same direction, and individual differences among annotators emerge only in the selection and magnitude of the cue's effect on prosodic rating.

We offer a mildly subjective quantitative test of the prediction here, by establishing a lower threshold on the peak effect of a predictor for considering the predictor as a cue—i.e., as boosting the probability of boundary (or prominence) marking. Specifically, if the overall estimated probability (controlling for the other predictors) rose above .1 across any range of a predictor, the participant was coded as showing a positive effect for the predictor as a cue for boundary (or prominence) marking. A value of .1 on the y-axis means that predictor values in that region increase the probability of prosodic marking by 10% or more. This pattern is indicated with a “+” in **Tables 4** and **5** below. If there was no range of the predictor values that reached or surpassed the .1 value for boosting prosodic marking, the participant was coded as *not* showing an effect for that predictor, indicated as “–” in the tables below. For example, **Figure 16** displays the GAMM visualization for the estimated effect of Max F0 on boundary marking for participant 22 (left panel) and participant 23 (right panel). For participant 22 there are multiple estimated values of Max F0 for which the probability of boundary marking exceeds .1, while for participant 23 there are no estimated values that exceed .1. Correspondingly, participant 22 is coded in **Table 5** as “+” for Max F0 as a boundary cue, while participant 23 is coded as “–” for the same cue.

In **Table 5**, the coding for each participant is applied for the selection of predictors as cues of prosodic boundaries in the data, as described above. Participants are grouped in the table to highlight shared patterns of cue selection. There is a subset of 6 untrained participants (annotators) who are using the same cues as the trained annotators (#33). The most widely used cue selected for boundary marking is Post-pause Duration, which is

Table 5: Shared effect across participants for the selection of predictors as cues to boundary marking. Selection criteria as described in text above.

Participant	Boundary					
	Intensity	Max F0	Word Frequency	Word Phone-rate	Post-pause Duration	Part of Speech
Trained 33	+	+	+	+	+	+
Untrained 8	+	+	+	+	+	+
Untrained 10	+	+	+	+	+	+
Untrained 12	+	+	+	+	+	+
Untrained 16	+	+	+	+	+	+
Untrained 22	+	+	+	+	+	+
Untrained 28	+	+	+	+	+	+
Untrained 6	+	-	+	+	+	+
Untrained 26	+	-	-	+	+	+
Untrained 24	+	-	-	+	+	-
Untrained 30	+	-	-	+	+	-
Untrained 31	+	-	-	+	+	-
Untrained 1	-	-	-	+	+	-
Untrained 2	-	-	-	+	+	-
Untrained 4	-	-	-	+	+	-
Untrained 5	-	-	-	+	+	-
Untrained 7	-	-	-	+	+	-
Untrained 9	-	-	-	+	+	-
Untrained 11	-	-	-	+	+	-
Untrained 14	-	-	-	+	+	-
Untrained 15	-	-	-	+	+	-
Untrained 17	-	-	-	+	+	-
Untrained 19	-	-	-	+	+	-
Untrained 20	-	-	-	+	+	-
Untrained 21	-	-	-	+	+	-
Untrained 23	-	-	-	+	+	-
Untrained 25	-	-	-	+	+	-
Untrained 29	-	-	-	+	+	-
Untrained 32	-	-	-	+	+	-
Untrained 3	-	-	-	-	+	-
Untrained 13	-	-	-	-	+	-
Untrained 18	-	-	-	-	+	-
Untrained 27	-	-	-	-	+	-

selected by every participant. Word Phone-rate is the second most widely used cue. The least used cues are Max F0, Word frequency, and Part of Speech, with Intensity used by 12 annotators.

Table 5 also illustrates a grouping of participants according to the cues they select for boundary marking. For instance, the ToBI annotation on the first row and the first 6 participants listed under it show the same pattern, as already noted. There is another, larger cluster of 17 participants in the middle of the table who select only Word Phone-rate

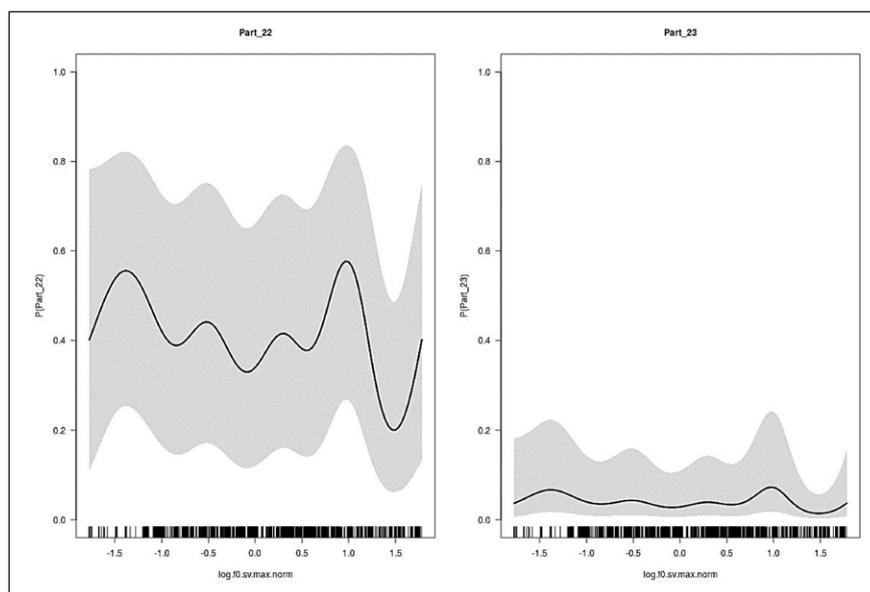


Figure 16: Example of Schema for Tables 4 and 5. For Max F0, “+” on the left (participant 22) and “-” on the right (participant 23). A value of .1 or higher on the y-axis means that predictor values in that region increase the probability of prosodic marking by 10% or more for that participant, in which case the predictor is coded with a “+” for that participant in Table 5. If there are no estimated values that exceed .1, the predictor is coded with a “-” for that participant in Table 5.

and Post-pause Duration as cues, with another 4, at the bottom of the table, who select only Post-pause Duration. The grouping of participants reveals a limited pattern of individual differences in cue selection, with the biggest distinction between participants who select *only* durational cues, and those who select one or more of the remaining cues.

Another very interesting finding that emerges from this predictor grouping is that there is an implicational hierarchy in cue selection. All annotators who use Max F0 as a cue for boundary annotation also use Frequency. Annotators who use Frequency also use Part of Speech. The full implicational hierarchy of cue selection is as follows: Max F0 → Frequency → PoS → Intensity → Phone-rate → post-pause duration. The hierarchy, to our knowledge, has not been reported in previous literature.

Table 6 displays the grouping of participants according to the cues they use for prominence marking. This table includes the same predictors shown in **Table 5**, with the addition of Boundary marking (by the same annotator) as a possible cue for prominence. The trained ToBI annotation is again listed in the top row and uses all available cues for prosodic marking. For prominence marking, we see a larger cluster of participants, 14 in all, who cluster with the trained ToBI annotators in selecting all predictors as cues for prominence marking. The most widely used cue for prominence marking is Part of Speech, followed closely by Word Frequency. The least used cues for prominence marking are Intensity, Max F0, and Post-pause Duration, in that order, though even these cues are used by many participants.

Compared to the cues selected for boundary marking, these listeners rely on a greater number of cues, and exhibit more individual differences in cue selection for prominence marking. Additional smaller clusters can be identified for combinations of two and three cues, but no striking patterns stand out here. The overall finding from this grouping of predictors for prominence marking is that there is greater variation in the patterning of individual differences in cue selection for prominence across participants, compared to boundary marking.

Table 6: Shared effect across participants for the selection of predictors as cues to prominence marking. Selection criteria as described in text above.

Participant	Prominence						
	Intensity	Max F0	Word Frequency	Word Phone-rate	Post-pause Duration	Part of Speech	Boundary
Trained 33	+	+	+	+	+	+	+
Untrained 5	+	+	+	+	+	+	+
Untrained 7	+	+	+	+	+	+	+
Untrained 8	+	+	+	+	+	+	+
Untrained 9	+	+	+	+	+	+	+
Untrained 10	+	+	+	+	+	+	+
Untrained 12	+	+	+	+	+	+	+
Untrained 14	+	+	+	+	+	+	+
Untrained 16	+	+	+	+	+	+	+
Untrained 17	+	+	+	+	+	+	+
Untrained 22	+	+	+	+	+	+	+
Untrained 28	+	+	+	+	+	+	+
Untrained 29	+	+	+	+	+	+	+
Untrained 31	+	+	+	+	+	+	+
Untrained 32	+	+	+	+	+	+	+
Untrained 6	-	+	+	+	+	+	+
Untrained 11	-	+	+	+	+	+	+
Untrained 21	-	+	+	+	+	-	-
Untrained 4	-	-	+	+	-	+	+
Untrained 18	-	-	+	+	-	+	+
Untrained 20	-	-	+	+	-	+	+
Untrained 25	-	-	+	+	-	+	-
Untrained 26	-	-	+	-	+	+	-
Untrained 24	-	-	+	-	-	+	+
Untrained 1	-	-	+	-	-	+	-
Untrained 15	-	-	-	+	-	-	+
Untrained 30	-	-	-	+	-	-	-
Untrained 23	-	-	-	-	-	+	-
Untrained 27	-	-	-	-	-	+	-
Untrained 13	-	-	-	-	-	-	+
Untrained 19	-	-	-	-	-	-	+
Untrained 2	-	-	-	-	-	-	-
Untrained 3	-	-	-	-	-	-	-

Finally, we observe an implicational hierarchy for cue selection with prominence rating similar to what was observed for boundary rating, though extending over only four cues in prominence rating: Intensity → Max F0 → Post-pause duration → Word Frequency.

An alternative approach to the analysis presented above would be a more purely data-driven and objective cluster analysis; however, generalized additive models do not readily lend themselves to such an approach. The next steps in this research are to develop quantitative techniques that incorporate the estimated smooths into a clustering algorithm

with a larger dataset to investigate precisely how shared cues are implemented, what groups emerge from this analysis, and patterns of how prosody is marked with respect to these cues.

5. Conclusion

Prosodic annotation is a noisy measure of the underlying prosody of an utterance, due in part to ambiguity of the cues to boundaries and prominences available to the listener, which in turn results from variability in speakers' production of prosody at both the phonological and phonetic levels. The noise in prosodic annotation can be measured in the rate of disagreement among independent annotators, including trained and untrained. Variability in annotators' identification of prosodic features is also matched by widely observed variability in the production of prosodic features as a function of the linguistic context of a word, and in the phonetic expression of prosodic features. Both types of variability—in the speaker's production and in the listener's perception of prosody—represent statistical noise for the analysis of prosody.

In this paper, we have examined individual differences in prosody ratings as a source of information about the factors that influence prosodic annotation, using GAMMs to examine patterns of convergence in listeners' judgments of prosody in relation to the source of those judgments in acoustic, lexical, frequency-based, and phonological contextual cues in the perception of prosodic boundaries and prominences. GAMMs of annotator ratings of boundaries and prominences provide robust evidence of individual differences. Despite the observed differences in the factors that cue prosody for untrained annotators, we find uniformity among annotators in the shape and relative magnitude of the effects for significant predictors. If an annotator's prosodic rating of a word is influenced by a particular acoustic cue or by a contextual factor, then the influence is in the same direction as for other annotators, and similarly matches that of trained annotators. Generally speaking, then, the observed variability is limited to the selection and magnitude of the factors that influence prosodic rating, and the relative weighting among those factors.

A related point here is that the factors that influence prosodic ratings for untrained annotators are the same factors reported in prior studies as measured correlates of prosodic features elicited from speakers under experimental control. Specifically, Post-pause Duration is strongly predictive of boundary rating, along with Word Phone-rate—the two durational measures among the predictors considered here. The cues for prominence ratings, on the other hand, appear to be more diffuse—with more individual variation in which cues are influencing listeners' prominence ratings.

Throughout our analyses we have demonstrated that non-linearity exists in the conditioning of prosodic marking and that this non-linearity can be amplified by some groups of untrained annotators and trained annotators (as clearly revealed in the visualization of individual predictor effects in Appendix A). The findings point to several questions for further research:

- Are the annotator groups we formed impressionistically, based on an effect minimum threshold we selected, statistically distinct from one another?
- Are the patterns of cue selection consistent within annotator groups, when tested against new data or new annotators?
- Are the patterns of cue selection and weighting consistent for individuals, when tested against new data from the same speech sample?
- Is the pattern of cue selection and weighting variable for individual annotators for ratings of different speech samples, or under different listening conditions?
- Is there a relationship between a listener's use of a cue in perception and in their own production?

Additional Files

The additional files for this article can be found as follows:

- **Appendix A.** R code and all visualizations of the data. DOI: <https://doi.org/10.5334/labphon.108.s1>
- **Appendix B.** Part of Speech Distribution for the Corpus. DOI: <https://doi.org/10.5334/labphon.108.s2>
- **Appendix C.** Bootstrapping Code in R. DOI: <https://doi.org/10.5334/labphon.108.s3>

Acknowledgements

We thank our RPT annotators and our colleagues in attendance at ETAP3 (Experimental and Theoretical Workshop on Prosody, UIUC, May 2015), where this work was first presented. We also thank José Hualde, Chris Eager, and Suyeon Im for contributions to our work with RPT data collection and for stimulating discussion on all matters related to prosody. We would also like to thank the issue editor, Chigusa Kurumada, and the four anonymous reviewers for their constructive and detailed feedback that helped us to improve the manuscript. The research reported here was made possible in part thanks to NSF grants BCS 12-51343 and SMA 14-16791 to the second author.

Competing Interests

The authors have no competing interests to declare.

References

- Baayen, R. H., Davidson, D. J. and Bates D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4), 390–412. DOI: <https://doi.org/10.1016/j.jml.2007.12.005>
- Baumann, S. and Riester, A. 2012. Referential and lexical givenness: Semantic, prosodic and cognitive aspects. In: Elordieta, G. and Prieto, P. (eds.), *Prosody and Meaning*, 25, 119–162. Berlin: Walter de Gruyter.
- Beckman, M. E. and Ayers, G. 1997. Guidelines for ToBI labelling. *The OSU Research Foundation*. Available at: http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/ (accessed 21 February 2017).
- Bishop, J. 2017. Focus projection and prenuclear accents: Evidence from lexical processing. *Language, Cognition and Neuroscience*, 32(2), 236–253. DOI: <https://doi.org/10.1080/23273798.2016.1246745>
- Breen, M., Dilley, L., Kraemer, J. and Gibson, E. 2012. Inter-transcriber reliability for two systems of Prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory*, 8(2), 277–312. DOI: <https://doi.org/10.1515/cllt-2012-0011>
- Breheny, P. and Burchett, W. 2017. visreg: Visualization of regression models. R package version 2.4-1. <http://CRAN.R-project.org/package=visreg>.
- Büring, D. 2016. *Intonation and Meaning*. Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780199226269.001.0001>
- Calhoun, S. 2010. How does informativeness affect prosodic prominence? *Language and Cognitive Processes*, 25, 1099–1140. DOI: <https://doi.org/10.1080/01690965.2010.491682>
- Cangemi, F., Krüger, M. and Grice, M. 2015. Listener-specific perception of speaker-specific production in intonation. In: Fuchs, S., Pape, D., Petrone, C. and Perrier, P. (eds.), *Individual Differences in Speech Production and Perception*, 123–145. Frankfurt am Main: Peter Lang.

- Chafe, W. 1987. Cognitive constraints on information flow. In: Tomlin, R. (ed.), *Coherence and grounding in discourse*, 20–51. Amsterdam: John Benjamins. DOI: <https://doi.org/10.1075/tsl.11.03cha>
- Chen, K. and Hasegawa-Johnson, M. 2004. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model. In: *Proceedings of ICASSP*, 509–512. Montreal, Canada: IEEE Signal Processing Society. DOI: <https://doi.org/10.1109/ICASSP.2004.1326034>
- Chihara, L. M. and Hesterberg, T. C. 2012. *Mathematical statistics with resampling and R*. John Wiley & Sons.
- Cole, J. 2015. Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1–2), 1–31. DOI: <https://doi.org/10.1080/23273798.2014.963130>
- Cole, J., Kim, H., Choi, H. and Hasegawa-Johnson, M. 2007. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics*, 35, 180–209. DOI: <https://doi.org/10.1016/j.wocn.2006.03.004>
- Cole, J., Mo, Y. and Baek, S. 2010a. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes*, 25, 1141–1177. DOI: <https://doi.org/10.1080/01690960903525507>
- Cole, J., Mo, Y. and Hasegawa-Johnson, M. 2010b. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1, 425–452. DOI: <https://doi.org/10.1515/labphon.2010.022>
- Dilley, L. C. and Heffner, C. 2013. The role of F0 alignment in distinguishing intonation categories: evidence from American English. *Journal of Speech Science*, 3(1), 3–67.
- D’Imperio, M., Cangemi, F. and Grice, M. 2016. Introducing Advancing Prosodic Transcription. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 7(1), 4. DOI: <https://doi.org/10.5334/labphon.32>
- Féry, C. 2011. German sentence accents and embedded prosodic phrases. *Lingua*, 121, 1906–1922. DOI: <https://doi.org/10.1016/j.lingua.2011.07.005>
- Féry, C. 2013. Focus as prosodic alignment. *Natural Language & Linguistic Theory*, 31(3), 683–734. DOI: <https://doi.org/10.1007/s11049-013-9195-7>
- Gamer, M., Lemon, J., Fellows, I. and Singh, P. 2012. IRR: Various coefficients of interrater reliability and agreement. R package version 0.84. CRAN: <http://www.r-project.org>.
- Gamer, M., Lemon, J., Fellows, I. and Singh, P. 2014. IRR: Various coefficients of interrater reliability and agreement. R package version 0.84. CRAN: <http://www.r-project.org>.
- Godfrey, J., Holliman, E. and McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP-92*, 517–20. San Francisco, CA. DOI: <https://doi.org/10.1109/ICASSP.1992.225858>
- Grabe, E. 2004. Intonational variation in urban dialects of English spoken in the British Isles. In: Gilles, P. and Peters, J. (eds.), *Regional Variation in Intonation*, 9–31. Linguistische Arbeiten, Tuebingen, Niemeyer.
- Gries, S. 2015. The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora*, 10(1), 95–125. DOI: <https://doi.org/10.3366/cor.2015.0068>
- Halliday, M. A. K. 1967. Notes on transitivity and theme in English, part 2. *Journal of Linguistics*, 3, 199–244. DOI: <https://doi.org/10.1017/S0022226700016613>
- Hirschberg, J. 1993. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1–2), 305–340. DOI: [https://doi.org/10.1016/0004-3702\(93\)90020-C](https://doi.org/10.1016/0004-3702(93)90020-C)

- Ito, K. and Speer, S. R. 2008. Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, 58(2), 541–573. DOI: <https://doi.org/10.1016/j.jml.2007.06.013>
- Jun, S. A. 2005. Prosodic typology: The phonology of intonation and phrasing, 1. Oxford, New York: Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780199249633.001.0001>
- Jun, S. A. 2014. Prosodic typology: By prominence type, word prosody, and macro-rhythm. In: Jun, S. A. (ed.), *Prosodic typology II: The phonology of intonation and phrasing*, 520–539. Oxford Scholarship Online. DOI: <https://doi.org/10.1093/acprof:oso/9780199567300.001.0001>
- Jun, S. A. and Fletcher, J. 2014. Methodology of studying intonation: From data collection to data analysis. In: Jun, S. A. (ed.), *Prosodic Typology II: The phonology of intonation and phrasing*, 493–519. Oxford Scholarship Online. DOI: <https://doi.org/10.1093/acprof:oso/9780199567300.001.0001>
- Ladd, D. R. 1980. *The structure of intonational meaning: Evidence from English*. Bloomington: Indiana University Press.
- Ladd, D. R. 2008. *Intonational phonology*. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511808814>
- Ladd, D. R., Turnbull, R., Browne, C., Caldwell-Harris, C., Ganushchak, L., Swoboda, K., Woodfield, V. and Dediu, D. 2013. Patterns of individual differences in the perception of missing-fundamental tones. *Journal of Experimental Psychology: Human Perception and Performance*, 39(5), 1386. DOI: <https://doi.org/10.1037/a0031261>
- Lehiste, I. 1973. Phonetic disambiguation of syntactic ambiguity. *The Journal of the Acoustical Society of America*, 53, 380–380. DOI: <https://doi.org/10.1121/1.1982702>
- Mahrt, T. 2016. LMEDS: Language markup and experimental design software. <https://github.com/timmahrt/LMEDS>.
- Möhler, G. and Conkie, A. 1998. Parametric modeling of intonation using vector quantization. *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.
- Pan, S. and McKeown, K. R. 1999. Word informativeness and automatic pitch accent modeling. In: *Proceedings of EMNLP/VLC*.
- Peppé, S., Maxim, J. and Wells, B. 2000. Prosodic variation in southern British English. *Language and Speech*, 43, 309–334. DOI: <https://doi.org/10.1177/00238309000430030501>
- Pfitzinger, H. R. 1998. Local speech rate as a combination of syllable and phone rate. *Proceedings of ICSLP*, 1087–1090.
- Pitrelli, J. F., Beckman, M. E. and Hirschberg, J. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. *Proceedings of ICSLP*.
- Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E. 2007. Buckeye corpus of conversational speech (2nd release). Columbus, OH: Department of Psychology, Ohio State University. DOI: <https://doi.org/10.1016/j.specom.2004.09.001>
- R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- Redi, L. and Shattuck-Hufnagel, S. 2001. Variation in the realization of glottalization in normal speakers. *Journal of Phonetics*, 29, 407–429. DOI: <https://doi.org/10.1006/jpho.2001.0145>
- Reichel, U. D. 2014. Linking bottom-up intonation stylization to discourse structure. *Computer Speech & Language*, 28(6), 1340–1365. DOI: <https://doi.org/10.1016/j.csl.2014.03.005>

- Selkirk, E. O. 1995. Sentence prosody: Intonation, stress and phrasing. In: Goldsmith, J. (ed.), *The Handbook of Phonological Theory*, 550–569. Oxford: Blackwell.
- Selkirk, E. O. 2011. The syntax-phonology interface. In: Goldsmith, J., Riggle, J. and Yu, A. (eds.), *The Handbook of Phonological Theory, 2nd Edition*, 435–484. DOI: <https://doi.org/10.1002/9781444343069.ch14>
- Sluijter, A. and Van Heuven, V. J. 1996. Acoustic correlates of linguistic stress and accent in Dutch and American English. *Fourth International Conference on Spoken Language Processing*. DOI: <https://doi.org/10.1109/ICSLP.1996.607440>
- Speer, S. R., Warren, P. and Schafer, A. J. 2011. Situationally independent prosodic phrasing. *Laboratory Phonology*, 2, 35–98. DOI: <https://doi.org/10.1515/labphon.2011.002>
- Steedman, M. 2008. Information-structural semantics for English intonation. In: Lee, C., Gordon, M. and Büring, D. (eds.), *Topic and focus*, 245–264. Netherlands: Springer.
- Syrdal, A. K., Hirschberg, J., McGory, J. and Beckman, M. 2001. Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication*, 33(1), 135–151. DOI: [https://doi.org/10.1016/S0167-6393\(00\)00073-X](https://doi.org/10.1016/S0167-6393(00)00073-X)
- Syrdal, A. K. and McGory, J. T. 2000. Inter-transcriber reliability of ToBI prosodic labeling. *Proceedings of Interspeech*, 235–238.
- Veilleux, N., Shattuck-Hufnagel, S. and Brugos, A. 2006. Transcribing prosodic structure of spoken utterances with ToBI. January IAP. (Massachusetts Institute of Technology: MIT OpenCourseWare), <http://ocw.mit.edu>. License: Creative Commons BY-NC-SA.
- Vogel, A. P., Maruff, P., Snyder, P. J. and Mundt, J. C. 2009. Standardization of pitch-range settings in voice acoustic analysis. *Behavior Research Methods*, 41, 318–324. DOI: <https://doi.org/10.3758/BRM.41.2.318>
- Wagner, M. and Watson, D. G. 2010. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25, 905–945. DOI: <https://doi.org/10.1080/01690961003589492>
- Wood, S. 2006. Generalized additive models: An introduction with R. CRC press. DOI: <https://doi.org/10.18637/jss.v016.b03>
- Wood, S. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73, 3–36. DOI: <https://doi.org/10.2307/41057423>
- Yoon, T. 2010. Speaker consistency in the realization of prosodic prominence in the Boston University Radio Speech Corpus. *Proceedings of Speech Prosody*.
- Yoon, T., Chavarria, S., Cole, J. and Hasegawa-Johnson, M. 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. *Proceedings of Interspeech*.

How to cite this article: Roy, J., Cole, J. and Mahrt, T. 2017 Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8(1): 22, pp. 1–36, DOI: <https://doi.org/10.5334/labphon.108>

Submitted: 07 August 2017 **Accepted:** 07 August 2017 **Published:** 08 September 2017

Copyright: © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 